Journal of Machine Learning Research 25 (2024) 1-38 Submitted 10/21; Revised 1/24; Published 3/24

A Semi-parametric Estimation of Personalized Dose-response Function Using Instrumental Variables

Wei Luo WEILUO@ZJU.EDU.CN
Center for Data Science
Zhejiang University
Hangzhou, P.R.China

Yeying Zhu YEYING.ZHU@UWATERLOO.CA
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON N2L 3G1, Canada

Xuekui Zhang XUEKUI@UVIC.CA
Department of Mathematics and Statistics
University of Victoria
Victoria, BC V8P 5C2, Canada

Lin Lin L.LIN@DUKE.EDU
Department of Biostatistics and Bioinformatics
Duke University
Durham, NC 27710, USA

Editor: Ilya Shpitser

Abstract

In the application of instrumental variable analysis that conducts causal inference in the presence of unmeasured confounding, invalid instrumental variables and weak instrumental variables often exist which complicate the analysis. In this paper, we propose a model-free dimension reduction procedure to select the invalid instrumental variables and refine them into lower-dimensional linear combinations. The procedure also combines the weak instrumental variables into a few stronger instrumental variables that best condense their information. We then introduce the personalized dose-response function that incorporates the subject's personal characteristics into the conventional dose-response function, and use the reduced data from dimension reduction to propose a novel and easily implementable nonparametric estimator of this function. The proposed approach is suitable for both discrete and continuous treatment variables, and is robust to the dimensionality of data. Its effectiveness is illustrated by the simulation studies and the data analysis of ADNI-DoD study, where the causal relationship between depression and dementia is investigated.

Keywords: causal inference, central mean subspace, Mendelian randomization, SCAD, sufficient dimension reduction

. Data used in preparation of this article were obtained from the Alzheimer Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1. Introduction

In observational studies, often the research interest is to estimate the causal effect of a treatment variable on a response variable. When the treatment variable is binary or multi-valued, the causal effect is often characterized by the average causal effect between two specific treatment arms (Rubin, 1974; Holland, 1986). When the treatment variable is continuous, the causal effect is often characterized by the dose-response function (Imai and Van Dyk, 2004; Zhu et al., 2015). In the causal inference literature, a variety of methods have been proposed to estimate the causal effect, such as propensity score based matching and regression adjustment (Schafer and Kang, 2008). These methods commonly require the ignorability assumption (Rosenbaum and Rubin, 1983) that the researchers have collected all the confounders for establishing the causal relationship; otherwise, they will lead to biased results.

Under the concern of potential unmeasured confounding, analysis of instrumental variables (IV) has become a popular alternative in many applications of causal inference, such as economics and epidemiology, etc. (Angrist et al., 1996; Greenland, 2000). An instrumental variable, or briefly called an instrument, is a random variable that is associated with the treatment variable but is not associated with any unmeasured confounders. In addition, an instrumental variable must affect the response variable only through the treatment variable; otherwise, it is called an invalid instrumental variable. In other words, an invalid instrumental variable is directly associated with the outcome in the presence of the treatment. Generally, the spirit of IV analysis is to incorporate the instrumental variables, and sometimes the invalid instrumental variables, into the joint modeling of the treatment and response variables appropriately, so that the causal effect can be embedded into the joint model and recovered from the corresponding results; see a detailed example below.

In the literature, a major application scenario of IV analysis is epidemiological research, where the interest is often to investigate the causal effect of an exposure variable on a certain disease. Since a genetic variant, such as a single-nucleotide polymorphism (SNP), is determined at conception, it is not related to any environmental factors or other unmeasured confounders. Thus, a genetic variant is a good instrument if it is closely linked to the exposure but has no direct effect on the disease (Didelez and Sheehan, 2007). The study of genetic variants as candidate instrumental variables, commonly known as Mendelian randomization, has been discussed extensively (Lawlor et al., 2008; Burgess et al., 2017).

In Mendelian randomization, researchers often find two phenomena that complicate IV analysis. First, a number of genetic variants can serve as valid instruments, but each of them only has a somewhat weak bond with the exposure variable. In the literature, the weak bond is commonly revealed by a small R^2 or equivalently a small F statistic under the linear model assumptions (Staiger and Stock, 1997; Sheehan and Didelez, 2011; Burgess et al., 2017), although only a heuristic cutoff of the F statistic, usually 10, has been widely used to define the "weak instrument" (Lee et al., 2021). The existence of weak instruments makes the causal effect estimator unstable and biased in the presence of even minor unmeasured confounding (Pierce et al., 2011), which harms the reliability of the causal conclusion (Staiger and Stock, 1997; Burgess et al., 2017; Lee et al., 2021). Pierce et al. (2011) proposed a remedy that linearly combines the weak instruments into a stronger one and studied its empirical consistency, which however requires either adequate prior knowledge about the effect of each individual weak instrument or a naive assumption on the equality of these effects. Second, datasets may contain a small set of genetic variants that are directly associated with the disease, which, as mentioned above, are invalid instruments. Such genetic variants would jeopardize the consistency of IV analysis if not distinguished from the rest. These two phenomena are also observed in other applications of IV analysis.

Let Y be the response variable, T be the treatment variable, and X be a p-dimensional vector consisting of candidate instruments, all with zero mean without loss of generality. Again, in Mendelian randomization, these variables correspond to certain measurements of a specific disease, the exposure variable, and the subject's characteristics including the genetic variants, respectively. The existence of a few invalid instruments in X naturally suggests estimating the causal effect by regressing Y on (X, T) with the aid of the variable selection technique. This was proposed in Kang et al. (2016), with the linear model assumption

$$Y = a_0 T + \gamma_0^{\mathsf{T}} X + \epsilon, \tag{1}$$

and the causal effect specified as the linear dose-response function $a_0 T$. The error term ϵ is potentially associated with T as it may include the effect of unmeasured confounders, but it can be safely assumed to be independent of X in Mendelian randomization, as X carries the personal characteristics determined at conception and cannot be contaminated by any unmeasured confounder. Because a component of X that corresponds to a nonzero component of γ_0 must affect Y in the presence of T, it is by definition an invalid instrument. Thus, the set of nonzero components of γ_0 exactly indexes the set of invalid instruments, and it is assumed to be sparse following the observations above.

Under the independence assumption between ϵ and X, Model (1) implies

$$E\{(Y - a_0 T - \gamma_0^{\mathsf{T}} X) X^{\mathsf{T}}\} = 0. \tag{2}$$

Because (2) holds for any element in $\{(a_0 + b, \gamma_0 - b\Sigma_X^{-1} E(XT)) : b \in \mathbb{R}\}$, where Σ_X denotes the covariance matrix of X, additional assumptions must be adopted to make (a_0, γ_0) identifiable. For this purpose, Kang et al. (2016) strengthened the sparsity assumption on γ_0 to that γ_0 is the uniquely sparsest among all that solve (2), i.e. with the least number of nonzero components. Accordingly, they estimated a_0 and γ_0 by minimizing

$$E_n\{(Y - aT - \gamma^{\mathsf{T}} X) X^{\mathsf{T}}\} E_n\{(Y - aT - \gamma^{\mathsf{T}} X) X^{\mathsf{T}}\} + \textstyle\sum_{i=1}^{p} \phi_\lambda(|\gamma_i|) \tag{3}$$

over $a \in \mathbb{R}$ and $\gamma \in \mathbb{R}^p$, where $E_n(\cdot)$ denotes the sample mean, $\phi_\lambda(\cdot)$ is a penalty function with tuning parameter λ, and γ_i denotes the ith component of γ. When $\phi_\lambda(\cdot)$ is appropriately chosen, (3) delivers a consistent and sparse estimator of γ_0, as well as a consistent estimator of the causal effect $a_0 T$.

The linear model (1) in Kang et al. (2016) can be regarded as a set of two assumptions: first, it imposes a low-dimensional structure in the data that X must affect Y through a linear combination of X in the presence of T; second, the joint effect of T and this linear combination of X on Y must convey a linear pattern. The former is violated if multiple linear combinations of X are uniquely informative to Y in the presence of T. The latter, which adopts a parametric model on $Y|(T, X)$, is violated if the effect of T or the effect of X on Y is nonlinear or if these two effects interact. In view of these concerns, we generalize

Kang et al.'s work into a multi-index and model-free manner by assuming

$$Y = g(T, \gamma_0^{\mathsf{T}} X) + \epsilon, \tag{4}$$

where γ_0 is a $p \times d$-dimensional matrix with only a few rows being nonzero, and the functional form of $g(\cdot, \cdot)$ is completely unspecified. Similarly to (1), a nonzero row of γ_0 corresponds to a component of X that affects Y in the presence of T, which by definition is an invalid instrument. Thus, the model allows the invalid instruments to affect the response through multiple linear combinations, and these invalid instruments can be selected by recovering the sparsity of γ_0. By allowing full freedom on $g(\cdot, \cdot)$, (4) can handle the potentially complex effect of $(T, \gamma_0^{\mathsf{T}} X)$ on Y, particularly any form of interaction between T and $\gamma_0^{\mathsf{T}} X$ that implies a heterogeneous causal effect varying with subject's characteristics. A semiparametric causal effect estimator has also been proposed in Li and Guo (2020), which we will compare in detail later. Same as in (1), we allow the error term ϵ in (4) to be associated with T, and we slightly relax the independence assumption between ϵ and X to

$$E(\epsilon \mid X) = 0, \tag{5}$$

which permits the variance of ϵ to vary with X. Referring to (2), this assumption catches the essence of using instrumental variables.

Before conducting effective estimation, two issues need to be addressed due to the potential dependence between ϵ and T and the possible presence of weak instruments. First, similar to (1), neither γ_0 nor $g(T, \gamma_0^{\mathsf{T}} X)$ in Model (4) are identifiable: for any matrix β of p rows and any measurable function $f(T, \beta^{\mathsf{T}} X)$ with $E\{f(T, \beta^{\mathsf{T}} X)|X\} = 0$, this model can always be rewritten as

$$Y = \{g(T, \gamma_0^{\mathsf{T}} X) + f(T, \beta^{\mathsf{T}} X)\} + \{\epsilon - f(T, \beta^{\mathsf{T}} X)\} \equiv g^*(T, (\gamma_0, \beta)^{\mathsf{T}} X) + \epsilon^*,$$

where ϵ^* also satisfies (5). The issue of identifiability is further complicated by the free form of $g(\cdot, \cdot)$, as one can always rewrite $g(T, \gamma_0^{\mathsf{T}} X)$ as $g[T, A^{-1}\{(\gamma_0 A^{\mathsf{T}})^{\mathsf{T}} X\}]$ for any d-dimensional invertible matrix A, making $\gamma_0 A^{\mathsf{T}}$ a valid substitute of γ_0. Second, even if $g(T, \gamma_0^{\mathsf{T}} X)$ is uniquely defined, its estimation induced from a natural generalization of (2) will suffer from the presence of weak instruments. Namely, without appropriate procedures that handle the weak instruments *a priori*, any function $\phi(X)$ that characterizes $g(T, \gamma_0^{\mathsf{T}} X)$ as the unique solution to

$$E_n[\{Y - g(T, \gamma_0^{\mathsf{T}} X)\}\phi(X)\} = 0 \tag{6}$$

would inevitably include functions of X that are weakly bonded with T, whose corresponding estimating equations in (6) would contribute little to the estimation of $g(T, \gamma_0^{\mathsf{T}} X)$ but generate more bias. The same phenomenon persists in the linear model (1) in Kang et al. (2016). A remedy that is robust to weak instruments under (1) has been proposed in Kang et al. (2022), but focuses on testing whether a_0 equals a prefixed value rather than estimating the causal effect $a_0 T$.

To address the issue of identifiability, we will reparametrize γ_0 with the aid of sufficient dimension reduction (SDR), a mainstream of model-free dimension reduction techniques in the statistical literature, as well as an appropriate strengthening of sparsity of γ_0 that resembles Kang et al.'s spirit. In light of the empirical findings in Pierce et al. (2011), SDR will also be adopted on $T|X$ to merge all the instruments appropriately into a few stronger ones, so as to alleviate the issue of weak instruments. Using the reduced predictor, i.e. the estimates of both the strengthened instruments and the strengthened invalid instruments $\gamma_0^\intercal X$ (subject to certain equivalence classes due to the reparametrization of γ_0), we then develop an innovative nonparametric estimator of $g(T, \gamma_0^\intercal X)$ that complies with the existence of unmeasured confounding and meanwhile avoids futile estimating equations. This two-step nonparametric estimation procedure is robust against the number of instruments, especially the weak instruments, which is another advantage over Kang et al.'s method in addition to the modeling flexibility.

Throughout the article, we follow the literature (Rosenbaum and Rubin, 1983; Luo et al., 2017) to adopt the common support condition for $\gamma_0^\intercal X$:

$$\Omega(\gamma_0^\intercal X \mid T = t) \equiv \Omega(\gamma_0^\intercal X) \text{ for all } t \in \Omega(T), \tag{7}$$

where $\Omega(\cdot)$ denotes the support of a distribution. This condition permits averaging $g(t, \gamma_0^\intercal X)$ over the marginal distribution of $\gamma_0^\intercal X$ for any fixed outcome t of T. The result $E\{g(t, \gamma_0^\intercal X)\}$ averages the causal effect of T on Y over the population and thus is the aforementioned dose-response function. In particular, it reduces to the linear dose-response function $a_0 T$ in Kang et al. (2016) when $g(\cdot, \cdot)$ is linear. Because $g(t, \gamma_0^\intercal x)$ identifies the causal effect specific for the subject's characteristics, we call it the personalized dose-response function. For clarity, we call the conventional dose-response function, i.e. $E\{g(t, \gamma_0^\intercal X)\}$, the marginal dose-response function. We will estimate both functions in this paper.

As mentioned above, Li and Guo (2020) also proposed a semi-parametric estimator of essentially the personalized dose-response function, in the presence of invalid instruments. They assumed that, given the unmeasured confounders, (T, X) affect Y through a single index $a_0 T + \gamma_0^\intercal X$ in a model-free manner, which however precludes any interaction effect of T and X. To address the issue of unmeasured confounders, Li and Guo (2020) also assumed the sparsity of γ_0 to make the causal effect identifiable and employed SDR to facilitate the estimation, and they additionally adopted a linear model on $T|X$ with an independent error. The latter is crucial to Li and Guo's method as the error term therein serves as a control variable and eases the causal effect estimation. Because Kang et al. (2016) imposed a linear model on $Y|(T, X)$ but is model-free on $T|X$, Li and Guo's method can be regarded as a conjugate to Kang et al.'s method. By contrast, using natural SDR regulations and innovative nonparametric techniques, our method is model-free on both $Y|(T, X)$ and $T|X$. Therefore, it is uniquely applicable if both $Y|(T, X)$ and $T|X$ convey nonlinear patterns, or if T and X interact in affecting Y, or if X affects Y through a multi-index manner, in the presence of unmeasured confounders.

The rest of the paper is organized as follows. We briefly review the literature of SDR in Section 2, using which we re-parameterize γ_0 and refine the instruments in Section 3. A sparse estimator of γ_0, which additionally selects the set of invalid instruments, is proposed in Section 4. Based on the reduced data, Section 5 regulates the identifiability of $g(T, \gamma_0^\intercal X)$ under a general assumption related to the instrument strength, and Section 6 introduces the new nonparametric estimators of both the personalized and the marginal dose-response functions, as well as a diagnosis procedure for the assumption in Section 5. Simulation studies and a real data application are presented in Section 7 and Section

8, respectively. Throughout the theoretical development of the paper, we assume that X consists of continuous random variables and that both Y and T are univariate.

2. A Review of SDR

As mentioned above, SDR is a widely applied family of model-free dimension reduction methods. For predictor X and a general response variable W, SDR assumes a low-dimensional structure on $W|X$ such that X affects W only through a low-dimensional linear combination $\beta_W^{\mathsf{T}} X$, that is,

$$W \perp\!\!\!\perp X \mid \beta_W^{\mathsf{T}} X \tag{8}$$

where $\perp\!\!\!\perp$ denotes the independence between two random elements. As no parametric assumptions are adopted on $W|\beta_W^{\mathsf{T}} X$, $\beta_W^{\mathsf{T}} X$ can serve as the working predictor in the subsequent analysis that permits full freedom of modeling. Such analysis will be both reliable and accurate, due to the sufficiency and the low dimensionality of $\beta_W^{\mathsf{T}} X$.

As (8) holds under an arbitrary invertible column transformation of β_W, it is a characterization of $\mathcal{S}(\beta_W)$, the column space of β_W. For identifiable parametrization, Cook (1998) introduced the central subspace $\mathcal{S}_{W|X}$ as the uniquely smallest subspace of $\mathbb{R}^p$ that satisfies (8) and is meanwhile included in any other space that also satisfies (8). The existence of the central subspace requires fairly general conditions on X regardless of the nature of W, so we assume it throughout the article. For simplicity, we still use β_W to denote an arbitrary basis matrix of $\mathcal{S}_{W|X}$. The reduced predictor of SDR is then $\beta_W^{\mathsf{T}} X$ or any of its invertible linear transformation.

Although $\mathcal{S}_{W|X}$ is the SDR parameter of interest, its non-Euclidean nature urges the necessity of introducing an intermediate Euclidean parameter for ease of estimation. Naturally, this intermediate parameter is β_W, or, better yet, a uniquely defined matrix Γ_W whose column space $\mathcal{S}(\Gamma_W)$ coincides with $\mathcal{S}_{W|X}$. For a major family of SDR methods, which include the popularly used sliced inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991), and directional regression (Li and Wang, 2007), etc., such Γ_W (denoted by M in these papers) is commonly constructed using the moments of $X|W$, and the coincidence between $\mathcal{S}(\Gamma_W)$ and $\mathcal{S}_{W|X}$ is guaranteed by mild conditions on X and general regularity conditions on $W|X$. Given an estimator $\widehat{\Gamma}_W$, which is typically $n^{1/2}$-consistent, $\mathcal{S}_{W|X}$ is commonly estimated by the linear span of the leading left singular vectors of $\widehat{\Gamma}_W$. The number of these vectors, which is the same as the dimension of $\mathcal{S}_{W|X}$, can be determined by the Bayesian information criterion (BIC; Zhu et al., 2006), the ladle estimator (Luo and Li, 2016), and the predictor augmentation estimator (PAE; Luo and Li, 2021), etc.

When the research interest is specified to regression, i.e. estimating $E(W|X)$, SDR can be adjusted to only detect the low-dimensional structure of $E(W|X)$; that is, it instead assumes the existence of a low-dimensional $\beta_W^{\mathsf{T}} X$ such that

$$E(W \mid X) = E(W \mid \beta_W^{\mathsf{T}} X), \tag{9}$$

where we abuse the notation β_W in (8) if no ambiguity is caused. Similar to $\mathcal{S}_{W|X}$ above, the identifiable parameter for (9) is the central mean subspace $\mathcal{S}_{E(W|X)}$ (Cook and Li, 2002) that satisfies (9) with minimal dimension, and the existing estimators of $\mathcal{S}_{E(W|X)}$ typically construct a uniquely defined matrix-valued intermediate parameter Γ_W that spans $\mathcal{S}_{E(W|X)}$; see, for example, the principal Hessian directions (pHd; Li, 1992) and the minimum average variance estimation (MAVE; Xia et al., 2002). Because any matrix β that satisfies (8) must also satisfy (9), $\mathcal{S}_{E(W|X)}$ is always a subspace of $\mathcal{S}_{W|X}$.

In some applications, a part of the predictor is prefixed to be used in the regression and does not participate in the SDR procedure. This is the case in our setting, where T must be included in fitting the personalized dose-response function. Accordingly, SDR for regression is adjusted to partial SDR, which assumes

$$E(W \mid T, X) = E(W \mid T, \beta_W^{\intercal} X). \tag{10}$$

Here, we still use the general response W for consistency of notations. Similarly to the above, the identifiable parameter for (10) is called the partial central mean subspace and denoted by $\mathcal{S}_{E(W|X)}^{(T)}$ (Chiaromonte et al., 2002), and is commonly regarded as $\mathcal{S}(\Gamma_W)$ for some uniquely defined matrix-valued intermediate parameter Γ_W. The existing estimators of $\mathcal{S}_{E(W|X)}^{(T)}$ are omitted here as they are inapplicable in our setting due to the unmeasured confounding in the data; see more details in Section 3.

Generally, an estimator of the central SDR subspace (i.e. $\mathcal{S}_{W|X}$, $\mathcal{S}_{E(W|X)}$, or $\mathcal{S}_{E(W|X)}^{(T)}$) has non-sparse basis matrices. This limits both the estimation consistency and the interpretability of the SDR result, especially when p is relatively large compared with the sample size. To address this issue, SDR can be adjusted to sparse SDR, where only a few components of X are assumed informative to W. Because X affects W through $\beta_W^{\intercal} X$, an informative component of X must correspond to a nonzero row of β_W or a nonzero row of the aforementioned unique matrix Γ_W. The equivalence between the latter two can be easily seen from the fact that any matrices with the identical column space must also share the same index of nonzero rows. Hence, sparse SDR implies the row-wise sparsity of Γ_W, and the level of sparsity of the central SDR subspace can be quantified by the number of zero rows of Γ_W. This resembles the transition from the sparsity of invalid instruments to the row-wise sparsity of γ_0 discussed below (4) in the Introduction, and will be revisited in the next section. The existing sparse SDR estimators, which truly select all the nonzero rows of Γ_W, include the coordinate independent sparse estimator (CISE; Chen et al., 2010) and lasso SIR (Lin et al., 2019), etc.

3. Regulation of Dimension Reduction

Using SDR, we now give an identifiable re-parametrization of γ_0 in (4) that simultaneously permits effective subsequent IV analysis. Given a specific $g(\cdot,\cdot)$, the arbitrariness of γ_0 mentioned in the Introduction can be readily addressed by the partial SDR theory, if we regard $g(T, \gamma_0^{\intercal} X)$ as W in (10) and use the corresponding $\mathcal{S}_{E(W|X)}^{(T)}$ as the parameter of interest. To guarantee the identifiability of this parametrization under potential arbitrariness of $g(\cdot,\cdot)$, we next follow the discussion below (4) to seek for effective additional regulations with the aid of the sparsity of invalid instruments, i.e. the row-wise sparsity of γ_0. Again, as reviewed in Kang et al. (2016), this sparsity is commonly observed in Mendelian randomization in practice. For ease of presentation, we introduce the notation $\mathcal{S}_{\mathrm{PDRF}}$ for the resulting identifiable $\mathcal{S}_{E(W|X)}^{(T)}$ before giving its formal definition; the subscript refers to the personalized dose-response function. We also use γ_0 to denote an arbitrary basis matrix of $\mathcal{S}_{\text{PDRF}}$ and use J_0 to denote the index set of its zero rows. As reviewed at the end of Section 2 above, J_0 is invariant of the arbitrariness of γ_0 and thus is uniquely defined; similar arguments will be omitted from the rest of the article.

To formulate the sparsity of $\mathcal{S}_{\text{PDRF}}$, we additionally adopt the SDR assumption (8) on $T|X$, with $\mathcal{S}_{T|X}$ being d_T-dimensional for some $d_T < p$ and spanned by some β_T. Referring to the empirical study in Pierce et al. (2011) mentioned in the Introduction, this also addresses the issue of weak instruments. Namely, let L_0 be the index set of nonzero rows of β_T, which is $\{1, \ldots, p\}$ if $\mathcal{S}_{T|X}$ is non-sparse. By the definition of $\mathcal{S}_{T|X}$, L_0 must index the components of X that are uniquely informative to T, which includes all the instruments and possibly some invalid instruments. Together with the interpretation of γ_0 above, $L_0 \cap J_0$ indexes the set of instruments in X, and the part of $\beta_T^{\mathsf{T}} X$ formed by these instruments, denoted by $\beta_{T,L_0 \cap J_0}^{\mathsf{T}} X_{L_0 \cap J_0}$, is their optimal linear combination in terms of preserving and condensing their signal in explaining T in the presence of invalid instruments. As seen later, this linear combination will be estimated in a nonparametric manner without assuming equal effect of instruments or requiring prior knowledge, for which it is advantageous compared with those discussed in Pierce et al. (2011). To avoid the extreme case that X is independent of T, which would preclude any link of the effect of X on Y to the causal effect of T on Y, we assume

$$d_T \geq 1. \tag{11}$$

Referring to the interpretation of $\mathcal{S}_{T|X}$ above, this can also be regarded as a prerequisite for the existence of instruments.

Since $E(Y|X)$ is equal to $E\{g(T, \gamma_0^{\mathsf{T}} X)|X\}$ under (5), it is measurable with respect to $(\beta_T, \gamma_0)^{\mathsf{T}} X$. This implies a low-dimensional $\mathcal{S}_{E(Y|X)}$, whose arbitrary basis matrix β_Y satisfies (9) if Y serves as W in the latter. By definition, we have

$$\mathcal{S}_{E(Y|X)} \subseteq \mathcal{S}(\mathcal{S}_{T|X}, \mathcal{S}_{\text{PDRF}}), \tag{12}$$

where $\mathcal{S}(\cdot, \cdot)$ denotes the space spanned by the union of two spaces. Because IV analysis hinges on (5), no hypothetical direction in $\mathcal{S}_{\text{PDRF}}$ that falls outside of $\mathcal{S}(\mathcal{S}_{T|X}, \mathcal{S}_{E(Y|X)})$ can be detected. Thus, we strengthen (12) to assume

$$\mathcal{S}_{\text{PDRF}} \in \mathcal{G} \equiv \{\mathcal{S}(\gamma) \subset \mathbb{R}^p : \mathcal{S}(\mathcal{S}_{T|X}, \gamma) = \mathcal{S}(\mathcal{S}_{T|X}, \mathcal{S}_{E(Y|X)})\}. \tag{13}$$

We next build the identifiability of $\mathcal{S}_{\text{PDRF}}$ among all in $\mathcal{G}$, using its aforementioned sparsity.

Recall that L_0 indexes the nonzero rows of β_T and J_0 indexes the zero rows of γ_0. Let q be the minimal number of nonzero entries for any nonzero vector in $\mathcal{S}_{T|X}$. By simple algebra, q must be less than or equal to the cardinality of L_0. Let $L_0 \backslash J_0$ be the set of elements in L_0 that are not in J_0, which by definition indexes the invalid instruments that are as well uniquely informative to T. Similar to *Corollary 1* in Kang et al. (2016), we regulate the sparsity of $\mathcal{S}_{\text{PDRF}}$ by

Assumption 1 *The cardinality of $L_0 \backslash J_0$ is less than $q/2$.*

Under Assumption 1, the cardinality of $L_0 \cap J_0$ must be greater than $q/2$, or equivalently that there are more than $q/2$ instruments. If $\mathcal{S}_{T|X}$ is not sparse, that is, if every invalid instrument contributes to the modeling of T, then Assumption 1 can be read as there are less than $q/2$ invalid instruments. Together, these require that more instruments than invalid instruments exist in the data, which is more restrictive than the existence of instruments conventionally adopted in IV analysis. The same has been assumed in Li and Guo (2020). This is the price we pay for allowing the existence of invalid instruments and allowing free form of their effects. Note that Assumption 1 does not impose any restriction on the components of X that fall out of L_0, which include those who are uninformative to (Y, T) as well as those invalid instruments who are additionally uninformative to T.

To illustrate how Assumption 1 identifies $\mathcal{S}_{\text{PDRF}}$, consider a special case where X affects T through $X_1+X_2+X_3$ and it affects Y through $2X_1+X_2+X_3$, X_i being the ith component of X for $i = 1, \ldots, p$. Then β_T is $(1, 1, 1, 0, \ldots, 0)^\intercal$ and β_Y is $(2, 1, 1, 0, \ldots, 0)^\intercal$, both up to multiplicative scalars. Under Assumption 1, any basis vector of $\mathcal{S}_{\text{PDRF}}$ must have at most one nonzero entry among its first three entries, so $\mathcal{S}\{(1, 0, \ldots, 0)^\intercal\}$ is the only choice for $\mathcal{S}_{\text{PDRF}}$ among all in $\mathcal{G}$. The next theorem justifies this identifiability in general.

Theorem 1 *Under Assumption 1, $\mathcal{S}_{\text{PDRF}}$ spanned by γ_0 is the uniquely sparsest space in $\mathcal{G}$ and has the smallest possible dimension; that is, any other space $\mathcal{S}(\gamma) \in \mathcal{G}$ must have an equal or larger dimension, and γ must have more nonzero rows than γ_0.*

Proof We first show that any $\mathcal{S}(\gamma) \in \mathcal{G}$ must have an equal or larger dimension compared with $\mathcal{S}(\gamma_0)$. Since $\mathcal{S}(\beta_T, \gamma_0) = \mathcal{S}(\beta_T, \gamma)$, γ must fall in $\mathcal{S}(\beta_T, \gamma_0)$; that is, there exist matrices A and B such that

$$\gamma = \beta_T A + \gamma_0 B. \tag{14}$$

If $\mathcal{S}(\gamma)$ is lower-dimensional than $\mathcal{S}(\gamma_0)$, then there must exist some $\beta \in \mathcal{S}(\gamma_0)$ that is orthogonal to γ. Similarly to (14), we have $\beta = \beta_T C + \gamma D$. The orthogonality between β and γ then implies $D = 0$, which means that $\beta = \beta_T C$. However, this contradicts Assumption 1, which means that $\mathcal{S}(\gamma)$ must have at least equal dimension as $\mathcal{S}(\gamma_0)$.

Now suppose $\mathcal{S}(\gamma)$ differs from $\mathcal{S}(\gamma_0)$, which means that A is nonzero in (14). We next show that γ must have more nonzero rows than γ_0. Let $Q(\beta_T) = I_p - \beta_T(\beta_T^\intercal \beta_T)^{-1}\beta_T$, i.e. the projection matrix onto the orthogonal complement of $\mathcal{S}(\beta_T)$. Since $\mathcal{S}(\gamma_0), \mathcal{S}(\gamma) \in \mathcal{G}$, we must have

$$\mathcal{S}(Q(\beta_T)\gamma) = \mathcal{S}(Q(\beta_T)\gamma_0). \tag{15}$$

For any $i \in \{1, \ldots, p\}$, let γ_i, $\beta_{T,i}$, and $\gamma_{0,i}$ be the ith row of γ, β_T, and γ_0, respectively. If $i \notin L_0$, then since $\beta_{T,i} = 0$, the ith row of $Q(\beta_T)$ must coincide with the ith row of I_p, which implies the identity between the ith row of $Q(\beta_T)\gamma$ and γ_i, as well as the identity between the ith row of $Q(\beta_T)\gamma_0$ and $\gamma_{0,i}$. Thus, (15) implies that γ_i and $\gamma_{0,i}$ must be either both zero or both nonzero, or equivalently that the sparsity of γ and γ_0 may differ only in their rows indexed by L_0. Since any direction in $\mathcal{S}(\beta_{T,L_0})$ must have at least q nonzero entries and any direction in $\mathcal{S}(\gamma_{0,L_0})$ must have less than $q/2$ nonzero entries, a nonzero A means that any direction in $\mathcal{S}(\beta_{T,L_0}A + \gamma_{0,L_0}B)$ or equivalently $\mathcal{S}(\gamma_{L_0})$ must have more than $q/2$ nonzero entries. This completes the proof. □

By Theorem 1, $\mathcal{S}_{\mathrm{PDRF}}$ is the uniquely sparsest as well as lowest dimensional among all in $\mathcal{G}$. This again conforms to the nature of sparsity of invalid instruments, and it also delivers maximal dimension reduction. An interesting question raised by a Referee is that how $\mathcal{S}_{\mathrm{PDRF}}$ will change if one removes an invalid instrument from X. This roughly depends on the complexity of data, and will be addressed in Appendix C. Let d be the dimension of $\mathcal{S}_{\mathrm{PDRF}}$, and let d_Y be the dimension of $\mathcal{S}_{E(Y|X)}$. By (13), there exists coefficient matrices $A_0 \in \mathbb{R}^{d_T \times d_Y}$ and $B_0 \in \mathbb{R}^{d \times d_Y}$ such that

$$\beta_Y = \beta_T A_0 + \gamma_0 B_0. \tag{16}$$

The minimality of d further implies the full row rank of B_0, unless $\mathcal{S}_{\mathrm{PDRF}}$ is trivially zero-dimensional.

In view of (16), the proposed regulation of $\mathcal{S}_{\mathrm{PDRF}}$ can be regarded as a semi-parametric generalization of Kang et al.'s work. Suppose both $\mathcal{S}_{T|X}$ and $\mathcal{S}_{E(Y|X)}$ are one-dimensional. Then, without assuming linear models on $T|X$ or $Y|X$, Li and Duan (1989) showed that $\Sigma_X^{-1}E(XT)$ equals $a\beta_T$ and $\Sigma_X^{-1}E(XY)$ equals $c\beta_Y$ for some scalars a and c, under some mild condition on X. The former equates (2) with

$$E[\{Y - (a\beta_T + \gamma_0)^{\intercal} X\} X^{\intercal}] = 0; \tag{17}$$

the latter means $E\{(Y - c\beta_Y^{\intercal} X)X^{\intercal}\} = 0$, which further equates (17) with (16) for $A_0 = a/c$ and $B_0 = 1/c$ if c is nonzero. Thus, (16) includes (2) as a special case without assuming any parametric model on $Y|(T, X)$. Again, the generality of our work at the dimension reduction stage is two-fold: first, we allow for multiple linear combinations of invalid instruments to affect Y in the presence of T; second and more importantly, we do not adopt any modeling assumptions when reducing data, permitting full freedom in the subsequent estimation of the personalized and the marginal dose-response functions.

4. Estimation of $\mathcal{S}_{\mathrm{PDRF}}$

By applying the existing SDR methods, both $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{T|X}$ can be consistently estimated. Using these estimates, the unknown terms in (16) are A_0, B_0, and γ_0. Thus, to estimate $\mathcal{S}_{\mathrm{PDRF}}$, it is natural to introduce an objective function based on (16).

As mentioned in Section 3, the coefficient matrix B_0 has full row rank, which means that $\gamma_0 B_0$ must also span $\mathcal{S}_{\mathrm{PDRF}}$. Because our parameter of interest is $\mathcal{S}_{\mathrm{PDRF}}$ rather than its basis matrix γ_0, we regard $\gamma_0 B_0$, denoted by $\Gamma_0 \in \mathbb{R}^{p \times d_Y}$, as the intermediate parameter in (16). This complies with the literature of SDR methods that introduces a matrix-valued intermediate parameter Γ_W that spans the central SDR subspace (see Section 2), and, more importantly, it simplifies (16) to the linear constraint

$$\beta_Y = \beta_T A_0 + \Gamma_0, \tag{18}$$

which facilitates the downstream implementation. To ease the theoretical development, we further impose the uniqueness of Γ_0 by requiring so for both β_Y and β_T, for which we restrict both the first d_Y nonzero rows of β_Y and the first d_T nonzero rows of β_T to form the identity matrix. Given an estimator of Γ_0, denoted by $\widehat{\Gamma}$, $\mathcal{S}_{\mathrm{PDRF}}$ can be estimated by the linear span of the leading left singular vectors of $\widehat{\Gamma}$. The uniqueness of Γ_0 is not essential:

the consistency results below will still hold in general if we build them directly for $\mathcal{S}_{\rm PDRF}$ rather than for Γ_0.

Let $\widehat{\beta}_Y$ and $\widehat{\beta}_T$ be the unique basis matrices of the consistent estimators of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{T|X}$, respectively, under the regulation above that their first few significantly nonzero rows must form the identity matrix. To tackle the sparsity of $\mathcal{S}_{\rm PDRF}$ under Assumption 1, we also incorporate a penalty function of certain rows of Γ. These together lead to the objective function

$$\widehat{s}(A,\Gamma) = \mathrm{tr}\{(\widehat{\beta}_Y - \widehat{\beta}_T A - \Gamma)^{\intercal}(\widehat{\beta}_Y - \widehat{\beta}_T A - \Gamma)\} + \textstyle\sum_{i\in\mathcal{I}} \phi_\lambda(\|\Gamma_i\|_2), \tag{19}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a square matrix, $\phi_\lambda(\cdot)$ is the penalty function introduced in (3), and Γ_i denotes the ith row of Γ for $i = 1, \ldots, p$. As an illustration, we set $\phi_\lambda(\cdot)$ as the smooth clapped absolute distance penalty (SCAD, Fan and Li, 2001) in this article. The index set $\mathcal{I}$ is $\{1, \ldots, p\}$ in general, but it can be reduced to exclude rows of Γ_0 that are surely nonzero or equivalently indicate the invalid instruments. In practice, these rows can be either presumed *a priori* or, if both $\mathcal{S}_{T|X}$ and $\mathcal{S}_{E(Y|X)}$ are sparse, detected by the intersection of zero rows of $\widehat{\mathcal{S}}_{T|X}$ and nonzero rows of $\widehat{\mathcal{S}}_{E(Y|X)}$ under (18). The overall gain in the latter case, however, is questionable, as sparse estimators of $\mathcal{S}_{T|X}$ and $\mathcal{S}_{E(Y|X)}$ are usually derived by penalized estimation and thus are also more biased; see the end of the section for some relative discussion.

Borrowing from the rich literature of penalized least square estimation, $\widehat{s}(A,\Gamma)$ can be readily minimized by an iterative algorithm that updates A and Γ alternatively. Namely, at each iteration, we first regard Γ as fixed, by which $\widehat{s}(A,\Gamma)$ is a quadratic function of A and can be easily minimized to update A; we then regard A as fixed, by which $\widehat{s}(A,\Gamma)$ is a penalized quadratic function of Γ and again can be easily minimized to update Γ. The iteration stops when a prefixed threshold is met. The details of this algorithm are presented in the following. Following Fan and Li (2001), a in Step 1 is fixed at 3.7, and λ is tuned by a five-fold cross validation. To set an initial value of the algorithm, we use the lasso penalty (Tibshirani, 1996) in (19), as for which $\widehat{s}(\cdot,\cdot)$ is a convex function and can be easily minimized. By the theory of lasso regression, such initial value will also approximate to (A_0, Γ_0) subject to appropriate tuning procedure, which speeds up the algorithm.

Algorithm 1 Algorithm for the estimation of $\mathcal{S}_{\rm PDRF}$

Step 0. Calculate the initial value of $\widetilde{A}$ using the lasso penalty in (19).

Step 1. Given $\widetilde{A}$, calculate $\check{\Gamma} = \widehat{\beta}_Y - \widehat{\beta}_T\widetilde{A}$. Let $\widetilde{\Gamma} = \check{\Gamma}$, but, for each $i \in \mathcal{I}$, modify $\widetilde{\Gamma}_i$ to be

$$\widetilde{\Gamma}_i = \begin{cases} (\check{\Gamma}_i/\|\check{\Gamma}_i\|_2)\max\{\|\check{\Gamma}_i\|_2 - \lambda, 0\} & \text{if } 0 < \|\check{\Gamma}_i\|_2 \le 2\lambda \\ (\check{\Gamma}_i/\|\check{\Gamma}_i\|_2)\,[\{(a-1)\|\check{\Gamma}_i\|_2 - a\lambda\}/(a-2)] & \text{if } 2\lambda < \|\check{\Gamma}_i\|_2 \le a\lambda \\ \check{\Gamma}_i & \text{if } \|\check{\Gamma}_i\|_2 > a\lambda \text{ or } \|\check{\Gamma}_i\|_2 = 0 \end{cases}$$

Step 2. Given $\widetilde{\Gamma}$, calculate $\widetilde{A} = (\widehat{\beta}_T^{\intercal}\widehat{\beta}_T)^{-1}\widehat{\beta}_T^{\intercal}(\widehat{\beta}_Y - \widetilde{\Gamma})$.

Step 3. Iterate between Step 1 and Step 2 until a convergence threshold is met. The most updated results are $\widehat{\Gamma}$ and $\widehat{A}$, respectively.

Denote the minimizer of (19) by $(\widehat{A}, \widehat{\Gamma})$. Referring to the discussion below (18), to consistently estimate $\mathcal{S}_{\rm PDRF}$ from $\widehat{\Gamma}$, we must consistently estimate the unknown dimension

d of $\mathcal{S}_{\text{PDRF}}$ first. Nonetheless, for the continuity of the presentation, we tentatively assume d to be known *a priori*, under which we estimate $\mathcal{S}_{\text{PDRF}}$ by the linear span of the leading d left singular vectors of $\widehat{\Gamma}$, denoted by $\widehat{\mathcal{S}}_{\text{PDRF}}$. As justified in Lemma 1 in Appendix A, this simplification does not change any downstream asymptotic results and thus is valid. A consistent estimator $\widehat{d}$ of d, which is used to implement $\widehat{\mathcal{S}}_{\text{PDRF}}$ in practice and can be readily derived by the existing methods based on $\widehat{\Gamma}$, is deferred to the end of the section.

In an oracle situation where the set of invalid instruments is known, one can estimate Γ_0 by minimizing the first term of $\widehat{s}(\cdot,\cdot)$ over $\{\Gamma \in \mathbb{R}^{p\times d_Y} : \Gamma_i = 0 \text{ for all } i \in J_0\}$. Denote this minimizer by $(\widehat{A}^{ora}, \widehat{\Gamma}^{ora})$. The linear span of the leading d left singular vectors of $\widehat{\Gamma}^{ora}$ is clearly the benchmark for all the sparse estimators of $\mathcal{S}_{\text{PDRF}}$. The following theorem shows that the proposed $\widehat{\mathcal{S}}_{\text{PDRF}}$ is not only asymptotically consistent, but also enjoys the strong oracle property that it tends to exactly coincide with this benchmark.

Theorem 2 *Suppose $\mathcal{S}(\widehat{\beta}_T)$ is a n^s-consistent estimator of $\mathcal{S}_{T|X}$ and $\mathcal{S}(\widehat{\beta}_Y)$ is a n^v-consistent estimator of $\mathcal{S}_{E(Y|X)}$. If $\lambda \to 0$ and $n^{\min\{s,v\}}\lambda \to \infty$, then $\widehat{\Gamma}$ satisfies $\|\widehat{\Gamma}-\Gamma_0\|_2 = O_P(n^{-\min\{s,v\}})$, $P(\widehat{\Gamma}_i = 0) \to 1$ for all $i \in J_0$, and $P(\widehat{\Gamma} = \widehat{\Gamma}^{ora}) \to 1$.*

Proof We use the notations in the iterative algorithm mentioned above, but we denote $\check{\Gamma}$ by $\check{\Gamma}(A)$ and denote $\widetilde{\Gamma}$ by $\widetilde{\Gamma}(A)$ to clarify the dependence of these terms on A. Let H_0 be the index set of nonzero rows of γ_0, i.e. the complement of J_0 with respect to $\{1,\ldots,p\}$. Denote the submatrices of β_T and β_Y consisting of rows indexed by J_0 by β_{T,J_0} and β_{Y,J_0}, respectively. By simple algebra, Assumption 1 implies that β_{T,J_0} must have full column rank (otherwise, there would exist a zero column of β_{T,J_0} after appropriate column transformation, making q not more than the cardinality of $L_0\backslash J_0$), and $(\widehat{A}^{ora}, \widehat{\Gamma}^{ora})$ has the closed form

$$\widehat{A}^{ora} = (\widehat{\beta}_{T,J_0}^{\intercal}\widehat{\beta}_{T,J_0})^{-1}\widehat{\beta}_{T,J_0}^{\intercal}\widehat{\beta}_{Y,J_0}, \quad \widehat{\Gamma}_i^{ora} = \widehat{\beta}_{Y,i} - \widehat{\beta}_{T,i}\widehat{A}^{ora} \text{ for all } i \in H_0. \tag{20}$$

Since $\widehat{\beta}_Y = \beta_Y + O_P(n^{-v})$ and $\widehat{\beta}_T = \beta_T + O_P(n^{-s})$, we have $\widehat{A}^{ora} = A + O_P(n^{-\min\{s,v\}})$, which means $\|\check{\Gamma}(\widehat{A}^{ora}) - \Gamma_0\| = O_P(n^{-\min\{s,v\}}) = o_P(\lambda)$. By the definition of $\widetilde{\Gamma}(A)$ in Algorithm 1, we have, with probability tending to one, $\|\widetilde{\Gamma}_i(\widehat{A}^{ora}) - \Gamma_i\| = O_P(n^{-\min\{s,v\}})$ for all $i \in H_0$, $\widetilde{\Gamma}_i(\widehat{A}^{ora}) = 0$ for all $i \in J_0$, and $\widetilde{\Gamma}(\widehat{A}^{ora}) = \widehat{\Gamma}^{ora}$. For simplicity of notations, we denote $\widehat{A}^{ora}$ by $\bar{A}$ and $\widehat{\Gamma}^{ora}$ by $\bar{\Gamma}$, and do not distinguish between $\bar{\Gamma}$ and $\widetilde{\Gamma}(\bar{A})$. The proof will be complete if we can show that $(\bar{A}, \bar{\Gamma})$ minimizes $\widehat{s}(A,\Gamma)$ with probability converging to one.

By simple algebra, for any fixed A, $\widetilde{\Gamma}(A)$ minimizes $\widehat{s}(A,\Gamma)$. Thus, let $\widetilde{s}(A)$ denote $\widehat{s}(A, \widetilde{\Gamma}(A))$. We only need to show that $\bar{A}$ minimizes $\widetilde{s}(A)$ with probability converging to one. By construction, we have

$$\begin{aligned}\widetilde{s}(A) &= \textstyle\sum_{i=1}^p \{\|\widehat{\beta}_{Y,i} - \widehat{\beta}_{T,i}A - \widetilde{\Gamma}_i(A)\|^2 + \phi_\lambda(\|\widetilde{\Gamma}_i(A)\|)\} \\ &= \textstyle\sum_{i=1}^p \{\|\check{\Gamma}_i(A) - \widetilde{\Gamma}_i(A)\|^2 + \phi_\lambda(\|\widetilde{\Gamma}_i(A)\|)\} \equiv \sum_{i=1}^p \widetilde{s}_i(A).\end{aligned}$$

For each $i = 1,\ldots,p$, if $\|\check{\Gamma}_i(A)\| < 2\lambda$, then we have

$$\begin{aligned}\widetilde{s}_i(A) &= \{\check{\Gamma}_i(A) - \widetilde{\Gamma}_i(A)\}^2 + \lambda\|\widetilde{\Gamma}_i(A)\| \\ &= \|\check{\Gamma}_i(A)\|^2 I(\|\check{\Gamma}_i(A)\| \le \lambda) + [\lambda^2 + \lambda\{\|\check{\Gamma}_i(A)\| - \lambda\}]I(\|\check{\Gamma}_i(A)\| > \lambda).\end{aligned} \tag{21}$$

In this case, the minimum value of $\widetilde{s}_i(A)$ is reached only when $\|\check{\Gamma}_i(A)\| \le \lambda$. If $\|\check{\Gamma}_i(A)\| \in [2\lambda, a\lambda)$, then we have

$$\begin{aligned}\widetilde{s}_i(A) &= \{\|\check{\Gamma}_i(A)\| - a\lambda\}^2/(a-2)^2 - \{\|\widetilde{\Gamma}_i(A)\| - a\lambda\}^2/\{2(a-1)\} + (a+1)\lambda^2/2 \\ &= (3-a)\{\|\check{\Gamma}_i(A)\| - a\lambda\}^2/\{2(a-2)^2\} + (a+1)\lambda^2/2, \end{aligned} \tag{22}$$

which has minimum value $\min\{2, (a+1)/2\}\lambda^2$. If $\|\check{\Gamma}_i(A)\| \geq a\lambda$, then we have

$$\widetilde{s}_i(A) = \{\check{\Gamma}_i(A) - \check{\Gamma}_i(A)\}^2 + (a+1)\lambda^2/2 = (a+1)\lambda^2/2. \tag{23}$$

Let r be the number of nonzero rows of Γ_0, i.e. the cardinality of H_0. By (21), (22), (23), $\lambda \to 0$, $n^{-\min\{s,t\}}\lambda \to \infty$, and the consistency of $\bar{\Gamma}$, we have

$$\widetilde{s}(\bar{A}) = r(a+1)\lambda^2/2 + \textstyle\sum_{i\in J_0} \|\widehat{\beta}_{Y,i} - \widehat{\beta}_{T,i}\bar{A}\|^2 = r(a+1)\lambda^2/2 + o_P(\lambda^2). \tag{24}$$

For any $A \in \mathbb{R}^{d_T \times d_Y}$, let $H_\lambda(A) = \{i = 1, \ldots, p : \|\check{\Gamma}_i(A)\| > a\lambda\}$ be the index set of rows of $\check{\Gamma}(A)$ whose norms are greater than $a\lambda$, and let $N_\lambda(A)$ be its cardinality. Let $\mathcal{A}_\lambda = \{A \in \mathbb{R}^{d_T\times d_Y} : N_\lambda(A) > r\}$, we have, for any $A \in \mathcal{A}_\lambda$, $\widetilde{s}(A) \geq (r+1)(a+1)\lambda^2/2$. By (24), we have $P(\min_{A\in\mathcal{A}} \widetilde{s}(A) > \widetilde{s}(\bar{A})) \to 1$ as $\lambda \to 0$, which means that, without loss of generality, we can minimize $\widetilde{s}(A)$ within $\mathcal{A}_\lambda^c \equiv \{A \in \mathbb{R}^{d_T\times d_Y} : N_\lambda(A) \leq r\}$.

We next show that for all small λ, $\mathcal{A}_\lambda^c = \{A \in \mathbb{R}^{d_T\times d_Y} : H_\lambda(A) = H_0\}$, denoted by $\mathcal{B}_\lambda$. Let $G_\lambda = (\|\widehat{\beta}_T - \beta_T\| < C_1\lambda) \cap (\|\widehat{\beta}_Y - \beta_Y\| < C_1\lambda)$ for a positive constant C_1. Since $\widehat{\beta}_T = \beta_T + o_P(\lambda)$ and $\widehat{\beta}_Y = \beta_Y + o_P(\lambda)$, we have $P(G_\lambda) \to 1$ as $\lambda \to 0$. For $i = 1, \ldots, p$, given G_λ, $\|\check{\Gamma}_i(A)\| < a\lambda$ implies $\|\beta_{Y,i} - \beta_{T,i}A\| < C_2\lambda$ for some constant $C_2 > 0$. Thus, given G_λ, $N_\lambda(A) \leq r$ implies $\#\{i = 1, \ldots p : \|\beta_{Y,i} - \beta_{T,i}A\| < C_2\lambda\} \geq q - r$. By Theorem 1, for all small λ, the only set that satisfies the latter is J_0. Thus, for all small λ, we have $\mathcal{B}_\lambda = \mathcal{A}_\lambda^c$ with probability tending to one.

Hence, without loss of generality, we can minimize $\widetilde{s}(A)$ within $\mathcal{B}_\lambda$. By (21), (22), and (23), for any $A \in \mathcal{B}_\lambda$, we have

$$\widetilde{s}(A) \geq r(a+1)\lambda^2/2 + \textstyle\sum_{i\in J_0}\{\|\check{\Gamma}_i(A)\|^2 I(\|\check{\Gamma}_i(A)\| < \lambda) + \lambda^2 I(\|\check{\Gamma}_i(A)\| \geq \lambda)\},$$

which is clearly minimized at $A = \bar{A}$. This completes the proof. □

Typically, $\mathcal{S}(\widehat{\beta}_T)$ and $\mathcal{S}(\widehat{\beta}_Y)$ are $n^{1/2}$-consistent. In this case, Theorem 2 justifies the $n^{1/2}$-consistency of $\widehat{\mathcal{S}}_{\text{PDRF}}$ in estimating $\mathcal{S}_{\text{PDRF}}$. The corresponding requirements for the tuning parameter λ in the SCAD penalty are $\lambda \to 0$ and $n^{1/2}\lambda \to \infty$, which conform to the results in Fan and Li (2001). Theorem 2 also applies when $\mathcal{S}_{T|X}$ and $\mathcal{S}_{E(Y|X)}$ are estimated by SDR methods with slower convergence rate, e.g. MAVE as mentioned in Section 2, in which case the range of appropriate λ needs to be adjusted accordingly.

In practice, the consistency of $\mathcal{S}(\widehat{\beta}_T)$ can be compromised and subsequently harm the consistency of $\widehat{\mathcal{S}}_{\text{PDRF}}$, if $\beta_T^\intercal X$ has a weak effect on T. Referring to the literature review in the Introduction, this complies with the common concern about the inconsistency of IV analysis in the presence of weak instruments. However, it should be less worrisome due to, first, the use of strengthened instruments $\beta_{T,L_0\cap J_0}^\intercal X_{L_0\cap J_0}$ rather than the individual instruments in $\beta_T^\intercal X$ (see the discussion above (11)), second, the nonparametric nature of SDR that allows a nonlinear effect of $\beta_T^\intercal X$ on T, and, third, the potential presence of invalid instruments in $\beta_T^\intercal X$ that intensifies this effect. A detailed discussion is deferred to Section 5 later. To assess the consistency of $\widehat{\mathcal{S}}_{\text{PDRF}}$ in practice, one can use the bootstrap method to approximate its variation, say measured by $E\{\|\Pi(\widehat{\mathcal{S}}_{\text{PDRF}}) - \Pi(\mathcal{S}_{\text{PDRF}})\|_2\}$ where $\Pi(\cdot)$ denotes the usual projection matrix of a linear space, details omitted.

When the research interest is extended to detect all the instruments, $\widehat{\mathcal{S}}_{\rm PDRF}$ can also serve the purpose in conjunction with a consistent sparse estimator of $\mathcal{S}_{T|X}$, the latter achievable by the existing sparse SDR methods, e.g. lasso SIR mentioned in Section 2. In details, recall that the set of instruments is indexed by $L_0 \cap J_0$, where L_0 is the set of nonzero rows of β_T and J_0 is the set of zero rows of γ_0; the consistent selection of $L_0 \cap J_0$ immediately follows that of J_0 by $\widehat{\mathcal{S}}_{\rm PDRF}$ and that of L_0 by the sparse estimation of $\mathcal{S}_{T|X}$.

To determine d, the dimension of $\mathcal{S}_{\rm PDRF}$, multiple existing methods mentioned in Section 2 can be used. For example, if the estimators of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{T|X}$ are asymptotically normal, which, by Delta method and the strong oracle property in Theorem 2, imply the asymptotic normality of the nonzero rows of $\widehat{\Gamma}$, then the ladle estimator (Luo and Li, 2016) is applicable. We also recommend using BIC (Zhu et al., 2006) and PAE (Luo and Li, 2021) if asymptotic normality is not guaranteed in the estimation of $\mathcal{S}_{E(Y|X)}$ or $\mathcal{S}_{T|X}$.

5. Regulation of the personalized dose-response function

Given the well-defined reduced predictor $\gamma_0^{\mathsf{T}} X$, it is now eligible to regulate the personalized dose-response function $g(t, \gamma_0^{\mathsf{T}} x)$ towards identifiability with the aid of additional assumptions. Recall that we have assumed the existence of instruments in Assumption 1, which means that $\gamma_0^{\mathsf{T}} X$ does not carry all the information in X about modeling T. Thus, it is fairly general to adopt

Assumption 2 *For any non-degenerate $f(T, \gamma_0^{\mathsf{T}} X)$ in $L_2(T, \gamma_0^{\mathsf{T}} X)$, $E\{f(T, \gamma_0^{\mathsf{T}} X)|X\}$ is also non-degenerate.*

That is, there is a one-to-one correspondence between each candidate $f(T, \gamma_0^{\mathsf{T}} X)$ and its conditional mean given X. Under this assumption, $g(T, \gamma_0^{\mathsf{T}} X)$ is clearly the unique function of $(T, \gamma_0^{\mathsf{T}} X)$ that satisfies (4) and (5). A similar assumption can be found in Newey and Powell (2003) (see their Proposition 2.1), which served the same purpose of identification. Because any $E\{f(T, \gamma_0^{\mathsf{T}} X)|X\}$ reduces to $E\{f(T, \gamma_0^{\mathsf{T}} X)|\beta_T^{\mathsf{T}} X, \gamma_0^{\mathsf{T}} X\}$ with the aid of SDR on $T|X$, (5) can be rewritten as

$$E\{Y - g(T, \gamma_0^{\mathsf{T}} X) \mid \beta_T^{\mathsf{T}} X, \gamma_0^{\mathsf{T}} X\} = 0, \tag{25}$$

and, under Assumption 2, $g(T, \gamma_0^{\mathsf{T}} X)$ can be estimated nonparametrically by solving this inverse problem without triggering the "curse of dimensionality".

Despite its theoretical generality, however, the effectiveness of Assumption 2 in practice hinges on how much $\beta_T^{\mathsf{T}} X$, particularly the strengthened instrument $\beta_{T, L_0 \cap J_0}^{\mathsf{T}} X_{L_0 \cap J_0}$, is associated with T. If this effect is weak, then there will exist some non-degenerate $f(T, \gamma_0^{\mathsf{T}} X)$ such that $E\{f(T, \gamma_0^{\mathsf{T}} X)|X\}$ is practically negligible, which adds noise to (25) and delivers biased and unstable estimation of $g(T, \gamma_0^{\mathsf{T}} X)$. In this sense, Assumption 2 is a nonparametric analog, as well as a relaxation, of the common requirement on the instrument strength in the conventional linear IV analysis (Sheehan and Didelez, 2011; Burgess et al., 2017). The relaxation is three-fold. First, instead of the individual instruments, Assumption 2 is only associated with their optimal linear combination $\beta_{T, L_0 \cap J_0}^{\mathsf{T}} X_{L_0 \cap J_0}$. Second, with the aid of the invalid instruments, the effect of $\beta_T^{\mathsf{T}} X$ on T can still be strong if the effect of $\beta_{T, L_0 \cap J_0}^{\mathsf{T}} X_{L_0 \cap J_0}$ is weak. Third, as no parametric models are specified on $T|\beta_T^{\mathsf{T}} X$, $\beta_T^{\mathsf{T}} X$ is allowed to have a weak linear effect as long as it has an otherwise, e.g. symmetric, strong effect on T.

Based on these relaxations, we speculate that Assumption 2 regulates the instrument strength in the most general way; that is, no consistent estimation of $g(T, \gamma_0^\intercal X)$ will be feasible if it fails. As seen in Section 6 later, the effectiveness of Assumption 2 also plays a central role in the reliability of the estimation of $g(T, \gamma_0^\intercal X)$, which urges the necessity to develop a corresponding diagnosis procedure.

Recall from (12) that $\mathcal{S}_{E(Y|X)}$ is always a subspace of $\mathcal{S}(\mathcal{S}_{T|X}, \mathcal{S}_{\mathrm{PDRF}})$, which implies that $\beta_Y^\intercal X$ is a linear combination of $(\beta_T^\intercal X, \gamma_0^\intercal X)$, and that $E\{Y|\beta_T^\intercal X, \gamma_0^\intercal X\}$ in (25) is identical to $E\{Y|\beta_Y^\intercal X\}$. When $\beta_Y^\intercal X$ is lower dimensional than $(\beta_T^\intercal X, \gamma_0^\intercal X)$, (25) can be refined to

$$E\{Y - g(T, \gamma_0^\intercal X) \mid \beta_Y^\intercal X\} = 0, \tag{26}$$

which has a reduced dimensionality that benefits the corresponding estimation of $g(T, \gamma_0^\intercal X)$. However, the uniqueness of $g(T, \gamma_0^\intercal X)$ as the solution to (26) requires strengthening Assumption 2 to the one-to-one correspondence between each non-degenerate $f(T, \gamma_0^\intercal X)$ and $E\{f(T, \gamma_0^\intercal X)|\beta_Y^\intercal X\}$, which subtly restricts the effect of $\beta_T^\intercal X$ on T in addition to its overall strength and can be easily violated in practice. For example, as easily seen from (18), it fails in the simple case that both $T|X$ and $Y|(T, X)$ convey a homoscedastic linear regression model. Under this concern, we choose to estimate $g(T, \gamma_0^\intercal X)$ based on (25) rather than (26) for the widest applicability of the proposed IV analysis, although with the price of compromised estimation efficiency in certain cases.

6. Estimation of the personalized dose-response function

Let $\widehat{\gamma}$ be an arbitrary orthonormal basis matrix of $\widehat{\mathcal{S}}_{\mathrm{PDRF}}$ derived in Section 4. Using the reduced predictor $\widehat{\gamma}^\intercal X$, we now estimate $g(T, \gamma_0^\intercal X)$ based on (25) under Assumption 2. Due to the potential confounding between T and ϵ in (4), it is generally infeasible to develop a simple nonparametric estimator of $g(T, \gamma_0^\intercal X)$ that resembles the local polynomial regression, e.g. the NW estimator, with well-developed asymptotic properties. As mentioned before, one choice is to construct a proxy for the unmeasured confounders based on a homoscedastic linear model on $T|X$ (Li and Guo, 2020), which however introduces additional modeling risk. Another choice is to impose a parametric model for $g(T, \gamma_0^\intercal X)$ whose flexibility grows with the sample size, under which a generalization of the conventional two-stage least squares estimator can be developed; see, for example, Newey (1990) and Newey and Powell (2003). However, the inevitable use of multiple regulation terms (Newey and Powell, 2003) in such estimation would complicate both the theoretical development and the implementation. For these reasons, we will estimate $g(T, \gamma_0^\intercal X)$ following the spirit of the reproducing kernel Hilbert space (RKHS)-based methods.

The RKHS-based methods have a wide application in nonparametric statistics and machine learning research. The essence of these methods is to first approximate the functional parameter by an element in an appropriate functional linear space, commonly known as the kernel trick, and then estimate this element using essentially the (functional) least squares method. We refer to Fukumizu's seminal work (Fukumizu et al., 2007, 2008; Sriperumbudur et al., 2010) for a detailed review of the relative literature.

Let $K : \mathbb{R} \mapsto \mathbb{R}^+$ be a kernel density function satisfying Condition (C.1) in Appendix B. For any random element R and its sample copies $(R^{[1]}, \ldots, R^{[n]})$, let $K_h(R) = K(\|R/h\|_2)$

with bandwidth h and

$$K_{h,i}(R) = K_h(R - R^{[i]}) / \sum_{k=1}^n K_h(R - R^{[k]}), \quad i = 1, \ldots, n.$$

Let $\mathcal{K}_h$ be the linear space spanned by $K_{h,1}(T, \widehat{\gamma}^\intercal X), \ldots, K_{h,n}(T, \widehat{\gamma}^\intercal X)$, and let $\mathcal{H}_b$ be that spanned by $K_{b,1}(\widehat{\beta}_T^\intercal X, \widehat{\gamma}^\intercal X), \ldots, K_{b,n}(\widehat{\beta}_T^\intercal X, \widehat{\gamma}^\intercal X)$ for some other bandwidth b, where $\widehat{\beta}_T$ spans a consistent estimator of $\mathcal{S}_{T|X}$ as mentioned in Section 4 above. When R is square integrable, we approximate $E(R \mid T, \gamma_0^\intercal X)$, i.e. the projection of R onto $L_2(T, \gamma_0^\intercal X)$, by the projection of R onto $\mathcal{K}_h$. Similarly, we approximate $E(R \mid \beta_Y^\intercal X)$ by the projection of R onto $\mathcal{H}_b$. Because $g(T, \gamma_0^\intercal X)$ falls in $L_2(T, \gamma_0^\intercal X)$ and satisfies (5), it can be naturally characterized as, first, it must approximate to its projection onto $\mathcal{K}_h$; second, by (25), its projection onto $\mathcal{H}_b$ must approximate to the projection of Y onto $\mathcal{H}_b$.

Generally, the projection of any random element onto an RKHS is derived by applying Pythagorean theorem with the aid of a ridge regularity term. To ease the calculation, here we approximate the projection of any R onto $\mathcal{K}_h$ by the simple Nadaraya-Watson (NW) estimator $\sum_{i=1}^n K_{h,i}(T, \widehat{\gamma}^\intercal X) R^{[i]}$, and likewise approximate the projection of R onto $\mathcal{H}_b$ by $\sum_{i=1}^n K_{b,i}(\widehat{\beta}_T^\intercal X, \widehat{\gamma}^\intercal X) R^{[i]}$. The consistency of these approximations, which is based on the well-developed asymptotic properties of the NW estimator (Fan and Gijbels, 2018), is justified in Theorem 3 later. Let $\mathbb{K}_h$ and $\mathbb{H}_b$ be matrices in $\mathbb{R}^{n\times n}$ whose (i,j)th entries are $K_{h,j}(T^{[i]}, \widehat{\gamma}^\intercal X^{[i]})$ and $K_{b,j}(\widehat{\beta}_T^\intercal X^{[i]}, \widehat{\gamma}^\intercal X^{[i]})$, respectively, and let $\mathbb{Y} = (Y^{[1]}, \ldots, Y^{[n]})^\intercal$. Using the characterization of $g(T, \gamma_0^\intercal X)$ in the previous paragraph, we estimate the sample copies of $g(T, \gamma_0^\intercal X)$, denoted as $V_n = \{g(T^{[1]}, \gamma_0^\intercal X^{[1]}), \ldots, g(T^{[n]}, \gamma_0^\intercal X^{[n]})\}^\intercal$, by minimizing

$$\widehat{\Psi}(v) = \|\mathbb{H}_b v - \mathbb{H}_b \mathbb{Y}\|_2^2 + \tau \|\mathbb{K}_h v - v\|_2^2 \tag{27}$$

over $v \in \mathbb{R}^n$. Here, τ is a prefixed positive constant that balances the two losses. Depending on the nature of data, other loss functions can be used for the two terms of $\widehat{\Psi}(\cdot)$. For example, if $(T, \gamma_0^\intercal X)$ and $(\beta_T^\intercal X, \gamma_0^\intercal X)$ have a tight support so that the boundary effect is not worrisome in the nonparametric estimation, then the L_∞ loss defined as $\|(v_1, \ldots, v_n)^\intercal\|_\infty = \max_{i=1,\ldots,n} |v_i|$ can be used; by contrast, if these random elements have heavily tailed distributions, then the L_1 loss should be considered under the concern of robustness. The loss functions for the two terms of $\widehat{\Psi}(\cdot)$ can also differ from each other.

Benefited from the use of L_2 losses, the minimizer of $\widehat{\Psi}(\cdot)$ has an analytic form; that is, subject to that $(I - \mathbb{K}_h, \mathbb{H}_b)$ has full row rank, which is implied by Assumption 2 as justified later, $\widehat{\Psi}(\cdot)$ has the unique minimizer

$$\widehat{V}_n = \mathbb{W}_\tau^{-1} \mathbb{H}_b^\intercal \mathbb{H}_b \mathbb{Y}, \tag{28}$$

where $\mathbb{W}_\tau = \tau(I - \mathbb{K}_h)^\intercal (I - \mathbb{K}_h) + \mathbb{H}_b^\intercal \mathbb{H}_b$. The consistency of $\widehat{V}_n$ is readily implied by the construction of $\widehat{\Psi}(\cdot)$. It can also be intuitively explained as, first, since $\widehat{\Psi}(\cdot)$ is always nonnegative and $\widehat{\Psi}(V_n)$ is negligible, $\widehat{\Psi}(\widehat{V}_n)$ must also be negligible by definition and thus is close to $\widehat{\Psi}(V_n)$; then, as long as all the eigenvalues of the Hessian matrix of $\widehat{\Psi}(\cdot)$, i.e. $\mathbb{W}_\tau$, are non-negligible, $\widehat{V}_n$ must be close to V_n.

Given $\widehat{V}_n$, we follow the formulations above to estimate $g(T, \gamma_0^\intercal X)$ by a smoothing procedure; that is, for any $(t, x) \in \Omega_{T,X}$, we use the NW estimator

$$\widehat{g}(t, \gamma_0^\intercal x) = \sum_{i=1}^n K_{h,1}(t, \widehat{\gamma}^\intercal x) \widehat{V}_n^{[i]}, \tag{29}$$

where $\widehat{V}_n^{[i]}$ denotes the ith component of $\widehat{V}_n$. By its nature, we call this estimator the semiparametric personalized instrumental variable estimator (SPIVE). Its consistency, which naturally follows the arguments above, is formulated in the following theorem. The proof is deferred to Appendix A. Let $\lambda_{\min}(\mathbb{W}_\tau)$ be the smallest eigenvalue of $\mathbb{W}_\tau$. Because $\mathbb{W}_\tau$ is always positive semi-definite, the desired non-negligibility of $\mathbb{W}_\tau$ above can be expressed as $1/\lambda_{\min}(\mathbb{W}_\tau) = O_P(1)$.

Theorem 3 *Under the assumptions in Theorem 2, Assumption 2, and the regularity conditions (C.1-C.4) in Appendix B, we have* $1/\lambda_{\min}(\mathbb{W}_\tau) = O_P(1)$, *and*

$$n^{-1/2}\|\widehat{V}_n - V_n\|_2 = O_P\{r_n(h, d+1) + r_n(b, d+d_T) + n^{-\min\{s,v\}}\}.$$

where $r_n(a,c) = a^2 + n^{-1/2}a^{-c/2}$ *for any scalars* $a, c > 0$. *In addition, for an independently generated copy of* (X,T), *denoted by* $(\widetilde{X}, \widetilde{T})$, *we have*

$$\widehat{g}(\widetilde{T}, \gamma_0^\top \widetilde{X}) - g(\widetilde{T}, \gamma_0^\top \widetilde{X}) = O_P\{r_n(h, d+1) + r_n(b, d+d_T) + n^{-\min\{s,v\}}\}. \tag{30}$$

Referring to the results in Theorem 2, the term $n^{-\min\{s,v\}}$ in Theorem 3 represents the cost of estimating $\mathcal{S}_{\text{PDRF}}$ and $\mathcal{S}_{E(Y|X)}$ from the SDR stage. When the bandwidths h and b are proportional to $n^{-1/(d+5)}$ and $n^{-1/(d+d_T+4)}$, respectively, $\widehat{g}(T, \gamma_0^\top X)$ reaches its optimal convergence rate $n^{-\min\{2/(d+d_T+4),s,t\}}$, which is reasonably fast as long as both $\mathcal{S}_{\text{PDRF}}$ and $\mathcal{S}_{T|X}$ are low-dimensional and the SDR estimations are sharp enough. Here, we measure the estimation accuracy of $\widehat{V}_n$ by the popularly used mean squared error, which complies with the nature of $\widehat{\Psi}(\cdot)$ in (27) as a L_2 loss function and also with the literature of RKHS methods (Fukumizu et al., 2007; Sriperumbudur et al., 2010; Kim and Scott, 2012; Li and Song, 2017). For the same reason, the consistency of $\widehat{g}(T, \gamma_0^\top X)$ is formulated in a probabilistic sense for a new observation, rather than being pointwise. A pointwise consistent estimator can be derived if we instead use the aforementioned L_∞ loss in $\widehat{\Psi}(\cdot)$.

From an omitted simulation study, the performance of SPIVE is robust to the choice of τ, and can be optimized if the bandwidths h and b fall in appropriate ranges. Thus, we use $\tau = 1$ in practice for simplicity, and we recommend using a grid point search with five-fold cross validation to tune h and b. Due to the existence of unmeasured confounding, such cross validation must not use the conventional mean squared error to evaluate the goodness of fit in the testing set. Instead, for any estimate $\widehat{R}$ of $g(T, \gamma_0^\top X)$ derived from the training set, we recommend using

$$[1 - \text{dCor}\{\widehat{R}, (T, \widehat{\gamma}^\top X)\}] + E\{E^2(Y - \widehat{R} \mid \widehat{\beta}_T^\top X, \widehat{\gamma}^\top X)\}/\log(n) \tag{31}$$

in the testing set, where $\text{dCor}(\cdot,\cdot)$ denotes the distance correlation that measures the dependency between two random elements, and the conditional mean of the residual $Y - \widehat{R}$ is approximated by the NW estimator conducted solely based on the testing set. As the first term in (31) is minimized if and only if $\widehat{R}$ is a measurable function of $(T, \widehat{\gamma}^\top X)$ (Székely et al., 2007), and the second term is minimized if and only if the residual $Y - \widehat{R}$ has negligible mean conditional on $(\beta_T^\top X, \gamma_0^\top X)$, they together punish any deviation of $\widehat{R}$ from $g(T, \gamma_0^\top X)$ under Assumption 2. The weight $\log(n)$ is employed to address the issue that the second term tends to vary more dramatically than the first term in practice.

To polish the finite-sample performance of SPIVE, another trick is to use a smaller bandwidth ℓ than h in (29), which satisfies $n^{-1/2}\ell^{(1-d)/2} + \ell^2 \to 0$. The reason is that, with $\widehat{V}_n$ being the response instead of $\mathbb{Y}$ in the NW estimator (29), the error term ϵ in (4) that causes the majority of the variance of the estimator has been smoothed out. Thus, a smaller bandwidth ℓ can reduce the bias of the estimator, while bringing little additional variance. A detailed explanation is attached at the end of the proof of Theorem 3 in Appendix A. Following this logic, $g(T, \gamma_0^\intercal X)$ can also be estimated by the K-nearest neighbors method given $\widehat{V}_n$, details omitted.

Compared with the conventional RKHS-based methods, a clear advantage of SPIVE is that the matrix inversion in its implementation, i.e. $\mathbb{W}_\tau^{-1}$, can be calculated properly without involving additional regularity terms. The invertibility of $\mathbb{W}_\tau$ is also crucial to the reliability of SPIVE: if the smallest eigenvalues of $\mathbb{W}_\tau$ are practically negligible, then the largest eigenvalues of $\mathbb{W}_\tau^{-1}$ will be excessive and unstable, making $\widehat{V}_n$ biased and vary dramatically by minor data disturbance. From the proof of Theorem 3 (see Appendix A), this occurs exactly when Assumption 2 is nearly void, that is, if there exists non-degenerate $f(T, \gamma_0^\intercal X)$ with negligible mean conditional on X. Therefore, $\lambda_{\min}(\mathbb{W}_\tau)$ delineates the essential role of Assumption 2 to the consistency of SPIVE, and we use it to diagnose the effectiveness of this assumption: a smaller value being stronger opposing evidence.

Generally, it is difficult to characterize the null distribution of $\lambda_{\min}(\mathbb{W}_\tau)$, particularly as it will be elevated if one uses a larger bandwidth h or a smaller b in $\mathbb{W}_\tau$. Fortunately, our simulation experience shows that, for a wide range of choices of h and b, there is a clear gap between the supports of $\lambda_{\min}(\mathbb{W}_\tau)$ when Assumption 2 holds and when it fails. Thus, we use a rule-of-thumb

$$\lambda_{\min}(\mathbb{W}_\tau) > n^{-3/4} \tag{32}$$

to determine whether Assumption 2 holds to a reasonable extent. The usefulness of this rule is supported by both the simulation studies in Section 7 and a complementary simulation result in Appendix D, which respectively suggest that (32) holds consistently when Assumption 2 is effective and that it fails consistently when Assumption 2 is ineffective. As mentioned in Section 5, the effectiveness of Assumption 2 relies on the strength of the effect of $\beta_T^\intercal X$ on T, so it can also be inferred by evaluating the standard error of $\widehat{\mathcal{S}}_{\mathrm{PDRF}}$ or more directly the distance correlation between T and $\beta_T^\intercal X$, etc. These approaches are omitted here as they are less explicitly related to the consistency of SPIVE than $\lambda_{\min}(\mathbb{W}_\tau)$.

To estimate the marginal dose-response function, we modify SPIVE to the semiparametric marginal instrumental variable estimator (SMIVE)

$$\widehat{E}\{g(t, \gamma_0^\intercal X)\} = E_n\{\widehat{g}(t, \gamma_0^\intercal X)\}, \quad t \in \Omega(T).$$

The asymptotic consistency of this estimator is readily implied by Theorem 3 under the common support condition (7). To reduce the boundary effect in the estimation, we suggest transforming T and $\gamma_0^\intercal X$ in the presence of heavy tails, as well as truncating the support of $(T, \gamma_0^\intercal X)$ when the data cloud does not convey a (hyper-)rectangular shape, i.e. approaching $\Omega(T) \times \Omega(\gamma_0^\intercal X)$. If the truncation is conducted, the interpretation of the fitted marginal dose-response function must be adjusted accordingly.

7. Simulation Studies

We now use simulation models to illustrate the effectiveness of the proposed method in selecting the invalid instrumental variables and in estimating the personalized and the marginal dose-response functions. For reference, we also record the performance of sisVIVE proposed by Kang et al. (2016), on both variable selection and estimation of the dose-response functions.

We generate X from a standard multivariate normal distribution unless otherwise specified, and generate ϵ from $N(0,(1/3)^2)$. In Appendix D, we also consider another case where all the components of X are generated independently from the Bernoulli distribution with mean equal to 0.5, in order to evaluate the effectiveness of the proposed method for discrete X. The identical ϵ is used as the error term in generating both Y and T, so there exist unmeasured confounders in the observed data. Under these settings, we study the following six models. Let $\beta_{\mathrm{I}}, \beta_{\mathrm{II}} \in \mathbb{R}^p$ be $(0.3, 0.5, 0.7, 0, \ldots, 0)^{\top}$ and $(0.1, 0.2, 0.3, 0.4, 0.5, 0, \ldots, 0)^{\top}$, respectively.

Model 1: $T = \beta_{\mathrm{I}}^{\top} X + \epsilon$, $Y = T + X_1 + \epsilon$.

Model 2: $T = 3\sin(\beta_{\mathrm{I}}^{\top} X) + \epsilon$, $Y = 0.5 + T + 0.25(X_1+2)^2 + 3\epsilon$.

Model 3: $T = 3\sin(\beta_{\mathrm{II}}^{\top} X - 0.5) + \epsilon$, $Y = 2T(0.5X_1 + 0.5X_2 - 1) + 3\epsilon$.

Model 4: $T = \beta_{\mathrm{II}}^{\top} X + 2 + \epsilon$, $Y = 2\sin(0.5T) + |0.5X_1 + 0.5X_2 + 1| + \epsilon$.

Model 5: same as Model 4 but the components of X are generated independently from the uniform distribution on $(-2, 2)$.

Model 6: $T = |\beta_{\mathrm{I}}^{\top} X| + |0.5X_4 + 0.9X_5| + 0.6\epsilon$, $Y = T + 2|X_1| + \epsilon$.

Among all these models, the effect of X on T is linear in Models 1, 4, and 5, and is symmetric in Model 6; the effect of X on Y is linear in Model 1, and is symmetric in Model 6. The joint effect of T and X on Y is linear in Model 1, which best favors sisVIVE, and it is nonlinear but still additive in Models 2, 4, 5, and 6, and includes an interaction term in Model 3. Because X affects T through $\beta_{\mathrm{II}}^{\top} X$ in Models $3-5$ and through $\beta_{\mathrm{I}}^{\top} X$ in Models 1, 2, and 6, the forms of β_{I} and β_{II} indicate a generally weaker effect of the individual components of X on T in Models $3-5$. As Model 4 and Model 5 share the same conditional distribution $(Y,T) \mid X$, they together can examine the robustness of the proposed method to the distribution of X. In summary, these models provide a comprehensive overview of the various situations in practice.

With the sample size n fixed at 500, we first set $p = 10$, and generate 1000 independent copies of samples for each model. To implement the proposed method, we set $\tau = 1$ in (27), and, depending on whether a symmetric data pattern exists in the data, we use SIR (Li, 1991) or SAVE (Cook and Weisberg, 1991) to estimate $\mathcal{S}_{T|X}$ and $\mathcal{S}_{E(Y|X)}$. In real data analysis, such pattern can be detected or excluded in the stage of exploratory data analysis. For the reliability of the proposed method, we first use (32) to check Assumption 2, where the bandwidths of $\mathbb{W}_{\tau}$ are tuned by the cross-validation procedure mentioned in Section 6. Based on the 1000 runs, the estimated 1% and 5% quantiles of $\lambda_{\min}(\mathbb{W}_{\tau})$ for each model are recorded in the first part of Table 1. Because these values are much larger than the cutoff $n^{-3/4} \approx .0095$ in (32), they suggest the effectiveness of Assumption 2 for Models 1-6, which meets our theoretical anticipation.

With a safe adoption of Assumption 2, we now evaluate the performance of the proposed method in selecting invalid instruments, using sisVIVE as a reference. Such performance is

Table 1: The extreme sample quantiles of $\lambda_{\min}(\mathbb{W}_\tau)$ for each model, based on 1000 runs. Of a/b in each cell, a is the 1% sample quantile of $\lambda_{\min}(\mathbb{W}_\tau)$, and b is the 5% sample quantile of $\lambda_{\min}(\mathbb{W}_\tau)$. For the large p cases, p is set at 200 for Models 1-5 and set at 120 for Model 6.

p	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	.148/.243	.219/.294	.229/.406	.071/.100	.100/.129	.036/.051
large	.063/.110	.140/.218	.247/.304	.022/.057	.054/.087	.025/.043

Table 2: Performance of the methods in variable selection, based on 1000 runs. Of a/b in each cell, a stands for the average number of misspecified invalid instruments, and b stands for the average number of misspecified instruments. For the large p cases, p is set at 200 for Models 1-5 and set at 120 for Model 6.

p	Method	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	Proposed	.012/0	.002/0	0/.005	0/0	0/0	.107/0
	sisVIVE	.119/0	.082/0	.151/0	.139/0	.160/0	.024/.982
large	Proposed	0/0	0/0	0/.026	0/0	0/.001	.391/.002
	sisVIVE	23.3/0	7.06/0	.028/.013	.498/0	.120/0	0/1

measured by both the average number of misspecified invalid instruments and the average number of misspecified instruments. The results based on the 1000 runs are summarized in the first part of Table 2.

Clearly, compared with sisVIVE, the proposed method is less likely to misspecify the instruments to be invalid instruments in most models. In particular, this applies to Model 1 where the linear model assumption that sisVIVE adopts is exactly satisfied. This is plausibly due to the use of the lasso penalty in the objective function (3) of sisVIVE in contrast to the SCAD penalty in (19): the estimation bias caused by lasso triggers a tradeoff between optimal variable selection consistency and optimal model fitting, so a tuning parameter selection criterion based on the latter would inevitably cause more bias in variable selection. In general, both methods consistently truly specify the invalid instruments. The only exception is that sisVIVE almost always fails to do so in Model 6. This is because sisVIVE assumes a linear model on the regression of Y on (T, X), whereas X has a symmetric effect on Y in the presence of T in Model 6.

Next, we evaluate the proposed SPIVE in estimating the personalized dose-response function, again in comparison of sisVIVE. To measure the consistency of an estimate of the personalized dose-response function $\widetilde{g}(T, \gamma_0^\intercal X)$, we define

$$D\{\widetilde{g}(T, \gamma_0^\intercal X)\} = \frac{E_n^{1/2}\{\widetilde{g}(T, \gamma_0^\intercal X) - g(T, \gamma_0^\intercal X)\}^2}{E_n^{1/2}[E_n\{g(T, \gamma_0^\intercal X)\} - g(T, \gamma_0^\intercal X)]^2} \tag{33}$$

where the denominator, the sample standard deviation of $g(T, \gamma_0^\intercal X)$, serves as the normalizing constant. The sample mean and the sample standard deviation of this measure in each case are recorded in the first part of Table 3. Except for Model 1, SPIVE outperforms

Table 3: Performance of the methods in estimating the personalized dose-response function, based on 1000 runs. In each cell, $a(b)$ stands for the sample mean (sample standard deviation) of $D\{\widetilde{g}(T, \gamma_0^{\intercal} X)\}$ defined in (33). For the large p cases, p is set at 200 for Models 1-5 and set at 120 for Model 6.

p	Method	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	SPIVE	.097(.025)	.125(.010)	.145(.056)	.118(.027)	.125(.028)	.157(.011)
	sisVIVE	.051(.013)	.165(.014)	.544(.030)	.518(.036)	.540(.029)	.839(.077)
large	SPIVE	.104(.007)	.128(.010)	.184(.081)	.138(.040)	.143(.041)	.168(.048)
	sisVIVE	.198(.012)	.172(.018)	.560(.033)	.540(.040)	.560(.033)	.797(.023)

Table 4: Performance of the methods in estimating the marginal dose-response function, based on 1000 runs. The meaning of $a(b)$ in each cell resembles that in Table 3. For the large p cases, p is set at 200 for Models 1-5 and set at 120 for Model 6.

p	Method	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	SMIVE	.079(.081)	.107(.021)	.116(.024)	.239(.028)	.124(.024)	.119(.029)
	sisVIVE	.048(.026)	.052(.023)	.197(.040)	.660(.047)	.688(.039)	.420(.338)
large	SMIVE	.084(.021)	.095(.017)	.134(.029)	.238(.030)	.126(.026)	.129(.034)
	sisVIVE	.930(.047)	.142(.073)	.200(.038)	.660(.047)	.696(.042)	.201(.107)

sisVIVE in all the other models, which is no surprise as all these models are equipped with a nonlinear personalized dose-response function. Compared with Model 2, the advantage of SPIVE over sisVIVE is more substantial in Models $3-5$. Referring to the discussion about the model settings above, we speculate that this is because the weaker effect of the individual components of X on T harms the consistency of sisVIVE in Models $3-5$, whereas SPIVE is robust against this issue with the aid of SDR on $T|X$, as discussed in Section 5.

Using a similar measure to (33), we also record the performance of the proposed SMIVE and sisVIVE in estimating the marginal dose-response function; see the first part of Table 4. The results generally resemble those for the personalized dose-response function in Table 3, except that sisVIVE now outperforms SMIVE in Models 1 and 2, where T affects Y in a linear pattern marginally. Nonetheless, SMIVE is still consistent in these two models.

To simulate the high-dimensional cases, we next raise p to 200 for Models $1-5$, and raise it to 120 for Model 6. A smaller p is used in Model 6 because SAVE is used instead of SIR, which is more demanding on the sample size. The sample size n is still fixed at 500, and 1000 independent samples are again generated for each model. To check Assumption 2, the 1% and 5% sample quantiles of $\lambda_{\min}(\mathbb{W}_{\tau})$ are recorded in the second part of Table 1. Since they are well above the cutoff .0095 in (32), they comply with the theory to support the effectiveness of Assumption 2 for all the six models.

The second part of Table 2 records the performance of the proposed method and sisVIVE in variable selection. In connection with the case of $p = 10$, the proposed method is robust to the dimensionality of the data. By contrast, sisVIVE now misspecifies a much larger number of invalid instruments in Models 1 and 2. Referring to the discussion above about the performance of sisVIVE in Model 1 when $p = 10$, it is plausible that, as p grows, an exaggeration occurs to the tradeoff between the optimal variable selection consistency and the optimal model fitting caused by the lasso penalty.

The second parts of Table 3 and Table 4 record the performance of the methods in estimating the personalized and the marginal dose-response functions, respectively, in the high-dimensional cases. Generally, the proposed SPIVE and SMIVE deliver similar results to those for $p = 10$, indicating their robustness to the dimensionality to the data. A notable step-down occurs to the performance of sisVIVE in Model 1, making it suboptimal to the proposed in all the six models. Again, this illustrates the benefit of using the low-dimensional structure in $T|X$, which becomes more crucial in the high-dimensional cases.

8. Real Data Analysis

In this section, we analyze the data set of the ADNI-DoD study (Weiner et al., 2013), obtained from the Alzheimer's Disease Neuroimaging Initiative database (`https://adni.loni.usc.edu`). The original primary objective of the study is to investigate how the veterans' traumatic brain injury and post-traumatic stress disorder are associated with their symptoms of Alzheimer's disease while aging. Our interest is instead on the causal relationship between the geriatric depression and dementia among veterans. While both diseases commonly occur in this population, dementia is more often recorded in their late lives, and depression is frequently observed during the early and middle stages of depression (Muliyala and Varghese, 2010; Byers and Yaffe, 2011). In the literature, some researchers (Byers and Yaffe, 2011) believed that depression can occur much earlier in veterans' lives than dementia, and, as such, the former precedes and potentially causes the latter. We adopt the same assumption here, under which depression is assigned to be the treatment variable T and dementia is assigned to be the outcome variable Y. Because this assumption lacks scientific justification, which hypothetically rules out the possibility of a bi-directional temporal order of depression and dementia, it is the limitation of our analysis that one must be aware of.

To quantify the severity of depression, the study uses the total score of Geriatric Depression Scale (GDS) (Brink et al., 2008), a self-report scale designed to identify depression symptoms in the elderly. The scale consists of 15 questions with yes or no answers based on how they felt over the past week. Each question contributes one point to the GDS total score. A total score in the range $0-4$ is considered normal; $5-8$ is mild depression; $9-11$ is moderate depression; and $12-15$ is severe depression. In addition, the development of dementia is quantified by the Sum of Box score in Clinical Dementia Rating (CDR-SOB) (O'Bryant et al., 2008). The CDR-SOB is a numeric score ranging from 0 to 18. A higher CDR-SOB score indicates more severe symptoms of dementia. CDR-SOB score describes five degrees of impairment in performance on each of six categories of cognitive functioning, including memory, orientation, judgement and problem solving, community affairs, home and hobbies, and personal care. When a patient has multiple observations of these scores, we take their average. In the data set, the GDS total score ranges from 0 to 13, and the range of CDR-SOB is from 0 to 4.5.

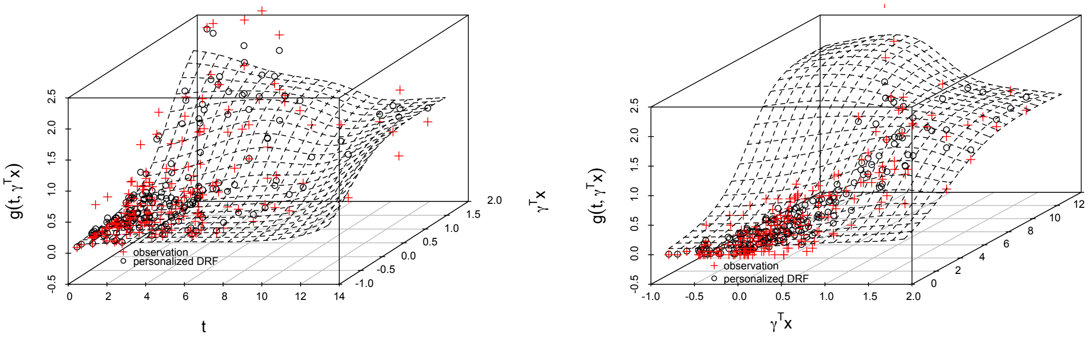

Figure 1: The fitted personalized dose-response function $\widehat{g}(T, \gamma_0^\intercal X)$ versus T and $\widehat{\gamma}^\intercal X$: "∘" marks the observations and "+" marks $\widehat{g}(T, \gamma_0^\intercal X)$ for the observed (T, X).

Using the Illumina HumanOmniExpress BeadChip, 713,014 target SNPs were genotyped from peripheral blood samples of 204 ADNI-DoD participants (Saykin et al., 2015). Due to the existence of possible unmeasured confounders in the study, such as the subject's underlying disease status (e.g., hypertension), we utilize the SNP data as the potential instruments. After applying an initial screening procedure, 1903 SNPs are retained that are possibly associated with either GDS or CDR-SOB score. We also include the subject's baseline characteristics, i.e., age, gender and educational level. By reformulating the categorical variables, i.e. the SNPs and the educational level, into binary indicators based on the dummy coding, the dimension of X is 3809 in total.

Based on the information above, we aim to estimate the personalized dose-response function that captures the personalized causal effect of depression on dementia. To address the high-dimensional nature of the data set, we assume sparsity of the instruments and use lasso SIR (Lin et al., 2019) to estimate both $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{T|X}$. By the ladle estimator, the resulting estimates of $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{T|X}$ are one-dimensional, and as is the proposed $\widehat{\mathcal{S}}_{\mathrm{PDRF}}$. After appropriate tuning of the bandwidths, $\lambda_{\min}(\mathbb{W}_\tau)$ is .85, which is well above the cutoff value $204^{-0.75} \approx .02$ in (32). Thus, it is sensible to adopt Assumption 2 and apply the proposed SPIVE to this data set.

Figure 1 illustrates the fitted personalized dose-response function by SPIVE, from two angles. Clearly, for the main cloud of (X, T), the fitted surface consists of monotone increasing curves along with the directions of both T and the strengthened invalid instruments $\widehat{\gamma}^\intercal X$, and the dynamic trend among these curves suggests the existence of an interaction term in the personalized dose-response function. To diagnose the fit, in the left panel of Figure 2 we plot the residual $Y - \widehat{g}(T, \gamma^\intercal X)$ against the univariate $\widehat{\beta}_Y^\intercal X$, which, by the definition of $\mathcal{S}_{E(Y|X)}$, assesses the relationship between ϵ in Model (4) and X. The fact that the residuals are randomly scattered around zero in this plot suggests that Assumption (5) that regulates the error term ϵ is satisfied.

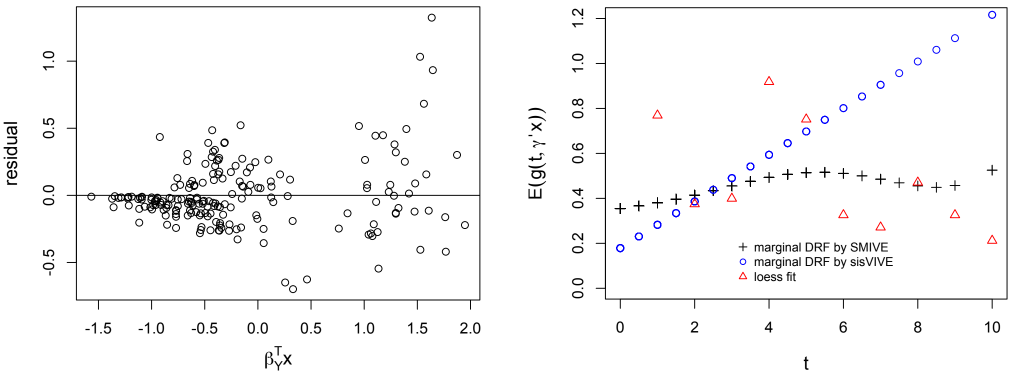

Figure 2: The left panel plots the residual versus $\widehat{\beta}_Y^{\mathrm{T}} X$; the right panel depicts the fits of the marginal dose-response function from SMIVE and sisVIVE, and the loess fit of $E(Y|T)$, marked by "+", "○", and "△", respectively.

The right panel of Figure 2 illustrates the fitted marginal dose-response function from the proposed SMIVE. To avoid extrapolation, it is plotted for the interval between the 2.5% and 97.5% quantiles of the severity of depression. The fitted curve deviates from the fitted loess curve of $E(Y|T)$, which however meets the theoretical anticipation and pinpoints the motivation of causal analysis: due to the existence of both the unmeasured confounders and the invalid instruments, a direct model fit of the observed (Y, T) cannot reveal their causal relationship. The fitted curve of SMIVE suggests that depression has an overall positive causal effect on dementia, especially when the depression level increases from zero to mild. From the discussion in the previous paragraph, this effect also varies with subject's characteristics. These conclusions are supported by independent evidence from other studies (Byers and Yaffe, 2011).

As pointed out by a Referee, the overall increasing pattern of the marginal dose-response function in Figure 2 suggests that the linear model (1) adopted in sisVIVE may also capture the causal effect of depression on dementia. To apply sisVIVE, which works only if $p < n$, we first use the proposed variable selection method to reduce $X \in \mathbb{R}^{3809}$ to 88 variables. The linear estimate of the marginal dose-response function from sisVIVE is depicted in the right panel of Figure 2 as well, with a positive slope equal to .104 that again suggests a positive causal effect of depression on dementia. However, the modification of this effect by subject's characteristics, as observed in Figure 1, cannot be captured by sisVIVE due to the intrinsic limitation of the linear model.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD

ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Luo's research is supported by the National Science Foundation of China (12131006, 12222117); Zhu's research is supported by the National Sciences and Engineering Research Council of Canada (RGPIN- 2017-04064); and Zhang's research is supported by Michael Smith Health Research BC Scholar (SCH-2022-2553) and NRC Digital Health and Geospatial Analytics Program (DHGA 121-1).

Appendix A. Complementary proofs

We first give a lemma that justifies the use of d in place of $\widehat{d}$ in the asymptotic studies of all the relative statistics that involve $\widehat{d}$, e.g. $\widehat{\mathcal{S}}_{\text{PDRF}}$ and $\widehat{g}(T, \gamma_0^{\intercal} X)$.

Lemma 1 *Suppose $\widehat{d}$ is an arbitrary consistent estimator of d. For any statistic R_n that involves $\widehat{d}$, let S_n be the statistic constructed in the same way as R_n but with $\widehat{d}$ replaced by d. Then we have, for any $r > 0$, $S_n - R_n = o_P(n^{-r})$; that is, S_n is always asymptotically equivalent to R_n.*

Proof Since the support of $\widehat{d}$ is discrete, i.e. $\{0, 1, 2, \ldots\}$, the consistency of $\widehat{d}$ means $P(\widehat{d} = d) \to 1$ as $n \to \infty$. Since $R_n = S_n$ when $\widehat{d} = d$, we have, for any $r > 0$ and $\delta > 0$,

$$\begin{aligned} P\{n^r(R_n - S_n) < \delta\} &= P\{n^r(R_n - S_n) < \delta | \widehat{d} = d\} P(\widehat{d} = d) \\ &\quad + P\{n^r(R_n - S_n) < \delta | \widehat{d} \neq d\} P(\widehat{d} \neq d) \\ &\geq P(0 < \delta | \widehat{d} = d) P(\widehat{d} = d) = P(\widehat{d} = d) \to 1, \end{aligned}$$

which means $R_n - S_n = o_P(n^{-r})$. This completes the proof. □

Next, We prove Theorem 3 from Section 6, before which we present two more lemmas that will be useful for the proof.

Lemma 2 *Let $\|\cdot\|_S$ denote the spectral norm of a matrix. Under Conditions (C.1), (C.3), and (C.4) in Appendix B, we have $\|\mathbb{H}_b\|_S = O_P(1)$.*

Proof Let $\mathbb{G}$ be the symmetric matrix in $\mathbb{R}^{n\times n}$ whose (i,j)th entry is $K_b((\widehat{\beta}_T, \widehat{\gamma})^\intercal (X^{[j]} - X^{[i]}))$. Let $\mathbb{D}$ be the diagonal matrix in $\mathbb{R}^{n\times n}$ whose ith diagonal entry is $\sum_{j=1}^n K_b((\widehat{\beta}_T, \widehat{\gamma})^\intercal (X^{[j]} - X^{[i]}))$. Here we omit the bandwidth in the subscript if no ambiguity is caused. Clearly, we have $\mathbb{H}_b = \mathbb{D}^{-1}\mathbb{G}$. For simplicity, we assume that appropriate regulations, such as those below (18) in the main text, have been made such that both β_T and γ_0 are unique, and $\|\widehat{\beta}_T - \beta_T\|_2 = O_P(n^{-s})$ and $\|\widehat{\gamma} - \gamma_0\|_2 = O_P(n^{-\min\{s,v\}})$.

We first show $\|\mathbb{D}\|_S = O_P(n)$ and $\|\mathbb{D}^{-1}\|_S = O_P(n^{-1})$. Under Conditions (C.1), (C.3), and (C.4), we have, uniformly for $i = 1, \ldots, n$ and $\{\beta \in \mathbb{R}^{p\times(d+d_T)} : \|\beta - (\beta_T, \gamma_0)\| < Cn^{-\min\{s,v\}}\}$,

$$\begin{aligned}
&\textstyle\sum_{j=1}^n K_b(\beta^\intercal X^{[j]} - \beta^\intercal X^{[i]}) = n\{f(\beta^\intercal X^{[i]}) + O_P(b^2 + n^{-1/2} b^{-(d+d_T)/2})\} \\
&= n\{f((\beta_T, \gamma_0)^\intercal X^{[i]}) + O_P(b^2 + n^{-1/2} b^{-(d+d_T)/2} + n^{-s}) = n\{f((\beta_T, \gamma_0)^\intercal X^{[i]}) + o_P(1)\}.
\end{aligned}$$

By Condition (C.1) and $\|\widehat{\beta}_T - \beta_T\|_2 = O_P(n^{-s})$ and $\|\widehat{\gamma} - \gamma_0\|_2 = O_P(n^{-\min\{s,v\}})$, this implies $\|\mathbb{D}\|_S = O_P(n)$ and $\|\mathbb{D}^{-1}\|_S = O_P(n^{-1})$.

Next, we show that $\mathbb{D} - \mathbb{G}$ is positive semi-definite. For simplicity, denote the (i,j)th entry of $\mathbb{G}$ by G_{ij} for $i, j = 1, \ldots, n$. Since $\mathbb{G}$ is symmetric, we have $G_{ij} = G_{ji}$. For any square matrices M_1 and M_2, we write $M_1 \geq M_2$ if $M_1 - M_2$ is positive semi-definite. For any $v = (v_1, \ldots, v_n)^\intercal$, we have

$$\begin{aligned}
v^\intercal(\mathbb{D} - \mathbb{G})v &= \textstyle\sum_{i=1}^n (\sum_{j=1}^n G_{ij}) v_i^2 - \sum_{j,k=1,\ldots,n} G_{jk} v_j v_k \\
&\geq \textstyle\sum_{i=1}^n (\sum_{j=1}^n G_{ij}) v_i^2 - \sum_{j,k=1,\ldots,n} G_{jk} (v_j^2 + v_k^2)/2 \\
&= \textstyle\sum_{i=1}^n (\sum_{j=1}^n G_{ij}) v_i^2 - \sum_{j=1}^n (\sum_{k=1}^n G_{jk}) v_j^2/2 - \sum_{k=1}^n (\sum_{j=1}^n G_{jk}) v_k^2/2 \\
&= \textstyle\sum_{i=1}^n (\sum_{j=1}^n G_{ij}) v_i^2 - \sum_{j=1}^n (\sum_{k=1}^n G_{jk}) v_j^2/2 - \sum_{j=1}^n (\sum_{k=1}^n G_{kj}) v_j^2/2 \\
&= \textstyle\sum_{i=1}^n (\sum_{j=1}^n G_{ij}) v_i^2 - \sum_{j=1}^n (\sum_{k=1}^n G_{jk}) v_j^2 \\
&= 0.
\end{aligned}$$

Thus $\mathbb{D} \geq \mathbb{G}$, which further implies $\mathbb{D}^{-1} \geq \mathbb{D}^{-1}\mathbb{G}\mathbb{D}^{-1}$. Let $\{w_n : n = 1, \ldots\}$ be a sequence of positive constants that diverge to infinity, and let $A_n = \{n^2 w_n I \geq \mathbb{D}^2\}$. Conditional on A_n, we have

$$\begin{aligned}
\|\mathbb{D}^{-1}\mathbb{G}\mathbb{G}\mathbb{D}^{-1}\|_S &= \|(\mathbb{D}^{-1}\mathbb{G}\mathbb{D}^{-1})\mathbb{D}^2(\mathbb{D}^{-1}\mathbb{G}\mathbb{D}^{-1})\|_S \\
&\leq \|(\mathbb{D}^{-1}\mathbb{G}\mathbb{D}^{-1})(n^2 w_n I)(\mathbb{D}^{-1}\mathbb{G}\mathbb{D}^{-1}\|_S = n^2 w_n \|\mathbb{D}^{-1}\mathbb{G}\mathbb{D}^{-1}\|_S^2 \\
&\leq n^2 w_n \|\mathbb{D}^{-1}\|_S^2 = O_P(w_n),
\end{aligned}$$

where the last equality is due to $\|\mathbb{D}^{-1}\|_S = O_P(n^{-1})$. Since $\|\mathbb{D}\|_S = O_P(n)$, we have $P(A_n) \to 1$. Hence, marginally we also have $\|\mathbb{D}^{-1}\mathbb{G}\mathbb{G}\mathbb{D}^{-1}\|_S = O_P(w_n)$, which indicates $\|\mathbb{H}_b\|_S = \|\mathbb{D}^{-1}\mathbb{G}\|_S = O_P(w_n^{1/2})$. By the arbitrariness of w_n, we have $\|\mathbb{H}_b\|_S = O_P(1)$. This completes the proof. □

Lemma 3 *Suppose the assumptions in Theorem 2, Assumption 2, and Conditions (C.1), (C.3), (C.4) hold. For any sequence $\{v_n \in \mathbb{R}^n : n = 1, \ldots\}$ such that $\|v_n\|_2 = 1$, if $\|\mathbb{H}_b\mathbb{K}_h v_n\|_2 = O_P(n^{-u})$ for some $u > 0$, then $\|\mathbb{K}_h v_n\|_2 = O_P(r_n(h, d+1) + r_n(b, d+d_T) + n^{-\min\{s,v\}} + n^{-u})$.*

Proof Without loss of generality, suppose $r_n(h, d+1) + r_n(b, d+d_T) + n^{-\min\{s,v\}} = O(n^{-r})$. For ease of presentation, we follow the proof of Lemma 2 to assume that appropriate column transformations have been made so that both β_T and γ_0 are unique, and $\|\widehat{\gamma} - \gamma_0\|_2 = O_P(n^{-\min\{s,v\}})$ and $\|\widehat{\beta}_T - \beta_T\|_2 = O_P(n^{-s})$. Let $B_w = \{\beta \in \mathbb{R}^{p\times(d+d_T)} : \|\beta - (\beta_T, \gamma_0)\|_2 \leq wn^{-\min\{s,v\}}\}$, $\Gamma_w = \{\gamma \in \mathbb{R}^{p\times d} : \|\gamma - \gamma_0\|_2 \leq wn^{-\min\{s,v\}}\}$, and, within the set of twice-differentiable functions, let $\mathcal{G}_w = \{g(T, \gamma^\intercal X) : \gamma \in \Gamma_w, E\{g(T, \gamma^\intercal X)\}^2 \leq w\}$. For each $g(T, \gamma^\intercal X) \in \mathcal{G}_w$, denote

$$V_{g,n} = (g(T^{[1]}, \gamma^\intercal X^{[1]}), \ldots, g(T^{[n]}, \gamma^\intercal X^{[n]}))^\intercal.$$

By simple algebra, we have, under Conditions (C.1), (C.3), and (C.4),

$$\begin{aligned}
&\textstyle\sum_{j=1}^n K_{b,i}(\beta^\intercal X^{[j]}) g(T^{[j]}, \gamma^\intercal X^{[j]}) / \sum_{j=1}^n K_{b,i}(\beta^\intercal X^{[j]}) \\
&= E\{g(T, \gamma^\intercal X) | \beta^\intercal X = \beta^\intercal X^{[i]}\} + O_P(b^2 + n^{-1/2} b^{-(d+d_T)/2}) \\
&= E\{g(T, \gamma_0^\intercal X) | (\beta_T, \gamma_0)^\intercal X = (\beta_T, \gamma_0)^\intercal X^{[i]}\} + O_P(b^2 + n^{-1/2} b^{-(d+d_T)/2} + n^{-\min\{s,v\}}) \\
&= E\{g(T, \gamma_0^\intercal X) | (\beta_T, \gamma_0)^\intercal X = (\beta_T, \gamma_0)^\intercal X^{[i]}\} + O_P(n^{-r})
\end{aligned}$$

uniformly for $i = 1, \ldots, n$, $\beta \in B_w$, and $g(T, \gamma^\intercal X) \in \mathcal{G}_w$, which indicates

$$n^{-1/2} \|\mathbb{H}_b V_{g,n}\|_2 = E^{1/2}[E^2\{g(T, \gamma_0^\intercal X) | (\beta_T, \gamma_0)^\intercal X\}] + O_P(n^{-r}) \tag{34}$$

uniformly for $\beta \in B_w$ and $g(T, \gamma^\intercal X) \in \mathcal{G}_w$. Following the same arguments as in the proof of Lemma 2, we have $\|K_h\|_S = O_P(1)$. Since $\|v_n\|_2 = 1$, this means $\|K_h v_n\|_2 = O_P(1)$, or equivalently

$$\lim_{w\to\infty} P\{n^{1/2} \textstyle\sum_{i=1}^n v_{n,i} K_{h,i}^N(T, \widehat{\gamma}^\intercal X) \in \mathcal{G}_w\} = 1,$$

where $K_{h,i}^N(T, \widehat{\gamma}^\intercal X)$ denotes $K_{h,i}(T, \widehat{\gamma}^\intercal X) / \sum_{k=1}^n K_{h,k}(T, \widehat{\gamma}^\intercal X)$, the superscript N for "normalized". Hence, with $w \to \infty$, (34) implies

$$\|\mathbb{H}_b\mathbb{K}_h v_n\|_2 = E^{1/2}[E^2\{n^{1/2} \textstyle\sum_{i=1}^n v_{n,i} K_{h,i}^N(T, \gamma_0^\intercal X) | (\beta_T, \gamma_0)^\intercal X\}] + O_P(n^{-r}). \tag{35}$$

Since $\|\mathbb{H}_b\mathbb{K}_h v_n\|_2 = O_P(n^{-u})$, (35) implies

$$E^{1/2}[E^2\{n^{1/2} \textstyle\sum_{i=1}^n v_{n,i} K_{h,i}^N(T, \gamma_0^\intercal X) | (\beta_T, \gamma_0)^\intercal X\}] = O(n^{-\min\{u,r\}}). \tag{36}$$

For any $w < \min\{u, r\}$, without loss of generality, suppose $n^{w+1/2} \sum_{i=1}^n v_{n,i} K_{h,i}^N(t, \gamma_0^\intercal x)$ converges to $\phi(t, \gamma_0^\intercal x)$ on $\Omega(T, \gamma_0^\intercal X)$ and is uniformly bounded by $\pm\Phi(t, \gamma_0^\intercal x)$ which satisfies $E\{|\Phi(T, \gamma_0^\intercal X)|\} < \infty$, both in probability. Then (36) implies $E\{\phi(T, \gamma_0^\intercal X) | (\beta_T, \gamma_0)^\intercal X\} = 0$ almost surely, which, by Assumption 2, indicates $\phi(T, \gamma_0^\intercal X) = 0$ almost surely. By letting $w \to \min\{u, r\}$, we have

$$E^{1/2}\{n^{1/2} \textstyle\sum_{i=1}^n v_{n,i} K_{h,i}^N(T, \gamma_0^\intercal X)\}^2 = O(n^{-\min\{u,r\}}). \tag{37}$$

Following the same arguments as in (34), we then have

$$\begin{aligned}\|\mathbb{K}_h v_n\|_2 &= n^{1/2}E^{1/2}\{\textstyle\sum_{i=1}^n v_{n,i}K_{h,i}^N(T,\gamma_0^\intercal X)\}^2 + O_P(h^2 + n^{-1/2}h^{-(d+1)/2} + n^{-\min\{s,v\}})\\ &= O_P(n^{-r} + n^{-u}),\end{aligned}$$

where the last equality is derived from (37). This completes the proof. □

Theorem 3 Under the assumptions in Theorem 2, Assumption 2, and the regularity conditions (C.1-C.4) in Appendix B, we have $1/\lambda_{\min}(\mathbb{W}_\tau) = O_P(1)$, and

$$n^{-1/2}\|\widehat{V}_n - V_n\|_2 = O_P\{r_n(h,d+1) + r_n(b,d+d_T) + n^{-\min\{s,v\}}\}.$$

where $r_n(h,d+1) = h^2 + n^{-1/2}h^{-(d+1)/2}$ and $r_n(b,d+d_T) = b^2 + n^{-1/2}b^{-(d+d_T)/2}$. In addition, for an independently generated copy of (X,T) denoted by $(\widetilde{X},\widetilde{T})$, we have

$$\widehat{g}(\widetilde{T},\gamma_0^\intercal\widetilde{X}) - g(\widetilde{T},\gamma_0^\intercal\widetilde{X}) = O_P\{r_n(h,d+1) + r_n(b,d+d_T) + n^{-\min\{s,v\}}\}.$$

Proof We first prove the consistency of $\widehat{V}_n$. For ease of presentation, denote $r_n(h,d+1) + r_n(b,d+d_T) + n^{-\min\{s,v\}}$ by n^{-r} for some $r > 0$. Suppose the following statements hold:

(a) $n^{-1/2}\|\{\tau(I-\mathbb{K}_h)^\intercal(I-\mathbb{K}_h) + \mathbb{H}_b^\intercal\mathbb{H}_b\}V_n - \mathbb{H}_b^\intercal\mathbb{H}_b\mathbb{Y}\|_2 = O_P(n^{-r})$,

(b) $\min\{v^\intercal\{\tau(I-\mathbb{K}_h)^\intercal(I-\mathbb{K}_h) + \mathbb{H}_b^\intercal\mathbb{H}_b\}v : v \in \mathbb{R}^n, v^\intercal v = 1\} = O_P^+(1)$,

where $O_P^+(1)$ denotes a sequence of random variables that are bounded below from zero (Luo and Li, 2016). Then (b) is equivalent to the statement $1/\lambda_{\min}(\mathbb{W}_\tau) = O_P(1)$ in this theorem, and it immediately implies that $(I-\mathbb{K}_h,\mathbb{H}_b)$ has full row-rank with probability tending to one, and that

$$\|\{\tau(I-\mathbb{K}_h)^\intercal(I-\mathbb{K}_h) + \mathbb{H}_b^\intercal\mathbb{H}_b\}^{-1}\|_S = O_P(1). \tag{38}$$

By (a) and (25) in the main text, we have

$$n^{-1/2}\|\{\tau(I-\mathbb{K}_h)^\intercal(I-\mathbb{K}_h) + \mathbb{H}_b^\intercal\mathbb{H}_b\}(\widehat{V}_n - V_n)\|_2 = O_P(n^{-r}).$$

Together with (38), we have

$$n^{-1/2}\|\widehat{V}_n - V_n\|_2 \le \|\{\tau(I-\mathbb{K}_h)^\intercal(I-\mathbb{K}_h) + \mathbb{H}_b^\intercal\mathbb{H}_b\}^{-1}\|_S O_P(n^{-r}) = O_P(n^{-r}).$$

Hence, it suffices to show (a) and (b).

To show (a), we will show the following stronger statements:

(a.1) $n^{-1/2}\|(I-\mathbb{K}_h)^\intercal(I-\mathbb{K}_h)V_n\|_2 = O_P(n^{-r})$,

(a.2) $n^{-1/2}\|\mathbb{H}_b^\intercal\mathbb{H}_b(\mathbb{Y} - V_n)\|_2 = O_P(n^{-r})$.

For each $w > 0$, let Γ_w be the same as defined in the proof of Lemma 3. For each $\gamma \in \Gamma_w$, let $\widetilde{\mathbb{G}}_\gamma$ be the n-dimensional matrix with (i,j)th entry being $K_h(T^{[j]} - T^{[i]}, \gamma^\intercal X^{[j]} - \gamma^\intercal X^{[i]})$, and let $\widetilde{\mathbb{D}}_\gamma$ be the n-dimensional diagonal matrix with ith diagonal entry being $\sum_{j=1}^n K_h(T^{[j]} - T^{[i]}, (\gamma^\intercal X^{[j]} - \gamma^\intercal X^{[i]}))$. Let $V_{\gamma,n} = (g(T^{[1]},\gamma^\intercal X^{[1]}),\ldots,g(T^{[n]},\gamma^\intercal X^{[n]}))^\intercal$. Similarly to the proof of Lemma 2, we have $\|\widetilde{D}_\gamma\|_S = O_P(n)$ and $\|\widetilde{D}_\gamma^{-1}\|_S = O_P(n^{-1})$ uniformly on Γ_w, and $\widetilde{\mathbb{D}}_\gamma - \widetilde{\mathbb{G}}_\gamma \geq 0$. Under Conditions (C.1-C.4), we have

$$\begin{aligned} n^{-1}\{\textstyle\sum_{j=1}^n K_{h,i}(T^{[j]}, \gamma^\intercal X^{[j]})\}g(T^{[i]}, \gamma^\intercal X^{[i]}) - \textstyle\sum_{j=1}^n \{K_{h,i}(T^{[j]}, \gamma^\intercal X^{[j]}) g(T^{[j]}, \gamma^\intercal X^{[j]})\} \\ = O_P(h^2 + n^{-1/2}h^{-(d+1)/2}) \end{aligned}$$

uniformly for $i = 1, \ldots, n$ and $\gamma \in \Gamma_w$, which means

$$n^{-3/2}\|(\widetilde{\mathbb{D}}_\gamma - \widetilde{\mathbb{G}}_\gamma)V_{\gamma,n}\|_2 = O_P(h^2 + n^{-1/2}h^{-(d+1)/2}). \tag{39}$$

In addition, since $\|\widehat{\gamma} - \gamma_0\| = O_P(n^{-\min\{s,v\}})$, and $\|\widetilde{\mathbb{D}}_\gamma\|_S = O_P(n)$ uniformly on Γ_w, we have, under Conditions (C.1) and (C.2),

$$n^{-3/2}\|(\widetilde{\mathbb{D}}_\gamma - \widetilde{\mathbb{G}}_\gamma)(V_{\gamma,n} - V_n)\|_2 = O_P(n^{-\min\{s,v\}}) \tag{40}$$

uniformly on Γ_w. (39) and (40) together imply

$$n^{-3/2}\|(\widetilde{\mathbb{D}}_{\widehat{\gamma}} - \widetilde{\mathbb{G}}_{\widehat{\gamma}})V_n\|_2 = O_P(h^2 + n^{-1/2}h^{-(d+1)/2} + n^{-\min\{s,v\}}) = O_P(n^{-r}),$$

which further indicates

$$\begin{aligned} \|(I - \mathbb{K}_h)^\intercal(I - \mathbb{K}_h)V_n\|_2 &= \|(\widetilde{\mathbb{D}}_{\widehat{\gamma}} - \widetilde{\mathbb{G}}_{\widehat{\gamma}})\widetilde{\mathbb{D}}_{\widehat{\gamma}}^{-2}(\widetilde{\mathbb{D}}_{\widehat{\gamma}} - \widetilde{\mathbb{G}}_{\widehat{\gamma}})V_n\|_2 \\ &\leq \|\widetilde{\mathbb{D}}_{\widehat{\gamma}} - \widetilde{\mathbb{G}}_{\widehat{\gamma}}\|_S\|\widetilde{\mathbb{D}}_{\widehat{\gamma}}^{-1}\|_S^2\|(\widetilde{\mathbb{D}} - \widetilde{\mathbb{G}})V_n\|_2 \\ &\leq 2\|\widetilde{\mathbb{D}}_{\widehat{\gamma}}\|_S\|\widetilde{\mathbb{D}}_{\widehat{\gamma}}^{-1}\|_S^2\|(\widetilde{\mathbb{D}} - \widetilde{\mathbb{G}})V_n\|_2 \\ &\leq n^{-1}n^{3/2}O_P(n^{-r}) = n^{-1/2}O_P(n^{-r}). \end{aligned}$$

Thus (a.1) holds. Since $\mathbb{Y} - V_n = (\epsilon^{[1]}, \ldots, \epsilon^{[n]})^\intercal$, by similar arguments to (34) in the proof of Lemma 3, we have $n^{-1/2}\|\mathbb{H}_b(\mathbb{Y} - V_n)\|_2 = O_P(b^2 + n^{-1/2}b^{-(d+d_T)/2} + n^{-\min\{s,v\}})$. By Lemma 2, we have $\|\mathbb{H}_b\|_S = O_P(1)$. These together imply

$$\begin{aligned} n^{-1/2}\|\mathbb{H}_b^\intercal\mathbb{H}_b(\mathbb{Y} - V_n)\|_2 &\leq \|\mathbb{H}_b^\intercal\|_S\{n^{-1/2}\|\mathbb{H}_b(\mathbb{Y} - V_n)\|_2\} \\ &= O_P(b^2 + n^{-1/2}b^{-(d+d_T)/2} + n^{-\min\{s,v\}}) = O_P(n^{-r}). \end{aligned}$$

Hence (a.2) holds and consequently (a) holds.

To prove (b), assume there exists $\{v_n :\in \mathbb{R}^n\}$ that satisfies $\|v_n\|_2 \equiv 1$ and $\|(I - \mathbb{K}_h, \mathbb{H}_b)v\|_2 = O_P(n^{-\delta})$ for some $\delta \in (0, r)$. Then we have $\|v - \mathbb{K}_h v_n\|_2 = O_P(n^{-\delta})$ and $\|\mathbb{H}_b v_n\|_2 = O_P(n^{-\delta})$. By Lemma 2, we have $\|\mathbb{H}_b\|_S = O_P(1)$. These together imply

$$\begin{aligned} \|\mathbb{H}_b\mathbb{K}_h v_n\|_2 &\leq \|\mathbb{H}_b v_n\|_2 + \|\mathbb{H}_b(v_n - \mathbb{K}_h v_n)\|_2 \leq \|\mathbb{H}_b v_n\|_2 + \|\mathbb{H}_b\|_S\|v_n - \mathbb{K}_h v_n\|_2 \\ &= O_P(n^{-\delta}), \end{aligned}$$

which, by Lemma 3, means $\|\mathbb{K}_h v_n\|_2 = O_P(n^{-r} + n^{-\delta}) = O_P(n^{-\delta})$. Thus, we have

$$\|v_n\|_2 \leq \|\mathbb{K}_h v_n\|_2 + \|v_n - \mathbb{K}_h v_n\|_2 = O_P(n^{-\delta}),$$

which contradicts the setting $\|v_n\|_2 = 1$. Hence we have

$$\min\{\|(I - \mathbb{K}_h, \mathbb{H}_b)v\|_2 : v \in \mathbb{R}^n, v^\intercal v = 1\} = O_P^+(1),$$

where, again, the concept of $O_P^+(1)$ can be seen in Luo and Li (2016).

We next prove the convergence of $\widehat{g}(\widetilde{T}, \gamma_0^\top \widetilde{X})$. For the most generality, we use ℓ as the bandwidth in (29), which, as discussed below Theorem 3, can differ from h for the optimal sample performance. Again, we assume that appropriate regulations have been made so that γ_0 is unique and, by Theorem 2, $\|\widehat{\gamma} - \gamma_0\| = O_P(n^{-\min\{s,v\}})$. In addition, for $i = 1, \ldots, n$, let $K_{\ell,\mathrm{sum}}(\widetilde{T}, \gamma^\top \widetilde{X}) = (K_\ell(\widetilde{T} - T^{[1]}, \gamma^\top \widetilde{X} - \gamma^\top X^{[1]}), \ldots, K_\ell(\widetilde{T} - T^{[n]}, \gamma^\top X - \gamma^\top \widetilde{X}^{[n]}))$ for any $\gamma \in \mathbb{R}^{p\times d}$. We first decompose $\widehat{g}(\widetilde{T}, \gamma_0^\top \widetilde{X})$ into two parts,

$$\begin{aligned}
\widehat{g}(\widetilde{T}, \gamma_0^\top \widetilde{X}) - g(\widetilde{T}, \gamma_0^\top \widetilde{X}) &= K_{\ell,\mathrm{sum}}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X})\widehat{V}_n / \{K_{\ell,\mathrm{sum}}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X})1_n\} - g(\widetilde{T}, \gamma_0^\top \widetilde{X}) \\
&= K_{\ell,\mathrm{sum}}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X})(\widehat{V}_n - V_n) / \{K_{\ell,\mathrm{sum}}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X})1_n\} \\
&\quad + [K_{\ell,\mathrm{sum}}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X})V_n / \{K_{\ell,\mathrm{sum}}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X})1_n\} - g(\widetilde{T}, \gamma_0^\top \widetilde{X})] \\
&\equiv \mathrm{I} + \mathbb{II},
\end{aligned}$$

where 1_n denotes the n-dimensional vector with all the elements being one. We next show $\mathrm{I} = O_P(n^{-r})$ and $\mathbb{II} = O_P(r(\ell, d-1) + n^{-r})$ where $r_n(\ell, d-1) = \ell^2 + n^{-1/2}\ell^{(1-d)/2}$.

For any $C > 0$, under Conditions (C.1), (C.2), and $r_n(\ell, d+1) = O_P(1)$, we have, uniformly for $\|\gamma - \gamma_0\| \le Cn^{-\min\{s,v\}}$, $(t, x) \in \Omega(T, X)$, and $i = 1, \ldots, n$,

$$\begin{aligned}
&E\{K_\ell(\widetilde{T} - T^{[i]}, \gamma^\top \widetilde{X} - \gamma^\top X^{[i]}) | T^{[i]}, X^{[i]}\} = f(T^{[i]}, \gamma^\top X^{[i]}) + O_P(\ell^2 + n^{-1/2}\ell^{-(d+1)/2}), \\
&n^{-1}K_{\ell,\mathrm{sum}}(t, \gamma^\top x)1_n = f(t, \gamma^\top x) + O_P(\ell^2 + n^{-1/2}\ell^{-(d+1)/2}).
\end{aligned} \tag{41}$$

Together with $\|\widehat{\gamma} - \gamma_0\| = O_P(n^{-\min\{s,v\}})$, this implies, uniformly for $i = 1, \ldots, n$,

$$E\{K_{\ell,i}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X}) | \mathbb{X}, \mathbb{T}, \mathbb{Y}\} = n^{-1}\{f(T^{[i]}, \widehat{\gamma}^\top X^{[i]}) / E\{f(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X})\} + O_P(r_n(\ell, d+1))\}.$$

Thus, by Condition (C.1) which regulates the lower and upper bound of $f(T, X)$, we have $\max_{i=1,\ldots,n} E\{K_\ell(\widetilde{T} - T^{[i]}, \widehat{\gamma}^\top \widetilde{X} - \widehat{\gamma}^\top X^{[i]}) | \mathbb{X}, \mathbb{T}, \mathbb{Y}\} = O_P(n^{-1})$, which means

$$\begin{aligned}
E(|\mathrm{I}|) &\le E\{\textstyle\sum_{i=1}^n E(K_\ell(\widetilde{T} - T^{[i]}, \widehat{\gamma}^\top \widetilde{X} - \widehat{\gamma}^\top X^{[i]}) | \mathbb{X}, \mathbb{T}, \mathbb{Y}) |\widehat{V}_{n,i} - V_{n,i}|\} \\
&= O\{n^{-1} \textstyle\sum_{i=1}^n E(|\widehat{V}_{n,i} - V_{n,i}|)\} \le O\{n^{-1/2} E(\|\widehat{V}_n - V_n\|_2)\} = O(n^{-r}),
\end{aligned}$$

where the last inequality is an application of the Cauchy-Schwarz inequality. By Markov's Inequality, this means $\mathrm{I} = O_P(n^{-r})$.

For $\mathbb{II}$, we first show that the impact of using $K(T, \widehat{\gamma}^\top X)$ instead of $K(T, \gamma_0^\top X)$ is asymptotically negligible. Under Conditions (C.1) and (C.3), we have, uniformly for $i = 1, \ldots, n$,

$$K_\ell(\widetilde{T} - T^{[i]}, \widehat{\gamma}^\top \widetilde{X} - \widehat{\gamma}^\top X^{[i]}) = K_\ell(\widetilde{T} - T^{[i]}, \gamma_0^\top \widetilde{X} - \gamma_0^\top X^{[i]}) + O_P(n^{-\min\{s,v\}}). \tag{42}$$

Since $\sum_{i=1}^n K_\ell(\widetilde{T} - T^{[i]}, \gamma_0^\top \widetilde{X} - \gamma_0^\top X^{[i]}) = n\{f(\widetilde{T}, \gamma_0^\top \widetilde{X}) + O_P(r_n(\ell, d+1))\}$, which, by (C.1) and (C.4), is $nf(\widetilde{T}, \gamma_0^\top \widetilde{X})\{1 + o_P(1)\}$ and further is $O_P^+(n)$, (42) implies

$$\begin{aligned}
\mathbb{II} &= \frac{K_{\ell,\mathrm{sum}}(\widetilde{T}, \gamma_0^\top \widetilde{X})\{V_n - g(\widetilde{T}, \gamma_0^\top \widetilde{X})1_n\}}{nf(\widetilde{T}, \gamma_0^\top \widetilde{X})\{1 + o_P(1)\}} + \frac{\{K_{\ell,\mathrm{sum}}(\widetilde{T}, \widehat{\gamma}^\top \widetilde{X}) - K_{\ell,\mathrm{sum}}(\widetilde{T}, \gamma_0^\top \widetilde{X})\}V_n}{nf(\widetilde{T}, \gamma_0^\top \widetilde{X})\{1 + o_P(1)\}} \\
&= \frac{K_{\ell,\mathrm{sum}}(\widetilde{T}, \gamma_0^\top \widetilde{X})\{V_n - g(\widetilde{T}, \gamma_0^\top \widetilde{X})1_n\}}{nf(\widetilde{T}, \gamma_0^\top \widetilde{X})\{1 + o_P(1)\}} + O_P(n^{-1}n^{1/2}n^{-\min\{s,v\}}\|V_n\|_2) \\
&= \frac{K_{\ell,\mathrm{sum}}(\widetilde{T}, \gamma_0^\top \widetilde{X})\{V_n - g(\widetilde{T}, \gamma_0^\top \widetilde{X})1_n\}}{nf(\widetilde{T}, \gamma_0^\top \widetilde{X})\{1 + o_P(1)\}} + O_P(n^{-\min\{s,v\}}).
\end{aligned} \tag{43}$$

In addition, let $\widetilde{R}$ denote $g(\widetilde{T}, \gamma_0^\intercal \widetilde{X})$ and $R^{[i]}$ denote $g(T^{[i]}, \gamma_0^\intercal X^{[i]})$, we have

$$\begin{aligned}
&E[[K_{\ell,\mathrm{sum}}(\widetilde{T}, \gamma_0^\intercal \widetilde{X})\{V_n - \widetilde{R}1_n\}]^2] \\
&\quad = \textstyle\sum_{i=1}^n E[K_\ell^2(T - T^{[i]}, \gamma_0^\intercal X - \gamma_0^\intercal X^{[i]})\{R^{[i]} - R\}^2] \\
&\qquad + \textstyle\sum_{i=1}^n \sum_{j\neq i} E[K_\ell(\widetilde{T} - T^{[i]}, \gamma_0^\intercal \widetilde{X} - \gamma_0^\intercal X^{[i]})\{R^{[i]} - \widetilde{R}\}K_\ell(\widetilde{T} - T^{[j]}, \gamma_0^\intercal \widetilde{X} - \gamma_0^\intercal X^{[j]})\{R^{[j]} - \widetilde{R}\}] \\
&\quad = O_P(n\ell^{-(d+1)}\ell^2 + n^2\ell^4),
\end{aligned}$$

where the last equality can be easily derived from the conventional nonparametric theory (Fan and Gijbels, 2018) and the mutual independence between $(\widetilde{T}, \gamma_0^\intercal \widetilde{X})$, $(T^{[i]}, \gamma_0^\intercal X^{[i]})$, and $(T^{[j]}, \gamma_0^\intercal X^{[j]})$ for any $i \neq j$. This means $K_{\ell,\mathrm{sum}}(T, \gamma_0^\intercal X)\{V_n - R1_n\} = O_P(n^{1/2}\ell^{(1-d)/2} + n\ell^2)$, which, together with (43) and Condition (C.1), implies

$$\mathrm{I\!I} = O_P\{n^{-1}(n^{1/2}\ell^{(1-d)/2} + n\ell^2)\}\{1 + o_P(1)\} + O_P(n^{-\min\{s,v\}}) = O_P(r_n(\ell, d-1) + n^{-r}).$$

Together with $\mathrm{I} = O_P(n^{-r})$, we have $\widehat{g}(T, \gamma_0^\intercal X) - g(T, \gamma_0^\intercal X) = O_P(r_n(\ell, d-1) + n^{-r})$. By simple algebra, an ℓ that is proportional to $n^{-1/(d+3)}$ will minimize $r_n(\ell, d-1)$ and meet the requirement $r_n(\ell, d+1) = O_P(1)$ above. This completes the proof. □

Appendix B. The Regularity Conditions

(C.1) (T, X) has is a compact support $\Omega(T, X)$, and its distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{p+1}$. The density function $f(T, X)$ is bounded away from infinity and zero, i.e. $P(a < f(T, X) < b) = 1$ for some $a, b > 0$, and $f(T, X)$ is twice differentiable almost surely on $\Omega(T, X)$.
(C.2) $g(\cdot, \cdot)$ is Lipschitz continuous and twice differentiable almost surely on $\{(t, \gamma^\intercal x) : (t, x) \in \Omega(T, X), \gamma \in \mathbb{R}^{p\times d}, \|\gamma - \gamma_0\|_2 \leq \epsilon\}$ for some $\epsilon > 0$. ϵ in (1) of the main text satisfies $E(\epsilon^4) < \infty$.
(C.3) The univariate density function $K(\cdot)$ satisfies $\int_{\mathbb{R}} K(x)dx = 1$, $\int_{\mathbb{R}} xK(x)dx = 0$, $\int_{\mathbb{R}} x^2K(x)dx < \infty$, and has a compact support.
(C.4) The bandwidths satisfy $b \to 0$, $bn^{1/d_Y} \to \infty$, $h \to 0$, and $hn^{1/(d+1)} \to \infty$.

Appendix C. Change of $\mathcal{S}_{\mathrm{PDRF}}$ with Removal of Invalid Instruments

An interesting question raised by a Referee is that how $\mathcal{S}_{\mathrm{PDRF}}$ under the proposed regulation will change if an invalid instrument is removed from X, assuming for simplicity that the components of X are mutually independent. In general, this can dramatically change $\mathcal{S}_{\mathrm{PDRF}}$, including both its dimension and the corresponding J_0.

To see this, note that an invalid instrument must induce either a nonzero row of β_Y that spans $\mathcal{S}_{E(Y|X)}$ or a nonzero row of β_T that spans $\mathcal{S}_{T|X}$ under the proposed regulation of $\mathcal{S}_{\mathrm{PDRF}}$: otherwise, (16) in the main text would imply that the corresponding row of γ_0 is zero, which contradicts the definition of an invalid instrument. Suppose an invalid instrument induces a nonzero row of β_Y, which means that it is uniquely informative to Y given the rest of X, then removing it may fundamentally change the dependence structure between Y and the rest of X, and thus fundamentally change β_Y, which subsequently changes $\mathcal{S}_{\mathrm{PDRF}}$ under the proposed regulation. For example, if $X = (X_1, X_2, X_3, X_4)^\intercal$ has mutually independent components with zero mean, and if

$$T = X_1 + X_2 + X_3 + X_4 + \epsilon, \quad Y = T + X_1 X_2 + \epsilon, \tag{44}$$

where ϵ is independent of X, then β_T is $(1,1,1,1)^\intercal$ and β_Y is $(\beta_T, (1,0,0,0)^\intercal, (0,1,0,0)^\intercal)$ up to invertible column transformations. Thus, $\mathcal{S}_{\text{PDRF}}$ is spanned by $((1,0,0,0)^\intercal, (0,1,0,0)^\intercal)$ under Assumption (13) and Assumption 1 in the main text, which means $J_0 = \{3,4\}$ and that both X_1 and X_2 are invalid instruments. We now remove X_1 from X, and denote X_1 by ϵ_1 for clarity. Since $Y = T + \epsilon_1 X_2 + \epsilon$, we have $E(Y|X) = E(T|X)$ and thus $\mathcal{S}_{E(Y|X)} = \mathcal{S}_{T|X}$ for the reduced X. Consequently, $\mathcal{S}_{\text{PDRF}}$ becomes the trivial origin $\{(0,0,0)^\intercal\}$ rather than a two-dimensional space, and J_0 is $\{2,3,4\}$ rather than $\{3,4\}$, again under Assumption (13) and Assumption 1 in the main text. The same applies if an invalid instrument induces a nonzero row of β_T.

Nonetheless, if an invalid instrument X_i satisfies the following assumption, then removing X_i from X will simply modify $\mathcal{S}_{\text{PDRF}}$ by removing the ith (nonzero) row of γ_0. Accordingly, J_0 will be invariant, and the dimension of $\mathcal{S}_{\text{PDRF}}$ will be either invariant or reduced by one. We next prove this statement, and discuss when the following assumption holds afterwards. Let β_Y^{-i} be the submatrix of β_Y with its ith row removed, and let $\beta_Y^{(-i)}$ be the new β_Y after X_i is removed from X. Define β_T^{-i} and $\beta_T^{(-i)}$ likewise. As discussed above, $\beta_Y^{(-i)}$ can differ from β_Y^{-i} and $\beta_T^{(-i)}$ can differ from β_T^{-i} in general.

Assumption 3 *Given the mutual independence of the components of X, we have $\beta_Y^{(-i)} = \beta_Y^{-i}$ and $\beta_T^{(-i)} = \beta_T^{-i}$ up to invertible column transformations.*

Recall from the main text that Γ_0 spans $\mathcal{S}_{\text{PDRF}}$ and satisfies Equation (18) with the uniquely smallest number of nonzero rows. For $i = 1, \ldots, p$, let β_Y^i, β_T^i, and Γ_0^i be the ith row of β_Y, β_T, and Γ_0, respectively, and let Γ_0^{-i} be the rest of Γ_0. With an invalid instrument X_i removed from X, we can resemble (18) to have

$$\beta_Y^{(-i)} = \beta_T^{(-i)} A_0^{(-i)} + \Gamma_0^{(-i)} \tag{45}$$

for $\Gamma_0^{(-i)}$ that spans $\mathcal{S}_{\text{PDRF}}$ for the reduced X, where $A_0^{(-i)}$ is some appropriate matrix. Under Assumption 3, if we insert $\beta_Y^i - \beta_T^i A_0^{(-i)}$ into $\Gamma_0^{(-i)}$ as its ith row, then the augmented matrix from $\Gamma_0^{(-i)}$ clearly satisfies (18). By the definition of Γ_0, the augmented matrix from $\Gamma_0^{(-i)}$ must have an equal or larger number of nonzero rows compared with Γ_0. Together with $\Gamma_0^i \neq 0$ induced from that X_i is an invalid instrument, $\Gamma_0^{(-i)}$ must have an equal or larger number of nonzero rows compared with Γ_0^{-i}. In addition, under Assumption 3 again, Γ_0^{-i} must satisfy (45) with $A_0^{(-i)}$ being the rest of A_0 with its ith row removed. Hence, Γ_0^{-i} satisfies (45) with an equal or smaller number of nonzero rows compared with $\Gamma_0^{(-i)}$. Since $\Gamma_0^{(-i)}$ by definition satisfies (45) with the uniquely smallest number of nonzero rows, we must have $\Gamma_0^{(-i)} = \Gamma_0^{-i}$; that is, Γ_0 is changed simply by its ith row removed after X_i is removed from X, which implies the same for γ_0.

We next show that Assumption 3 holds for all $i = 1, \ldots, p$ regardless of whether or not X_i is an invalid instrument, if both $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{T|X}$ can be fully recovered by SIR (Li, 1991). Using $\mathcal{S}_{E(Y|X)}$ as an example, SIR regards β_Y as the set of the eigenvectors of

$$M \equiv \Sigma_X^{-1} E\{E(X|Y) - E(X)\}^{\otimes 2} \Sigma_X^{-1}$$

associated with nonzero eigenvalues, where $v^{\otimes 2}$ denotes $vv^{\intercal}$ for any matrix v. For $i = 1, \ldots, p$, let $X_{(-i)}$ be the rest of X after X_i is removed, and let $\Sigma_{(-i)}$ be its covariance matrix. Then $M_{(-i)} \equiv (\Sigma_{(-i)})^{-1} E\{E(X_{(-i)}|Y) - E(X_{(-i)})\}^{\otimes 2} (\Sigma_{(-i)})^{-1}$ is M for SIR applied to $(X_{(-i)}, Y)$, whose eigenvectors associated with nonzero eigenvalues form $\beta_Y^{(-i)}$. Since Σ_X^{-1} is diagonal induced from the mutual independence of the components of X, $M_{(-i)}$ is clearly equal to the submatrix of M formed by its rows and columns indexed by the complement of $\{i\}$. Therefore, we have $\beta_Y^{(-i)} = \beta_Y^{-i}$ up to invertible column transformations.

In practice, the full recovery of $\mathcal{S}_{E(Y|X)}$ by SIR requires an asymmetric effect of X on Y and that Y has at least $d_Y + 1$ possible outcomes if Y is discrete, where d_Y denotes the dimension of $\mathcal{S}_{E(Y|X)}$. These requirements make SIR ineffective when the data have a complex structure such as Model (44) above. In these cases, other more complex SDR methods such as SAVE (Cook and Weisberg, 1991) should be used instead. However, Assumption 3 no longer holds for these complex SDR methods in general.

To summarize, $\mathcal{S}_{\mathrm{PDRF}}$ is specific for the working predictor X, and, considering the potential complexity of the underlying distribution of (Y, T, X), it should be considered as a different space once some invalid instruments are removed from X. An exception is that, if both $\mathcal{S}_{E(Y|X)}$ and $\mathcal{S}_{T|X}$ can be fully recovered by SIR and if the components of X are mutually independent, then removing any invalid instrument will deliver simple modification of $\mathcal{S}_{\mathrm{PDRF}}$. This however requires additional regulations of the underlying data structure.

Appendix D. Additional Simulation Studies

In this subsection, we report some complementary simulation study results. First, we evaluate the sensitivity of (32) when Assumption 2 is ineffective. Using the same notations β_{I} and β_{II} and the same distribution of (X, ϵ) as in Section 7 of the main text, the following three models are generated, where $\beta_T^{\intercal} X$ has a weak effect on T. Compared with the model settings in Section 7 of the main text, these models differ from Model 1, Model 3, and Model 4 there, respectively, only in the conditional distribution $T|\beta_T^{\intercal} X$.

Model 1*: $T = 0.05\beta_{\mathrm{I}}^{\intercal} X + 3\epsilon$, $Y = T + X_1 + \epsilon$.

Model 3*: $T = 0.3\sin(\beta_{\mathrm{II}}^{\intercal} X - 0.5) + 4\epsilon$, $Y = 2T(0.5X_1 + 0.5X_2 - 1) + 3\epsilon$.

Model 4*: $T = 0.05\beta_{\mathrm{II}}^{\intercal} X + 2 + 4\epsilon$, $Y = 2\sin(0.5T) + |0.5X_1 + 0.5X_2 + 1| + \epsilon$.

Same as in the main text, we set $n = 500$, and set $p = 10$ and $p = 200$ sequentially. Again, when implementing $\lambda_{\min}(\mathbb{W}_\tau)$, we set τ at one and tune the bandwidths in $\mathbb{W}_\tau$ by the proposed cross-validation procedure. Based on 1000 independent runs, the 99% and 95% sample quantiles of $\lambda_{\min}(\mathbb{W}_\tau)$ for each model are recorded in Table 5. Since these values are mostly well below the cutoff $n^{-0.75} \approx .0095$, with the 99% quantile of $\lambda\min(\mathbb{W}_\tau)$ for Model 4* being the only exception, (32) suggests the ineffectiveness of Assumption 2 for these models, which complies with the theory.

We next report the simulation results for the same model settings as in Section 7 of the main text, except that X is now discrete and generated by a p-tuple of independent Bernoulli distributions with mean equal to 0.5. Model 4 is the same as Model 5 in this case. These results suggest the high specificity of the proposed diagnosis procedure (Table 6), the variable selection consistency of the proposed method (Table 7), and the estimation consistency of SPIVE (Table 8) and SMIVE (Tables 9), when X is discrete in the data.

Table 5: The extreme sample quantiles of $\lambda_{\min}(\mathbb{W}_\tau)$ based on 1000 runs. The meanings of numbers in each cell follow those in Table 1 of the main text.

p	Model 1*	Model 3*	Model 4*
10	.005/.004	.004/.004	.012/.006
200	.005/.004	.006/.004	.015/.007

Table 6: The extreme sample quantiles of $\lambda_{\min}(\mathbb{W}_\tau)$ based on 1000 runs, for discrete X. The meanings of numbers in each cell follow those in Table 1 of the main text.

p	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	.075/.127	.205/.268	.110/.221	.048/.121	.090/.146	.030/.041
large	.017/.073	.235/.274	.147/.175	.022/.059	.039/.074	.040/.046

Table 7: Performance of the methods in variable selection for discrete X based on 1000 runs. The meanings of numbers follow those in Table 2 of the main text.

p	Method	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	Proposed	.006/0	.184/0	.003/.006	0/0	0/0	0/0
	sisVIVE	.130/0	.107/0	.179/.017	.218/0	.218/0	.074/0
large	Proposed	0/0	.008/0	16.4/.473	.393/.017	.393/.017	0/0
	sisVIVE	9.82/0	10.8/0	26.0/.381	17.6/0	17.6/0	56.5/0

Table 8: Performance of the methods in estimating the personalized dose-response function for discrete X based on 1000 runs. The meanings of numbers follow those in Table 3 of the main text.

p	Method	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	SPIVE	.140(.064)	.149(.041)	.260(.036)	.189(.032)	.189(.032)	.058(.012)
	sisVIVE	.085(.020)	.142(.037)	.599(.025)	.522(.029)	.522(.029)	.059(.015)
large	SPIVE	.174(.020)	.261(.043)	.525(.074)	.268(.077)	.268(.077)	.074(.010)
	sisVIVE	.385(.015)	.467(.041)	.636(.028)	.760(.041)	.760(.041)	.110(.035)

Table 9: Performance of the methods in estimating the marginal dose-response function for discrete X based on 1000 runs. The meanings of numbers in each cell follow those in Table 4 of the main text.

p	Method	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
10	SMIVE	.165(.099)	.229(.052)	.502(.040)	.408(.051)	.408(.051)	.109(.032)
	sisVIVE	.072(.029)	.089(.040)	.208(.057)	.830(.037)	.830(.037)	.060(.037)
large	SMIVE	.198(.035)	.212(.072)	.721(.072)	.555(.160)	.555(.160)	.165(.028)
	sisVIVE	.978(.014)	1.53(.109)	.932(.174)	2.51(.213)	2.51(.213)	.243(.296)

References

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91(434): 444–455, 1996.

T L Brink, Jerome A Yesavage, Owen Lum, Philip H Heersema, Michael Adey, and Terrence L Rose. Screening Tests for Geriatric Depression. Clinical Gerontologist, 1(1): 37–43, October 2008.

Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for Mendelian randomization. Statistical Methods in Medical Research, 26(5): 2333–2355, 2017.

Amy L Byers and Kristine Yaffe. Depression and risk of developing dementia. Nature Reviews Neurology, 7(6):323–331, 2011.

Xin Chen, Changliang Zou, and R Dennis Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. Annals of Statistics, 38:3696–3723, 2010.

Francesca Chiaromonte, R Dennis Cook, and Bing Li. Sufficient dimension reduction in regressions with categorical predictors. Annals of Statistics, pages 475–497, 2002.

R Denis Cook. Regression Graphics. Wiley, New York, 1998.

R Denis Cook and S Weisberg. Discussion of "Sliced inverse regression for dimension reduction". Journal of the American Statistical Association, 86:316–342, 1991.

R Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. The Annals of Statistics, 30:455–474, 2002.

Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. Statistical Methods in Medical Research, 16(4):309–330, 2007.

Jianqing Fan and Irene Gijbels. Local polynomial modelling and its applications: monographs on statistics and applied probability. Routledge, 2018.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456):1348–1360, 2001.

Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. Journal of Machine Learning Research, 8:361–383, 2007.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In Advances in neural information processing systems, pages 489–496, 2008.

Sander Greenland. An introduction to instrumental variables for epidemiologists. International Journal of Epidemiology, 29(4):722–729, 2000.

Paul W Holland. Statistics and causal inference. Journal of the American Statistical Association, 81(396):945–960, 1986.

Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association, 99 (467):854–866, 2004.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. Journal of the American Statistical Association, 111(513):132–144, 2016.

Hyunseung Kang, Youjin Lee, T Tony Cai, and Dylan S Small. Two robust tools for inference about causal effects with invalid instruments. Biometrics, 78(1):24–34, 2022.

JooSeuk Kim and Clayton D. Scott. Robust kernel density estimation. Journal of Machine Learning Research, pages 2529–2565, 2012.

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Statistics in Medicine, 27(8):1133–1163, 2008.

David S Lee, Justin McCrary, Marcelo J Moreira, and Jack R Porter. Valid t-ratio inference for iv. Technical report, National Bureau of Economic Research, 2021.

Bing Li and Jun Song. Nonlinear sufficient dimension reduction for functional data. Annals of Statistics, 45:1059–1095, 2017.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. Journal of the American Statistical Association, 35:2143–2172, 2007.

Ker-Chau Li. Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association, 86:316–342, 1991.

Ker-Chau Li. On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. Journal of the American Statistical Association, 87(420):1025–1039, 1992.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. The Annals of Statistics, 17:1009–1052, 1989.

Sai Li and Zijian Guo. Causal inference for nonlinear outcome models with possibly invalid instrumental variables. arXiv preprint arXiv:2010.09922, 2020.

Qian Lin, Zhigen Zhao, and Jun S Liu. Sparse sliced inverse regression via lasso. Journal of The American Statistical Association, 114(528):1726–1739, 2019.

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

Wei Luo and Bing Li. On order determination by predictor augmentation. Biometrika, 108: 557–574, 2021.

Wei Luo, Yeying Zhu, and Debashis Ghosh. On estimating regression-based causal effects using sufficient dimension reduction. Biometrika, 104(1):51–65, 2017.

Krishna Prasad Muliyala and Mathew Varghese. The complex relationship between depression and dementia. Annals of Indian Academy of Neurology, 13(Suppl2):S69, 2010.

Whitney K Newey. Efficient instrumental variables estimation of nonlinear models. Econometrica, 58:809–837, 1990.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. Econometrica, 71(5):1565–1578, 2003.

Sid E O'Bryant, Stephen C Waring, C Munro Cullum, James Hall, Laura Lacritz, Paul J Massman, Philip J Lupo, Joan S Reisch, Rachelle Doody, and Texas Alzheimer's Research Consortium. Staging Dementia Using Clinical Dementia Rating Scale Sum of Boxes Scores: A Texas Alzheimer's Research Consortium Study. Archives of Neurology, 65(8): 1091–1095, 2008.

Brandon L Pierce, Habibul Ahsan, and Tyler J VanderWeele. Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. International Journal of Epidemiology, 40(3):740–752, 2011.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5):688–701, 1974.

Andrew J Saykin, Li Shen, Xiaohui Yao, Sungeun Kim, Kwangsik Nho, Shannon L Risacher, Vijay K Ramanan, Tatiana M Foroud, Kelley M Faber, Nadeem Sarwar, Leanne M Munsie, Xiaolan Hu, Holly D Soares, Steven G Potkin, Paul M Thompson, John S K Kauwe, Rima Kaddurah-Daouk, Robert C Green, Arthur W Toga, Michael W Weiner, and Alzheimer's Disease Neuroimaging Initiative. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. Alzheimer's & Dementia, 11(7):792–814, 2015.

Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: A practical guide and simulated example. Psychological Methods, 13(4):279–313, 2008.

Nuala A Sheehan and Vanessa Didelez. Commentary: Can ‘many weak’instruments ever be ‘strong’? International Journal of Epidemiology, 40(3):752–754, 2011.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. On the relation between universality, characteristic kernels and rkhs embedding of measures. In AISTATS, 2010.

Douglas O Staiger and James H Stock. Instrumental variables regression with weak instruments. Econometrica, 65:557–586, 1997.

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6):2769–2794, 2007.

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Statistical Methodology), 58(1):267–288, 1996.

Michael W Weiner, Karl E Friedl, Anthony Pacifico, Julie C Chapman, Michael S Jaffee, Deborah M Little, Geoffrey T Manley, Ann McKee, Ronald C Petersen, Roger K Pitman, Kristine Yaffe, Henrik Zetterberg, Robert Obana, Lisa J Bain, and Maria C Carrillo. Military risk factors for Alzheimer’s disease. Alzheimer’s & Dementia, 9(4):445–451, 2013.

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

Lixing Zhu, Baiqi Miao, and Heng Peng. On sliced inverse regression with high-dimensional covariates. Journal of the American Statistical Association, 101:630–642, 2006.

Yeying Zhu, Donna L Coffman, and Debashis Ghosh. A boosting algorithm for estimating generalized propensity scores with continuous treatments. Journal of Causal Inference, 3 (1):25–40, 2015.