Published in final edited form as:

J Multivar Anal. 2018 November ; 168: 48–62. doi:10.1016/j.jmva.2018.07.005.

Joint sufficient dimension reduction for estimating continuous treatment effect functions

Ming-Yueh Huang^{*,a} and Kwun Chuen Gary Chan^b

^aInstitute of Statistical Science, Academia Sinica, Taiwan

^bDepartment of Biostatistics, University of Washington, U.S.A.

Abstract

The estimation of continuous treatment effect functions using observational data often requires parametric specification of the effect curves, the conditional distributions of outcomes and treatment assignments given multi-dimensional covariates. While nonparametric extensions are possible, they typically suffer from the curse of dimensionality. Dimension reduction is often inevitable and we propose a sufficient dimension reduction framework to balance parsimony and flexibility. The joint central subspace can be estimated at a $n^{1/2}$ -rate without fixing its dimension in advance, and the treatment effect function is estimated by averaging local estimates of a reduced dimension. Asymptotic properties are studied. Unlike binary treatments, continuous treatments require multiple smoothing parameters of different asymptotic orders to borrow different facets of information, and their joint estimation is proposed by a non-standard version of the infinitesimal jackknife.

Keywords

Central subspace; Cross-validation; Dose-response; Infinitesimal jackknife; Optimal bandwidth

1. Introduction

It is common in medical and social studies to collect data on treatments or exposures and a primary goal in these studies is to investigate the causal effects of the treatments on an outcome. In the case of continuous treatments, the effects are naturally described by functions, and model-free estimation typically involves smoothing to borrow information from adjacent treatment levels. While observational data are often the only data available in practice, direct smoothing of observed responses across different treatment levels usually results in biased estimates of causal treatment effects due to confounding.

In the literature, there are two major approaches for estimating treatment effects under confoundedness; they are based on outcome regression models and generalized propensity

*corresponding author: myh0728@stat.sinica.edu.tw.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

score models. Outcome regression models describe how the response relates to the treatment and covariates [9, 14], while generalized propensity score models extend the classical propensity score models for binary treatments to account for the conditional density of treatment given covariates [4, 10, 13]. Doubly robust estimation was recently proposed in [17] to unify these two approaches. It combines outcome regression and generalized propensity score models through estimation. Model-based applications require at least one correctly specified parametric or semiparametric model for the outcome regression functions or generalized propensity scores. However, in practice, it is often hard to check if these models are adequate without prior knowledge. In contrast, one can use fully nonparametric smoothing techniques to replace the modeling procedures as in [4, 17]. When the dimension of covariates is large, the model-free approaches suffer from the curse of dimensionality and can lead to unstable estimates for the treatment effects. Some middle ground between current model-based and model-free approaches would thus be desirable.

In this work, we study a nonparametric method that is more parsimonious and suffers less from the curse of dimensionality than the fully nonparametric smoothing estimator, while retaining robustness to potential model misspecification. Sufficient dimension reduction for treatment effect estimation has recently been studied in [11, 19] for binary treatment variables. To extend these ideas to continuous treatments, we consider a joint sufficient dimension reduction model [11] to capture all the information of the outcome regression and generalized propensity score. In contrast, Luo et al. [19] considered separate dimension reduction for different subgroups, which is more difficult to generalize to continuous treatments.

An advantage of sufficient dimension reduction is that, unlike generalized propensity scores which cannot be estimated nonparametrically at a $n^{1/2}$ convergence rate, the central subspace can be estimated with a $n^{1/2}$ -rate, which turns out to be crucial in guaranteeing a desirable theoretical and practical performance of the treatment effects estimator. The estimation is complicated by the need to employ multiple smoothing parameters, one for borrowing information across different treatment levels, and another for undersmoothing with respect to marginalizing the observed covariate distribution. The smoothing parameters are chosen in a data-adaptive manner by minimizing an estimator of the asymptotic mean squared error via a non-standard application of the infinitesimal jackknife.

The rest of this article is organized as follows. Section 2 introduces the considered dimension reduction model and proposes an estimation procedure to estimate the joint central subspace and the treatment effects of interest. The results of a series of simulation studies are reported in Section 3 and an application on food patterns data analysis is described in Section 4. Some concluding remarks are given in Section 5.

2. The proposed methodology

2.1. Joint sufficient dimension reduction model

Let $Y(t)$ be the potential outcome associated with each treatment level t , where $t \in \mathcal{T}$, and $\mathcal{T}$ is a connected subset in $\mathbb{R}$, the set of all real numbers. Also, let T be the continuous treatment variable. Then the observed outcome is $Y = Y(T)$. In addition, a vector $\mathbf{X} = (X_1,$

..., X_p) of covariates is observed for each subject. The main goal of this work is to estimate the continuous treatment effect $\mu(t) = E\{Y(t)\}$, $t \in \mathcal{T}$, based on a random sample $(Y_1, T_1, \mathbf{X}_1), \dots, (Y_n, T_n, \mathbf{X}_n)$. Since this parameter is defined in terms of potential outcomes which are not observed directly, some assumptions are needed to identify the quantity of interest by the observed data.

Assumption 1 (Positivity). $f_T(t | \mathbf{x}) > 0$ for all $t \in \mathcal{T}$ and $\mathbf{x} \in \mathcal{X}$, where $f_T(t | \mathbf{x})$ is the conditional density of T given $\mathbf{X} = \mathbf{x}$, and $\mathcal{X}$ is the support of $\mathbf{X}$.

Assumption 2 (Ignorability). $Y(t) \perp\!\!\!\perp T | \mathbf{X}$ for all $t \in \mathcal{T}$, where $\perp\!\!\!\perp$ denotes independence.

The positivity assumption states that every subject has a non-zero chance of receiving treatment level t . The ignorability assumption says that the potential outcomes and the treatment are unconfounded when the covariates are given. That is, all the confoundedness is captured by the vector of covariates. According to these assumptions, a simple bias removal strategy is to use the property

$$E\{Y(t)\} = E[E\{Y(t) | \mathbf{X}\}] = E[E\{Y(t) | T = t, \mathbf{X}\}] = E\{E(Y | T = t, \mathbf{X})\}. \quad (1)$$

Thus, an estimator for $\mu(t)$ can be obtained by averaging an estimator of the conditional effects $E(Y | T = t, \mathbf{X})$. In practice, the number of covariates is usually large and the nonparametric estimators for the conditional effects typically suffer from the curse of dimensionality. In order to find a lower dimensional score to attain dimension reduction while retaining a property similar to (1), we first note that the likelihood of the observed variables is

$$f_{Y, T, \mathbf{X}}(y, t, \mathbf{x}) = f_Y(y | t, \mathbf{x}) f_T(t | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) = f_{Y(t)}(y | \mathbf{x}) f_T(t | \mathbf{x}) f_{\mathbf{X}}(\mathbf{x})$$

based on Assumptions 1 and 2. One can see that all the information is conveyed through the conditional distributions of $Y(t)$ and T given $\mathbf{X}$. Thus, we propose a joint sufficient dimension reduction model to summarize both potential outcomes and treatment assignment, viz.

$$\forall_{t \in \mathcal{T}} Y(t) \perp\!\!\!\perp \mathbf{X} | B^T \mathbf{X}, \quad T \perp\!\!\!\perp \mathbf{X} | B^T \mathbf{X} \quad (2)$$

for some $p \times d$ full-rank parameter matrix B with $d \leq p$. Since model (2) has a nested structure in d , we can focus on searching the joint sufficient dimension reduction subspace $\text{span}(B)$ with the smallest dimension, which is called the joint central subspace and the corresponding basis matrix is denoted by B_0 with dimension d_0 . Related discussion on the model structure and identification is given in [11] for binary treatments. Model (2) leads to the following result on the ignorability of treatment assignment.

Proposition 1. Under Assumption 2 and model (2), the joint central subspace $\text{span}(B_0)$ satisfies $Y(t) \perp\!\!\!\perp T \mid B_0^\top \mathbf{X}$ for all $t \in \mathcal{T}$.

Thus, the bias-removal strategy in (1) can still be applied with $\mathbf{X}$ replaced by $B_0^\top \mathbf{X}$. That is,

$$\mu(t) = E\{Y(t)\} = E[E\{Y(t) \mid \mathbf{X}\}] = E\{E(Y \mid T = t, \mathbf{X})\} = E\{E(Y \mid T = t, B_0^\top \mathbf{X})\},$$

which forms a basis for our estimator of $\mu(t)$. Let $m(t, \mathbf{x}; B) = E(Y \mid T = t, B^\top \mathbf{X} = B^\top \mathbf{x})$. If B_0 is known, we can estimate $\mu(t)$ by $n^{-1} \sum_{i=1}^n \hat{m}(t, \mathbf{X}_i; B_0)$, with a locally smoothed estimator

$$\hat{m}(t, \mathbf{x}; B) = \frac{\sum_{i=1}^n Y_i K_{h_1}(T_i - t) \mathcal{K}_{h_2}\{B^\top(\mathbf{X}_i - \mathbf{x})\}}{\sum_{i=1}^n K_{h_1}(T_i - t) \mathcal{K}_{h_2}\{B^\top(\mathbf{X}_i - \mathbf{x})\}},$$

where $K_{h_1}(t) = K(t/h_1)/h_1$, $\mathcal{K}_{h_2}(\mathbf{u}) = \prod_{k=1}^d K(u_k/h_2)/h_2$ with $\mathbf{u} = (u_1, \dots, u_d)^\top$, h_1, h_2 are positive bandwidths, and K is an r th-order kernel function. The choice of r will be discussed in Section 2.3. When B_0 is unknown and we have an estimator $\hat{B}$, we can estimate $\mu(t)$ by $\hat{\mu}(t) = n^{-1} \sum_{i=1}^n \hat{m}(t, \mathbf{X}_i; \hat{B})$. The following result is proven for any estimator $\hat{B}$ that converges in distribution at a $n^{1/2}$ -rate under mild assumptions.

Theorem 1. Suppose that Conditions A1–A4 in the Appendix are satisfied. Then, as $n \rightarrow \infty$,

$$(nh_1)^{1/2} \{\hat{\mu}(t) - \mu(t) - h_1^r \mathcal{B}(t)\} \rightsquigarrow \mathcal{N}\{0, \mathcal{V}(t)\}$$

where

$$\mathcal{B}(t) = \frac{\mu_{r,K}}{r!} E \left[\frac{\partial_t^r \{m(t, \mathbf{X}; B_0) f_T(t \mid B_0^\top \mathbf{X})\} - m(t, \mathbf{X}; B_0) \partial_t^r f_T(t \mid B_0^\top \mathbf{X})}{f_T(t \mid B_0^\top \mathbf{X})} \right],$$

$$\mathcal{V}(t) = \sigma_K^2 E \left[\frac{\text{var}(Y \mid T, \mathbf{X}) + \{m(T, \mathbf{X}; B_0) - m(t, \mathbf{X}; B_0)\}^2}{f_T^2(t \mid B_0^\top \mathbf{X})} \right],$$

$$\mu_{r,K} = \int u^r K(u) du, \text{ and } \sigma_K^2 = \int K^2(u) du.$$

Remark 1. Different bandwidths h_1 and h_2 are needed because $\hat{\mu}(t)$ averages over the covariate distribution but not the treatment distribution. Therefore, undersmoothing is required in h_2 , but not in h_1 , to minimize the mean squared error of the estimator. Details are given in Sections 2.3 and 2.4.

Remark 2. Galvao and Wang [4] proposed an estimator for the treatment effects based on a conditional density adjustment. The main idea comes from the property

$$E\{Y(t) - \mu(t)\} = E\left[\{Y - \mu(t)\} \frac{f_{T(t) | Y, \mathbf{X}}}{f_{T(t) | \mathbf{X}}}\right]$$

under Assumption 2 and some regularity conditions for changing the order of limits and integral. Since Proposition 1 shows that the conditional independence in Assumption 2 also holds for $B_0^T \mathbf{X}$, we can deduce that

$$E\{Y(t) - \mu(t)\} = E\left[\{Y - \mu(t)\} \frac{f_{T(t) | Y, B_0^T \mathbf{X}}}{f_{T(t) | B_0^T \mathbf{X}}}\right].$$

Thus, an estimator for $\mu(t)$ can be

$$\hat{\mu}_{SE}(t) = \frac{\sum_{i=1}^n Y_i \hat{f}_{T(t) | Y_i, \hat{B}^T \mathbf{X}_i} / \hat{f}_{T(t) | \hat{B}^T \mathbf{X}_i}}{\sum_{i=1}^n \hat{f}_{T(t) | Y_i, \hat{B}^T \mathbf{X}_i} / \hat{f}_{T(t) | \hat{B}^T \mathbf{X}_i}}$$

with

$$\begin{aligned} \hat{f}_{T(t) | y, B^T \mathbf{x}} &= \frac{\sum_{i=1}^n K_{\zeta}(T_i - t) K_{\zeta}(Y_i - y) \mathcal{K}_{\zeta}\{B^T(\mathbf{X}_i - \mathbf{x})\}}{\sum_{i=1}^n K_{\zeta}(Y_i - y) \mathcal{K}_{\zeta}\{B^T(\mathbf{X}_i - \mathbf{x})\}}, \\ \hat{f}_{T(t) | B^T \mathbf{x}} &= \frac{\sum_{i=1}^n K_{\zeta}(T_i - t) \mathcal{K}_{\zeta}\{B^T(\mathbf{X}_i - \mathbf{x})\}}{\sum_{i=1}^n \mathcal{K}_{\zeta}\{B^T(\mathbf{X}_i - \mathbf{x})\}}. \end{aligned}$$

Note that this estimator requires a fully nonparametric estimation for the density $f_{T, Y, B^T \mathbf{X}}(t, y, B^T \mathbf{x})$, which has a slower rate of convergence than that of $\hat{m}(t, \mathbf{x}; B)$. In our simulations, we found that this estimator usually has a poorer performance than the proposed estimator.

Remark 3. In the literature of causal inference with binary treatments, doubly robust estimation is often appealing. It combines the outcome regression and propensity score models and requires only one of them to be correctly specified. As in cases with continuous treatments, Kennedy et al. [17] proposed a kernel-based approach based on an influence function for a kernel-weighted projection parameter, which requires only mild smoothness conditions and still allows for double robustness. The key proposition of their approach is

$$\mu(t) = E\left\{\frac{Y - m(T, \mathbf{X})}{f_T(T | \mathbf{X}) / \omega(T)} + \mu(T) \mid t = t\right\}, \quad (3)$$

where $m(t, \mathbf{x}) = E(Y | T = t, \mathbf{X} = \mathbf{x})$ and $\omega(t) = E\{f_T(t | \mathbf{X})\}$. Under model (2), it can be easily seen that the equality in (3) still holds when $m(t, \mathbf{x})$ and $f_T(t | \mathbf{X})$ are replaced with $m(t, B_0^T \mathbf{x})$ and $f_T(t | B_0^T \mathbf{X})$, respectively. Thus, an alternative estimator for $\mu(t)$ is

$$\hat{\mu}_{\text{DR}}(t) = \frac{\sum_{i=1}^n \hat{\xi}_i K_{\xi}(T_i - t)}{\sum_{i=1}^n K_{\xi}(T_i - t)} \quad \text{with} \quad \hat{\xi}_i = \frac{Y_i - \hat{m}(T_i, \hat{B}^T \mathbf{X}_i)}{\hat{f}_{T_i}(T_i | \hat{B}^T \mathbf{X}_i) / \hat{\omega}(T_i)} + \hat{\mu}(T_i),$$

where $\hat{\omega}(t) = n^{-1} \sum_{i=1}^n \hat{f}_{T_i}(t | \hat{B}^T \mathbf{X}_i)$. This estimator can be viewed as an adjusted version of $\hat{\mu}(t)$. However, $\hat{f}_{T_i}(t | \hat{B}^T \mathbf{X}_i)$ requires one more smoothing parameter than the proposed method. In our simulations, we found that $\hat{\mu}_{\text{DR}}(t)$ has a similar performance as $\hat{\mu}(t)$ in finite samples.

Remark 4. We may also provide more general assumptions like those in [4, 17] to allow non-kernel-based estimators such as splines for estimating $m(t, \mathbf{x}; B)$. However, in some cases the conditions become more stringent. For example, Galvao and Wang [4] require the estimate of $f_T(t | y, \mathbf{x}) / f_T(t | \mathbf{x})$ to converge at the rate $n^{-1/4}$, which is hardly satisfied for multiple covariates $\mathbf{X}$ or even the lower-dimensional $B_0^T \mathbf{X}$. In addition, the conditions of [4, 17] do not directly lead to results for the mean squared error, which can only be computed case by case for different smoothers, and will typically be different from our results. To provide a complete estimation procedure including proper smoothing parameter selection based on asymptotically optimal mean squared error, we focus on the kernel-based estimator and assume more specialized smoothness conditions for the consistency and the weak convergence to hold in this work.

2.2. Estimation of joint central subspace

Estimation of B_0 does not follow immediately from model (2) because $Y(t)$ is only partially observed. We need the following identification result using the observed data.

Proposition 2. *Suppose that Assumptions 1–2 hold. Then, model (2) is equivalent to $(Y, T) \perp\!\!\!\perp \mathbf{X} | B^T \mathbf{X}$.*

Therefore, we can employ existing methodologies for sufficient dimension reduction of a bivariate outcome to estimate B_0 . While different methods of sufficient dimension reduction have been proposed in the literature (see, e.g., [20, 28, 29]), we use a generalization of the estimator of Huang and Chiang [12] for multivariate response (Y, T) in all numerical examples because it can estimate simultaneously the structural dimension, the basis matrix, and a rate-optimal smoothing parameter under a single objective function, while avoiding certain linearity assumptions of covariate distribution required by inverse regression [18].

Let

$$\hat{F}(y, t \mid B^T \mathbf{x}) = \frac{\sum_{j=1}^n \mathbf{1}(Y_j \leq y, T_j \leq t) \mathcal{K}_h(B^T(\mathbf{X}_j - \mathbf{x}))}{\sum_{j=1}^n \mathcal{K}_h(B^T(\mathbf{X}_j - \mathbf{x}))}$$

be a kernel smoothing estimator for $F(y, t \mid B^T \mathbf{x}) = \Pr(Y \leq y, T \leq t \mid B^T \mathbf{X} = B^T \mathbf{x})$, where $\mathbf{1}$ denotes the indicator function, h is another positive bandwidth, and K is a q th-order kernel function. In practice, we suggest to take $q = 4 \vee 2\lfloor(d+6)/4\rfloor$, where $\lfloor \cdot \rfloor$ is the floor function. Further, let $(T^0, Y^0, \mathbf{X}^0)$ be a future run independent of current data $(Y_1, T_1, \mathbf{X}_1), \dots, (Y_n, T_n, \mathbf{X}_n)$ and define the prediction risk

$$\mathbb{E}[\{\mathbf{1}(Y^0 \leq y, T^0 \leq t) - \hat{F}(y, t \mid B^T \mathbf{X}^0)\}^2]. \quad (4)$$

A simple calculation shows that the prediction risk in (4) can be decomposed into

$$\sigma_0^2 + b_0^2(B) + \text{MISE}_B(h) + C(B, h), \quad (5)$$

where

$$\begin{aligned} \sigma_0^2 &= \mathbb{E}[\{\mathbf{1}(Y^0 \leq y, T^0 \leq t) - F(y, t \mid B_0^T \mathbf{X}^0)\}^2], \quad b_0^2(B) = \mathbb{E}[\{F(y, t \mid B_0^T \mathbf{X}^0) - F(y, t \mid B^T \mathbf{X}^0)\}^2], \\ \text{MISE}_B(h) &= \mathbb{E}[\{F(y, t \mid B^T \mathbf{X}^0) - \hat{F}(y, t \mid B^T \mathbf{X}^0)\}^2] \text{ and} \\ C(B, h) &= 2\mathbb{E}[\{F(y, t \mid B_0^T \mathbf{X}^0) - F(y, t \mid B^T \mathbf{X}^0)\}\{F(y, t \mid B^T \mathbf{X}^0) - \hat{F}(y, t \mid B^T \mathbf{X}^0)\}]. \end{aligned}$$

When $h \rightarrow 0$ and $nh^d \rightarrow \infty$, it is shown in the Appendix that the last two summands in (5) converge to zero and are dominated by the first two terms. Note that $b_0^2(B) \geq 0$ and the equality holds when $\text{span}(B)$ is a joint sufficient dimension reduction subspace. Thus, the minimum of the prediction risk must occur when $\text{span}(B) \supseteq \text{span}(B_0)$. Moreover, since our model has a nested structure, the prediction risk will decrease when the working dimension increases as long as the working dimension is less than the structural dimension. In contrast, once the working dimension d is equal to, or larger than, the structural dimension d_0 and $\text{span}(B) \supseteq \text{span}(B_0)$, the prediction risk has an asymptotic order of $\sigma_0^2 + O_p\{n^{-2q/(2q+d)}\}$, which will start to increase in d . Therefore, for large enough n , the minimum of prediction risk occurs at the joint central subspace B_0 . A formal result is given as follows.

Proposition 3. *Under Assumptions 1–2 and model (2), the joint central subspace B_0 and the optimal bandwidth $h_0 = c_{d_0} n^{-1/(2q+d_0)}$ minimize the prediction risk in (4) as $h \rightarrow 0$, $nh^d \rightarrow \infty$, and $n \rightarrow \infty$, where the constant c_{d_0} is given in the Appendix.*

According to Proposition 3, we can estimate B_0 by minimizing a criterion that converges to the prediction risk (4) uniformly in probability. Such a criterion function can be obtained by leave-one-out cross-validation. However, the minimization of such a criterion is complicated

by the unknown dimension of B . Fortunately, according to Proposition 3 and the fact that the prediction risk is asymptotically convex in d , we implement the following forward selection algorithm which asymptotically minimizes the prediction risk (4):

Step 1. For $d = 0$, compute

$$\widetilde{CV}(0) = \frac{1}{n} \sum_{i=1}^n \int \int \{1(T_i \leq t, Y_i \leq y) - \hat{F}_{Y,T}(y, t)\}^2 d\hat{F}_{Y,T}(y, t),$$

where $\hat{F}_{Y,T}(y, t)$ is the empirical distribution associated with $(Y_1, T_1), \dots, (Y_m, T_m)$.

Step 2. For $d = 1$, let $(\hat{B}_d, \hat{h}_d)$ be the minimizer of

$$CV(B, h) = \frac{1}{n} \sum_{i=1}^n \int \int \{1(Y_i \leq y, T_i \leq t) - \hat{F}^{-i}_{(Y, T)}(y, t | B^T \mathbf{X}_i)\}^2 d\hat{F}_{Y,T}(y, t),$$

where B is a $p \times d$ matrix, and the superscript $-i$ indicates the estimator based on a sample with i th subject being deleted. Then, compute $\widetilde{CV}(d) = CV(\hat{B}_d, \hat{h}_d)$.

Step 3. Repeat Step 2 until $d = \hat{d}$ with $\widetilde{CV}(\hat{d} + 1) > \widetilde{CV}(\hat{d})$ and the proposed estimator is $(\hat{B}, \hat{h}) = (\hat{B}_{\hat{d}}, \hat{h}_{\hat{d}})$.

Based on the notations and assumptions in the Appendix, we can prove the asymptotic property of $(\hat{B}, \hat{h})$ as follows.

Theorem 2. Suppose that Conditions B1–B5 in the Appendix are satisfied. Then,

$$\Pr(\hat{d} = d_0) \rightarrow 1, \hat{h} = O_p\{n^{-1/(2q+d_0)}\}, \text{ and}$$

$$n^{1/2} \text{vec}(\hat{B} - B_0) \mathbf{1}(\hat{d} = d_0) \rightsquigarrow \mathcal{N}_{pd_0}[0, V^{-1}(B_0) E\{S^{\otimes 2}(B_0)\} V^{-1}(B_0)]$$

as $n \rightarrow \infty$, where $S(B)$ and $V(B)$ are defined in the Appendix.

Without fixing d_0 in advance, we allow model (2) to be flexible and to contain the fully nonparametric model [11]. Another important consideration is that this estimator of the central subspace is $n^{1/2}$ -consistent under mild assumptions [12], which is fast enough to keep the large-sample properties of $\hat{\mu}(t)$ as if the true joint central subspace is known. This property will not hold if we use a nonparametric estimator of generalized propensity score in place of $\hat{B}^T \mathbf{x}$, because the convergence rates of nonparametric density estimators are slower than $n^{1/2}$.

Remark 5. From the proof steps of Theorem 2, we can show that $\hat{h}_d = O_p\{n^{-1/(2q+d)}\}$ for fixed $d \in \{0, \dots, p\}$. Coupled with Condition B3, the order of kernel function should satisfy

$q > 2 \vee (d + 2)/2$. Since we always use symmetric kernel functions, whose orders are all even, and require the order as small as possible, a simple choice is to take $q = 4 \vee 2 \lfloor (d + 6)/4 \rfloor$ for each working dimension d .

Remark 6. For binary treatments, Huang and Chan [11] used a conditional distribution version of Proposition 2 for identification, and the estimation of B_0 is performed through a weighted average of objective functions based on stratified subsamples by treatment statuses. For continuous treatments, such a strategy would require an additional smoothing parameter to borrow information locally for each value of T . Estimation based on Proposition 2 is more straightforward for continuous treatments and is therefore employed.

Remark 7. For regression of a univariate outcome, a single-index model is another popular dimension reduction method. In fact, it can be viewed as a special case of sufficient dimension reduction with $d_0 = 1$. Using a single-index model for multivariate outcomes, however, is prone to misspecification if different combinations of covariate values are related to each outcome. In contrast, if a single-index model is indeed true for a multivariate outcome, efficiency will be lost and nearly collinear indices will be obtained if marginal estimation is performed separately for each outcome. Since the joint sufficient dimension reduction model is defined by multiple outcomes, we do not fix the structural dimension a priori, and the method of [12] is employed to simultaneously estimate the structural dimension and the basis matrix.

2.3. Asymptotic mean squared error

Following [24], we can rewrite $\hat{\mu}(t) - \mu(t) = A_n + B_n + C_n + O_p(\|\hat{B} - B_0\|)$, where

$$\begin{aligned} A_n &= \frac{1}{n} \sum_{i=1}^n \frac{Y_i - m(t, \mathbf{X}_i; B_0)}{f_T(t | \mathbf{X}_i)} K_{h_1}(T_i - t), \\ B_n &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\hat{m}(t, \mathbf{X}_i; B_0) - \mu(t) - \frac{Y_i - m(t, \mathbf{X}_i; B_0)}{f_T(t | \mathbf{X}_i)} K_{h_1}(T_i - t) \mid \mathcal{D}^{-i} \right], \\ C_n &= \frac{1}{n} \sum_{i=1}^n \left[\hat{m}(t, \mathbf{X}_i; B_0) - \mu(t) - \frac{Y_i - m(t, \mathbf{X}_i; B_0)}{f_T(t | \mathbf{X}_i)} K_{h_1}(T_i - t) - \mathbb{E} \left[\hat{m}(t, \mathbf{X}_i; B_0) - \mu(t) - \frac{Y_i - m(t, \mathbf{X}_i; B_0)}{f_T(t | \mathbf{X}_i)} K_{h_1}(T_i - t) \mid \mathcal{D}^{-i} \right] \right], \end{aligned}$$

and $\mathcal{D}^{-1} = \{(Y_j, T_j, \mathbf{X}_j) : j \neq i\}$. From standard derivation of kernel smoothing estimation, we have $A_n = O_p\{h_1^r + (nh_1)^{-1/2}\}$. Similar to the argument for Theorem 2.1 of [2], we can further show that $B_n = O_p(h_2^r)$ and $C_n = O_p\{1 / (n^2 h_1 h_2^{d_0})\}$, which are both dominated by A_n . Thus, the mean squared error $\mathbb{E}[\{\hat{\mu}(t) - \mu(t)\}^2]$ is asymptotically equivalent to that of A_n , which equals $h_1^{2r} \mathcal{B}^2(t) + \mathcal{V}(t) / (nh_1)$ asymptotically, and the optimal bandwidth is

$$h_{1, \text{opt}} = [\mathcal{V}(t) / \{2r \mathcal{B}^2(t)\}]^{1 / (2r + 1)} n^{-1 / (2r + 1)}.$$

Moreover, the remainder terms involving h_2 satisfy $E(B_n) = O(h_2^r)$ and

$\text{var}(C_n) = O\{1/n^2 h_1 h_2^{d_0}\}$. Thus, when $h_1 = O\{n^{-1/(2r+1)}\}$ and $h_2 = O\{n^{-(4r+1)/\{(2r+1)(2r+d_0)\}}\}$, the mean squared error attains the optimal rate $O\{n^{-2r/(2r+1)}\}$ asymptotically.

According to these theoretical results, when we take the bandwidths in optimal rate $h_1 = O\{n^{-1/(2r+1)}\}$ and $h_2 = O\{n^{-(4r+1)/\{(2r+1)(2r+d_0)\}}\}$, the order of kernel function should be taken as $r > \{3d_0 + (9d_0^2 + 8d_0)^{1/2}\} / 4$ to ensure Condition A3 and the optimal rate of mean squared error. However, if we only require the mean squared error to converge to zero at the optimal rate, $r > (d-1)/2$ for each working dimension d is enough to ensure this property. Thus, we suggest to take $r = 2\lceil d/4 \rceil$ in practice, where $\lceil \cdot \rceil$ is the ceiling function.

2.4. Bandwidth selection

To select the bandwidths (h_1, h_2) , one could minimize the estimated mean squared error of $\hat{\mu}(t)$. Here we follow the idea of [1] and use an infinitesimal version of the generalized jackknife statistic to estimate the variance and bias of $\hat{\mu}(t)$. Set $(h_1, h_2) = (h_{1,0}n^{-1/(2r+1)}, h_{2,0}n^{-(4r+1)/\{(2r+1)(2r+d_0)\}})$, which is suggested by the optimal rates given in Section 2.3 and then our main goal turns to selecting the optimal $(h_{1,0}, h_{2,0})$. Since the bias of $\hat{\mu}(t)$ is now of order $O\{n^{-2r/(2r+1)}\}$, Schucany et al. [25] suggested using the generalized jackknife pseudovalues defined, for all $i \in \{1, \dots, n\}$, by

$$g_{(i)}(t) = \frac{n^{2r/(2r+1)}\hat{\mu}(t) - (n-1)^{2r/(2r+1)}\hat{\mu}^{-i}(t)}{n^{2r/(2r+1)}\hat{\mu}(t) - (n-1)^{2r/(2r+1)}\hat{\mu}(t)},$$

to form a generalized jackknife estimate $g_{(\cdot)}(t) = n^{-1} \sum_{i=1}^n g_{(i)}(t)$. The bias can be estimated by $\hat{\mu}(t) - g_{(\cdot)}(t)$ and the variance estimator is given by

$$\frac{n-1}{n} \sum_{i=1}^n \{\hat{\mu}^{-i}(t) - \bar{\mu}(t)\}^2 = \left\{ \frac{R_n}{(1-R_n)(n-1)} \right\}^{-2} \times \frac{1}{n(n-1)} \sum_{i=1}^n \{g_{(i)}(t) - g_{(\cdot)}(t)\}^2,$$

where $R_n = \{(n-1)/n\}^{2r/(2r+1)}$ and $\bar{\mu}(t) = n^{-1} \sum_{i=1}^n \hat{\mu}^{-i}(t)$. Related discussion can be found in [5, 26]. Although these estimators have closed forms which can be used in practice, it is time-consuming to obtain $\hat{\mu}^{-i}(t)$ for each $i \in \{1, \dots, n\}$. To reduce the computational burden, we further apply the infinitesimal jackknife of Jaeckel [15] to the generalized jackknife estimates. Following his derivation, we obtain

$$\text{var}\{\hat{\mu}(t)\} = \frac{1}{n^2} \sum_{i=1}^n \hat{D}_i^2(t), \quad \text{bias}\{\hat{\mu}(t)\} = \frac{R_n}{(1-R_n)(n-1)} \times \frac{1}{2n^2} \sum_{i=1}^n \hat{D}_{it}(t),$$

where

$$\hat{D}_i(t) = \hat{m}(t, \mathbf{X}_i; \hat{B}) - \hat{\mu}(t) + \sum_{k=1}^n \frac{\{Y_i - \hat{m}(t, \mathbf{X}_k; \hat{B})\} K_{h_1}(T_i - t) \mathcal{K}_{h_2}\{\hat{B}^\top(\mathbf{X}_i - \mathbf{X}_k)\}}{\sum_{j=1}^n K_{h_1}(T_j - t) \mathcal{K}_{h_2}\{\hat{B}^\top(\mathbf{X}_j - \mathbf{X}_k)\}}$$

$$\hat{D}_{ii}(t) = \frac{2n\{Y_i - \hat{m}(t, \mathbf{X}_i; \hat{B})\} K_{h_1}(T_i - t) \mathcal{K}_{h_2}^{(0)}}{\sum_{j=1}^n K_{h_1}(T_j - t) \mathcal{K}_{h_2}\{\hat{B}^\top(\mathbf{X}_j - \mathbf{X}_i)\}} - 2n \sum_{k=1}^n \frac{\{Y_i - \hat{m}(t, \mathbf{X}_k; \hat{B})\}^2 K_{h_1}^2(T_i - t) \mathcal{K}_{h_2}^2\{\hat{B}^\top(\mathbf{X}_i - \mathbf{X}_k)\}}{[\sum_{j=1}^n K_{h_1}(T_j - t) \mathcal{K}_{h_2}\{\hat{B}^\top(\mathbf{X}_j - \mathbf{X}_k)\}]^2} - 2\hat{D}_i(t)$$

Thus, $E[\{\hat{\mu}(t) - \mu(t)\}^2]$ can be estimated by $\text{var}\{\hat{\mu}(t)\} + \text{bias}^2\{\hat{\mu}(t)\}$ and $(\hat{h}_{1,0}, \hat{h}_{2,0})$ is suggested to be the minimizer of

$$\int [\text{var}\{\hat{\mu}(t)\} + \text{bias}^2\{\hat{\mu}(t)\}] d\hat{F}_T(t),$$

where $\hat{F}_T(t)$ is the empirical distribution of T .

3. Simulation studies

In this section, we use Monte Carlo simulations to demonstrate the finite-sample performance of our proposal. Specifically, we simulate data from the following models conditional on $\mathbf{X} = (X_1, \dots, X_{10}) \sim \mathcal{N}_{10}(0, I_{10})$:

M1: $T = X_1 + \varepsilon_1$ with $\varepsilon_1 \sim \mathcal{N}(0, 1)$, $Y(t) = t - X_1 + X_2 + \varepsilon_2$ with $\varepsilon_2 \sim \mathcal{N}(0, 0.1)$,

M2: $T = \sum_{k=1}^{10} X_k + \varepsilon_1$ with $\varepsilon_1 \sim \mathcal{N}(0, 1)$, $Y(t) = \sum_{k=1}^{10} X_k + \varepsilon_2$ with $\varepsilon_2 \sim \mathcal{N}(0, 0.1)$,

M3: $T \mid \mathbf{X} = \mathbf{x} \sim 20\mathcal{B}\{\lambda(\mathbf{x}), 20 - \lambda(\mathbf{x})\}$, $\lambda(\mathbf{x}) = 20\text{expit}(0.5x_1 + 0.5x_3 - 0.5x_4 - 0.5x_5)$,
 $Y(t) = (x_2 + x_3 - x_4 - x_5)^2/4 - (t/4 - 5/2)^2 + \varepsilon_1$ with $\varepsilon_1 \sim \mathcal{N}(0, 1/8^2)$,

M4: $T = X_1 X_2 + \varepsilon_1$ with $\varepsilon_1 \sim \mathcal{N}(0, 1)$, $Y(t) = X_1 X_2 + \varepsilon_2$ with $\varepsilon_2 \sim \mathcal{N}(0, 0.1)$,

where ε_1 , ε_2 , and $\mathbf{X}$ are mutually independent. For M1, one can directly see that the basis of the joint central subspace is $B_0 = (\mathbf{e}_1, \mathbf{e}_2)$, where $\{\mathbf{e}_1, \dots, \mathbf{e}_{10}\}$ is the standard basis of $\mathbb{R}^{10}$, and the true treatment effect is $\mu(t) = t$. The basis of the joint central subspace for M2 is $\text{span}\{(1, \dots, 1)^\top\}$ and the true treatment effect is the zero function. The basis of the joint central subspace for M3 is $(\mathbf{e}_1 + \mathbf{e}_3 - \mathbf{e}_4 - \mathbf{e}_5, \mathbf{e}_2 + \mathbf{e}_3 - \mathbf{e}_4 - \mathbf{e}_5)$ with structural dimension 2 and the true treatment effect curve is $\mu(t) = 1 - (t/4 - 5/2)^2$. The basis of the joint central subspace for M4 is $(\mathbf{e}_1, \mathbf{e}_2)$ and the true treatment effect is also the zero function.

The performance of the estimated joint central subspace is evaluated through the error measure $|\hat{B}(\hat{B}^\top \hat{B})^{-1} \hat{B}^\top - B_0(B_0^\top B_0)^{-1} B_0^\top|$. To evaluate the estimated treatment effects, we use the mean integrated squared error

$$E\left[\int \{\hat{\mu}(t) - \mu(t)\}^2 d\hat{F}_T(t)\right]$$

as an accuracy measure. All the results are based on 1000 simulations with sample size $n \in \{100, 200, 400\}$.

To investigate the practicality of the proposed joint central subspace on the estimation of continuous treatment effects, we consider four estimators: $\hat{\mu}(t)$, $\hat{\mu}_{\text{DR}}(t)$, $\hat{\mu}_{\text{SE}}(t)$, as discussed in Remarks 2–3, and the marginal regression plug-in estimator based on the generalized propensity score

$$\hat{\mu}_{\text{GPS}}(t) = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^n Y_j K_{h_1}(T_j - t) K_{h_3} \{\hat{f}_T(t | \mathbf{X}_j) - \hat{f}_T(t | \mathbf{X}_i)\}}{\sum_{i=1}^n K_{h_1}(T_j - t) K_{h_3} \{\hat{f}_T(t | \mathbf{X}_j) - \hat{f}_T(t | \mathbf{X}_i)\}}.$$

We compare them with their counterparts in which the proposed balancing score $\hat{B}^T \mathbf{X}$ is replaced by the original high-dimensional $\mathbf{X}$. The simulation results are displayed in Tables 1 and 2. From Table 1, we can see that the proportion of correct structural dimension selected by our proposed cross-validation criterion approaches 1 as sample size increases while the error measure tends to zero. As for the estimators of the treatment effect function, we can see from Table 2 that the estimators based on joint sufficient dimension reduction always have smaller mean integrated squared errors, which confirms the advantage of our proposal.

The estimator $\hat{\mu}_{\text{GPS}}$ performed poorly because it is an empirical version of $E[E\{Y|T=t, f_T(t|\mathbf{X})\}]$, but $f_T(t|\mathbf{X})$ cannot be estimated at the $n^{1/2}$ -rate. The merit of the proposed estimator, as we have shown theoretically, is that B_0 can be estimated at the $n^{1/2}$ -rate, and the effect of the estimated central subspace on the final estimator is asymptotically negligible because $\mu(t)$ is estimated at a rate slower than $n^{1/2}$. The estimator $\hat{\mu}_{\text{SE}}$ did not perform well for the reason discussed in Remark 2, namely that an estimator of the joint density $f_{T,Y,B^T \mathbf{X}}(t, y, B^T \mathbf{x})$ is needed which typically has a slower convergence rate. The poor performance of these two estimators is more apparent for model M2, which corresponds to a situation of strong confounding. In this scenario, the true treatment effect is $\mu(t) = 0$ but the observed (Y, T) are highly correlated. The estimators $\hat{\mu}_{\text{GPS}}$ and $\hat{\mu}_{\text{SE}}$ can have very poor performance for large t while the proposed estimator still has good performance as $E\{Y(t) | \sum_{k=1}^{10} X_k\}$ can still be accurately estimated even for large t .

4. Application

We illustrate the proposed method by analyzing the food patterns dataset of [27], to investigate the effects of obesity on the systolic blood pressure. In the literature, researchers have found that there is a positive relationship between systolic blood pressure and body mass index [3, 16]. However, the effects can possibly be confounded by diet. The dataset contains 1073 Nepalese adults who are 18 years or older participating in the baseline survey of the Dhulikhel Heart Study, which collects the responses from Food Frequency Questionnaires to identify the food patterns of the participants. The diet variables include whole grains, refined grains, lentils, vegetable oils, solid fats, fatty foods, vegetables, fruits,

potatoes, nuts, poultry, red meat, fish, milk, milk product, fast food, processed cereal, noodles, salty snacks, soda drinks, tea/coffee, sweets, and alcohol. Other variables also include age and sex. All the covariates are centralized and standardized in this analysis.

The estimated joint central subspace is one-dimensional and the corresponding index coefficients with standard errors are listed in Table 3. Figure 1 displays the estimated treatment effects with 95% point-wise confidence intervals using the proposed method, and point estimates based on unadjusted analysis and the estimators of [4, 17]. Compared to the unadjusted estimates, the estimated average systolic blood pressure is smaller for the overweight individuals when age, sex and diet are adjusted. Although the proposed estimator is similar to the estimators of [4, 17], they have different sensitivities to small changes in the bandwidth parameter h_2 that smooths over the covariates. We perturbed the bandwidth by multiplicative factors and computed the L_p -distance between the perturbed functional estimators and the original ones. The results given in Table 4 indicate that the dimension reduction approach is more stable than other nonparametric estimators.

5. Discussion

Nonparametric estimation of treatment effect functions received considerable attention recently, because it does not require a priori assumption on the shape of the function which may not be reasonably assumed. Two advanced approaches were recently proposed. Kennedy et al. [17] developed a doubly robust estimator which requires two sets of modeling assumptions and is valid when one of the two is correct. Galvao and Wang [4] proposed a fully nonparametric estimator that may suffer from the curse of dimensionality. The pros and cons of the two approaches are discussed extensively in the literature and they are favored by different groups of statisticians. Doubly robust methods are prevalent in the biostatistics literature, whereas the fully nonparametric approach is popular in econometrics. The proposed sufficient dimension reduction approach is a middle ground between the two and preserves the merit of both methods. On one hand, the sufficient dimension reduction model contains the fully nonparametric model and therefore cannot be mis-specified, thus more robust than the doubly robust estimators. On the other hand, when the underlying data generating mechanism has a lower-dimensional structure, the proposed method can greatly reduce the dimensionality and hence attain a better convergence rate than fully nonparametric methods. The complexity is controlled by the effective dimension of the central subspace, for which we can estimate consistently using the proposed method. Therefore, our approach is both flexible and parsimonious.

The asymptotic result in Theorem 1 holds for any $n^{1/2}$ -consistent estimator $\hat{B}$ of the central subspace. Thus, one may use other existing sufficient dimension reduction approaches in lieu of the proposed method given in Section 2.2. We use the proposed method because the effective dimension and the central subspace are being estimated by a single criterion, which results in a more stable estimator in finite samples, and the asymptotic result can be proven for an estimated dimension. Other methods require separate criteria to estimate the dimension and the basis matrix, and their theoretical properties are often investigated separately and may not be easily combined. Comparisons between different sufficient dimension reduction techniques are reported elsewhere; see, e.g., [12]. Our primary interest

in this paper is the treatment effect function, and sufficient dimension reduction is a means to an end rather than an end in itself.

For a binary treatment, Luo et al. [19] proposed a separated dimension reduction model for $E(Y_0|\mathbf{X})$ and $E(Y_1|\mathbf{X})$, viz.

$$E(Y_0|\mathbf{X}) \perp\!\!\!\perp \mathbf{X} \mid B_{M,0}^\top \mathbf{X}, \quad E(Y_1|\mathbf{X}) \perp\!\!\!\perp \mathbf{X} \mid B_{M,1}^\top \mathbf{X}.$$

Under a positivity condition on the original covariates and a measurable condition on the conditional variance of potential outcomes with respect to reduced covariates, these authors showed that their proposed estimator can be super-efficient. While this modeling approach may be extended to continuous treatment by assuming $E\{Y(t) \mid \mathbf{X}\} \perp\!\!\!\perp \mathbf{X} \mid B_{M,t}^\top \mathbf{X}$ for each t , $B_{M,t}$ is typically not estimable at $n^{1/2}$ -rate based on the observational data, since for most values of t there is only one or even no observation of outcomes. Also, since all the existing estimators for continuous treatment effect function are irregular, we cannot compare the asymptotic variance directly and there is no guarantee that alternative methods will lead to better efficiency. In fact, the limiting distributions also involve the bias part, whose rates relate to the smoothness of underlying distribution functions instead of the dimension of balancing score only. In our proposal, we focus on using $B^\top \mathbf{X}$ to preserve all the information from the observed data, i.e., the conditional distribution of (Y, T) given $\mathbf{X}$, where B is estimable at $n^{1/2}$ -rate. Comparisons of the resulting asymptotic mean squared errors between our proposed model and other dimension reduction models require future investigation.

It should be noted that sufficient dimension is not the only cure to the curse of dimensionality. A natural limitation is that it reduces the dimension via linear indices, which might not be appropriate in some applications. When the underlying structural covariates live on a general manifold instead of hyperplanes, a dimension reduction method based on non-linear functionals may be needed to solve this problem. In contrast, when there are irrelevant covariates involved, an additional variable selection will lead to better performance by reducing the number of coefficients to be estimated [6]. Such more complicated situations are commonly seen in practice and need further investigation.

In this work we consider the case where the conditional independence assumption holds, which requires the collection of all confounders. However, in many applications researchers often encounter the non-ignorable selection of exposure, in which there are unobserved confounding variables. In those cases, instrumental variables are often used to identify certain partial treatment effects under different sets of assumptions. The structure of the sufficient dimension reduction and the corresponding estimation will then be quite different from the present paper.

Acknowledgment

The authors thank the Editor-in-Chief, an Associate Editor and three reviewers for constructive comments. Both authors are partially supported by the United States National Institutes of Health grant R01HL122212, and the second author is partially supported by the United States National Science Foundation grant DMS1711952. We also thank the Kathmandu University Hospital and Drs. Rajendra Koju, Biraj Karmacharya and Archana Shrestha for providing the Dhulikhel Heart Study data.

Appendix

A.1. Proof of Proposition 1

Proof. For each fixed $t \in \mathcal{T}$ and any measurable functions f_1, f_2 , we have

$$\begin{aligned} \mathbb{E}[f_1\{Y(t)\}f_2(T) \mid B_0^\top \mathbf{X}] &= \mathbb{E}(\mathbb{E}[f_1\{Y(t)\}f_2(T) \mid \mathbf{X}] \mid B_0^\top \mathbf{X}) = \mathbb{E}(\mathbb{E}[f_1\{Y(t)\} \mid \mathbf{X}]\mathbb{E}[f_2(T) \mid \mathbf{X}] \mid B_0^\top \mathbf{X}) \\ &= \mathbb{E}(\mathbb{E}[f_1\{Y(t)\} \mid B_0^\top \mathbf{X}]\mathbb{E}[f_2(T) \mid B_0^\top \mathbf{X}] \mid B_0^\top \mathbf{X}) = \mathbb{E}[f_1\{Y(t)\} \mid B_0^\top \mathbf{X}]\mathbb{E}[f_2(T) \mid B_0^\top \mathbf{X}]. \end{aligned}$$

Thus, we have $Y(t) \perp\!\!\!\perp T \mid B_0^\top \mathbf{X}$ for all $t \in \mathcal{T}$. $\square$

A.2. Proof of Proposition 2

Proof. Let $\mathcal{S}_{(Y,T) \mid \mathbf{X}}$ denote the central subspace of model (3). First note that the density function can be decomposed into $f_{Y,T}(y, t \mid \mathbf{x}) = f_{Y(t),T}(y, t \mid \mathbf{x}) = f_{Y(t)}(y \mid \mathbf{x}) f_T(t \mid \mathbf{x})$. Thus, we have $\mathcal{S}_{(Y,T) \mid \mathbf{X}} \subseteq \text{span}(B_0)$. Conversely, from the facts

$$f_T(t \mid \mathbf{x}) = \int f_{Y,T}(y, t \mid \mathbf{x}) dy \quad \text{and} \quad f_{Y(t)}(y \mid \mathbf{x}) = \frac{f_{Y,T}(y, t \mid \mathbf{x})}{\int f_{Y,T}(y, t \mid \mathbf{x}) dy},$$

one can easily see that $\text{span}(B_0) \subseteq \mathcal{S}_{(Y,T) \mid \mathbf{X}}$. $\square$

A.3. Proof of Proposition 3

Proof. By paralleling the proof steps of [7, 8], one can show that $\text{MISE}_B(h) = \text{AMISE}_B(h)\{1 + o_p(1)\}$, where

$$\begin{aligned} \text{AMISE}_B(h) &= \int \int \int h^{2q} \mathcal{B}^2(y, t \mid B^\top \mathbf{x}) + (nh^d)^{-1} \mathcal{V}(y, t \mid B^\top \mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}) dF_{Y,T}(y, t), \\ \mathcal{B}(y, t \mid B^\top \mathbf{x}) &= \frac{\mu_{q,K}}{q!} \times \frac{D^q_{B^\top \mathbf{x}} \{F(y, t \mid B^\top \mathbf{x}) f_{B^\top \mathbf{x}}(B^\top \mathbf{x})\} - F(y, t \mid B^\top \mathbf{x}) D^q_{B^\top \mathbf{x}} f_{B^\top \mathbf{x}}(B^\top \mathbf{x})}{f_{B^\top \mathbf{x}}(B^\top \mathbf{x})} (1_d, \dots, 1_d), \\ \mathcal{V}(y, t \mid B^\top \mathbf{x}) &= \frac{\sigma_K^{2d} F(y, t \mid B^\top \mathbf{x}) \{1 - F(y, t \mid B^\top \mathbf{x})\}}{f_{B^\top \mathbf{x}}(B^\top \mathbf{x})}, \end{aligned}$$

$F_{\mathbf{X}}(\mathbf{x})$ is the marginal distribution function of $\mathbf{X}$, $u_{q,K} = \int v^q K(v) dv$, $\sigma_K^2 = \int K^2(v) dv$, D^q denotes the q th order differentiation in tensor form, and $1_d = (1, \dots, 1)^\top$ with dimension d . Thus, when $h \rightarrow 0$ and $nh^d \rightarrow \infty$, the last two terms of (5) converge to zero and are dominated by the first two terms. Since $b_0^2(B) \geq 0$ and the equality holds when $\text{span}(B)$ is a joint sufficient dimension reduction subspace, the minimum of predicted risk occurs when $\text{span}(B) \supseteq \text{span}(B_0)$ for large enough n . Finally, when $h = c_d n^{-1/(2q+d)}$ with

$$c_d = \left[d \int \int \int \mathcal{V}(y, t | B^T \mathbf{x}) dF_{\mathbf{x}}(\mathbf{x}) dF_{Y, T}(y, t) / \left\{ 2q \int \int \int \mathcal{B}^2(y, t | B^T \mathbf{x}) dF_{\mathbf{x}}(\mathbf{x}) dF_{Y, T}(y, t) \right\} \right]^{1/(2q+d)},$$

AMISE $_B(h)$ has minimum

$$\int \int c_d^{2q} \mathcal{B}^2(y, t | B^T \mathbf{x}) + c_d^{-d} \mathcal{V}(y, t | B^T \mathbf{x}) dF_{\mathbf{x}}(\mathbf{x}) dF_{Y, T}(y, t) n^{-2q/(d+2q)},$$

which is increasing in d as $n \rightarrow \infty$, it is then obvious that the prediction risk attains its minimum when $B = B_0$ and $h = c_{d_0} n^{-1/(2q+d_0)}$, which is the optimal bandwidth that minimize AMISE $_{B_0}(h)$. $\square$

A.4. Proof of Theorem 1

To derive the large-sample properties of $\hat{\mu}(t)$, we need the following assumptions:

A1. $\partial_{(t, B_0^T \mathbf{x})}^r m(t, \mathbf{x}; B_0)$, $\partial_{(t, B_0^T \mathbf{x})}^r f_T(t | B_0^T \mathbf{x})$, and $\partial_{\mathbf{u}^T B^T \mathbf{x}}^r f_{B^T \mathbf{x}}(\mathbf{u})$ are Lipschitz continuous in t

with the Lipschitz constant independent of $\mathbf{x}$.

A2. $\inf_{(t, \mathbf{x})} f_T(t | B_0^T \mathbf{x}) > 0$.

A3. $h_1 \rightarrow 0$, $nh_1 h_2^{2r} \rightarrow 0$, and $nh_1 h_2^{2d_0} \rightarrow \infty$ as $n \rightarrow \infty$.

A4. $\Pr(\hat{d} = d_0) \rightarrow 1$ and $\| \hat{B} - B_0 \| = o_p\{h_1^{2r} + h_2^{2r} + (nh_1 h_2^{d_0})^{-1}\}$.

Assumptions A1–A3 are similar to those in [17], and A4 is satisfied by the estimator of [12].

Proof. First we establish some asymptotic properties of the kernel smoothing estimators. For $k \in \{0, 1\}$, let

$$\begin{aligned} \hat{m}_k(t, \mathbf{x}; B) &= \frac{1}{n} \sum_{i=1}^n Y_i^k K_{h_1}(T_i - t) \mathcal{K}_{h_2}\{B^T(\mathbf{X}_i - \mathbf{x})\}, \\ m_k(t, \mathbf{x}; B) &= E(Y^k | T = t, B^T \mathbf{X} = B^T \mathbf{x}) f_T(t | B^T \mathbf{x}) f_{B^T \mathbf{x}}(B^T \mathbf{x}). \end{aligned}$$

By Assumption A1 and an application of Theorem II.37 of [23] shows that

$$\sup_{t, \mathbf{x}, B} | \hat{m}_k(t, \mathbf{x}; B) - m_k(t, \mathbf{x}; B) | = O(h_1^r) + O(h_2^r) + o\left\{ \frac{\ln n}{(nh_1 h_2^d)^{1/2}} \right\} \quad (\text{A.1})$$

almost surely. Now we decompose $\hat{\mu}(t) - \mu(t)$ into

$$\frac{1}{n} \sum_{i=1}^n \{\widehat{m}(t, \mathbf{X}_i; \widehat{B}) - \widehat{m}(t, \mathbf{X}_i; B_0)\} + \frac{1}{n} \sum_{i=1}^n \{\widehat{m}(t, \mathbf{X}_i; B_0) - m(t, \mathbf{X}_i; B_0)\}. \quad (\text{A.2})$$

By using a Taylor expansion and Theorem 4 in [12], we see that the first term in (A.2) is

$$\frac{1}{n} \sum_{i=1}^n \partial_{\text{vec}(B)} \widehat{m}(t, \mathbf{X}_i; B_0) \text{vec}(\widehat{B} - B_0) + O_p(\|\widehat{B} - B_0\|^2) = O_p(\|\widehat{B} - B_0\|). \quad (\text{A.3})$$

Also, by using the Taylor expansion and (A.1), we have

$$\begin{aligned} \widehat{m}(t, \mathbf{x}; B) - m(t, \mathbf{x}; B) &= \frac{\widehat{m}_1(t, \mathbf{x}; B) - m_1(t, \mathbf{x}; B)}{m_0(t, \mathbf{x}; B)} \\ &\quad - \frac{m(t, \mathbf{x}; B) \{\widehat{m}_0(t, \mathbf{x}; B) - m_0(t, \mathbf{x}; B)\}}{m_0(t, \mathbf{x}; B)} + O_p\left(h_1^{2r} + h_2^{2r} + \frac{1}{nh_1 h_2^d}\right). \end{aligned}$$

Thus, the second term in (A.2) can be further written as

$$\begin{aligned} &\frac{1}{n^2} \sum_{i,j=1}^n \frac{Y_j K_{h_1}(T_j - t) \mathcal{K}_{h_2}\{B_0^\top(\mathbf{X}_j - \mathbf{X}_i)\} - m_1(t, \mathbf{x}; B_0)}{m_0(t, \mathbf{x}; B_0)} - \frac{1}{n^2} \sum_{i,j=1}^n \frac{m(t, \mathbf{x}; B_0)}{m_0(t, \mathbf{x}; B_0)} [K_{h_1}(T_j - t) \\ &\quad \mathcal{K}_{h_2}\{B_0^\top(\mathbf{X}_j - \mathbf{X}_i)\} - m_0(t, \mathbf{x}; B_0)] + O_p\left(h_1^{2r} + h_2^{2r} + \frac{1}{nh_1 h_2^d}\right) \triangleq U_1 + U_2 \\ &\quad + O_p\left(h_1^{2r} + h_2^{2r} + \frac{1}{nh_1 h_2^d}\right). \end{aligned} \quad (\text{A.4})$$

Note that U_1 and U_2 are second-order U-statistics. A direct calculation shows that the first-order Hájek projections are

$$\widetilde{U}_1 = \frac{1}{n} \sum_{i=1}^n \frac{Y_i K_{h_1}(T_i - t) f_{B_0^\top \mathbf{x}}(B_0^\top \mathbf{x}_i)}{m_0(t, \mathbf{X}_i; B_0)} + O_p\left(\frac{h_1^r}{n^{1/2}}\right) + O_p(h_2^r), \quad (\text{A.5})$$

$$\tilde{U}_2 = \frac{1}{n} \sum_{i=1}^n \frac{m(t, \mathbf{X}_i; B_0) K_{h_1}(T_i - t) f_{B_0^\top \mathbf{X}}(B_0^\top \mathbf{X}_i)}{m_0(t, \mathbf{X}_i; B_0)} + O_p\left(\frac{h_1^r}{n^{1/2}}\right) + O_p(h_2^r). \quad (\text{A.6})$$

Further, from Theorem 6 of [21], the remainder terms between the U-statistics and their first-order Hájek projections are bounded by $O_p\{1 / (nh_1 h_2^d)\}$. Thus, by substituting (A.3)–(A.6) into (A.2), we have

$$\hat{\mu}(t) - \mu(t) = \frac{1}{n} \sum_{i=1}^n \frac{Y_i - m(t, \mathbf{X}_i; B_0)}{f_{T(t|B_0^\top \mathbf{X}_i)}} K_{h_1}(T_i - t) + O_p(|\hat{B} - B_0|) + O_p\left(h_1^{2r} + h_2^{2r} + \frac{1}{nh_1 h_2^d}\right).$$

Together with the asymptotic mean $h_1^r \mathcal{B}(t)$ and variance $\mathcal{V}(t) / (nh_1)$ of the iid representation above and the Central Limit Theorem, Theorem 1 is established. $\square$

A.5. Proof of Theorem 2

The corresponding large-sample properties rely on the smoothness of the following parameter functions, defined for $m \in \{0, 1, 2\}$ by

$$\begin{aligned} f^{[m]}(\mathbf{x}; B) &= \frac{\partial^m}{B^\top \mathbf{X}} [\mathbb{E}\{(\mathbf{X} - \mathbf{x})^{\otimes m} \mid B^\top \mathbf{X} = B^\top \mathbf{x}\} f_{B^\top \mathbf{X}}(B^\top \mathbf{x})], \\ G^{[m]}(y, t \mid \mathbf{x}; B) &= \frac{\partial^m}{B^\top \mathbf{X}} [\Pr(Y \leq t, T \leq t \mid B^\top \mathbf{X} = B^\top \mathbf{x}) \mathbb{E}\{(\mathbf{X} - \mathbf{x})^{\otimes m} \mid B^\top \mathbf{X} = B^\top \mathbf{x}\} f_{B^\top \mathbf{X}}(B^\top \mathbf{x})]. \end{aligned}$$

The partial derivatives $\partial_{\text{vec}(B)}^m \hat{F}(y, t \mid B^\top \mathbf{x})$ can be shown to converge uniformly to

$F^{[m]}(y, t \mid \mathbf{x}; B) = \sum_{\ell=0}^m G^{[\ell]}(y, t \mid \mathbf{x}; B) f^{\ell-m}(\mathbf{x}; B)$, where

$$f^{-1}(\mathbf{x}; B) = -\frac{f^{[1]}(\mathbf{x}; B)}{f_{B^\top \mathbf{X}}^2(B^\top \mathbf{x})}, \quad f^{[-2]}(\mathbf{x}; B) = \frac{2(f^{[1]}(\mathbf{x}; B))^2}{f_{B^\top \mathbf{X}}^3(B^\top \mathbf{x})} - \frac{f^{[2]}(\mathbf{x}; B)}{f_{B^\top \mathbf{X}}^2(B^\top \mathbf{x})}.$$

Based on these properties, we can define the corresponding score vector and information matrix of $\text{CV}(B, h)$, viz.

$$\begin{aligned} S(B) &= \int \int \{\mathbf{1}(Y \leq y, T \leq t) - F(y, t \mid B^\top \mathbf{X})\} F^{[1]}(y, t \mid \mathbf{X}; B) dF_{Y, T}(y, t), \\ V(B) &= \mathbb{E} \left[\int \int \{F^{[1]}(y, t \mid \mathbf{x}; B)\}^{\otimes 2} - \{\mathbf{1}(Y \leq y, T \leq t) - F(y, t \mid B^\top \mathbf{X})\} F^{[2]}(y, t \mid \mathbf{X}; B) dF_{Y, T}(y, t) \right], \end{aligned}$$

where $F_{Y, T}(y, t)$ is the joint distribution of (Y, T) .

To prove Theorem 2, we need the following regularity conditions:

$B1 \partial_{\mathbf{u}}^{q+2} \Pr(Y \leq y, T \leq t \mid B^T \mathbf{X} = \mathbf{u}), \partial_{\mathbf{u}}^{q+m} \mathbb{E}\{(\mathbf{X} - \mathbf{x})^{\otimes m} \mid B^T \mathbf{X} = \mathbf{u}\}$, and $\partial_{B^T \mathbf{x}}^{q+2} f_{B^T \mathbf{x}}(\mathbf{u})$ are Lipschitz continuous in $\mathbf{u}$ with the Lipschitz constants being independent of $(y, t, \mathbf{x}, B)$.

$B2 \inf_{(\mathbf{x}, B)} f_{B^T \mathbf{x}}(B^T \mathbf{x}) > 0$.

$B3$ For $d \geq 1$, there exist $\delta \in (1/(4q), 1/\max(2d+2, d+4))$ and positive constants $h_{\ell d}$ and $h_{u,d}$ such that h falls in the interval $H_{\delta,n} = [h_{\ell d} n^{-\delta}, h_{u,d} n^{-\delta}]$.

$B4 \inf_{B: d < d_0} b_0^2(B) > 0$ and $b_0^2(B) = 0$ if and only if $B = B_0$ when $d = d_0$.

$B5 V(B_{d0})$ is non-singular for $d = d_0$.

The proof steps of Theorem 2 are similar to those of [12]. Here we give an outline and indicate some key properties. As mentioned in [20], the uniqueness of B_0 requires a certain local coordinate system of the Grassmann manifold. Without loss of generality, we assume B is already parameterized in such a way. This is in fact a necessary condition of Assumption B5 and is not an additional assumption. First, we derive the convergence rates of the involved kernel smoothing estimators. Let

$$\begin{aligned} \hat{f}(B^T \mathbf{x}) &= \frac{1}{n} \sum_{i=1}^n \mathcal{K}_h\{\mathbf{X}_i - \mathbf{x}\}, \quad \hat{f}_c^{[m]}(\mathbf{x}; B) = \partial_{\text{vec}(B)}^m \hat{f}(B^T \mathbf{x}) - f^{[m]}(\mathbf{x}; B), \\ \hat{G}_c^{[m]}(y, t \mid \mathbf{x}; B) &= \partial_{\text{vec}(B)}^m \{\hat{F}(y, t \mid B^T \mathbf{x}) \hat{f}(B^T \mathbf{x})\} - G^{[m]}(y, t \mid \mathbf{x}; B). \end{aligned}$$

Moreover, we define the asymptotic linear representation for $\partial_{\text{vec}(B)}^m \hat{F}(y, t \mid B^T \mathbf{x})$ with $m \in \{0, 1\}$, as follows:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \xi_i(y, t \mid \mathbf{x}; B) &= \frac{\hat{G}_c^{[0]}(y, t \mid \mathbf{x}; B)}{f_{B^T \mathbf{x}}(B^T \mathbf{x})} - \frac{F(y, t \mid B^T \mathbf{x}) \hat{f}_c^{[0]}(\mathbf{x}; B)}{f_{B^T \mathbf{x}}(B^T \mathbf{x})}, \\ \frac{1}{n} \sum_{i=1}^n \xi_i^{[1]}(y, t \mid \mathbf{x}; B) &= \frac{\hat{G}_c^{[1]}(y, t \mid \mathbf{x}; B)}{f_{B^T \mathbf{x}}(B^T \mathbf{x})} - \frac{F(y, t \mid B^T \mathbf{x}) \hat{f}_c^{[1]}(\mathbf{x}; B)}{f_{B^T \mathbf{x}}(B^T \mathbf{x})} - \frac{f^{[1]}(\mathbf{x}; B) \hat{G}_c^{[0]}(y, t \mid \mathbf{x}; B)}{f_{B^T \mathbf{x}}^2(B^T \mathbf{x})} + \frac{\{2F(y, t \mid B^T \mathbf{x}) f^{[1]}(\mathbf{x}; B) - G^{[1]}(y, t \mid \mathbf{x}; B)\} \hat{f}_c^{[0]}(\mathbf{x}; B)}{f_{B^T \mathbf{x}}^2(B^T \mathbf{x})} \end{aligned}$$

By verifying the Euclidean class properties through Lemma 2.12 of [22], Lemma 2.14 of [22] and Theorem II.37 of [23], we get

$$\begin{aligned} \sup_{\mathbf{x}, B} \|\hat{f}_c^{[m]}(\mathbf{x}; B)\| &= O(h^q) + o\left\{\frac{\ln n}{n^{1/2} h^{(d+m)/2}}\right\} \text{ a.s.}, \text{ and } \sup_{\mathbf{x}, B} \|\hat{G}_c^{[m]}(y, t \mid \mathbf{x}; B)\| = O(h^q) \\ &+ o\left\{\frac{\ln n}{n^{1/2} h^{(d+m)/2}}\right\} \text{ a.s.}, \end{aligned}$$

for all $m \in \{0, 1, 2\}$. By applying the Taylor expansion theorem and the results above, we can further ensure from Conditions B2–B3 that

$$\sup_{\mathbf{x}, B} \left| \hat{F}(y, t | B^T \mathbf{x}) - \frac{1}{n} \sum_{i=1}^n \xi_i(y, t | \mathbf{x}; B) \right| = o_p(n^{-1/2}) \text{ and}$$

$$\sup_{\mathbf{x}, B} \left| \partial_{\text{vec}(B)} \hat{F}(y, t | B^T \mathbf{x}) - \frac{1}{n} \sum_{i=1}^n \xi_i^{[1]}(y, t | \mathbf{x}; B) \right| = o_p(n^{-1/2}).$$

The second step is to derive the uniform consistency of $\text{CV}(B, h)$ to

$\text{ECV}(B, h) = \sigma_0^2 + b_0^2(B) + \text{AMISE}_B(h)$. This can be done by paralleling the arguments of [12]

and we have

$$\begin{cases} \sup_{B, h} |\text{CV}(B, h) - \text{ECV}(B, h)| / \text{AMISE}_B(h) = o(1) \text{ a.s.} & \text{for } \text{span}(B) \supseteq \text{span}(B_0), \\ \sup_{B, h} |\text{CV}(B, h) - \text{ECV}(B, h)| / b_0(B) \text{AMISE}_B^{1/2}(h) = O(1) \text{ a.s.} & \text{for } \text{span}(B) \not\supseteq \text{span}(B_0). \end{cases}$$

(A.7)

Now let $\text{DCV}(B, h) = \text{CV}(B, h) - \text{ECV}(B, h)$. From the inequalities

$$\begin{aligned} 1 &= \Pr\{\text{CV}(\hat{B}, \hat{h}) \leq \text{CV}(B_0, h_0)\} \\ &\leq \Pr\{b_0^2(\hat{B}) < \varepsilon\} + \Pr\{b_0^2(\hat{B}) \geq \varepsilon, \text{DCV}(\hat{B}, \hat{h}) + \text{DCV}(B_0, h_0) \geq \text{ECV}(\hat{B}, \hat{h}) - \text{ECV}(B_0, h_0)\} \\ &\leq \Pr\{b_0^2(\hat{B}) < \varepsilon\} + \Pr\left\{b_0^2(\hat{B}) \geq \varepsilon, \frac{\text{DCV}(\hat{B}, \hat{h})}{b_0(\hat{B})} + \frac{\text{DCV}(B_0, h_0)}{\varepsilon^{1/2}} \geq \varepsilon^{1/2} - \frac{\text{AMISE}_{B_0}(h_0)}{\varepsilon^{1/2}}\right\} \end{aligned}$$

(A.8)

for any $\varepsilon > 0$, (A.7) further implies that, for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr\{b_0^2(\hat{B}) < \varepsilon\} = 1. \quad (\text{A.9})$$

At the same time, by taking $\varepsilon = \inf_{\{B: d < d_0\}} b_0^2(B) / 2$ and using Boole's inequality again, we have

$$\begin{aligned} & \Pr \left\{ b_0^2(\hat{B}) < \inf_{\{B: d < d_0\}} b_0^2(B) / 2 \right\} \leq \Pr \left\{ b_0^2(\hat{B}) < \inf_{\{B: d < d_0\}} b_0^2(B) / 2, \hat{d} < d_0 \right\} \\ & + \Pr \left\{ b_0^2(\hat{B}) < \inf_{\{B: d < d_0\}} b_0^2(B) / 2, \hat{d} \geq d_0 \right\} \leq \Pr(\hat{d} \geq d_0). \end{aligned}$$

Hence, (A.9) further implies that

$$\lim_{n \rightarrow \infty} \Pr(\hat{d} \geq d_0) = 1.$$

The third step is to derive the asymptotic properties of $\hat{B}_d$ for $d \rightarrow d_0$. Let $h_{d,0} = c_d n^{-2q/(2q+d)}$. By virtue of the minimizer $(\hat{B}_d, \hat{h}_d)$ of $\text{CV}(B, h)$ under fixed d and Boole's inequality, a similar inequality as (A.8) can also be derived as follows:

$$\begin{aligned} 1 & \leq \Pr\{b_0^2(\hat{B}_d) < \varepsilon\} + \Pr \left\{ b_0^2(\hat{B}_d) \geq \varepsilon, \frac{\text{DCV}(\hat{B}_d, \hat{h}_d)}{b_0(\hat{B}_d)} + \frac{\text{DCV}(B_{d,0}, h_{d,0})}{\varepsilon^{1/2}} \geq \varepsilon^{1/2} \right. \\ & \left. - \frac{\text{AMISE}_{B_{d,0}}(h_{d,0})}{\varepsilon^{1/2}} \right\} \quad (\text{A.10}) \end{aligned}$$

for any $\varepsilon > 0$. Also note that $b_0^2(\hat{B}_d) \geq \varepsilon$ implies that $\text{span}(\hat{B}_d) \not\supseteq \text{span}(B_0)$. Eq. (A.7) and

Condition B3 further imply that $\text{DCV}(\hat{B}_d, \hat{h}_d) / b_0(\hat{B}_d) = O_p\{\text{AMISE}_{\hat{B}_d}^{1/2}(\hat{h}_d)\} \rightarrow 0$ and

$\text{DCV}(B_{d,0}, h_{d,0}) / \varepsilon^{1/2} = o_p\{\text{AMISE}_{B_{d,0}}(h_{d,0})\} \rightarrow 0$ as $n \rightarrow \infty$. Coupled with the fact that $\text{AMISE}_{B_{d,0}}(h_{d,0}) \rightarrow 0$ as $n \rightarrow \infty$, the second inequality in the second term of (A.10) becomes $0 \leq \varepsilon^{1/2}$ as $n \rightarrow \infty$, and, hence, one has, for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr\{b_0^2(\hat{B}_d) < \varepsilon\} = 1.$$

Since $\text{span}(\hat{B}_d) \supseteq \text{span}(B_0)$ implies $b_0^2(\hat{B}_d) = 0$, we now consider the case when

$\text{span}(\hat{B}_d) \not\supseteq \text{span}(B_0)$ and, hence, $\hat{B}_d \xrightarrow{p} B_{d,0}$. By the first-order Taylor expansion of

$\text{vec}(B) \text{CV}(B, h)$ at $B = B_{d,0}$ and $\partial_{\text{vec}(B)} \text{CV}(\hat{B}_d, \hat{h}_d) = 0$, one has

$$[I_{pd} + V^{-1}(B_{d,0})\{\partial_{\text{vec}(B)}^2 \text{CV}(\hat{B}_d^*, \hat{h}_d) - V(B_{d,0})\}]n^{1/2} \text{vec}(\hat{B}_d - B_{d,0}) = n^{1/2} V^{-1}(B_{d,0}) \partial_{\text{vec}(B)} \text{CV}(B_{d,0}, \hat{h}_d)$$

where $\text{vec}(\hat{B}_d^*)$ lies on the line segment between $\text{vec}(\hat{B}_d)$ and $\text{vec}(B_{d,0})$. Coupled with the facts that

$$n^{1/2} \partial_{\text{vec}(B)} \text{CV}(B_{d,0}, \hat{h}_d) \rightsquigarrow \mathcal{N}[0, E\{S^{\otimes 2}(B_{d,0})\}] \quad \text{and} \quad \|\partial_{\text{vec}(B)}^2 \text{CV}(\hat{B}_d^*, \hat{h}_d) - V(B_{d,0})\| \xrightarrow{p} 0,$$

we have, as $n \rightarrow \infty$,

$$n^{1/2} \text{vec}(\hat{B}_d - B_{d,0}) \rightsquigarrow \mathcal{N}[0, V^{-1}(B_{d,0}) E\{S^{\otimes 2}(B_{d,0})\} V^{-1}(B_{d,0})]. \quad (\text{A.11})$$

Using a Taylor expansion technique and $\partial_{\text{vec}(B)} b_0^2(B_{d,0}) = 0$, we further deduce from Eq. (A.11) that

$$b_0^2(\hat{B}_d) = \frac{1}{2} \text{vec}(\hat{B}_d - B_{d,0})^T \partial_{\text{vec}(B)}^2 b_0^2(\hat{B}_d^*) \text{vec}(\hat{B}_d - B_{d,0}) = O_p(n^{-1})$$

for $d = d_0$ and hence $b_0^2(\hat{B}) = O_p(n^{-1})$. Coupled with Condition B3, it further implies that

$$b_0^2(\hat{B}) / \text{AMISE}_{\hat{B}}(\hat{h}) = o_p(1).$$

The next step is to show the consistency of $(\hat{B}, \hat{h})$. Let

$$\begin{aligned} E_0 &= \{b_0^2(\hat{B}) < \ln n / n, \hat{h} \in H_{1/(2q+\hat{d}), n}, \hat{d} = d_0\}, \quad E_1 = \{b_0^2(\hat{B}) \geq \ln n / n\}, \\ E_2 &= \{\hat{d} < d_0\}, E_3 = \{b_0^2(\hat{B}) < \ln n / n, \hat{d} \geq d_0, \hat{h} \in H_{\hat{d}, n} \text{ with } \hat{d} \neq 1/(2q+\hat{d})\}, \\ E_4 &= \{b_0^2(\hat{B}) < \ln n / n, \hat{h} \in H_{1/(2q+\hat{d}), n}, \hat{d} > d_0\}, \\ \text{and } E_{\text{con}} &= \{\text{DCV}(\hat{B}, \hat{h}) + \text{DCV}(B_0, h_0) \geq \text{ECV}(\hat{B}, \hat{h}) - \text{ECV}(B_0, h_0)\}. \end{aligned}$$

By the minimizer $(\hat{B}, \hat{h})$ of $\text{CV}(B, h)$ and Boole's inequality, one has

$$1 = \Pr\{\text{CV}(\hat{B}, \hat{h}) \leq \text{CV}(B_0, h_0)\} \leq \Pr(E_0) + \sum_{m=1}^4 \Pr(E_{\text{con}} \cap E_m).$$

The consistency follows by showing $\Pr(E_{\text{con}} \cap E_m) \rightarrow 0$ as $n \rightarrow \infty$ for all $m \in \{1, \dots, 4\}$, which implies that

$$\lim_{n \rightarrow \infty} \Pr(E_0) = 1.$$

Finally, the asymptotic normality is ensured by (A.12) and (A.11).

References

- [1]. Aerts M, Claeskens G, Hens N, Molenberghs G, Local multiple imputation, *Biometrika* 89 (2002) 375–388.
- [2]. Cheng PE, Nonparametric estimation of mean functionals with data missing at random, *J. Amer. Statist. Assoc* 89 (1994) 81–87.
- [3]. Dua S, Bhuker M, Sharma P, Dhall M, Kapoor S, et al., Body mass index relates to blood pressure among adults, *North Amer. J. Med. Sci* 6 (2014) 89.
- [4]. Galvao AF, Wang L, Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment, *J. Amer. Statist. Assoc* 110 (2015) 1528–1542.
- [5]. Gray HL, Schucany WR, *The Generalized Jackknife Statistic*, Marcel Dekker, Inc., New York, 1972.
- [6]. Hall P, Li Q, Racine JS, Nonparametric estimation of regression functions in the presence of irrelevant regressors, *The Review of Economics and Statistics* 89 (4) (2007) 784–789.
- [7]. Härdle WK, Hall P, Marron JS, How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc* 83 (1988) 86–101.
- [8]. Härdle WK, Marron JS, Optimal bandwidth selection in nonparametric regression function estimation, *Ann. Statist* 13 (1985) 1465–1481.
- [9]. Hill JL, Bayesian nonparametric modeling for causal inference, *J. Comput. Graph. Statist* 20 (2011) 217–240.
- [10]. Hirano K, Imbens GW, The propensity score with continuous treatments, in: *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, Wiley, Chichester, 2004, pp. 73–84.
- [11]. Huang M-Y, Chan KCG, Joint sufficient dimension reduction and estimation of conditional and average treatment effects, *Biometrika* 104 (2017) 583–596. [PubMed: 29430034]
- [12]. Huang M-Y, Chiang C-T, An effective semiparametric estimation approach for the sufficient dimension reduction model, *J. Amer. Statist. Assoc* 112 (2017) 1296–1310.
- [13]. Imai K, van Dyk DA, Causal inference with general treatment regimes: Generalizing the propensity score, *J. Amer. Statist. Assoc* 99 (2004) 854–866.
- [14]. Imbens GW, Nonparametric estimation of average treatment effects under exogeneity: A review, *The Review of Economics and Statistics* 86 (2004) 4–29.
- [15]. Jaeckel LA, *The infinitesimal jackknife*, Bell Telephone Laboratories, 1972.
- [16]. Kaufman JS, Asuzu MC, Mufunda J, Forrester T, Wilks R, Luke A, Long AE, Cooper RS, Relationship between blood pressure and body mass index in lean populations, *Hypertension* 30 (1997) 1511–1516. [PubMed: 9403575]
- [17]. Kennedy EH, Ma Z, McHugh MD, Small DS, Nonparametric methods for doubly robust estimation of continuous treatment effects, *J. R. Stat. Soc. Ser. B Stat. Methodol* 79 (2017) 1229–1245.
- [18]. Li K-C, Sliced inverse regression for dimension reduction, *J. Amer. Statist. Assoc* 86 (1991) 316–342.
- [19]. Luo W, Zhu Y, Ghosh D, On estimating regression-based causal effects using sufficient dimension reduction, *Biometrika* 104 (2017) 51–65.
- [20]. Ma Y, Zhu L, Efficient estimation in sufficient dimension reduction, *Ann. Statist* 41 (2013) 250–268.
- [21]. Nolan D, Pollard D, *U-processes: Rates of convergence*, *Ann. Statist* 15 (1987) 780–799.
- [22]. Pakes A, Pollard D, Simulation and the asymptotics of optimization estimators, *Econometrica* 57 (1989) 1027–1057.
- [23]. Pollard D, *Convergence of Stochastic Processes*, Springer, New York, 1984.
- [24]. Ruppert D, Wand MP, Multivariate locally weighted least squares regression, *Ann. Statist* 22 (1994) 1346–1370.
- [25]. Schucany WR, Gray H, Owen D, On bias reduction in estimation, *J. Amer. Statist. Assoc* 66 (1971) 524–533.

- [26]. Schucany WR, Sommers JP, Improvement of kernel type density estimators, J. Amer. Statist. Assoc 72 (1977) 420–423.
- [27]. Shrestha A, Koju RP, Beresford SA, Chan KCG, Karmacharya BM, Fitzpatrick AL, Food patterns measured by principal component analysis and obesity in the Nepalese adult, Heart Asia 8 (2016) 46–53.
- [28]. Yin X, Li B, Sufficient dimension reduction based on an ensemble of minimum average variance estimators, Ann. Statist 39 (2011) 3392–3416.
- [29]. Zhu L-P, Zhu L-X, Feng Z-H, Dimension reduction in regressions through cumulative slicing estimation, J. Amer. Statist. Assoc 105 (2010) 1455–1466.

Figure 1:

Estimated treatment effect functions. The bold solid line represents the proposed estimator and the thin solid lines represent 95% pointwise confidence intervals, the dotted line represents the naive estimator without adjusting for covariates, the dashed line represents the Kennedy et al. estimator, and the dot-dashed line represents the Galvao–Wang estimator.

Table 1:

The performance of the estimated joint central subspace.

Model	n	$\hat{d}$						Error	
		0	1	2	3	4	5	Mean	S.D.
M1	100	0.000	0.073	0.712	0.196	0.019	0.000	0.823	0.5701
	200	0.000	0.017	0.874	0.102	0.007	0.000	0.448	0.4082
	400	0.000	0.000	0.956	0.044	0.000	0.000	0.212	0.2376
M2	100	0.000	0.935	0.065	0.000	0.000	0.000	0.070	0.2471
	200	0.000	0.966	0.034	0.000	0.000	0.000	0.036	0.1815
	400	0.000	0.991	0.009	0.000	0.000	0.000	0.010	0.0946
M3	100	0.000	0.213	0.738	0.046	0.003	0.000	0.492	0.4543
	200	0.000	0.030	0.938	0.032	0.000	0.000	0.206	0.2593
	400	0.000	0.000	0.979	0.021	0.000	0.000	0.100	0.1587
M4	100	0.017	0.003	0.956	0.024	0.000	0.000	0.205	0.3056
	200	0.002	0.000	0.996	0.002	0.000	0.000	0.064	0.1023
	400	0.001	0.000	0.999	0.000	0.000	0.000	0.027	0.0634

Table 2:

The error measures of the estimated treatment effects.

Model	n	$\hat{\mu}(t)$		$\hat{\mu}_{\text{GPS}}(t)$		$\hat{\mu}_{\text{DR}}(t)$		$\hat{\mu}_{\text{SE}}(t)$	
		$\hat{B}^T \mathbf{X}$	$\mathbf{X}$	$\hat{B}^T \mathbf{X}$	$\mathbf{X}$	$\hat{B}^T \mathbf{X}$	$\mathbf{X}$	$\hat{B}^T \mathbf{X}$	$\mathbf{X}$
M1	100	0.739	0.764	0.806	0.816	0.761	0.805	0.911	0.972
	200	0.626	0.677	0.753	0.763	0.634	0.703	0.775	0.890
	400	0.516	0.579	0.671	0.689	0.508	0.590	0.590	0.758
M2	100	0.643	5.601	8.445	8.641	0.608	5.411	6.653	7.837
	200	0.564	3.837	7.765	8.021	0.539	3.725	6.546	7.790
	400	0.528	1.929	6.943	7.311	0.510	1.884	6.475	7.960
M3	100	0.515	0.619	0.623	0.623	0.499	0.617	0.484	0.643
	200	0.393	0.516	0.519	0.519	0.355	0.514	0.323	0.533
	400	0.367	0.458	0.461	0.461	0.317	0.456	0.230	0.469
M4	100	0.080	0.236	0.467	0.475	0.065	0.205	0.252	0.301
	200	0.051	0.204	0.419	0.445	0.039	0.177	0.210	0.270
	400	0.035	0.170	0.333	0.373	0.027	0.152	0.212	0.272

Table 3:

Estimated coefficients of the linear index for the food patterns dataset.

Covariate	Coefficient	S.E.	Covariate	Coefficient	S.E.
Age	1		Red meat	0.1059	0.02233
Sex	0.1958	0.02458	Fish	0.0835	0.01255
Whole grains	0.0170	0.02361	Milk	−0.0587	0.01713
Refined grains	0.2736	0.02433	Milk products	−0.0365	0.03479
Lentils	0.3886	0.02753	Fast food	−0.0829	0.02253
Vegetable oils	−0.0456	0.01799	Processed cereals	0.3975	0.02415
Solid fats	−0.1212	0.01038	Noodles	−0.1273	0.01906
Fatty foods	−0.0646	0.01823	Salty snacks	−0.0285	0.01709
Vegetables	−0.1257	0.02912	Soda drinks	0.0408	0.02827
Fruits	0.1533	0.02786	Tea/coffee	−0.1042	0.01775
Potatoes	−0.1912	0.02059	Sweets	−0.0971	0.02202
Nuts	0.2208	0.01808	Alcohol	0.2682	0.01339
Poultry	0.2460	0.01745			

Table 4:

L_p -distance between estimators with perturbed bandwidth parameters with the original estimators.

Method / L_p -distance	$0.9 \times \hat{h}_2$			$1.1 \times \hat{h}_2$		
	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Proposed	0.392	0.099	0.016	0.280	0.070	0.011
Galvao–Wang	1.707	0.454	0.303	2.179	0.564	0.466
Kennedy et al.	1.444	0.856	0.786	0.845	0.533	0.297

Method / L_p -distance	$0.5 \times \hat{h}_2$			$1.5 \times \hat{h}_2$		
	$p = 0$	$p = 1$	$p = 2$	$p = 0$	$p = 1$	$p = 2$
Proposed	4.427	1.384	3.476	1.130	0.390	0.210
Galvao–Wang	1.294	0.364	0.262	3.040	0.743	0.906
Kennedy et al.	31.539	18.743	372.69	3.526	1.690	3.313