Nonparametric estimation of mediation effects with a general treatment

Lukang Huang, Wei Huang, Oliver Linton, and Zheng Zhang

ABSTRACT
To investigate causal mechanisms, causal mediation analysis decomposes the total treatment effect into the natural direct and indirect effects. This article examines the estimation of the direct and indirect effects in a general treatment effect model, where the treatment can be binary, multi-valued, continuous, or a mixture. We propose generalized weighting estimators with weights estimated by solving an expanding set of equations. Under some sufficient conditions, we show that the proposed estimators are consistent and asymptotically normal. Specifically, when the treatment is discrete, the proposed estimators attain semiparametric efficiency bounds. Meanwhile, when the treatment is continuous, the convergence rates of the proposed estimators are slower than $N^{-1/2}$; however, they are still more efficient than those constructed from the true weighting function. A simulation study reveals that our estimators exhibit satisfactory finite-sample performance, while an application shows their practical value.

ARTICLE HISTORY
Received 22 January 2024
Accepted 22 January 2024

KEYWORDS
Covariate balancing; direct and indirect effects; general treatment; semiparametric efficiency.

1. Introduction

One essential goal of program evaluation and scientific research is to understand why and how a treatment variable affects the potential outcomes of interest, going beyond the estimation of the average treatment effects (ATE). For example, when assessing the effect of participating in an academic and vocational institute on criminal behavior, investigators might want to separate the program’s direct effects on criminal behavior, such as integrity or discipline, from its effects relayed through the employment chances, which may indirectly affect the criminal behavior. In this regard, causal mediation analysis plays an important role by decomposing the total treatment effect into the natural direct effect and the indirect effects mediated through an intermediate variable, called the mediator (Pearl, 2001; Robins and Greenland, 1992). Such an approach has been widely used in a number of disciplines in the medical and social sciences (see e.g., Baron and Kenny, 1986; Imai et al., 2011; VanderWeele, 2015).

A fundamental problem of treatment effect analysis is that, although a randomized trial is the golden standard to identify the treatment effects, it is often unavailable or even unethical. More usually, in practice, treatments are randomly assigned to the individuals based on some features that also relate to the mediator and outcome of interest, thus causing the confounding issue. Such features are called confounders. The literature on causal mediation analysis, which can identify the direct and indirect effects from data with confounding, has rapidly grown over the last decade, producing abundant approaches and extensions.
An important nonparametric identifiability condition called the sequential ignorability assumption, which establishes a minimum set of assumptions required to identify the direct and indirect treatment effects regardless of the statistical models used, is commonly imposed (Hue et al., 2018; Imai, Keele, and Tingley, 2010; Imai, Keele, and Yamamoto, 2010). Most of the approaches focus on the binary treatment where an individual either receives or does not receive the treatment (see, e.g., Chan et al., 2016; Hsu et al., 2018; Huber, 2014; Imai, Keele, and Yamamoto, 2010; Joffe et al., 2007; Liu et al., 2022; VanderWeele, 2009; VanderWeele and Vansteelandt, 2010). For example, based on the sequential ignorability assumption, Hsuet al. (2018) proposed nonparametric estimation of the binary natural direct and indirect effects by weighting observations using the inverse of estimated propensity scores.

While binary treatment effects have been extensively studied, the estimation of continuous treatment effects has also drawn considerable attention (see Ai et al., 2021; Colangelo and Lee, 2020; Dong et al., 2019; Fong et al., 2018; Galva and Wang, 2015; Hirano and Imbens, 2004; Huang et al., 2022; Imai and van Dyk, 2004; Kennedy et al., 2017 among others). However, all of the above studies focus on the total treatment effect rather than analyzing the causal mediation effects. In recent work, Huber et al. (2020) studied the identification and estimation of the natural direct and indirect effects when the treatment variable is continuous. Their estimators are constructed from weighting observations using the estimated marginal density of the treatment and the inverse of two estimated conditional densities, namely the conditional density of the treatment given the confounder and that given both the confounder and the mediator. Singh et al. (2021) identified the natural direct and indirect effects by the g-formula and estimated them using a Reproducing Kernel Hilbert Space approach. They showed their estimators are consistent and provided the rate of convergence, which achieves the minimax rate.

In many applications, the treatment variable can be complex, being a mixture of discrete and continuous elements, or even multidimensional, comprising both discrete and continuous components. For example, when evaluating the impact of an academic and vocational training program on criminal behavior, the causal effect might depend not only on participation in the program but also on the duration of this participation. Additionally, it could depend on multi-factors such as the duration and number of training courses undertaken.

The contribution of the work is three-fold. First, we propose a general framework for estimating the direct and indirect effects, which unifies the binary, multi-valued, and continuous treatments as well as the mixture of discrete and continuous treatments. It can also be easily adapted to multidimensional treatment. Specifically, we integrate the sequential ignorability assumption for discrete treatments from Hsu et al. (2018) with that for continuous treatments in Huber et al. (2020), and use a sieve method to estimate the treatment effect parameters. This synthesis results in a more general form of identifying and estimating direct and indirect effects, enabling nonparametric estimation for discrete, continuous, mixed, and multidimensional treatments.

Second, we demonstrate that the moment balancing method for estimating the stabilized weights, originally proposed for unconditional continuous treatment effects by Ai et al. (2021), can improve the mediation analysis in the existing framework. Specifically, Huber et al.’s (2020) mediation identification requires estimation of a weight that consists of the marginal density of the treatment and the inverse of two conditional densities, namely the conditional density of the treatment given the confounder and that given both the confounder and the mediator. Huber et al. (2020) estimate this weight using a ratio of kernel density estimators. However, it is known that the inverse probability weighting is sensitive to estimated densities in the denominator, potentially leading to extreme weights and unstable results (see e.g., Ai et al., 2021; Fong et al., 2018). By reformulating the weight estimation, we apply the moment balancing idea to estimate the weight integrally, significantly reducing the likelihood of extreme weights, and enhancing result accuracy.

Third, we verify that the proposed estimators of the direct and indirect effects are consistent and asymptotically converge to normal distributions after appropriate normalization. Specifically, the proposed estimators attain the semiparametric efficiency bounds when the treatment is discrete. Meanwhile, when the treatment is continuous, the convergence rates of the proposed nonparametric estimators are slower than \(N^{-1/2}\). However, we show that our proposed estimators achieve an asymptotic variance
smaller than or equal to those constructed from the true weighting function, provided that the tuning parameters are selected to ensure equivalent convergence rates. The equality is only attained when the conditional mean of the observed outcome given the observed treatment, mediators, and confounders is always zero.

The remainder of the article is structured as follows. Section 2 establishes the basic framework and Section 3 presents the estimation procedure. In Section 4, we derive the asymptotic properties of the proposed estimators and their efficiency results. Section 5 constructs the consistent estimators for the asymptotic variances based on plug-in approaches. In Section 6, we propose data-driven approaches for selecting the tuning parameters. Subsequently, Section 7 reports the results of a simulation study and Section 8 applies the proposed estimation method to analyze the effect of Job Corps, an educational and vocational training program, on criminal activity. Finally, Section 9 presents the study’s conclusions.

2. Basic framework

Let \( T \) denote the observed treatment variable with support \( \mathcal{T} \subseteq \mathbb{R}^1 \), where \( \mathcal{T} \) is either a discrete set, a continuum, or a mixture of discrete and continuum subsets. Let \( f_T(t) \) denote a probability distribution function of \( T \) at point \( t \); that is, it stands for the probability density if \( T \) is continuous at \( t \) and the probability mass if \( T \) is discrete at \( t \). Under the standard framework of causal inference, let \( M(t) \in \mathcal{M} \subseteq \mathbb{R}^s \), for some positive integer \( s \), denote a potential mediating variable that represents the value of the mediator if the treatment variable is equal to \( t \in \mathcal{T} \). Similarly, let \( Y(t, m) \in \mathbb{R} \) denote the potential outcome if one receives treatment \( t \) and mediator \( m \). The observed mediator and response are denoted by \( M := M(T) \) and \( Y := Y(T, M(T)) \), respectively. Define \( \mu(t, t') := \mathbb{E}[Y(t, M(t'))] \) for any \((t, t') \in \mathcal{T} \times \mathcal{T} \), then \( \mu(t, t) - \mu(t', t) \) is the average treatment effect (ATE) caused by the change in the treatment level from \( t' \) to \( t \). In particular, if \( \mathcal{T} = \{0, 1\} \), then \( \mu(1, 1) - \mu(0, 0) \) is the binary ATE studied by Hahn (1998), Hirano et al. (2003), Chan et al. (2016), Hsu and Lai (2018), and Ai et al. (2022) among others; if \( \mathcal{T} = \{0, 1, \ldots, J\} \) for some \( J \in \mathbb{N} \), then \( \mu(t, t) - \mu(t', t) \) is the multi-valued ATE studied by Cattaneo (2010) and Lee (2018) among others; and if the treatment variable is continuous, then \( \mu(t, t) - \mu(t', t') \) is the continuous ATE studied by Imai and van Dyk (2004), Hirano and Imbens (2004), Lee and Lemieux (2010), Colangelo and Lee (2020), and Ai et al. (2021) among others.

A primary goal of causal mediation analysis is to decompose the ATE into the average natural indirect effect and the average natural direct effect, for any \( t \neq t' \in \mathcal{T} \),

\[
\mu(t, t) - \mu(t', t') = \{\mu(t, t) - \mu(t', t)\} + \{\mu(t', t) - \mu(t', t')\},
\]

where \( \mu(t, t) - \mu(t', t) \) is the average natural direct effect representing the average difference if the treatment variable changes from \( t' \) to \( t \) while the mediator is held constant at \( M(t) \); \( \mu(t', t) - \mu(t', t') \) is the average natural indirect effect representing the average difference if the mediator value changes from \( M(t') \) to \( M(t) \) while holding the treatment variable constant at \( t' \). Therefore, this decomposition enables researchers to quantitatively explore the extent to which the treatment and mediator contribute to the treatment effect.

Note from (2.1) that all the ATE, average natural indirect effect, and the average natural direct effect depend only on \( \mu : \mathcal{T} \times \mathcal{T} \mapsto \mathbb{R} \). Our goal is then reduced to estimating \( \mu \) on \( \mathcal{T} \times \mathcal{T} \). Due to the confounding issue, the potential outcome \( Y(t, M(t')) \) and the potential mediator \( M(t) \) are not observed for all \((t, t') \in \mathcal{T} \times \mathcal{T} \). To address this identification problem, most studies impose a selection on the observable condition. Specifically, let \( X \in \mathcal{X} \subseteq \mathbb{R}^r \) be a vector of observed covariates, for some positive integer \( r \). We maintain the following sequential ignorability assumption imposed on the treatment and mediator assignment (Hsu et al., 2018; Huber, 2014; Huber et al., 2020; Imai, Keele, and Yamamoto, 2010).
In this section, we introduce a two-step nonparametric estimation of Sequential Ignorability.

Remark 1. Let \( \{Y(t', m), M(t)\} \perp T | X = x \) for all \((t, t', m, x) \in \mathcal{T} \times \mathcal{T} \times \mathcal{M} \times \mathcal{X}\);

\( Y(t', m) \perp M(t) | (T = t, X = x) \) for all \((t, t', m, x) \in \mathcal{T} \times \mathcal{T} \times \mathcal{M} \times \mathcal{X}\),

where the conditional probability functions (in the sense as \( f_T \)) satisfy \( f_{T|X}(t|x) > 0 \) and \( f_{T|M,X}(t|m, x) > 0 \) for all \((t, m, x) \in \mathcal{T} \times \mathcal{M} \times \mathcal{X}\).

When \( M \) is multidimensional, Assumption 1 needs to hold for each element in \( M \). Under Assumptions 1, we show in Appendix A that \( \mu(t, t') \) can be identified as follows:

\[
\mu(t, t') = \mathbb{E} \left[ \frac{\pi_{M,X}(T, M, X)}{\pi_{M,X}(T + \delta, M, X)} \cdot \pi_X(T + \delta, X) | Y = t \right],
\]

where, for \((t, t') \in \mathcal{T} \times \mathcal{T},

\[
\delta := t' - t, \quad \pi_Z(t, Z) := \frac{f_T(t)}{f_T(t|Z)} \text{ for } Z \in \{X, (M, X)\}.
\]

In the particular case of \( \delta = 0, \mu(t, t) = \mathbb{E} [\pi_X(T, X) | Y = t] \) is the dose-response function studied in Ai et al. (2021).

Let \( \{T_i, M_{i1}, X_i, Y_i\}_{i=1}^{N} \) denote an independent and identically distributed (i.i.d.) sample of observations drawn from the joint distribution of \((T, M, X, Y)\). If \( \pi_X(T, X) \) and \( \pi_{M,X}(T, M, X) \) were known, \( \mu(t, t') \) can be estimated by the nonparametric series regression (Newey, 1994, 1997):

\[
\tilde{\mu}(t, t') := \left[ \sum_{i=1}^{N} \frac{\pi_{M,X}(T_i, M_{i1}, X_i)}{\pi_{M,X}(T_i + \delta, M_i, X_i)} \cdot \pi_X(T_i + \delta, X_i) u_{K_0}(T_i)^{\top} \right] \times \left[ \sum_{i=1}^{N} u_{K_0}(T_i) u_{K_0}(T_i)^{\top} \right]^{-1} u_{K_0}(t),
\]

where \( u_{K_0}(T) = (u_{K_0,1}(T), \ldots, u_{K_0,K_0}(T))^{\top} \) is a prespecified basis function with dimension \( K_0 \in \mathbb{N} \). However, both \( \pi_X(T, X) \) and \( \pi_{M,X}(T, M, X) \) are unknown in practice and need to be replaced by some estimates.

Remark 1. When \( T \) is a continuous treatment variable, Huber et al. (2020) identify \( \mu(t, t') \) as

\[
\mu(t, t') = \mathbb{E}[Y(t, M(t'))] = \lim_{h \to 0} \mathbb{E} \left[ \frac{K_h(T - t)}{f_{T|M,X}(t|M, X)} \cdot \frac{f_{T|M,X}(t'|M, X)}{f_{T|X}(t'|X)} \cdot Y \right],
\]

where \( K_h(x) = K(x/h)/h \) with \( K(\cdot) \) a univariate kernel function. Our proposed identification differs in two significant ways. First, (2.2) does not depend on the kernel weighting, \( K_h(T - t) \), allowing integration with any nonparametric regression methods for discrete, continuous, or mixture \( T \). Indeed, our proposed sieve estimator in (2) accommodates all the scenarios. Second, we address the selection bias using three stabilized weights, \( \pi_{M,X}(T, M, X), \pi_{M,X}(T + \delta, M, X), \text{ and } \pi_X(T + \delta, M, X) \), rather than the conditional densities used in (2.4). This approach, leveraging the moment balancing idea from Ai et al. (2021) (see Section 3 for details), ensures stability, especially when the densities, \( f_{T|M,X}(t|M, X) \) and \( f_{T|X}(t'|X) \), are close to zero. However, Ai et al. (2021) does not consider the mediation and only requires one stabilized weight, \( \pi_X(T, X) \), to identify their causal parameter.

3. Estimation

In this section, we introduce a two-step nonparametric estimation of \( \mu(t, t') \) across all pairs \((t, t') \in \mathcal{T} \times \mathcal{T} \) based on the identification (2.2). In the first step, shown in Section 3.1, we propose a unified framework for estimating the weighting functions, \( \pi_X \) and \( \pi_{M,X} \). Subsequently, in the second step shown
in Section 3.2, we regress the estimated \( \frac{\pi_{M X}(T, M, X) \cdot \pi_X(T + \delta, X) \cdot Y}{\pi_{M X}(T + \delta, M, X)} \) against \( T \) to obtain our final estimator of \( \mu(t, t') \). For this estimation, we employ a sieve method suitable for discrete, continuous, or a mixture of discrete and continuous treatment types.

### 3.1. Unified framework for estimating the weighting functions

This section proposes the estimators for \( \pi_X \) and \( \pi_{M X} \). A naive estimator of \( \pi_X \) can be the ratio of some estimated \( f_T \) and \( f_{T|X} \). However, such a ratio estimator is very sensitive to small values of estimated \( f_{T|X} \) and lead to undesirable results. To mitigate this problem, Ai et al. (2021) propose estimating \( \pi_X \) as a whole.

We note that, using the definition of \( \pi_Z \) in (2.2), their approach can be applied to estimate \( \pi_X \) and \( \pi_{M X} \) in the same framework. Note that for \( Z \in \{ X, (M, X) \} \),

\[
\mathbb{E} [\pi_Z(T, Z) u(T) v(Z)] = \mathbb{E} [u(T)] \cdot \mathbb{E} [v(Z)],
\]

(3.1)

holds for any suitable functions \( u(T) \) and \( v(Z) \). Using Ai et al. (2021 Theorem 2), one can show that (3.1) identifies \( \pi_Z \). Eq. (3.1) suggests a possible way of estimating \( \pi_Z \); however, it implies an infinite number of equations, which is impossible to solve using a finite sample of observations. To overcome this difficulty, they approximate the infinite-dimensional function space using a sequence of finite-dimensional sieve spaces. Specifically, let \( u_k(T) = (u_{k1}(T), \ldots, u_{k_1 k_1}(T))^\top \) and \( v_k(Z) = (v_{k_1}(Z), \ldots, v_{k_2 k_2}(Z))^\top \) be specified basis with dimensions \( k_1 \in \mathbb{N} \) and \( k_2 \in \mathbb{N} \), respectively, and let \( K_Z := k_1 \cdot k_2 \). The functions \( u_k(T) \) and \( v_k(Z) \) are approximation sieves that can approximate any suitable functions \( u(T) \) and \( v(Z) \) arbitrarily well (see Chen, 2007 for discussions on the sieve approximation).

Then (3.1) implies that

\[
\mathbb{E} \left[ \pi_Z(T, Z) u_{k_1}(T) v_{k_2}(Z)^\top \right] = \mathbb{E} [u_{k_1}(T)] \cdot \mathbb{E} [v_{k_2}(Z)]^\top.
\]

(3.2)

Following Ai et al. (2021), \( \pi_Z(T_i, Z_i) \), for \( i = 1, \ldots, N \), can be estimated by the \( \hat{\pi}_i \)'s, which are the following maximizer of an entropy, subject to the sample analog of (3.2):

\[
\begin{aligned}
\{ \hat{\pi}_i \}_{i=1}^N &= \arg \max \left\{ -N^{-1} \sum_{i=1}^N \pi_i \log \pi_i \right\} \\
\text{subject to} \quad &\frac{1}{N} \sum_{i=1}^N \pi_i u_{k_1}(T_i) v_{k_2}(Z_i)^\top = \left\{ \frac{1}{N} \sum_{i=1}^N u_{k_1}(T_i) \right\} \\
&\left\{ \frac{1}{N} \sum_{i=1}^N v_{k_2}(Z_i)^\top \right\}.
\end{aligned}
\]

(3.3)

Note that by including a constant of one in the sieve bases \( u_{k_1}(T) \) and \( v_{k_2}(Z) \), (3.3) implies that \( N^{-1} \sum_{i=1}^N \hat{\pi}_i = 1 \). Moreover,

\[
\max \left\{ -N^{-1} \sum_{i=1}^N \pi_i \log \pi_i \right\} = -\min \left\{ \sum_{i=1}^N (N^{-1} \pi_i) \cdot \log \left( \frac{N^{-1} \pi_i}{N^{-1}} \right) \right\},
\]

that is, the entropy maximization problem is equivalent to the minimization of the Kullback-Leibler divergence between \( \{N^{-1} \pi_i\}_{i=1}^N \) and the uniform empirical distribution \( \{N^{-1}\}_{i=1}^N \). Such an entropy serves a reasonable metric in the sense that the \( \{N^{-1} \pi_i\}_{i=1}^N \) can be treated as a discrete probability distribution. This comes from the observation that \( \pi_Z(T_i, Z_i) \) is positive and \( \mathbb{E}(\pi_Z(T_i, Z_i)) = 1 \). Moreover, the formulation in (3.3) also guarantee the empirical counterparts, the \( \pi_i \)'s are all positive and \( \sum_{i=1}^N (N^{-1} \pi_i) = 1 \).

To estimate \( \pi_Z(T, Z) \), we use the dual solution to (3.3) showed by Ai et al. (2021):

\[
\hat{\pi}_{K_Z}(T, Z) := \rho' \left\{ u_{k_1}(T)^\top \Lambda_{k_1 \times k_2} v_{k_2}(Z) \right\},
\]

(3.4)
where \( \rho'(\nu) = \exp(-\nu - 1) \) is the first derivative of \( \rho(\nu) = -\exp(-\nu - 1) \), and \( \hat{\Lambda}_{k_1 \times k_Z} \) is the maximizer of the strictly concave function \( \hat{G}_{k_1 \times k_Z} \) defined by

\[
\hat{G}_{k_1 \times k_Z}(\Lambda) := \frac{1}{N} \sum_{i=1}^{N} \rho \left\{ u_{k_1}(T_i) \Lambda v_{k_Z}(Z_i) \right\} \left\{ \frac{1}{N} \sum_{i=1}^{N} u_{k_1}(T_i) \right\}^\top \Lambda \left\{ \frac{1}{N} \sum_{j=1}^{N} v_{k_Z}(Z_j) \right\} \tag{3.5}
\]

The first-order condition of (3.5) implies that \( (\hat{\pi}_{k_Z}(T_i, Z_i))_{i=1}^{N} \) satisfies the sample analog of (3.2). Such restrictions improve the robustness of the estimation, with extreme weights being unlikely to be obtained. The concavity of (3.5) enables us to easily obtain the solution via the Gauss-Newton algorithm. To ensure consistent estimation of \( \pi_{\mathbf{Z}}(T, \mathbf{Z}) \) for continuous \( T \) (resp. \( \mathbf{Z} \)), the dimension of the bases, \( k_1 \) (resp. \( k_Z \)), shall increase as the sample size increases.

### 3.2. Final estimator of \( \mu(t, t') \)

With (2.2) and the estimated weighting functions \( \hat{\pi}_{k_X} \) and \( \hat{\pi}_{k_M, X} \), we define the estimator of \( \mu(t, t') \) the sieve regression estimator:

\[
\hat{\mu}(t, t') := \left[ \sum_{i=1}^{N} \frac{\hat{\pi}_{k_M, X}(T_i, M_i, X_i)}{\hat{\pi}_{k_M, X}(T_i + \delta, M_i, X_i)} \cdot \hat{\pi}_{k_X}(T_i + \delta, X_i) Y_i u_{k_0}(T_i) \right]^\top u_{k_0}(t). \tag{3.6}
\]

**Remark 2.** When \( T \) is a binary treatment taking values in \( T = \{0, 1\}, u_{k_1}(T) = u_{k_0}(T) = (1(T = 0), 1(T = 1))^\top \) with \( k_1 = k_0 = 2 \). In this case, \( \{ \mu(t, t') : (t, t') \in T \times T \} \) is a finite discrete set and estimable at a rate of \( N^{-1/2} \). The semiparametric efficiency bounds and efficient estimation for \( \{ \mu(1, 1), \mu(0, 0) \} \) are presented in Hahn (1998), Hirano et al. (2003), and Chan et al. (2016). The semiparametric efficiency bounds and efficient estimation for the mediation causal effects \( \{ \mu(1, 0), \mu(0, 1) \} \) are presented in Tchetgen Tchetgen and Shpitser (2012) and Hsu et al. (2018). Specifically, Hsu et al. (2018) construct the estimator for \( \mu(1, 0) \) based on the following representation:

\[
\mu(1, 0) = \mathbb{E}[Y|\{1, M(0)\}] = \mathbb{E}\left[ \frac{1(T = 1)}{\mathbb{P}(T = 1|M, X)} \cdot \frac{\mathbb{P}(T = 0|M, X)}{\mathbb{P}(T = 0|X)} \cdot Y \right],
\]

with the generalized propensity score functions \( \mathbb{P}(T = 1|M, X) \) and \( \mathbb{P}(T = 1|X) \) estimated by the nonparametric series logit regression.

In the discrete setting, that is, \( T = \{0, 1, \ldots, J\} \) for some \( J \in \mathbb{N} \), our proposed estimators can be obtained by taking \( u_{k_1}(T) = u_{k_0}(T) = (1(T = 0), 1(T = 1), \ldots, 1(T = J))^\top \) with \( k_1 = k_0 = J + 1 \). We show in the next section that the proposed estimators have \( \sqrt{N} \)-asymptotic normality, provided some conditions on \( k_Z \), and attain the semiparametric efficiency bounds.

In a special case of a mixture of discrete and continuous treatment variable, where \( T = 0 \) with positive probability and \( T > 0 \) continuous with positive probability, our estimators can be obtained by taking \( u_{k_1}(T) = u_{k_0}(T) = (1(T = 0), u_{k_0}(T)|1 - (1(T = 0)|), \ldots, u_{k_0, k_0}(T)|1 - (1(T = 0)|))^\top \).

When \( T \) is continuous or a mixture, we show in Theorem 2 that our proposed estimator converges to normal distributions slower than \( N^{-1/2} \), but has an asymptotic variance smaller than or equal to that of \( \hat{\mu} \) in (2.2) that uses the true \( \pi_X \) and \( \pi_{M, X} \), provided the convergence rates are the same. The equivalence is only taken when \( \mathbb{E}(Y|T, M, X) = 0 \).

**Remark 3.** For continuous \( T \), Huber et al. (2020) propose a nonparametric weighted kernel (Nadaraya-Watson) type estimator for \( \mu(t, t') \) based on (2.4). The conditional density functions \( f_{T|M,X} \) and \( f_{T|X} \) are
estimated through the kernel method. Their estimators are shown to be asymptotically normal. In such a case, using our proposed estimators of the weights \( \pi_Z \), a kernel (Nadaraya-Watson) type estimator for \( \mu(t, t') \), an alternative to (3.6), can also be constructed:

\[
\hat{\mu}_h(t, t') = \frac{\sum_{i=1}^N \hat{R}_{KM,X}(T_i, T_i + \delta, M_i, X_i) \cdot \hat{\pi}_{KZ}(T_i + \delta, X_i) K_h(T_i - t) Y_i}{\sum_{i=1}^N \hat{R}_{KM,X}(T_i, T_i + \delta, M_i, X_i) \cdot \hat{\pi}_{KZ}(T_i + \delta, X_i) K_h(T_i - t)},
\]

(3.7)

where \( \hat{R}_{KM,X}(t, t', m, x) := \hat{\pi}_{KM,X}(t, m, x)/\hat{\pi}_{KM,X}(t', m, x) \) for \( (t, t', m, x) \in \mathcal{T} \times \mathcal{T} \times \mathcal{M} \times \mathcal{X} \). We establish its asymptotic properties in Appendix C, where we show that it attains a convergence rate as a standard Nadaraya-Watson estimator and it achieves a faster convergence rate and smaller asymptotic variance than Huber et al. (2020), unless Huber et al. (2020) use a same bandwidth for all the estimations of \( f_{T|M,X}, f_{T|X} \), and \( \mu(t, t') \). However, in practice, since these three quantities are conditional on different set of variables, namely \( \{M, X\}, X, \) and \( T \), respectively, using a same bandwidth does not guarantee good finite sample performance. Indeed, Huber et al. (2020) suggest using different bandwidths for these estimations, and thus their \( \mu(t, t') \) estimator cannot achieve the optimal convergence rate (see footnote 7 in Huber et al., 2020). Our proposed estimator can achieve a faster rate and smaller variance, because our proposed \( \hat{\pi}_Z \) efficiently incorporates the additional information available in the form of \( \pi_Z \) as per (3.1) (see more discussions in Hirano et al., 2003).

Remark 4. In addition to the natural direct effect \( \mathbb{E}[Y(t, M(t))] - \mathbb{E}[Y(t', M(t))] \) for \( t \neq t' \) studied in this article, the controlled direct effect (CDE) defined as \( \mathbb{E}[Y(t, m)] - \mathbb{E}[Y(t', m)] \) for \( t \neq t' \) and a fixed value \( m \) of the mediator is also of interest in the literature, see Goetgeluk et al. (2008), VanderWeele (2009), Hong et al. (2015) for examples. The proposed method in this article can also be adapted to estimate \( \mathbb{E}[Y(t, m)] \) as well as CDE; indeed, by Hong et al. (2015 Theorem 2), \( \mathbb{E}[Y(t, m)] \) can be identified as the following density ratio weighting form:

\[
\mathbb{E}[Y(t, m)] = \mathbb{E} \left[ \frac{f_{T|M}(T, M)}{f_{T|M,X}(T, M|X)} Y \bigg| T = t, M = m \right],
\]

which is similar to the identification of \( \mathbb{E}[Y(t, M(t'))] \) in Eq. (2.2). Therefore, we can apply the same procedure of estimating \( \pi_Z(T, Z) \) and \( \mathbb{E}[Y(t, M(t'))] \) to obtain the estimators of \( f_{T,M}(T, M)/f_{T|M,X}(T, M|X) \) and \( \mathbb{E}[Y(t, m)] \), respectively.

4. Large sample properties

This section studies the asymptotic properties of the proposed estimator \( \hat{\mu}(t, t') \). The convergence rates for \( \hat{\pi}_{KZ}(\cdot, Z), Z \in \{X, (M, X)\} \), are implied directly by the results in Ai et al. (2022). We also recall these results in Appendix B. To facilitate the presentation our main results, we introduce the following notations: \( \Phi_{K_{0} \times K_{0}} := \mathbb{E}[u_{K_0}(T) u_{K_0}^{T}(T)] \),

\[
d_{K_{0},j}(T, M, X, Y; \delta) := \text{IF}_{\pi_{X,i}} \left\{ \delta \frac{\pi_{MX}(T, M, X)}{\pi_{MX}(T + \delta, M, X)} \cdot u_{K_0}(T) Y \right\}
\]

\[
+ \text{IF}_{\pi_{MX,i}} \left\{ 0 \frac{\pi_{X}(T + \delta, X)}{\pi_{MX}(T + \delta, M, X)} \cdot u_{K_0}(T) Y \right\}
\]

\[
- \text{IF}_{\pi_{MX,i}} \left\{ \delta \frac{\pi_{MX}(T, M, X) \pi_{X}(T + \delta, X)}{\pi_{MX}^{2}(T + \delta, M, X)} \cdot u_{K_0}(T) Y \right\}
\]

\[
- \mathbb{E} \left[ \frac{\pi_{MX}(T_i, M_i, X_i) \pi_{X}(T_i + \delta, X_i)}{\pi_{MX}(T_i + \delta, M_i, X_i)} \cdot u_{K_0}(T_i) Y_i \bigg| T_i \right]
\]

\[
+ \mathbb{E} \left[ \frac{\pi_{MX}(T_i, M_i, X_i) \pi_{X}(T_i + \delta, X_i)}{\pi_{MX}(T_i + \delta, M_i, X_i)} \cdot u_{K_0}(T_i) Y_i \right],
\]

(4.1)
for any $\delta \geq 0$ and $i = 1, \ldots, N$, where the IF$_{Z,i}$’s are the i.i.d. mean zero influence functions for $\hat{F}_Z$, $Z \in \{X, (M, X)\}$, defined in (B.2) in Appendix B, and

$$V_{t'} := \mathbb{E} \left[ u_{K_0}(t) \Phi_{K_0 \times K_0}^{-1} d_{K_0,i}(T, M, X, Y; \delta) \right]^2$$

$$= u_{K_0}(t) \cdot \Phi_{K_0 \times K_0}^{-1} : \mathbb{E} \left[ d_{K_0,i}(T, M, X, Y; \delta) \cdot d_{K_0,i}^T(T, M, X, Y; \delta) \right] \cdot \Phi_{K_0 \times K_0}^{-1} \cdot u_{K_0}(t). \quad (4.2)$$

The following conditions are maintained throughout this article.

**Assumption 2.** For every fixed $t' \in \mathcal{T}$, there exist a $\gamma^* \in \mathbb{R}^{K_0}$ and a positive constant $\beta > 0$ such that

$$\sup_{t \in \mathcal{T}} |\mu(t, t') - (\gamma^*)^T u_{K_0}(t)| = O \left( K_0^{-\beta} \right).$$

**Assumption 3.** The eigenvalues of $\mathbb{E} \left[ u_{K_0}(T) u_{K_0}^T(T) \right]$ and $\mathbb{E} \left[ d_{K_0,i}(T, M, X, Y; \delta) \cdot d_{K_0,i}^T(T, M, X, Y; \delta) \right]$ are bounded away from zero and infinity uniformly with respect to $K_0 \in \mathbb{N}$.

Assumption 2 requires the sieve approximation error of the function $\mu(\cdot, t')$ to shrink at a polynomial rate. This condition is satisfied for a variety of sieve basis functions. For example, if $T$ is discrete, then the approximation error is zero for sufficiently large $K_0$; thus, in this case, Assumption 2 is satisfied with $\beta = +\infty$. If $T$ is continuous or a mixture, the polynomial rate $\beta$ depends positively on the smoothness of $\mu(t, t')$ in $t$; indeed, for power series and $B$-splines, $\beta$ is the smoothness of $\mu(t, t')$ in $t$ (Chen, 2007 Section 2.3.1). We show that the convergence rate of the estimated $\mu(t, t')$ is bounded by this polynomial rate. Assumption 3 essentially ensures the variance of the estimator is nondegenerate. Under these conditions and Assumptions 4–6 presented in Appendix B, we establish the following two theorems, which hold whenever the treatment is continuous, discrete, or mixed:

**Theorem 1.** Under Assumptions 1–3 and 4–6 presented in Appendix B, we have

1. (Convergence Rates)

$$\int_{\mathcal{T}} |\hat{\mu}(t, t') - \mu(t, t')|^2 dF_T(t) = O_p \left( \left\{ \frac{K_0}{N} + K_0^{-2\beta} \right\}^2 + \zeta(K_X)^2 \left\{ \frac{K_X}{N} + K_X^{-2\alpha_X} \right\} \right)$$

$$+ \zeta(K_{M,X})^2 \left\{ \frac{K_{M,X}}{N} + K_{M,X}^{-2\alpha_{M,X}} \right\},$$

$$\sup_{t \in \mathcal{T}} |\hat{\mu}(t, t') - \mu(t, t')| = O_p \left( \zeta_1(K_0) \left\{ \left( \sqrt{\frac{K_0}{N} + K_0^{-\beta}} \right) + \zeta(K_X) \left( \sqrt{\frac{K_X}{N} + K_X^{-\alpha_X}} \right) \right\} \right.$$

$$\left. + \zeta(K_{M,X}) \left( \sqrt{\frac{K_{M,X}}{N} + K_{M,X}^{-\alpha_{M,X}}} \right) \right\}.$$

hold for every fixed $t' \in \mathcal{T}$, where $\zeta(K_Z)$, for $Z = \{X, (M, X)\}$, and $\zeta_1(K_0)$ are defined in Assumption 6 of Appendix B.

2. (Asymptotic Normality) Suppose $\sqrt{N}K_0^{-\beta} \rightarrow 0$. Then, for any fixed $t, t' \in \mathcal{T}$,

$$\sqrt{N} \left\{ \hat{\mu}(t, t') - \mu(t, t') \right\} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \phi_{tt'}(Y_i, T_i, M_i, X_i; \delta) + o_p(1),$$

where

$$\phi_{tt'}(Y_i, T_i, M_i, X_i; \delta) = u_{K_0}^T(t) \Phi_{K_0 \times K_0}^{-1} d_{K_0,i}(T, M, X, Y; \delta).$$
Thus, we have
\[ \sqrt{N}V_{tt'}^{-1/2} \{ \hat{\mu}(t, t') - \mu(t, t') \} \xrightarrow{d} N(0, 1), \]
and \( V_{tt'} = \text{const} \times \| u_{K_0}(t) \|^2 = O(K_0) \) for some positive constant \( \text{const} \).

The proof of Theorem 1 is presented in Section S3 in supplemental material. In addition to the convergence rate and asymptotic normality, we also provide the asymptotic linear expansion of \( \sqrt{N} \{ \hat{\mu}(t, t') - \mu(t, t') \} \). This can help conduct statistical inference, as we can approximate the limiting distribution of our estimator by adopting the exchangeable bootstrap method (Chernozhukov et al., 2013; Donald and Hsu, 2014; Huang et al., 2022). Using Theorem 1, the asymptotic results of the estimated direct and indirect effects can be easily obtained. In particular, note from Remark 1 and Assumption 2 that if the treatment is discrete, \( K_0 \) is a constant independent of \( N \) and \( \beta = +\infty \); thus \( \hat{\mu}(t, t') - \mu(t, t') \) attains a \( \sqrt{N} \)-asymptotic normality. If the treatment is continuous or a mixture, and \( u_{K_0}(t) \) is a sieve series, then the convergence rate of \( \hat{\mu}(t, t') - \mu(t, t') \) is slower than \( N^{-1/2} \); see Remark 5 for more discussion on the curse of dimensionality that arises from \( M \) and \( X \).

The next theorem shows that our proposed estimator of mean potential outcome \( \hat{\mu}(t, t') \) is asymptotically more efficient than the oracle estimator \( \hat{\mu}(t, t') \) in (2) that uses the true \( \pi_X \) and \( \pi_{M,X} \).

**Theorem 2.** Suppose \( \sqrt{N}K_0^{-\beta} \rightarrow 0. \) Under Assumptions 1–3, for any fixed \( t, t' \in T \),
\[ \sqrt{N}V_{tt'}^{-1/2} \{ \hat{\mu}(t, t') - \mu(t, t') \} \xrightarrow{d} N(0, 1), \]
where \( \hat{V}_{tt'} \) is the asymptotic variance satisfying \( \hat{V}_{tt'} \geq V_{tt'} \).

It is known that in the estimation of average treatment effects with a binary treatment, the inverse probability weighting (IPW) estimator constructed from nonparametrically estimated propensity score is more efficient than that constructed by using the true one, see Hahn (1998), Hirano et al. (2003), and Chen et al. (2008). Theorem 2 establishes the similar result for the mediation effects with a general treatment. The proof of Theorem 2 is presented in Section S4 in supplemental material, where we also derive the asymptotic linear expansion of \( \sqrt{N}V_{tt'}^{-1/2} \{ \hat{\mu}(t, t') - \mu(t, t') \} \) and the detailed asymptotic variance \( \hat{V}_{tt'} \). Theorem 2 implies the efficiency of our estimator over the oracle one for discrete, continuous, and mixed treatments. Specifically, when the treatment variable is discrete, we further prove the corollary below that our proposed estimators attain the semiparametric efficiency bounds. The proof is in Section S5 in supplemental material.

**Corollary 1.** Suppose that the treatment variable \( T \) is discrete with values in \( T = \{0, 1, \ldots, J\} \), where \( J \geq 1 \) is an positive integer. Under Assumptions 1–3 and 4–6 presented in Appendix B, we have
\[ \sqrt{N} \{ \hat{\mu}(t, t') - \mu(t, t') \} \xrightarrow{d} N(0, V_{tt'}), \]
where \( V_{tt'} = \mathbb{E} \left[ S^2_{\mu(t,t')} \right] \) and
\[ S_{\mu(t,t')} = \frac{1 \{ T = t \} f_{M \mid T,X}(M \mid T = t', X)}{f_{T \mid X}(t \mid X) f_{M \mid T,X}(M \mid T = t, X)} \{ Y - \mathbb{E}(Y \mid X, M, T = t) \}
+ \frac{1 \{ T = t' \} \{ \mathbb{E}(Y \mid X, M, T = t) - \eta(t, t', X) \}}{f_{T \mid X}(t' \mid X)}
+ \eta(t, t', X) - \mu(t, t'), \]
where
\[ \eta(t, t', X) = \int \mathbb{E}(Y \mid X, M = m, T = t)f_{M \mid T,X}(m \mid T = t', X)dm, \]
and \( S_{\mu(t,t')} \) is equivalent to the efficient influence function in Hsu et al. (2018).
Remark 5. Note from Assumptions 5 and 6 in Appendix B that the more continuous or mixed components in the covariate set $X$ or the mediator set $M$, the harder our estimation of $\pi_Z$ would be, for $Z \in \{X, (M, X)\}$. This is a curse of dimensionality of our fully nonparametric estimation. To resolve this issue, it is common to reduce the dimension. A solution is to use an index model (see e.g., Fan and Gijbels, 1996; Lewbel and Linton, 2007). Specifically, we may assume that the random treatment assignment $T$ depends on $Z$ through a function known up to a finite number of parameters $\theta$, $s_{Z,\theta} : Z \mapsto \mathbb{R}$, in the sense that $f_{T|Z}(t|z) = f_{T|s_{Z,\theta}(Z)}$ for all $(t, z) \in T \times Z$, and for $Z \in \{X, (M, X)\}$ that has a large dimension of continuous or mixed components. Then $\pi_Z$ can be identified by

$$E[\pi_Z(T, Z)u(T)v(s_{Z,\theta}(Z))] = E[u(T)] \cdot E[v(s_{Z,\theta}(Z))]$$

for all suitable functions $u$ and $v$. However, the moment equation does not identify $\theta$, we need to first estimate $\theta$. One possible way is to estimate $\theta$ by the maximum likelihood (MLE) with a nonparametric estimator of $f_{T|s_{Z,\theta}(Z)}(t|s)$, $f_{T|s_{Z,\theta}(Z)}(t|s)$. That is,

$$\hat{\theta} := \arg \max_{\theta} \sum_{i=1}^{N} \log \left( f_{T|s_{Z,\theta}(Z)}(T_i|s_{Z,\theta}(Z_i)) \right).$$

For example, we can take the kernel density estimator

$$\hat{f}_{T|s_{Z,\theta}(Z)}(t|s) = \frac{\sum_{i=1}^{N} K_{h_1}(T_i - t)K_{h_2}(s - s_{Z,\theta}(Z_i))}{\sum_{i=1}^{N} K_{h_2}(s_{Z,\theta}(Z_i))},$$

where $K_{h}(x) = K(x/h)h$ with $K$ a pre-specified kernel function, and $h_1, h_2$ are the rule of thumb bandwidths for $T$ and $s_{Z,\theta}(Z)$, respectively. Then with $\hat{\theta}$ and the moment Eq. (4.3), we can estimate $\pi_Z$ using the same way as in Section 3. Since the index model $s_{Z,\theta}$ maps $Z$ to a univariate space, both the nonparametric estimator of $f_{T|s_{Z,\theta}(Z)}(t|s)$ and the sieve approximation of (4.3) are two-dimensional nonparametric estimators, reducing the curse of dimensionality.

5. Variance estimation

Using the expression (4.2), we propose a plug-in estimator for $V_{tt'}$. For $i = 1, \ldots, N$ and $\delta \geq 0$, let $\hat{d}_{K_0,i}$ defined in Section B.1 in Appendix be an estimator of $d_{K_0,i}$ defined in (4.1). Then, a consistent estimator of $V_{tt'}$ is given by

$$\hat{V}_{tt'} := u_{K_0}^\top(t) \hat{\Phi}_{K_0 \times K_0}^{-1} \left[ \frac{1}{N} \sum_{i=1}^{N} \hat{d}_{K_0,i}(T, M, X, Y; \delta) \hat{d}_{K_0,i}^\top(T, M, X, Y; \delta) \right] \cdot \hat{\Phi}_{K_0 \times K_0}^{-1} u_{K_0}^\top(t),$$

where $\hat{\Phi}_{K_0 \times K_0} = N^{-1} \sum_{i=1}^{N} u_{K_0}(T_i)u_{K_0}^\top(T_i)$. From Proposition 1 and Theorem 1, we have $\sup_{(t,z) \in T \times Z} |\pi_Z(t, z) - \hat{\pi}_Z(t, z)| = o_p(1)$ and $|\hat{\mu}(t, t) - \mu(t, t)| \to 0$. With these results, the consistency of $\hat{V}_{tt'}$ follows from standard arguments in Chen (2007).

6. Selecting the smoothing parameters

The proposed estimator $\hat{\mu}(t, t')$ (resp. $\hat{\mu}_h(t, t')$) in (3.6) (resp. (3.7)) involves tuning parameters $k_1, k_2, k_{M, X}$, and $K_0$ (resp. $h$). If the treatment $T$ is discrete, $k_1$ and $K_0$ can be determined in the way described in Remark 1, and the kernel-type estimator $\hat{\mu}_h(t, t')$ is not applicable. When the confounders $X$ or the mediators $M$ are discrete, the corresponding sieve bases and parameters $k_X$ and $k_{M, X}$ can also be determined in that way. Thus, in this section, we focus on proposing a data-driven method for choosing the parameters for continuous $T, X$ and $M$.

Note that our estimators $\hat{\mu}(t, t')$ and $\hat{\mu}_h(t, t')$ for continuous $T$ are nonparametric regression-type estimators. For such estimators, the smoothing parameters are usually selected by minimizing certain cross-validation (CV) criteria that approximates the mean squared error (MSE) of the estimator.
Although simultaneously selecting all the smoothing parameters using one CV criteria can better approximate the MSE, it is too time-consuming, given so many parameters. Thus, we propose to choose them separately.

First, we propose a method to choose $K_Z$ for $\hat{\pi}_{K_Z}$, where $Z \in \{X, (M, Z)\}$, inspired by the least square CV idea for choosing the bandwidth of a kernel density estimator (see e.g., Fan and Gijbels, 1996). Notice that a weighted integrated squared error (WISE) of the estimated weighting functions can be written in the following way:

$$
\int \{\hat{\pi}_{K_Z}(T, Z) - \pi_Z(T, Z)\}^2 f_{T,Z}(t, z) dt dz = \int \{\hat{\pi}_{K_Z}(T, Z)\}^2 f_{T,Z}(t, z) dt dz
$$

where the definition of $\pi_Z$, have $\pi_Z(t, z)f_{T,Z}(t, z) = f_T(t)f_Z(z)$. This result suggests that minimizing the WISE w.r.t. the number of sieve basis $k_1$ and $k_Z$ is equivalent to minimizing

$$
\int \{\hat{\pi}_{K_Z}(T, Z)\}^2 f_{T,Z}(t, z) dt dz - 2 \int \hat{\pi}_{K_Z}(t, z)f_T(t)f_Z(z) dt dz.
$$

With a finite sample, we can then choose $k_1$ and $k_Z$ by minimizing the following least square CV, which is an empirically fully accessible analog of the above criteria

$$CV(k_1, k_Z) = \frac{1}{N} \sum_{i=1}^{N} \{\hat{\pi}_{K_Z}(T_i, Z_i)\}^2 - \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i} \hat{\pi}_{K_Z}(T_i, Z_i),$$

over a candidate set of $k_1$ and $k_Z$.

To select the parameter $K_0$ in $\hat{\mu}(t, t')$, we rewrite $\hat{\mu}(t, t')$ as $\hat{\mu}_{K_0}(t, t')$ for now. Then, we obtain $K_0$ by using a leave-one-out CV (see Li and Racine (2006 Section 15.2)),

$$\hat{K}_0 = \arg \min_{K_0} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\hat{\pi}_{K_M,X}(T_i, M_i, X_i)\hat{\pi}_{K_M,X}(T_i + \delta, M_i, X_i)}{\hat{\pi}_{K_M,X}(T_i + \delta, M_i, X_i)} - \hat{\mu}_{K_0}^{(-i)}(T_i, T_i) \right\}^2 \right\},$$

where $\hat{K}_Z = \hat{K}_1 \cdot \hat{K}_Z$ for $Z \in \{X, (M, X)\}$, and $\hat{\mu}_{K_0}^{(-i)}(T_i, T_i + \delta)$ is computed as $\hat{\mu}_{K_0}(T_i, T_i + \delta)$ but without using $\{T_i, X_i, M_i, Y_i\}$. For $\hat{\mu}_h(t, t')$, we need to choose the bandwidth $h$. We can also apply a leave-one-out CV method for choosing $h$:

$$\hat{h} = \arg \min_{h} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{\hat{\pi}_{K_M,X}(T_i, M_i, X_i)\hat{\pi}_{K_M,X}(T_i + \delta, M_i, X_i)}{\hat{\pi}_{K_M,X}(T_i + \delta, M_i, X_i)} - \hat{\mu}_h^{(-i)}(T_i, T_i + \delta) \right\}^2 \right\},$$

where $\hat{\mu}_h^{(-i)}(T_i, T_i + \delta)$ is computed as the series estimator $\hat{\mu}_h(T_i, T_i + \delta)$ but without using $\{T_i, X_i, M_i, Y_i\}$.

Apart from the CV method, the rule of thumb bandwidth for kernel methods (see e.g. Silverman, 1986) is also available, which sacrifices a small amount of accuracy but is less time-consuming. We experimented with the rule of thumb bandwidth in our numerical studies and obtained satisfactory results. Specifically, the rule of thumb bandwidth $h = C \cdot \text{sd}(T) \cdot N^{-1/5}$, where $\text{sd}(T)$ is the standard deviation of $T$ and $C$ is a constant depending on the kernel function. For example, $C = 2.34$ for second-order Epanechnikov kernels, $C = 3.03$ for fourth-order Epanechnikov kernels, and $C = 1.06$ for the Gaussian kernel. However, Theorem 3 requires an undersmoothing bandwidth such that $Nh^5 \to 0$ as $N \to \infty$. Thus, following the suggestion of Huber et al. (2020), we take $h = C \cdot N^{-0.25}$.

### 7. Monte Carlo simulation

In this section, we conduct Monte Carlo simulations to evaluate the finite sample performance of our proposed estimators.
7.1. Continuous treatment

We consider data-generating processes (DGPs) similar to the designs in Huber et al. (2020). The confounder $X$ is drawn from the uniform distribution over $[-1.5, 1.5]$. We consider the continuous treatment variable generated from $T = 0.3X + \epsilon$, where $\epsilon$ is drawn from the uniform distribution over $[-2, 2]$. We further generate $U$ and $V$ from the uniform distribution over $[-2, 2]$ such that $\epsilon$, $U$, and $V$ are independent of each other. The mediator variable is then generated according to $M = 0.3T + 0.3X + V$. The outcomes in each scenario are given as follows:

- **Scenario I:** $Y = 0.3T + 0.3M + 0.5TM + 0.3X + U$.
- **Scenario II:** $Y = 0.3T + 0.3M + 0.3X + 0.25T^3 + U$.
- **Scenario III:** $Y = 0.3T + 0.3M + 0.5TM + 0.3X + 0.25T^3 + U$.

In Scenario I, there is an interaction effect between $T$ and $M$, with the outcome model being linear. Scenario II considers a nonlinear outcome model; however, there is no interaction effect, implying that the direct and indirect effects are homogeneous, that is, $\mu(t, t) - \mu(t', t) = \mu(t, t') - \mu(t', t)$ and $\mu(t, t) - \mu(t, t') = \mu(t', t) - \mu(t', t')$. In Scenario III, there is an interaction effect, with the outcome model being nonlinear. The true dose-response functions are

- **Scenario I:** $\mu(t, t') = 0.3t + 0.09t' + 0.15tt'$.
- **Scenario II:** $\mu(t, t') = 0.3t + 0.09t' + 0.25t^3$.
- **Scenario III:** $\mu(t, t') = 0.3t + 0.09t' + 0.15tt' + 0.25t^3$.

In all the scenarios, we set the sample size $N = 500$ and 1000. The Monte Carlo trials are repeated for 500 times. We set $t' = 0$ and let $t$ vary over the interval $[-1.5, 0] \cup (0, 1.5]$. Specifically, we set $t$ equal to the grid points $G_r := \{-1.5, -1.4, \ldots, -0.1, 0.1, \ldots, 1.4, 1.5\}$.

To evaluate the performance, we compare the proposed estimators, the covariate-balancing series (CBS) estimators in (3.6), and the covariate-balancing kernel (CBK) regression estimators in (3.7), with the alternative estimators established in the literature. Specifically, we estimate the weighting functions in both CBS and CBK estimators using the power series. The smoothing parameters are selected by the data-driven method described in Section 6.

The alternative competitors considered here are the nonparametric weighting kernel (NWK) and semiparametric weighting kernel (SWK) estimators produced by Huber et al. (2020) and the linear ordinary least squares (OLS) regression estimators used in the simulation study of Huber et al. (2020 Section 5). In particular, SWK (incorrectly) assumes the conditional distributions of $T$ given $X$ and $T$ given $(M, X)$ are both normal and estimates the distribution parameters using maximum likelihood. OLS estimates the direct and indirect treatment effects by linearly regressing the observed mediator on the observed treatment and covariates and linearly regressing the observed outcome on the observed mediator, treatment, and covariates, respectively. Thus, for nonlinear models, OLS is biased.

For each $t \in G_r$, we measure the performance of the estimators by the square root mean square errors (Rmse) over the 500 Monte Carlo trials. Table 1 reports the averages of Rmse (ARMSE) for the estimated direct and indirect effects under Scenarios I to III, where the averages are taken across all the treatment values $t \in G_r$. Specifically, for any estimator $\hat{\mu}(t)$ of a function $\mu(t)$ for $t \in G_r$, ARMSE($\hat{\mu}(t)$) = $|G_r|^{-1} \sum_{t \in G_r} \sqrt{\frac{1}{500} \sum_{j=1}^{500} (\hat{\mu}(t) - \mu(t))^2}/500$, where $|G_r|$ is the number of elements in the set $G_r$ and $\hat{\mu}(t)$ denotes the estimate calculated from the $j$th Monte Carlo trial.

Overall, our CBS estimator consistently outperforms the nonparametric NWK across all scenarios, and our proposed CBK method surpasses NWK in every instance except one – the average natural direct effects under Scenario I. These observations align well with our theories and validate the robust finite sample properties of our methods. Notably, CBS tends to perform better than CBK in most of the cases. The advantage is more obvious for the average natural direct effects, where the models involves
Table 1. \(10^3 \times \) ARMSE for the estimated direct and indirect effects under Scenarios I–III, with the smallest value and the smallest nonparametric value in each configuration highlighted in boldface and underline, respectively.

|                  | Average natural direct effects | Average natural indirect effects |
|------------------|--------------------------------|---------------------------------|
|                  | ARMSE\(\{\tilde{\mu}(t, t) - \tilde{\mu}(0, t)\}\) | ARMSE\(\{\tilde{\mu}(t, 0) - \tilde{\mu}(0, 0)\}\) | ARMSE\(\{\tilde{\mu}(t, t) - \tilde{\mu}(0, 0)\}\) | ARMSE\(\{\tilde{\mu}(0, t) - \tilde{\mu}(0, 0)\}\) |
|                  | I     | II   | III  | I     | II   | III  | I     | II   | III  | I     | II   | III  |
| N=500            |       |      |      |       |      |      |       |      |      |       |      |      |
| OLS              | 133.70 | 283.79 | 305.15 | 46.14 | 283.79 | 285.59 | 125.65 | 15.96 | 125.69 | 18.17 | 15.96 | 18.39 |
| SWK              | 100.79 | 122.31 | 130.11 | 95.14 | 123.63 | 127.52 | 32.85 | 31.17 | 39.48 | 22.68 | 24.25 | 24.82 |
| NWK              | 73.31  | 149.91 | 155.49 | 97.92 | 153.31 | 169.17 | 118.38 | 102.31 | 119.75 | 35.79 | 34.79 | 34.71 |
| CBS              | 64.08  | 61.16  | 70.24  | 57.16 | 54.25  | 62.44  | 32.09  | 27.00 | 33.99  | 28.02 | 27.81 | 28.65 |
| CBK              | 128.29 | 119.89 | 130.14 | 125.88 | 120.57 | 127.94 | 34.98 | 28.00 | 35.78  | 27.01 | 26.52 | 27.06 |
| n=1000           |       |      |      |       |      |      |       |      |      |       |      |      |
| OLS              | 129.72 | 281.73 | 302.32 | 34.78 | 281.73 | 283.04 | 124.79 | 11.05 | 124.84 | 12.48 | 11.05 | 12.85 |
| SWK              | 86.19  | 87.84  | 94.66  | 84.07 | 90.01  | 93.54  | 23.54 | 22.32 | 28.04  | 15.38 | 15.59 | 16.02 |
| NWK              | 59.97  | 116.40 | 120.46 | 90.29 | 121.97 | 139.49 | 107.25 | 91.49 | 107.44 | 29.74 | 29.05 | 29.12 |
| CBS              | 45.72  | 41.48  | 49.49  | 40.53 | 37.62  | 44.34  | 21.75  | 18.44 | 23.22  | 19.18 | 18.76 | 19.70 |
| CBK              | 107.37 | 100.23 | 108.27 | 105.57 | 100.14 | 106.27 | 24.97 | 20.75 | 25.34  | 20.19 | 20.22 | 20.21 |

Table 2. \(10^3 \times \) ARMSE for the estimated direct and indirect effects under the binary treatment model, with the smallest value in each configuration highlighted in boldface.

|                  | Average natural direct effects | Average natural indirect effects |
|------------------|--------------------------------|---------------------------------|
|                  | ARMSE\(\{\tilde{\mu}(1, 0) - \tilde{\mu}(0, 0)\}\) | ARMSE\(\{\tilde{\mu}(1, 1) - \tilde{\mu}(0, 1)\}\) | ARMSE\(\{\tilde{\mu}(1, 1) - \tilde{\mu}(1, 0)\}\) | ARMSE\(\{\tilde{\mu}(0, 1) - \tilde{\mu}(0, 0)\}\) |
|                  | I     | II   | III  | I     | II   | III  | I     | II   | III  | I     | II   | III  |
| n=500            |       |      |      |       |      |      |       |      |      |       |      |      |
| IPW              | 123.59 | 125.53 | 94.83  | 43.72 |
| CBS              | 122.58 | 122.95 | 91.53  | 38.84 |
| n=1000           |       |      |      |       |      |      |       |      |      |       |      |      |
| IPW              | 85.95  | 85.16  | 66.80  | 30.22 |
| CBS              | 81.03  | 80.71  | 66.66  | 28.88 |

Note that the results of SWK and NWK are consistent with those in Huber et al. (2020), where the misspecified SWK is better than NWK in most cases. This maybe because the conditional normal distribution assumption can capture the linear relationships of \(T\) with \(X\) and \((M, X)\) in the simulated models fairly well. However, our proposed CBS performs better than SWK in most of the cases.

In particular, under Scenario I, the direct and indirect effects are heterogeneous. Specifically, \(\mu(t, 0) - \mu(0, 0)\) and \(\mu(0, t) - \mu(0, 0)\) are reduced to linear models, whereas \(\mu(t, t) - \mu(0, t)\) and \(\mu(t, t) - \mu(t, 0)\) are nonlinear. It is thus not surprising that the OLS estimator gives the best estimation for \(\mu(t, 0) - \mu(0, 0)\) and \(\mu(0, t) - \mu(0, 0)\), but performs worst for the others two.

In Scenario II, the direct and indirect effects are homogeneous, and the direct effects models are nonlinear. The OLS estimators for the direct effects thus perform the worst and do not show any consistency. However, they estimate the indirect effects consistently, and the best since the true indirect effects are linear. Both the proposed CBS and CBK estimators produce smaller ARMSE than NWK. While CBK is comparable to SWK, CBS outperforms SWK.

Under Scenario III, our proposed CBS and CBK outperform the nonparametric NWK for all the effects. Moreover, the direct effects and indirect effect \(\mu(t, t) - \mu(t, 0)\) are nonlinear and the indirect effect \(\mu(0, t) - \mu(0, 0)\) is linear. The OLS estimators are not consistent except for \(\mu(0, t) - \mu(0, 0)\). In those nonlinear configurations, our proposed CBS estimator provides the best performance among all the estimators.
7.2. Binary treatment

We also conduct a simulation to investigate the finite sample performance when the treatment is binary, that is, \( T \in \{0, 1\} \). Following Hsu et al. (2018), we generate \( T \) from a conditional Bernoulli distribution with a success probability of \( \exp(X)/(1 + \exp(X)) \). The confounder \( X \) is generated from the uniform distribution over \([-1.5, 1.5]\), the mediator variable is generated through \( M = 0.3T + 0.3X + V \), and the outcome is generated through \( Y = 0.3T + 0.3M + 0.5TM + 0.3X + 0.25T^3 + U \), where \( U \) and \( V \) are independently drawn from the uniform distribution over \([-2, 2]\). The population means of the potential outcomes are \( \mu(0, 0) = 0, \mu(0, 1) = 0.09, u(1, 0) = 0.55, \) and \( \mu(1, 1) = 0.79 \).

We compare our proposed CBS estimator with the inverse probability weighting estimator (IPW) introduced by Hsu et al. (2018), where the generalized propensity score functions \( \mathbb{P}(T = t|M, X) \) and \( \mathbb{P}(T = t|X) \) are estimated using the nonparametric series logit regression. Table 2 reports the ARMSE for the estimated direct and indirect effects. The results show that both estimators improve with the sample size increases, whereas our CBS estimator performs slightly better in all configurations.
Figure 2. Estimated direct effects $\hat{\mu}_B(t, t) - \hat{\mu}_B(40, t)$ (top left) and $\hat{\mu}_B(t, 40) - \hat{\mu}_B(40, 40)$ (top right), and indirect effects $\hat{\mu}_B(t, t) - \hat{\mu}_B(t, 40)$ (bottom left) and $\hat{\mu}_B(40, t) - \hat{\mu}_B(40, 40)$ (bottom right) for $t \in \{100, 200, \cdots , 2000\}$, with the estimated 95% confidence bands (dashed lines).

8. Application

To evaluate the practical value of our method, we revisit the case study of Job Corps analyzed by Schochet et al. (2008) and Huber et al. (2020). Job Corps is a publicly funded U.S. training program that targets economically disadvantaged youths between the ages of 16 and 24 who are legal US residents. Participants received approximately 12:00 hours of vocational training and education, housing, and boarding over an average duration of 8 months. Previous analyses by Huber (2014) and Frölich and Huber (2017) focused on the effect of the program’s participation, as a binary treatment effect, on the health and earnings, respectively. Schochet et al. (2008) find that participation in the Job Corp program increases educational attainment and reduces criminal activity. Huber et al. (2020) apply a continuous treatment effect analysis to investigate how the total hours spent in either academic or vocational classes during the 12 months affects the participants’ criminal activities, mediated through the employment status after the training.

Similar to Huber et al. (2020), our research explores the training program’s effects on criminal activity, mediated by post-program employment. But different from them, we first apply our method to assess the direct and indirect effects of attending the program (binary treatment). Subsequently, we utilize our method to examine how these effects vary with the length of participation (continuous treatment) in
the program. Specifically, the mediator variable, denoted by $M$, is the proportion of weeks employed in the second year after the training. The outcome observation, denoted by $Y$, is the number of times an individual was arrested by the police in the fourth year after the program. The confounding variables, denoted by $X$, include a rich set of pre-treatment covariates: age, gender, ethnicity, language competency, education, marital status, household size, and income, previous receipt of social aid, family background (e.g., parents’ education), health and health-related behavior at baseline, the expectations about the Job Corps program, and the interaction with the recruiters.

Let $T$ denote the total hours a participant spent in either academic or vocational classes during the 12-months training program. To analyze the binary treatment mediation effect, we set the binary treatment $D = 1$ for all $T > 0$ and $D = 0$ for $T = 0$. The dataset contains 10,775 individuals, whose post-treatment variables $M$ and $Y$ were fully observed in the follow-up surveys after 2 and 4 years, respectively (see Huber et al., 2020 for a detailed description of the dataset and related statistics). The empirical results, as summarized in Table 3, indicate that both direct and indirect effects of this treatment on reducing criminal behavior are not significant. This lack of significance suggests that simply attending the training, regardless of the duration, does not have a causal effect on altering criminal behavior. Given the finding, it is important to consider not just the presence of the treatment (i.e., attending the training) but also its intensity (how many hours were spent in training).

We then analyze the 4,000 individuals, from the 10,775 ones in the original sample, who received a positive treatment intensity, that is, $T > 0$. We apply the proposed method to this dataset to analyze the mediation effects to help policymakers design more efficient intervention programs. The benchmark treatment level is fixed at $t' = 40$, that is, a rather small intensity of 40 hours. We compute the response curve estimators $\hat{\mu}(t, t')$ and $\hat{\mu}_h(t, t')$ proposed in (3.6) and (3.7), respectively. Then, we use them to obtain the estimates of the direct and indirect effects for varying $t \in \{100, 200, ..., 1900, 2000\}$. The smoothing parameters are determined using the data-driven approach described in Section 6.

Figure 1 reports the estimated direct and indirect effects based on our CBS $\hat{\mu}(t, t')$, while Figure 2 reports the estimated direct and indirect effects based on our CBK $\hat{\mu}_h(t, t')$. Figure 1 shows that the direct effects are significantly negative at the 5% level for all $t$ values considered, with the effects becoming larger as the training time increases. Additionally, all estimated indirect effects are insignificant.

The results in Figure 2 are similar to those of Huber et al. (2020). They show that small treatment intensities do not reduce the number of arrests. From 800 hours onward, the direct effects become significantly negative, the 95% confidence bands exclude zero, and the effect peaks occur around 1400 hours. The estimated indirect effects are mostly insignificant, consistent with the results presented in Figure 1. In conclusion, the Job Corps program has significant direct effects on the number of arrests in the fourth year. In contrast, for the investigated range of treatment intensities, the indirect effects of program-induced employment changes on arrests are close to zero.

Comparing Figures 1 and 2 and the results in Huber et al. (2020), our sieve regression estimator gives smoother results than the kernel-type ones. The kernel-type estimators seem to have some boundary effects that show unstable results on the boundaries of the support of the treatment values. Recall that in our simulation, with or without interaction for both linear and nonlinear models, our CBS gives better direct effects estimation than CBK. Thus, Figure 1 may provide a more reliable result. However, to confirm these analyses, some specification and monotonicity tests are needed, which can be an interesting future direction.

| Table 3. 10× Direct and indirect effects with confidence interval under the binary treatment. |
|-----------------------------------------------|
| Average natural direct effects | Average natural indirect effects |
|-----------------------------------------------|
| $\{\hat{\mu}(1,0) - \hat{\mu}(0,0)\}$ | $\{\hat{\mu}(1,1) - \hat{\mu}(0,1)\}$ | $\{\hat{\mu}(1,1) - \hat{\mu}(1,0)\}$ | $\{\hat{\mu}(0,1) - \hat{\mu}(0,0)\}$ |
| CBS $-0.075\,(-0.348,0.184)$ | $-0.082\,(-0.336,0.185)$ | $0.005\,(-0.041,0.051)$ | $0.011\,(-0.050,0.072)$ |
9. Discussion and conclusion

This study provides a novel approach for estimating causal mediation effects, which unifies the binary, multi-valued, and continuous treatments and the mixture of discrete and continuous treatments under a sequential ignorability condition. Furthermore, we establish the asymptotic normality for our proposed estimators. In particular, we show that our proposed estimators attain the semiparametric efficiency bounds when the treatment is discrete and asymptotically more efficient than the existing method when the treatment is continuous. Nevertheless, our study has some limitations that should be addressed by future research. First, an extension to allow for high-dimensional covariates is needed since the fully nonparametric estimation suffers from the curse of dimensionality. Second, our idea is readily applicable to panel data, which is commonly observed in practice; however, the asymptotic analysis could be more difficult and a worthwhile future study.

Acknowledgments

The authors sincerely thank the editor Esfandiar Maasoumi and the referees for their constructive suggestions and comments. Wei Huang's research is supported by the Professor Maurice H. Belz Fund of the University of Melbourne.

Supplementary Material

Supplementary Material for “Nonparametric Estimation of Mediation Effects with A General Treatment”: The supplementary material is only for online publication (pdf file). It contains the assumptions required to derive the asymptotic properties of $\hat{\pi}_Z$ and detailed discussions on the assumptions, the asymptotic results of $\hat{\pi}_Z$ and the proofs of Theorems 1, 2 and Corollary 1.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Lukang Huang is supported by the Fundamental Research Funds for the Central Universities, Nankai University (No.63231181) and the Natural Science Funds of Tianjin (No.22JCQNJC01660). Zheng Zhang is supported by the fund from the Natural Science Foundation of Beijing, China [grant number 1222007] and the fund for building world-class universities (disciplines) of Renmin University of China [project number KYGJC2023011].

References

Ai, C., Huang, L., Zhang, Z. (2022). A simple and efficient estimation of average treatment effects in models with unmeasured confounders. *Statistica Sinica* 32(3).
Ai, C., Linton, O., Motegi, K., Zhang, Z. (2021). A unified framework for efficient estimation of general treatment models. *Quantitative Economics* 12(3):779–816. doi:10.3982/QE1494.
Ai, C., Linton, O., Zhang, Z. (2022). Estimation and inference for the counterfactual distribution and quantile functions in continuous treatment models. *Journal of Econometrics* 228(1):39–61. doi:10.1016/j.jeconom.2020.12.009.
Andrews, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59(2):307. doi:10.2307/2938259.
Baron, R. M., Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* 51(6):1173–1182. doi:10.1037//0022-3514.51.6.11733806354.
Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* 155(2):138–154. doi:10.1016/j.jeconom.2009.09.023.
Chan, K., Imai, K., Yam, S., Zhang, Z. (2016). Efficient nonparametric estimation of causal mediation effects, arXiv preprint, arXiv:1601.03501.
Chan, K. C. G., Yam, S. C. P., Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 78(3):673–700. doi:10.1111/rssb.12129 27346982.
Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics* 6:5549–5632.

Chen, X., Hong, H., Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics* 36(2):808–843.

Chernozhukov, V., Fernández-Val, I., Melly, B. (2013). Inference on counterfactual distributions, *Econometrica* 81:2205–2268.

Colangelo, K., Lee, Y.-Y. (2020). Double debiased machine learning nonparametric inference with continuous treatments, arXiv preprint, arXiv:2004.03036.

Donald, S. G., Hsu, Y.-C. (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics* 178:383–397. doi:10.1016/j.jeconom.2013.03.010.

Dong, Y., Lee, Y.-Y., Gou, M. (2019). Regression discontinuity designs with a continuous treatment. *Journal of the American Statistical Association* 118:208–21. Available at SSRN 3167541.

Fan, J., Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*, Vol. 66. New York: CRC Press.

Fong, C., Hazlett, C., Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Annals of Applied Statistics* 12(1):156–177.

Frölich, M., Huber, M. (2017). Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79(5):1645–1666. doi:10.1111/rssb.12232.

Galvao, A. F., Wang, L. (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association* 110(512):1528–1542. doi:10.1080/01621459.2014.978005.

Goetgeluk, S., Vansteelandt, S., Goetghebeur, ELS. (2008). Estimation of controlled direct effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70(5):1049–1066. doi:10.1111/j.1467-9868.2008.00673.x.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2):315. doi:10.2307/2998560.

Hirano, K., Imbens, G. W. (2004). Bayesian modeling and causal inference from incomplete-data perspectives, In: Gelman, A., Meng, X.-L., eds., The Propensity Score with Continuous Treatments. John Wiley & Sons Ltd, pp. 73–84.

Hirano, K., Imbens, G. W., Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–1189. doi:10.1111/1468-0262.00442.

Hong, G., Deutsch, J., Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-BY-mediated interaction. *Journal of Educational and Behavioral Statistics* 40(3):307–340. doi:10.3102/1076998615583902.

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. *Journal of the American Statistical Association* 109(503):1029–1051. doi:10.1080/01621459.2013.802513.

Huber, M., Hsu, Y., Lee, Y., Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics* 35(7):814–840. doi:10.1002/jae.2765.

Imai, K., Keele, L., Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods* 15(4):309–334. doi:10.1037/a0020761.

Imai, K., Keele, L., Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105(4):765–789. doi:10.1017/S0003055411000414.

Imai, K., Keele, L., Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* 25(1):51–71.

Imai, K., Van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association* 99(467):854–866. doi:10.1198/016214504000001187.

Joffe, M. M., Small, D., Hsu, C.-Y. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science* 22(1):74–97.

Kennedy, E. H., Ma, Z., McHugh, M. D., Small, D. S. (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B, Statistical Methodology* 79(4):1229–1245. doi:10.1111/rssb.12212.

Lee, D. S., Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature* 48(2):281–355. doi:10.1257/jel.48.2.281.

Lee, Y.-Y. (2018). Efficient propensity score regression estimator of multivalued treatment effects for the treated. *Journal of Econometrics* 204(2):207–222. doi:10.1016/j.jeconom.2018.02.002.

Lewbel, A., Linton, O. (2007). Nonparametric matching and efficient estimators of homothetically separable functions. *Econometrica* 75(4):1209–1227. doi:10.1111/j.1468-0262.2007.00787.x.
Appendix

A. Proof of (2.2)

We rewrite the identification of $\mu(t, t')$ in Huber et al. (2020) as follows:

\[
\mu(t, t') = \mathbb{E} \left[ \frac{f_{T}(t)}{f_{T|M}(t|M, X)} \cdot \frac{f_{T|M|X}(t'|M, X)}{f_{T|M}(t'|X)} \cdot Y \mid T = t \right]
\]

\[
= \mathbb{E} \left[ \frac{f_{T}(t)}{f_{T|M}(t|M, X)} \cdot \frac{f_{T|M|X}(t'|M, X)}{f_{T}(t')} \cdot \frac{f_{T}(t)}{f_{T|M}(t'|X)} \cdot Y \mid T = t \right]
\]

\[
= \mathbb{E} \left[ \frac{f_{T}(T)}{f_{T|M}(T|M, X)} \cdot \frac{f_{T|M|X}(T + \delta|M, X)}{f_{T}(T + \delta)} \cdot \frac{f_{T}(T + \delta)}{f_{T|M}(T + \delta|X)} \cdot Y \mid T = t \right]
\]

\[
= \mathbb{E} \left[ \frac{\pi_{M|X}(T, M, X)}{\pi_{M}(T + \delta, M, X)} \cdot \pi_{X}(T + \delta, X) \mid Y \mid T = t \right].
\]

B. Some Preliminary Results

We recall some preliminary results on the convergence rates of $\hat{\pi}_{K_{Z}}(t, Z)$ that are directly implied by the results in Ai et al. (2021). We impose the following conditions based on those of Ai et al. (2021): For $Z \in \{X, (M, X)\}$, we assume

**Assumption 4.** (i) The support of $Z$, $\mathcal{Z}$ is a compact set. (ii) There exist two positive constants $\eta_1$ and $\eta_2$ such that

\[ 0 < \eta_1 \leq \pi_Z(t, z) \leq \eta_2 < \infty \quad \forall (t, z) \in T \times \mathcal{Z}. \]
Assumption 5. There exist $\Lambda_{k_1 \times k_Z} \in \mathbb{R}^{k_1 \times k_Z}$ and constant $\alpha_Z > 0$ such that

$$\sup_{(t,z) \in T \times Z} \left| \rho^{-1} \{ \pi_Z(t, z) \} - u_{k_1}(t)^\top \Lambda_{k_1 \times k_Z} \nu_{k_Z}(z) \right| = O(K_Z^{-\alpha_Z}),$$

where $\rho^{-1}(v) = -\log v - 1$.

Assumption 6. (i) The eigenvalues of $[\mathbb{E}[u_{k_1}(T)u_{k_1}(T)^\top], \mathbb{E}[\nu_{k_Z}(Z)\nu_{k_Z}(Z)^\top]]$ are bounded away from zero and infinity uniformly in $k_1$, $k_Z$. (ii) There are sequences of constants $\xi_1(k_1)$ and $\zeta_1(k_Z)$ satisfying $\sup_{t \in T} \| u_{k_1}(t) \| \leq \xi_1(k_1)$ and $\sup_{z \in Z} \| \nu_{k_Z}(z) \| \leq \zeta_1(k_Z)$ such that $\sqrt{N}K_Z^{-\alpha Z} \to 0$ and $\sigma(K_Z)\sqrt{K_Z^2/N} \to 0$ as $N \to \infty$, where $K_Z = k_1k_Z$ and $\zeta(K_Z) = \xi_1(k_1)\zeta_1(k_Z)$.

Assumption 4 (i) requires the covariates, the treatment variable, and the mediator to be bounded. This condition, despite being restrictive, is commonly imposed in the nonparametric regression literature. However, we can replace it with a restriction on the tail distribution of $(M, X, T)$. For example, Chen et al. (2008 Assumption 3) assume that the support of $X$ is the entire Euclidean space, but impose $\int_{\mathbb{R}^2}(1 + |x|^2)^{\omega}f_X(x)dx < \infty$ for some $\omega > 0$. Assumption 4 (ii) requires the weighting function to be bounded and bounded away from zero. We can relax Assumption 4 (ii) by allowing $\eta_1(\eta_2)$ to go to zero (infinity) slowly as $N \to \infty$. Notice that $u_{k_1}(t)^\top \Lambda \nu_{k_Z}(z)$ is a linear sieve approximation for $\rho^{-1}(\{\pi_Z(t, z)\})$. Assumption 5 requires the sieve approximation error to shrink to zero at a polynomial rate. A variety of sieve basis functions satisfy this condition. When both $T$ and $Z$ are discrete, $K_Z$ is a finite constant independent of $N$ and $\alpha_Z = +\infty$. When $(T, Z)$ has continuous or mixed components, $K_Z \to \infty$ as $N \to \infty$ and $\alpha_Z$ is positively affected by the smoothness of $\rho^{-1}(\cdot)$ and negatively affected by the number of continuous and mixed components. Assumption 6 (i) ensures that the sieve estimator is nondegenerate. This condition is common in the sieve regression literature (see Andrews, 1991; Newey, 1997). If the approximation error is nonzero, Assumption 6 (ii) imposes a restriction on the growth rate of the smoothing parameters $k_1$ and $k_Z$ to ensure undersmoothing. Under these conditions, the following are direct results from Ai et al. (2021 Propositions 1 and 2):

Proposition 1. Suppose that Assumptions 4–6 hold. For $Z \in \{X, (M, X)\}$, we have

$$\sup_{(t,m,x) \in T \times M \times X} |\tilde{\pi}_{K_Z}(t, z) - \pi_Z(t, z)| = O_p \left( \max \left\{ \zeta(K_Z)K_Z^{-\alpha_Z}, \zeta(K_Z)\sqrt{K_Z^2/N} \right\} \right),$$

and

$$\int_T \int_Z \left| \tilde{\pi}_{K_Z}(t, z) - \pi_Z(t, z) \right|^2 dF_T Z(t, z) = O_p \left( \max \left\{ K_Z^{-2\alpha_Z}, K_Z^2/N \right\} \right),$$

and

$$\frac{1}{N} \sum_{i=1}^N \left| \tilde{\pi}_{K_Z}(T_i, Z_i) - \pi_Z(T_i, Z_i) \right|^2 = O_p \left( \max \left\{ K_Z^{-2\alpha_Z}, K_Z^2/N \right\} \right).$$

Proposition 2. Assume that Assumptions 4–6 hold. For for any square-integrable random variable $\phi(T, M, X, Y) \in L^2$ and $Z \in \{X, (M, X)\}$, if there exist a $\Gamma_{k_1 \times k_Z} \in \mathbb{R}^{k_1 \times k_Z}$ and a constant $\gamma_Z > 0$, s.t.

$$\sup_{t \times z \in T \times Z} \left| \mathbb{E}[\phi(T, M, X, Y)|T = t, Z = z] - u_{k_1}(t)^\top \Gamma_{k_1 \times k_Z} \nu_{k_Z}(z) \right| = O(K_Z^{-\gamma_Z}),$$
then we have, for any $\delta \geq 0$,

$$
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[ \hat{\pi}_{KZ}(T_i + \delta, Z_i) \phi(T_i, M_i, X_i, Y_i) - \mathbb{E}[\pi_Z(T + \delta, Z) \phi(T, M, X, Y)] \right]
$$

$$
= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} IF_{\pi Z, i} \{ \delta, \phi(T, M, X, Y) \}
$$

$$
+ O_p \left( \sqrt{NK_Z^{\alpha_2}} \right) + O_p \left( K_Z^{\gamma_Z} \right) + O_p \left( \zeta(K_Z) \sqrt{\frac{K_Z^2}{N}} \right)
$$

(B.1)

where, for $i = 1, \ldots, N$,

$$
IF_{\pi Z, i} \{ \delta, \phi(T, M, X, Y) \} = \pi_Z(T_i + \delta, Z_i) \phi(T_i, M_i, X_i, Y_i)
$$

$$
- \pi_Z(T_i, Z_i) \frac{f_{T|Z}(T_i - \delta | Z_i)}{f_{T|Z}(T_i | Z_i)} \mathbb{E}[\phi(T_i - \delta, M_i, X_i, Y_i) | T_i, Z_i]
$$

$$
+ \mathbb{E}[\pi_Z(T_i, Z_i) \frac{f_{T|Z}(T_i - \delta | Z_i)}{f_{T|Z}(T_i | Z_i)} \phi(T_i - \delta, M_i, X_i, Y_i) | Z_i]
$$

$$
- \mathbb{E}[\pi_Z(T_i, Z_i) \frac{f_{T|Z}(T_i - \delta | Z_i)}{f_{T|Z}(T_i | Z_i)} \phi(T_i - \delta, M_i, X_i, Y_i) | T_i]
$$

$$
+ \mathbb{E}[\pi_Z(T_i, Z_i) \frac{f_{T|Z}(T_i - \delta | Z_i)}{f_{T|Z}(T_i | Z_i)} \phi(T_i - \delta, M_i, X_i, Y_i) | T_i]
$$

$$
- \mathbb{E}[\pi_Z(T_i, Z_i) \frac{f_{T|Z}(T_i - \delta | Z_i)}{f_{T|Z}(T_i | Z_i)} \phi(T_i - \delta, M_i, X_i, Y_i) | T_i]
$$

(B.2)

and $\mathbb{E}[IF_{\pi Z, i} \{ \delta, \phi(T, M, X, Y) \}] = 0$.

Using these results, we show the following lemma that is useful for deriving our main theorems. The proof can be found in Section S2 in supplemental material.

**Lemma 3.** Assume that Assumptions 4–6 hold. For any square-integrable random variable $\phi(T, M, X, Y) \in L^2$ and $Z \in \{X, (M, X)\}$, if there exist a $\Gamma_{k_1 \times k_Z} \in \mathbb{R}^{k_1 \times k_Z}$ and a constant $\gamma_Z > 0$, s.t.

$$
\sup_{t \times z \in T \times Z} \left| \mathbb{E}[\phi(T, M, X, Y) | T = t, Z = z] - u_{k_1}(t)^\top \Gamma_{k_1 \times k_Z} v_{k_Z}(z) \right| = O(K_Z^{-\gamma_Z}),
$$

then we have, for any $\delta \geq 0$,

$$
\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \hat{\pi}_{K_M}(T_i, M_i, X_i) \hat{\pi}_{K_X}(T_i + \delta, X_i) \phi(T_i, M_i, X_i, Y_i)
$$

$$
- \mathbb{E} \left[ \frac{\pi_{M,X}(T, M, X)}{\pi_{M,X}(T + \delta, M, X)} \pi_{X}(T + \delta, X) \phi(T, M, X, Y) \right] \right\}
$$

$$
= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ IF_{\pi M, i} \{ \delta, \frac{\pi_{M,X}(T, M, X)}{\pi_{M,X}(T + \delta, M, X)} \cdot \phi(T, M, X, Y) \}
$$

$$
+ IF_{\pi M, i} \{ 0, \frac{\pi_{X}(T + \delta, X)}{\pi_{M,X}(T + \delta, M, X)} \cdot \phi(T, M, X, Y) \}
$$

$$
- IF_{\pi M, i} \{ \delta, \frac{\pi_{M,X}(T, M, X)\pi_{X}(T + \delta, X)}{\pi_{M,X}(T + \delta, M, X)} \cdot \phi(T, M, X, Y) \} \right\}
$$
\begin{align*}
  &+ O_p \left( \sqrt{N} K^{-\alpha_X} \right) + O_p \left( K^{-\gamma_X} \right) + O_p \left( \zeta(K_\chi) \sqrt{\frac{K^2}{N}} \right) \\
  &+ O_p \left( \sqrt{N} K^{-\alpha_{MX}} \right) + O_p \left( K^{-\gamma_{MX}} \right) + O_p \left( \zeta(K_{MX}) \sqrt{\frac{K^2_{MX}}{N}} \right). \quad (B.3)
\end{align*}

**B.1. Estimating the \( d_{K_{0,i}} \)'s**

To estimate the asymptotic variance \( V_{tt'} \) in Section 5, for \( i = 1, \ldots, N \), we estimate \( d_{K_{0,i}} \) by

\[
\hat{d}_{K_{0,i}}(T, M, X, Y; \delta) := \hat{F}_{\pi, X, i} \left\{ \delta, \frac{\hat{F}_{K, M, X}(T, M, X)}{\hat{F}_{K, M, X}(T + \delta, M, X)} \cdot u_{K_0}(T) Y \right\} \\
+ \hat{F}_{\pi, M, X, i} \left\{ 0, \frac{\hat{F}_{K, X}(T + \delta, X)}{\hat{F}_{K, M, X}(T + \delta, M, X)} \cdot u_{K_0}(T) Y \right\} \\
- \hat{E} \left[ \frac{\hat{F}_{K, M, X}(T, M, X)}{\hat{F}_{K, M, X}(T + \delta, M, X)} \cdot u_{K_0}(T) Y_i \mid T_i \right] \\
+ \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{F}_{K, M, X}(T, M, X)}{\hat{F}_{K, M, X}(T + \delta, M, X)} \cdot u_{K_0}(T) Y_i, 
\]

where, for \( Z \in \{X, (M, X)\} \), \( \hat{F}_{\pi, Z, i} \) is an estimator of \( IF_{\pi, Z, i} \) defined in (B.2) in Appendix B, and \( \hat{E} \) denotes the least square regression. Specifically, we estimate \( \pi_Z, f_{T_i Z} \), the conditional expectations, and expectations in (B.2) by \( \hat{F}_{K, x} \), the conditional kernel density estimation, the least square regression of the estimated response variable on the corresponding sieve basis, and the sample average of the estimated variables, respectively. For example, \( \hat{E} \left[ \pi_X(T_i, X_i) u_{K_0}(T_i) Y_i \mid T_i, X_i \right] \) is estimated by the least square regression of \( \hat{F}_{K, X}(T_i, X_i) u_{K_0}(T_i) Y_i \) on a sieve basis \( w_{K, T, X}(T_i, X_i) \).

**C. Asymptotics of kernel estimators under continuous treatments**

To derive the asymptotic normality of our kernel regression estimators, the following assumptions are imposed.

**Assumption 7.** \( K(\cdot) \) is a univariate kernel function symmetric around the origin that satisfies (i) \( \int K(u) du = 1 \); (ii) \( \int u^2 K(u) du = \kappa_2 \in (0, \infty) \); (iii) \( \int K^2(u) du = \kappa_2 < \infty \); and (iv) \( \int |K(u)|^{2+\delta} du < \infty \), for some \( \delta > 0 \).

**Assumption 8.** As \( N \to \infty \), \( h \to 0 \), \( Nh \to \infty \) and \( Nh^5 \to 0 \).

Assumption 8 is common in the kernel regression literature (see Li and Racine, 2006).

**Theorem 4.** Suppose Assumptions 1, 4–8 hold. Then, we have

\[
\sqrt{Nh} \left[ \hat{\mu}(t, t') - \mu(t, t') \right] = \sqrt{\frac{h}{N}} \sum_{i=1}^{N} \psi_{tt'}(Y_i, T_i; M_i; X_i; h) + o_P(1) \overset{d}{\longrightarrow} N(0, V_{tt'}^h),
\]
where
\[
\psi_{tt'}(Y_t, T_t, M_t, X_t; h) = \frac{\pi_{M,X}(T_t, M_t, X_t)\pi_X(T_t + \delta, X_t)}{\pi_{M,X}(T_t + \delta, M_t, X_t) \cdot p_{t,h}} \kappa_h(T_t - t) \{ Y_t - \mathbb{E}[Y|T_t, M_t, X_t] \},
\]
\[
\quad - \frac{f_{T|X}(T_t - \delta|X_t)\pi_X(T_t, X_t)}{f_{T|X}(T_t|X_t) \cdot p_{t,h}} \kappa_h(T_t - t - \delta) \mathbb{E} \left[ \frac{\pi_{M,X}(T_t - \delta, M_t, X_t)Y_t}{\pi_{M,X}(T_t, M_t, X_t)} \right]_{T_t, X_t}
\]
\[
+ \frac{f_{T}(T_t - \delta)\pi_X(T_t, X_t)}{f_{T}(T_t) \cdot p_{t,h}} \kappa_h(T_t - t - \delta) \mathbb{E} \left[ Y_t|T_t, M_t, X_t \right],
\]
and
\[
V_{tt'}^h = \lim_{h \to 0} \frac{h}{2} \mathbb{E} \left[ \left( \psi_{tt'}(Y_t, T_t, M_t, X_t; h) \right)^2 \right]
\]
\[
= \frac{\kappa_{02}}{f_{T}(t)} \mathbb{E} \left[ \frac{\pi_{M,X}^2(T_t, M_t, X_t)}{\pi_{M,X}^2(T_t + \delta, M_t, X_t)} \pi_X^2(T_t + \delta, X_t) \{ Y_t - \mathbb{E}[Y|T_t, M_t, X_t] \}^2 \right]_{T_t = t}
\]
\[
+ \frac{\kappa_{02}f_{T}(t + \delta)}{f_{T}^2(t)} \mathbb{E} \left[ \pi_X(T_t, X_t)f_{T|X}(T_t - \delta|X_t) \mathbb{E} \left[ \frac{\pi_{M,X}(T_t - \delta, M_t, X_t)Y_t}{\pi_{M,X}(T_t, M_t, X_t)} \right]_{T_t, X_t} \right]
\]
\[
- \frac{\pi_X(T_t, X_t)f_{T}(T_t - \delta)}{f_{T}(T_t)} \mathbb{E} \left[ Y_t|T_t, M_t, X_t \right]_{T_t = t + \delta},
\]
and
\[
p_{t,h} = \mathbb{E} [\kappa_h(T - t)], \ k_{ij} = \int u^i K'(u) \, du.
\]

Following Theorem 3, we establish the asymptotics of the estimated direct and indirect effect.

Remark 6. Note that the asymptotic variances of \(\hat{\mu}_{NKW}(t, t)\) and \(\hat{\mu}_{NKW}(t, t')\) introduced by Huber et al. (2020) are
\[
V_{tt'}^{NKW} = \begin{cases} 
\kappa_{02} \mathbb{E} \left[ \text{Var} \left[ Y|T = t, X \right]/f_{T|X}(t|X) \right] & \text{if } h = h_1 = h_2 \text{ and } \kappa_{1,h_1}(\cdot) = \kappa_{2,h_2}(\cdot); \\
\kappa_{02} \mathbb{E} \left[ \left( Y - \mu(t, t) \right)^2|T = t, X \right]/f_{T|X}(t|X) \right] & \text{if } h = h_2 < h_1,
\end{cases}
\]
and
\[
V_{tt'}^{NKW} = \begin{cases} 
\kappa_{02} \mathbb{E} \left[ \text{Var} \left[ Y|T = t, M, X \right]/f_{T|M,X}(t|M,X)f_{T|X}(t|X) \right] + \mathbb{E} \left[ \text{Var} \left[ g(t, M, X)|T = t', X \right]/f_{T|X}(t'|X) \right] & \text{if } h = h_1 = h_2 \text{ and } \kappa_{1,h_1}(\cdot) = \kappa_{2,h_2}(\cdot); \\
\kappa_{02} \mathbb{E} \left[ \left( Y - \mu(t, t') \right)^2|T = t, M, X \right]/f_{T|M,X}(t|M,X)f_{T|X}(t|X) \right] & \text{if } h = h_2 < h_1,
\end{cases}
\]
where \(\kappa_{1,h_1}\) and \(\kappa_{2,h_2}\) are two prespecified kernel function used in the estimators of Huber et al. (2020), \(g(t, M, X) = \mathbb{E}[Y|T = t, M, X]\). We prove that \(V_{tt'}^h \leq V_{tt'}^{NKW}\) and \(V_{tt'}^{h} \leq V_{tt'}^{NKW}\) in Section S7 in supplementary material.