Instrumental Variable Estimation of Average Partial Causal Effects

Yuta Kawakami [1 2] **Manabu Kuroki** [1] **Jin Tian** [2]

Abstract

Instrumental variable (IV) analysis is a powerful tool widely used to elucidate causal relationships. We study the problem of estimating the average partial causal effect (APCE) of a continuous treatment in an IV setting. Specifically, we develop new methods for estimating APCE based on a recent identification condition via an integral equation. We develop two families of methods, nonparametric and parametric - the former uses the Picard iteration to solve the integral equation; the latter parameterizes APCE using a linear basis function model. We analyze the statistical and computational properties of the proposed APCE estimators and illustrate them on synthetic and real-world data.

1. Introduction

Instrumental variable (IV) analysis is a powerful tool used to elucidate causal relationships when a controlled experiment is not feasible or when a randomized experiment is not able to successfully treat each unit (Imbens, 2014; Angrist & Krueger, 2001). For example, consider a study of the effect of years of education (treatment variable X) on monthly wages (outcome variable Y) (Card, 1999; Angrist & Krueger, 1991). Since researchers cannot force people to attend or drop out of school, they use the mother's years of education (Z) as an instrumental variable with the understanding that the mother's education affects the subject's education but has no direct influence on the subject's wages. This setting is represented by the causal graph in Fig. 1, where $\boldsymbol{H}$ represents unmeasured confounders.

In general, further assumptions are needed to identify the causal effect of the treatment on the outcome. Assume

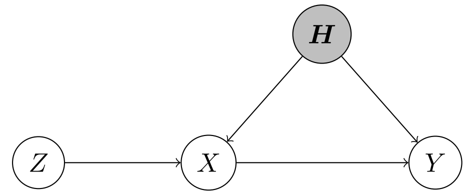

Figure 1: A causal graph representing the IV setting.

the causal relations are represented by structural equations $Y = f_Y(X, \boldsymbol{H}, \boldsymbol{u}_Y)$ and $X = f_X(Z, \boldsymbol{H}, \boldsymbol{u}_X)$. Typical assumptions include: the structure equations are linear (linearity); the structural equations are monotonic on certain arguments (monotonicity); the influence of the IV on the treatment is separable from that of the confounders, i.e., $f_X(Z, \boldsymbol{H}, \boldsymbol{u}_X) = f_{X1}(Z, \boldsymbol{u}_X) + f_{X2}(\boldsymbol{H}, \boldsymbol{u}_X)$ (separability I); or the influence of the treatment on the outcome is separable from that of the confounders, i.e., $f_Y(X, \boldsymbol{H}, \boldsymbol{u}_Y) = f_{Y1}(X, \boldsymbol{u}_Y) + f_{Y2}(\boldsymbol{H}, \boldsymbol{u}_Y)$ (separability Ⅱ). We note that separability I is testable but separability Ⅱ is not (Breusch & Pagan, 1979; Su et al., 2015).

One of the most widely used methods for estimating causal effects via IV is linear two-stage least squared (TSLS) (Stock, 2001) which assumes linearity and separability I and Ⅱ. By contrast, the two-stage predictor substitution (TSPS) method (Hausman, 1978; Terza et al., 2008) assumes separability I and Ⅱ but is applicable for nonlinear models. Methods not relying on the separability I assumption include the generalized method of moments (GMM) (Hansen, 1982; Baum et al., 2003), the nonparametric two-stage least squared estimator (NPTSLS) (Newey & Powell, 2003; Hartford et al., 2017; Singh et al., 2019), and the nonparametric conditional quantile estimation (CQE) (Chernozhukov et al., 2007; Imbens & Newey, 2009; Chen et al., 2014; Torgovitsky, 2015). GMM uses the semiparametric estimation framework and requires separability Ⅱ and the probability distribution of the IV. NPTSLS requires separability Ⅱ while CQE requires monotonicity of the function $f_Y(X, \boldsymbol{H}, \boldsymbol{u}_Y)$ with respect to $\boldsymbol{H}$. Both NPTSLS and CQE solve integral equations to identify the effect of the treatment on the outcome. However, these integral equations are ill-posed problems[1] and lead to severe estimation difficul-

[1]Department of Mathematics, Physics, Electrical Engineering and Computer Science, Yokohama National University, Yokohama, Kanagawa, JAPAN [2]Department of Computer Science, Iowa State University, Ames, Iowa, USA. Correspondence to: Yuta Kawakami <kawakami-yuta-yd@ynu.jp>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]A well-posed problem satisfies the following three properties (Tikhonov et al., 1995): existence, uniqueness, and stability of the

Table 1: The assumptions made by the related works: requiring probability distribution of the IV (DI), linearity (LI), separability I (SP I), separability II (SP II), monotonicity (MO) with respect to the hidden confounder. Check marks (✓) represent assumptions of the methods. Asterisks ($*$) denote an ill-posed problem.

Assumptions	DI	LI	SP I	SP II	MO
TSLS	✓	✓	✓	✓	
TSPS			✓	✓	
GMM	✓			✓	
NPTSLS$*$				✓	
CQE$*$					✓
This paper				✓	

ties. We summarize the assumptions made by these existing works in Table 1.

We study the causal effects of a continuous treatment variable in this paper. While historically, the majority of the previous work has focused on binary or categorical treatment variables (e.g., (Imbens & Angrist, 1994; Balke & Pearl, 1997; Wang & Tchetgen Tchetgen, 2018; Syrgkanis et al., 2019)), in recent years, there has been a growing interest in continuous treatment variables (Hirano & Imbens, 2005; Kennedy et al., 2017; Bahadori et al., 2022). In particular, Wong (2022) has recently introduced an integral equation for identifying the effect (more exactly, the average partial causal effect (APCE) (Wooldridge, 2005)) of a continuous treatment variable under the separability II assumption. Wong (2022) proved an identification condition but did not provide a method for actually solving the integral equation and estimating APCE from data samples.

In this paper, we recognize that the integral equation in (Wong, 2022) is well-posed in bounded domains and develop two families of methods, nonparametric and parametric, for estimating APCE from observed data. The nonparametric method solves the integral equation with the Picard iteration (Fridman, 1965; Diaz & Metcalf, 1970) using numerical integration and interpolation. The parametric method reduces the estimation problem to a linear regression problem by parameterizing APCE using a linear basis function model. We analyze the statistical and computational properties of the proposed methods. We illustrate them on synthetic data showing superior performance to the existing methods. Finally, we apply the proposed APCE estimators on a real-world dataset to analyze the effect of years of education on wages, which is of great interest in economics (Card, 1999; Angrist & Krueger, 1991).

solution. Problems where one or more of these conditions do not hold are called ill-posed problems.

2. Notation and Background

We represent each variable with a capital letter (X) and its realized value with a small letter (x). Let $\mathbb{1}_\Omega(x)$ be an indicator function, which is 1 if $x \in \Omega$; and 0 if $x \notin \Omega$. Denote Ω_X be the domain of X, $\mathbb{E}[Y]$ and $\mathbb{V}(Y)$ be the expectation and the variance of Y, and $\mathbb{P}_X[x] = \mathbb{P}(X \leq x)$ be the cumulative distribution function (CDF) of X. In addition, $\mathbb{E}[Y|X=x]$ and $\mathbb{P}_X[x|Z=z] = \mathbb{P}(X \leq x|Z=z)$ be the conditional expectation of Y given $X=x$ and the conditional CDF of X given $Z=z$. We write $g(x) = \mathcal{O}(h(x))$ as $x \to \infty$ if there exists a positive real number M and a real number δ such that $|g(x)| \leq Mh(x)$ for all $x \geq \delta$. In contrast, we write $g(x) = \mathcal{O}(h(x))$ as $x \to 0$ if there exists a positive real number M and a real number δ such that $|g(x)| \leq Mh(x)$ for all $0 \leq |x| \leq \delta$.

Functional Analysis. We explain the notations of functional analysis (Muscat, 2014). Let $\mathcal{H}$ be a Hilbert space, where an inner product is defined by $\langle a, b\rangle = \int_{\Omega_X} a(x)b(x)dx$ and a norm is $\|a\| = \langle a, a\rangle^{\frac{1}{2}}$ for all $a, b \in \mathcal{H}$. A sequence $\{a_n\} \in \mathcal{H}$ converges strongly to $a \in \mathcal{H}$ if $\|a_n - a\| \to 0$ as $n \to \infty$. Let an operator $\mathcal{L}$ be $(\mathcal{L}(a))(x) = \int_{\Omega_X} L(x', x)a(x')dx'$ for $a \in \mathcal{H}$, and $\|\mathcal{L}\|$ is an operator norm $\|\mathcal{L}\| = \sup\{\|\mathcal{L}(a)\| : \|a\| = 1 \text{ and } a \in \mathcal{H}\}$. When $\mathcal{L}$ possesses a countable set of positive eigenvalues, denote them $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq 0$, together with corresponding eigenvectors $v_1, v_2, v_3, \ldots$ in $\mathcal{H}$ such that $\mathcal{L}(v_i) = \lambda_i v_i$. The set $\{v_1, v_2, v_3, \ldots\}$ is an orthonormal basis.

Structural Causal Models. We use the language of Structural Causal Models (SCM) as our basic semantic and inferential framework (Pearl, 2009). An SCM $\mathcal{M}$ is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, \mathbb{P}_{\boldsymbol{U}}\rangle$, where $\boldsymbol{U}$ is a set of exogenous (unobserved) variables following a joint distribution $\mathbb{P}_{\boldsymbol{U}}$, and $\boldsymbol{V}$ is a set of endogenous (observable) variables whose values are determined by structural functions $\mathcal{F} = \{f_{V_i}\}_{V_i \in \boldsymbol{V}}$ such that $v_i := f_{V_i}(\mathbf{pa}_{V_i}, \boldsymbol{u}_{V_i})$ where $\mathbf{PA}_{V_i} \subseteq \boldsymbol{V}$ and $U_{V_i} \subseteq \boldsymbol{U}$. Each SCM $\mathcal{M}$ induces an observational distribution $\mathbb{P}_{\boldsymbol{V}}$ over $\boldsymbol{V}$, and a causal graph $G(\mathcal{M})$ over $\boldsymbol{V}$ in which there exists a directed edge from every variable in $\mathbf{PA}_{V_i}$ to V_i. An intervention of setting a set of endogenous variables $\boldsymbol{X}$ to constants $\boldsymbol{x}$, denoted by $do(\boldsymbol{x})$, replaces the original equations of $\boldsymbol{X}$ by the constants $\boldsymbol{x}$ and induces a *sub-model* $\mathcal{M}_{\boldsymbol{x}}$.

Average Partial Causal Effect (APCE). We denote the potential outcome Y under intervention $do(\boldsymbol{x})$ by $Y_{\boldsymbol{x}}(\boldsymbol{u})$, which is a solution of Y with $\boldsymbol{U} = \boldsymbol{u}$ in the sub-model $\mathcal{M}_{\boldsymbol{x}}$. Considering a continuous treatment X, we aim to estimate the APCE $\mathbb{E}[\partial_x Y_x] := \mathbb{E}_{\boldsymbol{U}}[\frac{\partial}{\partial x} Y_x(\boldsymbol{U})]$ (Chamberlain, 1984; Wooldridge, 2005; Graham & Powell, 2012), which is a real-valued function on $x \in \Omega_X$. APCE is a natural generalization of the average causal effect of a binary treatment

$\mathbb{E}_{\boldsymbol{U}}[Y_1(\boldsymbol{U})] - \mathbb{E}_{\boldsymbol{U}}[Y_0(\boldsymbol{U})]$.

Instrumental Variable (IV) Model. We consider the IV model represented by the causal graph in Fig 1, with the following SCM $\mathcal{M}_{IV}$:

$$\begin{cases} Y := f_Y(X, \boldsymbol{H}, \boldsymbol{u}_Y) \\ X := f_X(Z, \boldsymbol{H}, \boldsymbol{u}_X) \\ Z := f_Z(\boldsymbol{u}_Z) \end{cases} \tag{1}$$

We assume X, Y, and Z are continuous variables, and $\boldsymbol{u}_X$ and $\boldsymbol{u}_Y$ are exogenous noises, where $\boldsymbol{U} = \{\boldsymbol{H}, \boldsymbol{u}_X, \boldsymbol{u}_Y, \boldsymbol{u}_Z\}$ and $\boldsymbol{V} = \{Z, X, Y\}$

Conditions for identifying APCE. We explain Wong's (2022) conditions for identifying APCE from $\mathbb{P}(X, Y|Z)$.

Assumption 1. *For all $\boldsymbol{h} \in \Omega_{\boldsymbol{H}}$ under the SCM $\mathcal{M}_{IV}$,*

1. *Instrument relevance: the instrument Z has a causal effect on X, i.e., X_z is not a constant function by varying z for each subject.*
2. *Y_x is differentiable and bounded in $x \in \Omega_X$.*
3. *$\sup_{x,z} p(X_z = x) < \infty$, where p denotes the density function.*
4. *The set of distributions $\mathbb{P}(X|Z = z)$, induced by varying z, is a complete set.*

These assumptions are needed just to set up the model and are not restrictive. The second assumption means that there exists APCE for all units for $x \in \Omega_X$. The third assumption means the density function of X_Z is bounded. The fourth assumption implies that h is a zero function if $\mathbb{E}[h(X)|Z = z]$ does not depend on z.

Assumption 2 (Separability II)**.** *The function $f_Y(X, \boldsymbol{H}, \boldsymbol{u}_Y)$ is separable, i.e., it can be represented as $f_{Y_1}(X, \boldsymbol{u}_Y) + f_{Y_2}(\boldsymbol{H}, \boldsymbol{u}_Y)$.*

The following proposition holds (Wong, 2022):

Proposition 2.1. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1 and 2, APCE $\mathbb{E}[\partial_x Y_x]$ is identifiable via the following integral equation*

$$\mu(z) = \int_{\Omega_X} k(x, z)\mathbb{E}[\partial_x Y_x]dx, \tag{2}$$

where

$$\begin{aligned} \mu(z) &= \mathbb{E}[Y|Z = z_0] - \mathbb{E}[Y|Z = z] \\ k(x, z) &= \mathbb{P}_X[x|Z = z] - \mathbb{P}_X[x|Z = z_0] \end{aligned}, \tag{3}$$

and z_0 is a fixed value.

In this paper, we aim to estimate the APCE by solving the integral equation (2). We show that this is a well-posed problem, and develop nonparametric and parametric estimators.

3. Nonparametric Approach

In this section, we develop a nonparametric approach for estimating the APCE based on the Picard iteration method for solving integral equations. First, we assume

Assumption 3. *Ω_X and Ω_Z are bounded.*

Then, we show that solving the integral equation (2) is a well-posed problem (All proofs are given in Appendix A).

Proposition 3.1. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, and 3, solving the function $\mathbb{E}[\partial_x Y_x]$ via the integral equation (2) is a well-posed problem, that is, there exists a unique solution and the solution changes continuously with changes in the input functions.*

In contrast, the integral equations in NPTSLS (Newey & Powell, 2003) and CQE (Chernozhukov et al., 2007) for estimating $\mathbb{E}[Y_x]$ are ill-posed due to the use of a density function in the integral kernels instead of a CDF, which is bounded, in (2).

3.1. Nonparametric APCE estimator

The integral equation (2) is a "Fredholm Integral Equation of the First Kind" with k called an integral kernel (Bôcher, 1926). A Fredholm integral equation of the first kind is an integral equation of the form $b = \mathcal{L}(a)$ ($a, b \in \mathcal{H}$), for which Picard (1910) introduced a necessary and sufficient condition for the existence of a solution, called Picard's condition, shown in the following:

Picard's Condition. *Given an operator $\mathcal{L}$ and a function $b \in \mathcal{H}$, there is a function a such that $\mathcal{L}(a) = b$ if and only if $\sum_{i=1}^{\infty} \langle a, v_i \rangle^2 / \lambda_i^2 < \infty$, where v_i and λ_i are the eigenvalues and eigenvectors of $\mathcal{L}$, and $\langle a, v \rangle = 0$ for all v such that $\mathcal{L}(v) = 0$.*

We will construct a nonparametric estimator for APCE based on the Picard iteration scheme which is a powerful tool for solving integral equations (Fridman, 1965).

Picard Iteration Scheme. First, we make the following assumption:

Assumption 4. *$\Omega_X \subseteq \Omega_Z$.*

This means that the domain of Z includes the domain of X. This assumption is needed because the Picard iteration (5) is not defined in $\Omega_X \setminus \Omega_Z$. The assumption is often practical in the IV setting because the IV and treatment variables are often very similar variables and have the same domains. Denote an operator $\mathcal{K}$ be

$$(\mathcal{K}(a))(x) = \int_{\Omega_X} k(x', x)a(x')dx' \quad \text{for any } a \in \mathcal{H}. \tag{4}$$

Then the Picard iteration scheme for solving the integral equation (2) becomes

$$\theta_{t+1}(x) \leftarrow \theta_t(x) + \alpha\Big(\mu(x) - \int_{\Omega_X} k(x',x)\theta_t(x')dx'\Big) \quad (5)$$

for all $x \in \Omega_X$, where α is a real number representing a step size that satisfies $0 < \alpha < 2/\|\mathcal{K}\|$. We prove the following results to show that $\theta_t(x)$ converges to the APCE $\mathbb{E}[\partial_x Y_x]$. The following lemma holds:

Lemma 3.2. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, and 3, the operator $\mathcal{K}$ satisfies the following three properties:*

1. *$\mathcal{K}$ is a compact operator: $\mathcal{K}$ maps a bounded set into a compact set in the sense of strong convergence.*
2. *$\mathcal{K}$ is self-adjoint: $\langle \mathcal{K}(a), b\rangle = \langle a, \mathcal{K}(b)\rangle$ for $a, b \in \mathcal{H}$.*
3. *$\mathcal{K}$ is positive semi-definite: $\langle \mathcal{K}(a), a\rangle \geq 0$ for $a \in \mathcal{H}$.*

From Lemma 3.2, $\mathcal{K}$ possesses a countable set of positive eigenvalues, and the following lemma holds:

Lemma 3.3. *Under SCM $\mathcal{M}_{IV}$ and Assumption 1, $\mathcal{K}$ satisfies the Picard's condition.*

Finally, we obtain the following theorem:

Theorem 3.4. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, and 4, the Picard iteration scheme (5) converges strongly to the APCE, that is,* $\lim_{t\to\infty} \theta_t(x) = \mathbb{E}[\partial_x Y_x]$.

Nonparametric APCE (N-APCE) estimator. Next, we present our proposed nonparametric APCE estimator (named N-APCE) based on the Picard iteration scheme (5), shown in Algorithm 1.

We assume we have available a set of observations $\mathcal{D} = \{z^{(i)}, x^{(i)}, y^{(i)}\}_{i=1}^N$. The algorithm needs as inputs a stop threshold ϵ, a step size α, and an initial function $\theta_1(x)$. Let $\hat{\mathbb{E}}[Y|Z = z]$ be predictors of Y given $Z = z$, and $\hat{\mathbb{P}}_X[x|Z = z]$ be predictors of CDF of X given $Z = z$. These predictor functions will be learned using a supervised ML model from the observations $\mathcal{D}$ (Hastie et al., 2009). We denote $\hat{\mu}(x) = \hat{\mathbb{E}}[Y|Z = x_0] - \hat{\mathbb{E}}[Y|Z = x], \hat{k}(x', x) = \hat{\mathbb{P}}_X[x'|Z = x] - \hat{\mathbb{P}}_X[x'|Z = x_0]$, and an operator $\hat{\mathcal{K}}$ be $(\hat{\mathcal{K}}(a))(x) = \int_{\Omega_X} \hat{k}(x', x)a(x')dx$ for any $a \in \mathcal{H}$.

We approximate integrations by numerical integration. First, choose a finite set of values $\mathscr{X} = \{x_1, \ldots, x_R\} \in \Omega_X$ where $x_r < x_{r+1}$ for $r = 1, \ldots, R-1$. For example, we can use an equidistant interval division $x_r = \{\max(\Omega_X) - \min(\Omega_X)\} \times r/R + \min(\Omega_X)$. Note that X and Z share the values $\mathscr{X}$. Next, let $\mathcal{I}[a(x); \mathscr{X}]$ denote a numerical integration of the integration $\int_{\Omega_X} a(x)dx$ for any function $a \in \mathcal{H}$ given $\mathscr{X}$. $\mathcal{I}[a(x); \mathscr{X}]$ takes the form of (Burden et al., 2015)

$$\mathcal{I}[a(x); \mathscr{X}] = \sum_{q=1}^{R} I(x_q, x_{q+1})(x_{q+1} - x_q), \quad (6)$$

Algorithm 1 Nonparametric APCE (N-APCE) estimator

1: **Input:** A set of observations $\mathcal{D} = \{z^{(i)}, x^{(i)}, y^{(i)}\}_{i=1}^N$, a stop threshold ϵ, a step size α, and a set of X values $\mathscr{X} = (x_0, x_1, \ldots, x_R)$.
2: Learn two predictive functions $\hat{\mathbb{E}}[Y|Z = z]$ and $\hat{\mathbb{P}}_X[x|Z = z]$ from the observations $\mathcal{D}$ using a supervised ML method.
3: Initialize the function $\theta_1(x)$, and $t \leftarrow 1$.
4: Calculate $R + R^2$ values

$$\hat{\mu}(x_r) = \hat{\mathbb{E}}[Y|Z = x_0] - \hat{\mathbb{E}}[Y|Z = x_r]$$
$$\hat{k}(x_q, x_r) = \hat{\mathbb{P}}_X[x_q|Z = x_r] - \hat{\mathbb{P}}_X[x_q|Z = x_0]$$

for $q, r = 1, \ldots, R$.
5: **while** $\{\mathcal{I}[(\hat{\mu}(x) - \mathcal{I}[\hat{k}(x', x)\hat{\theta}_t(x'); \mathscr{X}])^2; \mathscr{X}]\}^{1/2} > \epsilon$
6: Update the function $\theta_{t+1}(x_r)$ by

$$\hat{\theta}_{t+1}(x_r) \leftarrow \hat{\theta}_t(x_r) + \alpha\Big(\hat{\mu}(x_r) - \mathcal{I}[\hat{k}(x', x_r)\hat{\theta}_t(x'); \mathscr{X}]\Big)$$

for $r = 1, \ldots, R$.
7: **end while**
8: **Return** a function $\hat{\theta}(x)$ as the N-APCE estimator by interpolating over the final step function $\hat{\theta}_T(x_r)$ values for $r = 1, \ldots, R$.

where $I(x_q, x_{q+1})$ can take different forms. For example, the left hand rule uses $I(x_q, x_{q+1}) = a(x_q)$. See Appendix B.1 for other options.

We introduce the following empirical risk to use as a stopping criterion:

$$J_N(\theta; \mathcal{D}) = \left\{\int_{\Omega_X}\left(\hat{\mu}(x) - \int_{\Omega_X}\hat{k}(x', x)\theta(x')dx'\right)^2 dx\right\}^{1/2}. \quad (7)$$

The empirical risk contains two integrations which will be approximated by numerical integration. Hence, the numerical empirical risk is computed as below:

$$\tilde{J}_N(\theta; \mathcal{D}) = \{\mathcal{I}[(\hat{\mu}(x) - \mathcal{I}[\hat{k}(x', x)\theta(x'); \mathscr{X}])^2; \mathscr{X}]\}^{1/2}. \quad (8)$$

To run the Picard iteration and compute the numerical empirical risk, we first calculate the following $R + R^2$ values

$$\hat{\mu}(x_r) = \hat{\mathbb{E}}[Y|Z = x_0] - \hat{\mathbb{E}}[Y|Z = x_r]$$
$$\hat{k}(x_q, x_r) = \hat{\mathbb{P}}_X[x_q|Z = x_r] - \hat{\mathbb{P}}_X[x_q|Z = x_0] \quad (9)$$

for $q, r = 1, \ldots, R$. At each iteration, update $\hat{\theta}_t$ as:

$$\hat{\theta}_{t+1}(x_r) \leftarrow \hat{\theta}_t(x_r) + \alpha\Big(\hat{\mu}(x_r) - \mathcal{I}[\hat{k}(x', x_r)\hat{\theta}_t(x'); \mathscr{X}]\Big) \quad (10)$$

for $r = 1, \dots, R$, from the initial function $\hat{\theta}_1(x) = \theta_1(x)$ until $\hat{\theta}_t$ satisfies the stop condition $\tilde{J}_N(\theta; \mathcal{D}) \le \epsilon$.

Finally, after the Picard iteration converges to $\hat{\theta}_T$, compute a function $\hat{\theta}(x)$ as the estimator of the APCE by interpolating over the function $\hat{\theta}_T(x_r)$ values for $r = 1, \dots, R$. We use the Lagrange interpolating polynomial (Jeffreys & Jeffreys, 1988). See Appendix B.2 for the details.

3.2. Properties of N-APCE estimator

The error in the N-APCE estimator due to the interpolation is well studied and understood in the field of numerical analysis (Burden et al., 2015). The convergence of the Picard iteration and the error in $\hat{\theta}_T(x_r)$ depend on the error in numerical integration and the error in estimating the predictor functions by ML methods. The former error is well understood in the field of numerical analysis (Burden et al., 2015). Thus, we will focus on the impacts of the ML error on the N-APCE estimator.

Consistency and Computational Complexity. First, we investigate the consistency and the algorithm complexity of the N-APCE estimator. We make the following assumption:

Assumption 5. *$\hat{\mu}$ and $\hat{k}$ learned by ML methods are consistent estimators of μ and k in (3).*

We assume the values in $\mathscr{X}$ satisfy $[x_0, x_R] = [\min(\Omega_X), \max(\Omega_X)]$ and $\lim_{R\to\infty} |x_{r+1} - x_r| = 0$ for $r = 1, \dots, R$. Then, we obtain the following result:

Theorem 3.5. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 5, taking limits $N \to \infty$, $R \to \infty$, and $t \to \infty$, $\hat{\theta}_t(x)$ is a pointwise consistent estimator of the APCE $\mathbb{E}[\partial_x Y_x]$ for $x \in \Omega_X$ almost everywhere.*

Next, we show the termination of Algorithm 1. We make the following assumption:

Assumption 6. *$\hat{\mathcal{K}}$ satisfies the Picard's condition.*

Then, the following result holds:

Theorem 3.6. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 6, taking the limit $R \to \infty$, Algorithm 1 stops after a finite number of iterations for any $\epsilon > 0$.*

We denote $\lim_{t\to\infty} \hat{\theta}_t(x) = \hat{\theta}_\infty(x)$. Then the following corollary holds:

Corollary 3.7. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 6, the sequence $\hat{\theta}_t$ converges linearly to $\hat{\theta}_\infty$.*

The complexity of building predictive models and calculating $R + R^2$ prediction values depends on the ML methods used. Assume that Algorithm 1 stops after T iterations, then numerical integration takes total $\mathcal{O}(T \times R^2)$ time, and the final interpolation takes $\mathcal{O}(N \times R^2)$ time. Long iterations T may provoke serious problems in the calculation.

Bias and Variance. We investigate the bias and the variance of the N-APCE estimator. The estimator contains an attenuation bias (Wooldridge, 2010), which is caused by the errors in $\hat{\mathcal{K}}$. We make the following assumption:

Assumption 7. *$\hat{\mu}$ and $\hat{k}$ learned by ML methods are unbiased estimators of μ and k in (3).*

We compute $\hat{\mu}$ and $\hat{k}$ by the conditional sample means in the experiments, which satisfy Assumptions 5 and 7.

Then, we obtain the following result:

Theorem 3.8. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, 6, and 7, letting $\hat{\mathcal{K}}^{-1} = \alpha \sum_{t=0}^{\infty} (I - \alpha\hat{\mathcal{K}})^t$, if $\|\hat{\mathcal{K}}^{-1}\|$ is bounded by M, then the expected absolute bias $\mathbb{E}[\|\hat{\theta}_\infty - \theta_\infty\|]$ is bounded by $M(A + \|\theta_\infty\| B)$, where*

$$A = \sqrt{\int_{\Omega_X} \mathbb{V}(\hat{\mu}(x)) dx}, \; B = \sqrt{\int_{\Omega_X} \mathbb{V}(\hat{k}(x, x)) dx}. \quad (11)$$

The conditional variance functions in A and B can be computed using the method in (Fan & Yao, 1998). The expected absolute bias decreases according to $O(g(N)^{1/2})$ if the conditional variance functions of ML estimation decrease according to $O(g(N))$. For example, the conditional sample means used in the experiments have a rate of $O(N^{-1})$, then the expected absolute bias decreases according to $O(N^{-1/2})$. Furthermore, the N-APCE estimator also has other biases due to numerical integration and interpolation.

Finally, we assess the variance of the N-APCE estimator at $X = x$. We obtain the following theorem:

Theorem 3.9. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 7, when Algorithm 1 stops at $t = T$, the upper bound of the variance of $\hat{\theta}_T(x)$ is $\alpha^2 (T-1)^2 \nu(x) + \mathcal{O}(\alpha^3)$ as $\alpha \to 0$ for $x \in \Omega_X$, where $\nu(x)$ is*

$$\begin{aligned} &\mathbb{V}(\hat{\mu}(x)) + \Big(\int_{\Omega_X} \sqrt{\mathbb{V}(\hat{k}(x', x))} |\theta_T(x')| dx' \Big)^2 \\ &\quad + 2\sqrt{\mathbb{V}(\hat{\mu}(x))} \Big(\int_{\Omega_X} \sqrt{\mathbb{V}(\hat{k}(x', x))} |\theta_T(x')| dx' \Big). \quad (12) \end{aligned}$$

Furthermore, the following corollary holds:

Corollary 3.10. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 7, when Algorithm 1 stops at $t = T$, the variance of $\hat{\theta}_T(x)$ is*

$$\begin{aligned} &\alpha^2 (T-1)^2 \mathbb{V}(\hat{\mu}(x)) \\ &\quad + \mathcal{O}(\alpha^3) + \mathcal{O}(\{\max_{x'} \{\mathbb{V}(\hat{k}(x', x))\}\}^{1/2}) \qquad (13) \end{aligned}$$

as $\alpha \to 0$, $\{\max_{x'} \{\mathbb{V}(\hat{k}(x', x))\}\}^{1/2} \to 0$ for $x \in \Omega_X$.

The limit $\{\max_{x'} \{\mathbb{V}(\hat{k}(x', x))\}\}^{1/2} \to 0$ means the ML estimator $\hat{\mathbb{E}}[Y|Z = z]$ exhibits small conditional variance.

4. Parametric Approach

In this section, we present a parametric approach for estimating the APCE.

4.1. Parametric APCE estimator

Linear basis function model. To solve the integral equation (2), we parameterize the APCE by a linear basis function model

$$\mathbb{E}[\partial_x Y_x] = \sum_{p=1}^{P} \theta_p \phi_p(x) \tag{14}$$

using the basis functions $\{\phi_p(x)\}_{p=1,\ldots,P}$ (Bishop, 2006), where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_P\}$ are the model parameters to be estimated from data. Then, the integral equation (2) becomes

$$\mu(z) = \sum_{p=1}^{P} \theta_p \int_{\Omega_X} k(x,z)\phi_p(x)dx. \tag{15}$$

Letting the anti-derivative of the basis functions be $\varphi_p(x) = \int \phi_p(x)dx$ for $p = 1, \ldots, P$, the integral equation becomes

$$\mu(z) = \sum_{p=1}^{P} \theta_p \{\mathbb{E}[\varphi_p(X)|Z=z] - \mathbb{E}[\varphi_p(X)|Z=z_0]\}. \tag{16}$$

Next, we show that the estimation problem reduces to a system of linear equations.

First, select a set of values $\{z_1, \ldots, z_R\} \in \Omega_Z$, where $z_r < z_{r+1}$ for $r = 1, \ldots, R-1$. Let $c_r = \mathbb{E}[Y|Z=z_r] - \mathbb{E}[Y|Z=z_0]$, $d_r^p = \mathbb{E}[\varphi_p(X)|Z=z_r] - \mathbb{E}[\varphi_p(X)|Z=z_0]$ for $r = 1, \ldots, R$ and $p = 1, \ldots, P$. Furthermore, denote $\boldsymbol{d}^p = (d_1^p, \ldots, d_R^p)^T$ for $p = 1, \ldots, P$, $\boldsymbol{c} = (c_1, \ldots, c_R)^T$, and $\mathbf{D} = (\boldsymbol{d}^1, \ldots, \boldsymbol{d}^P)$. Then, parameters $\boldsymbol{\theta}$ are given by solving $\boldsymbol{c} = \mathbf{D}^T\boldsymbol{\theta}$. Here, $\mathbf{D}^T$ denotes the transposed matrix of $\mathbf{D}$.

Parametric APCE estimator. Next, we present our proposed parametric APCE estimator (named P-APCE), shown in Algorithm 2.

We assume we have available observations $\mathcal{D} = \{z^{(i)}, x^{(i)}, y^{(i)}\}_{i=1}^N$. Let $\hat{\mathbb{E}}[Y|Z=z]$ be predictors of Y given $Z=z$, and $\hat{\mathbb{E}}[\varphi_p(X)|Z=z]$ be predictors of $\varphi_p(X)$ given $Z=z$ for $p = 1, \ldots, P$. These predictor functions will be learned using a supervised ML model from the observations $\mathcal{D}$ (Hastie et al., 2009). Then, calculate the following $R + R \times P$ values

$$\begin{aligned} \hat{c}_r &= \hat{\mathbb{E}}[Y|Z=z_r] - \hat{\mathbb{E}}[Y|Z=z_0] \\ \hat{d}_r^p &= \hat{\mathbb{E}}[\varphi_p(X)|Z=z_r] - \hat{\mathbb{E}}[\varphi_p(X)|Z=z_0] \end{aligned} \tag{17}$$

for $r = 1, \ldots, R$ and $p = 1, \ldots, P$. Denote $\hat{\boldsymbol{d}}^p = (\hat{d}_1^p, \ldots, \hat{d}_R^p)^T$ for $p = 1, \ldots, P$, $\hat{\boldsymbol{c}} = (\hat{c}_1, \ldots, \hat{c}_R)^T$, and $\hat{\mathbf{D}} = (\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^P)$. We obtain the estimator $\hat{\boldsymbol{\theta}}$ by minimizing the following empirical risk (Olive, 2017)

$$J_P(\boldsymbol{\theta}; \mathcal{D}) = \|\hat{\boldsymbol{u}} - \hat{\mathbf{D}}\boldsymbol{\theta}\|^2. \tag{18}$$

The solution is given as $(\hat{\mathbf{D}}^T\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}^T\hat{\boldsymbol{u}}$ if the matrix $\hat{\mathbf{D}}^T\hat{\mathbf{D}}$ is invertible; otherwise, it is solvable by the singular value decomposition (Mandel, 1982) or the regularization techniques (Hilt et al.).

Algorithm 2 Parametric APCE (P-APCE) estimator

1: **Input:** A set of observations $\mathcal{D} = \{z^{(i)}, x^{(i)}, y^{(i)}\}_{i=1}^N$, the basis functions $\{\phi_p(x)\}_{p=1,\ldots,P}$, and a set of values $\{z_1, \ldots, z_R\} \in \Omega_Z$.
2: Learn predictive functions $\hat{\mathbb{E}}[Y|Z=z]$ and $\hat{\mathbb{E}}[\varphi_p(X)|Z=z]$ for $p = 1, \ldots, P$ from the observations $\mathcal{D}$ using a supervised ML method.
3: Calculate $R + R \times P$ values

$$\begin{aligned} \hat{c}_r &= \hat{\mathbb{E}}[Y|Z=z_r] - \hat{\mathbb{E}}[Y|Z=z_0] \\ \hat{d}_r^p &= \hat{\mathbb{E}}[\varphi_p(X)|Z=z_r] - \hat{\mathbb{E}}[\varphi_p(X)|Z=z_0] \end{aligned}$$

for $r = 1, \ldots, R$ and $p = 1, \ldots, P$.
4: Letting $\hat{\boldsymbol{c}} = (\hat{c}_1, \ldots, \hat{c}_R)^T$, $\hat{\boldsymbol{d}}^p = (\hat{d}_1^p, \ldots, \hat{d}_R^p)^T$ for $p = 1, \ldots, P$, and $\hat{\mathbf{D}} = (\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^P)$, solve the optimization problem $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \|\hat{\mathbf{c}} - \hat{\mathbf{D}}\boldsymbol{\theta}\|^2$.
5: **Return** $\sum_{p=1}^{P} \hat{\theta}_p \phi_p(x)$ as the P-APCE estimator.

4.2. Properties of P-APCE estimator

Next, we show the properties of the P-APCE estimator. We make the following assumption:

Assumption 8. *$\hat{\boldsymbol{c}}$ and $\hat{\mathbf{D}}$ learned by ML methods are consistent estimators of $\boldsymbol{c}$ and $\mathbf{D}$.*

Then, the following theorem holds:

Theorem 4.1. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, and 8, the estimator $\hat{\boldsymbol{\theta}}$ given by Algorithm 2 is a pointwise consistent estimator of $\boldsymbol{\theta}$ in Eq. (15).*

The estimator $\hat{\theta}$ has a bias since this model can be considered as an errors-in-variables model (Söderström, 2007) or a measurement error model (Fuller, 2009). Since the norm of the bias of the errors-in-variables model decreases according to the inverse of the norm of $\hat{\mathbf{D}}^T\hat{\mathbf{D}}$ (Greene, 1997), the expected norm of bias decreases according to $\mathcal{O}(N^{-1/2})$ if the conditional variances of prediction values decrease according to $\mathcal{O}(N^{-1})$ as $N \to \infty$. As for computational complexity, the complexity of building predictive models and calculating $R + R \times P$ prediction values depends on the ML method. The inverse matrix is computed in $O(R^3)$ time.

Model Selection. We can use the empirical risk in equation (18) as a performance metric of the trained model with parameters $\hat{\boldsymbol{\theta}}$ given a separate test dataset $\mathcal{D}' = \{z'^{(i)}, x'^{(i)}, y'^{(i)}\}_{i=1}^{N'}$. Assume $\hat{\boldsymbol{c}}'$ and $\hat{\mathbf{D}}'$ are computed using $\mathcal{D}'$. Then, we can evaluate the trained model by the following test error:

$$J_P^{test}(\hat{\boldsymbol{\theta}}; \mathcal{D}') = \|\hat{\boldsymbol{c}}' - \hat{\mathbf{D}}'\hat{\boldsymbol{\theta}}\|^2. \tag{19}$$

Given a separate dataset, this performance metric can also be used for model selection from among various candidate basis functions or the number P of basis.

5. Experiments

In this section, we present numerical experiments to demonstrate the performance of the P-APCE and N-APCE estimators. We compare them with the parametric method TSPS (Terza et al., 2008) and the nonparametric method NPTSLS (Newey & Powell, 2003). Note that among the existing methods summarized in Table 1, GMM requires the distribution of the IV which we don't assume available, and CQE requires monotonicity. NPTSLS computes $\mathbb{E}[Y_x]$ which we differentiate to compute APCE $\mathbb{E}[\partial_x Y_x]$.

Settings. We consider the following SCM:

$$\begin{cases} X := \frac{1}{25}Z^2 + \frac{1}{5}Z + 0.5 + (\frac{Z}{3} + 0.1)U \\ Y := X^3 + X^2 + X + U + E \end{cases}. \tag{20}$$

This SCM has non-linear functions for both f_X and f_Y; it satisfies separability Ⅱ but not separability I; and it satisfies Assumption 1. Each realized value of U and E are i.i.d. and sampled from a uniform distribution $U[-1, 1]$.

We generated random samples using the SCM in (20) for 11 different values of Z in $(0, 0.3, \ldots, 2.7, 3)$. The sample size at each Z value is 10 and 100, respectively, for a total sample size of $N = 100$ and $N = 1000$. We compute the $\hat{\mu}$ and $\hat{k}$ in (3) by the conditional sample means, e.g, $\hat{\mathbb{E}}[Y|Z = z] = \{\sum_{i=1}^{N} y^{(i)}\mathbb{1}_{z^{(i)}=z}\}/\{\sum_{i=1}^{N} \mathbb{1}_{z^{(i)}=z}\}$. We also compute the $\hat{\boldsymbol{c}}$ and $\hat{\mathbf{D}}$ by the conditional means. The conditional means satisfy Assumptions 5 and 7. We conduct each simulation 100 times.

Settings of N-APCE (Algorithm 1) We let $\mathscr{X} = \{0, 0.3, \ldots, 2.7, 3\}$; and the N-APCE estimator at $X = 0$ is not defined since x_0 is 0. We calculate the numerical integration using the left-hand rule. We let the initial function $\hat{\theta}_1$ be a zero function, and the stop threshold ϵ be 10. We choose the step size as the smallest one from $(1, 0.5, 0.1, \ldots)$ when Algorithm 1 stops before 100 iterations, and the chosen step size α is 0.5.

Settings of P-APCE (Algorithm 2) We use the polynomial basis functions $\phi_p(x) = x^{p-1}$ for $p = 1, 2, \ldots$, and calculate the solution of the equation (18) by $(\hat{\mathbf{D}}^T\hat{\mathbf{D}})^{-1}\hat{\mathbf{D}}^T\hat{\boldsymbol{c}}$. To determine the best degree of the model, we separate the data set into training set $\mathcal{D}$ and validation set $\mathcal{D}'$, estimate $\hat{\theta}$ by the training set, and evaluate the trained model using the performance measure (19). We simulate 100 times and compute the performance measure. From the results (shown in Table 5 in the appendix), we decide that the highest degree of the polynomial functions in the P-APCE estimator will be 3, when the mean of the performance measure is the smallest. Due to overfitting, the validation errors gradually increase when the model degree is greater than 4. We let the degree of TSPS also be 3. For NPTSLS, we use the Hermite polynomial basis functions $h_0(X) = 1$, $h_1(X) = X$, $h_2(X) = X^2 - 1$.

Results. The means and standard deviations (SD) of the N-APCE estimator at different X values over 100 runs, and the approximate SD by the equation (13) are shown in Table 2. The means and the SD of the estimated coefficients by P-APCE and TSPS are shown in Table 3. The boxplots of the estimated values by N-APCE, P-APCE, TSPS, and NPTSLS at each point in $(0.3, 0.6, \ldots, 2.7, 3)$ are shown in Figure 2.

We have the following observations from the results. The P-APCE estimator performed superior to the TSPS - this may be because the underlying IV model does not satisfy separability I. The SD and biases of the P-APCE estimators are relatively large when $N = 100$, and the estimators have relatively small biases and SD when $N = 1000$. In contrast, the TSPS estimators have relatively large biases and SD for both $N = 100$ and $N = 1000$. The N-APCE estimator performed superior to NPTSLS. The interquartile range (IQR) of N-APCE is narrower than that of NPTSLS in Figure 2.

The N-APCE estimators have relatively small SD, and the means are close to the true APCE values. Compared to the P-APCE estimator, the N-APCE estimator is less likely to misrepresent the form of the function. The P-APCE estimators sometimes become an upward convex function at small X values, which misrepresents the characteristic of the function. In addition, the approximate SD by the equation (13) is close to the SD of the N-APCE estimates. Additional information about this experiment is given in Appendix C.1.

We have performed additional numerical experiments on non-monotonic (Appendix C.2) or non-polynomial (Appendix C.3) APCE functions, and experiments on a model satifying both separability I and Ⅱ (Appendix C.4). We have the following observations from the results. First, our N-APCE and P-APCE estimators work well in situations where the APCE is not monotonic. Second, in a situation where the APCE is not polynomial, the P-APCE estimator does not work well and the N-APCE estimator still works well; thus, the N-APCE estimator is superior to the parametric

Table 2: The means and standard deviations (SD) of the N-APCE over 100 runs, and the means of the approximate SD by the equation (13) (Approx SD) at $(0.3, 0.6, \ldots, 2.7, 3.0)$ when $N = 100$ and $N = 1000$.

$X = x$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True APCE	1.87	3.28	5.23	7.72	10.75	14.32	18.43	23.08	28.27	34
$(N = 100)$ Mean	2.519	4.416	6.907	10.001	13.588	18.721	24.918	30.223	35.955	44.400
SD	4.386	4.577	5.387	6.335	8.269	8.145	11.114	12.389	14.135	16.371
Approx SD	4.578	5.481	5.408	5.478	7.472	11.881	15.197	14.683	19.716	24.110
$(N = 1000)$ Mean	1.822	3.647	5.397	7.514	10.561	14.394	18.601	24.017	30.521	37.536
SD	1.365	1.180	1.161	1.415	1.971	2.373	3.094	3.813	4.472	5.004
Approx SD	1.246	1.414	1.655	2.010	2.468	2.987	3.817	4.789	5.843	6.824

Table 3: The means and standard deviations (SD) of the P-APCE, and the TSPS over 100 runs when $N = 100$ and $N = 1000$; "$D = m$" means "the estimated coefficient of the m-th degree term." The true coefficients are 1, 2, 3 for $D = 0, 1, 2$.

Results		$D = 0$	$D = 1$	$D = 2$
P-APCE	Mean	7.879	-11.128	8.525
$N = 100$	SD	10.516	20.884	9.214
TSPS	Mean	-9.749	29.489	-11.571
$N = 100$	SD	48.039	118.938	69.845
P-APCE	Mean	0.995	2.132	2.878
$N = 1000$	SD	4.951	9.477	3.997
TSPS	Mean	-1.590	8.403	1.531
$N = 1000$	SD	15.110	37.188	21.470

method when a reasonable model for the data is unknown. Finally, our N-APCE and P-APCE estimators are superior to the TSPS and NPTSLS in terms of the SD of the estimates even when the underlying IV model satisfies the separability I and Ⅱ.

6. Application in a Real-World Dataset

In this section, we present an application of our estimators to real-world data in economics.

Real-world Dataset. We take up an open dataset in the R package "wooldridge" (`https://cran.r-project.org/package=wooldridge`), which was analyzed by Griliches (1977) and Blackburn & Neumark (1992). The data source is the National Longitudinal Survey of Young Men, and the sample size is 935. We estimate the effect of years of education on monthly wages, which is of great interests in economics (Card, 1999; Angrist & Krueger, 1991). Since researchers cannot force people to attend or drop out of school, they use the mother's years of education as an instrumental variable. We take the subject's years of education as the treatment variable (X), their monthly wage as the outcome variable (Y), and their mother's years of education as the instrumental variable (Z). Here, X and Z are discretized continuous variables, and the domains of X and Z are $\{9, 10, \ldots, 18\}$, ranging from the 1st year of high school to the 2nd year of master's degree. We exclude samples where one of the three variables is NA. We estimate the conditional expectation using the conditional means. We determined the degree of polynomials in the P-APCE estimator by the test error (19), and chose linear functions for the candidates of the APCE. We evaluate the APCE by 1000 times bootstrapping method. To reduce the variance of the estimator, we regularize the matrix $\hat{\mathbf{D}}^T\hat{\mathbf{D}}$ by adding $0.1 \times \mathbf{I}$, where $\mathbf{I}$ is an identity matrix of size R.

Results. We show the basic bootstrapping statistical properties of the P-APCE estimators and the TSPS in Table 4. The N-APCE estimator did not converge. For the P-APCE estimator, the mean of the constant term is 192.491; the mean of the coefficient of the first degree term is -10.267. As for the TSPS, the mean of the constant term is 108.484; the mean of the coefficient of the first degree term is 0.073. Both P-APCE and TSPS predict that years of education increase the wages, which is consistent with the results of previous works (Blackburn & Neumark, 1992; Wooldridge, 2010). On the other hand, the result of the TSPS implies that the effect of years of education on wages is close to constant; however, the result of the P-APCE estimator implies that the effect of years of education on wages gets weaker from year to year. Our results from the P-APCE estimator suggest that education significantly affects wages at the compulsory school level, which coincides with (Angrist & Krueger, 1991); on the other hand, education has little effect at the college level. The increase in wages by getting a higher education at the college level seems to be due to the phenomenon described in Spence (1973) and Caplan (2018) that people with an academic degree earn higher incomes than people who don't have an academic degree, even if they possess the same skills, not the effect of education. This difference is called "sheepskin effect" which is described in Jaeger & Page (1996) as "*the difference in earnings between individuals possessing a diploma and those who do*

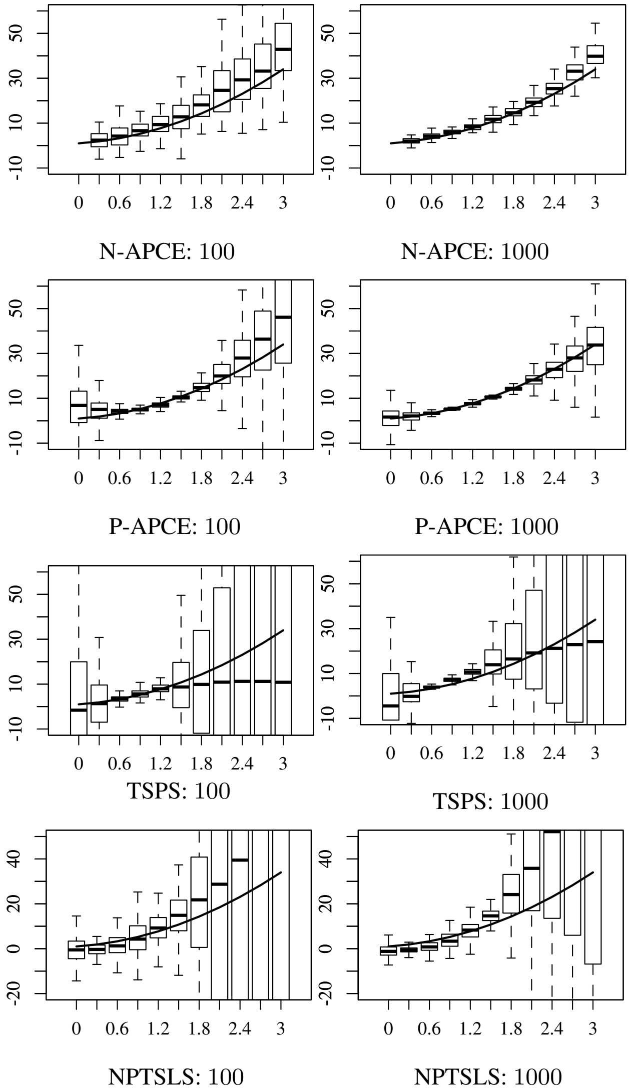

Figure 2: Boxplots of the estimated APCE values by the N-APCE, P-APCE, TSPS, and NPTSLS estimators at $X = (0, 0.3, \ldots, 2.7, 3.0)$. The X-axis is the value of the treatment variable X, and Y-axis is the value of the APCE. The black curves are the true APCE.

not conditional on years of schooling."

7. Conclusion

In this paper, we have developed two novel methods for estimating the APCE of a continuous treatment via an instrumental variable. We analyzed the properties of the proposed P-APCE and N-APCE estimators and demonstrated their applications on synthetic and real-world data. The performance of the parametric P-APCE estimators depends critically on the choice of the basis functions. Nonparametric N-APCE estimators do not have to make functional assumptions but are computationally expensive.

Table 4: The results of the P-APCE estimator and the TSPS for the real-world dataset.

	P-APCE estimator		TSPS	
	$D=0$	$D=1$	$D=0$	$D=1$
Min.	-664.970	-55.854	-990.480	-82.835
1st Qu.	73.724	-18.613	-106.334	-15.593
Median	186.969	-9.842	118.563	-0.657
3rd Qu.	312.781	-1.939	313.215	16.273
Max.	834.128	51.523	1216.730	82.870
Mean	192.491	-10.267	108.484	0.073
SD	182.698	13.0290	325.414	24.646

In contrast to the most existing work, the P-APCE and N-APCE estimators do not directly estimate the effect $\mathbb{E}[Y_x]$, but the APCE $\mathbb{E}[\partial_x Y_x]$. However, the main interest in causal inference is to infer the effects of the treatment under changing conditions (Pearl, 2010); thus, the APCE is often sufficient to reveal causal relationships. In particular, APCE enables us to evaluate a popular target, the average causal effect (ACE) of changing treatments from x' to x'', by $\mathbb{E}[Y_{x'}] - \mathbb{E}[Y_{x''}] = \int_{x''}^{x'} \mathbb{E}[\partial_x Y_x] dx$.

Acknowledgements

The authors thank the anonymous reviewers for their time and thoughtful comments. Yuta Kawakami was supported by JSPS KAKENHI Grant Number 22J21928. Manabu Kuroki was supported by JSPS KAKENHI Grant Number 19K11856 and 21H03504. Jin Tian was partially supported by NSF grant IIS-2231797.

References

Alexanderian, A. On compact operators. *Journal of Mathematical Analysis and Applications*, 18(2):5–36, 2013.

Angrist, J. D. and Krueger, A. B. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Angrist, J. D. and Krueger, A. B. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, December 2001.

Bahadori, M. T., Tchetgen, E. J. T., and Heckerman, D. E. End-to-end balancing for causal continuous treatment-effect estimation. In *ICML 2022, UAI 2022 Workshop on Advances in Causal Inference*, 2022.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Baum, C. F., Schaffer, M. E., and Stillman, S. Instrumental variables and gmm: Estimation and testing. *The Stata Journal*, 3(1):1–31, 2003.

Bishop, C. *Pattern Recognition and Machine Learning*. Springer, January 2006.

Blackburn, M. and Neumark, D. Unobserved ability, efficiency wages, and interindustry wage differentials. *The Quarterly Journal of Economics*, 107(4):1421–1436, 1992. ISSN 00335533, 15314650. URL `http://www.jstor.org/stable/2118394`.

Bôcher, M. *An introduction to the study of integral equations*. Number 10. University Press, 1926.

Breusch, T. S. and Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294, 1979.

Burden, R. L., Faires, J. D., and Burden, A. M. *Numerical analysis*. Cengage learning, 2015.

Caplan, B. *The Case against Education: Why the Education System Is a Waste of Time and Money*. Princeton University Press, 2018. ISBN 9780691174655. URL `http://www.jstor.org/stable/j.ctvc772xh`.

Card, D. Chapter 30 - the causal effect of education on earnings. volume 3 of *Handbook of Labor Economics*, pp. 1801–1863. Elsevier, 1999.

Chamberlain, G. Panel data. In Griliches†, Z. and Intriligator, M. D. (eds.), *Handbook of Econometrics*, volume 2, chapter 22, pp. 1247–1318. Elsevier, 1 edition, 1984. URL `https://EconPapers.repec.org/RePEc:eee:ecochp:2-22`.

Chen, X., Chernozhukov, V., Lee, S., and Newey, W. K. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809, 2014.

Chernozhukov, V., Imbens, G. W., and Newey, W. K. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4–14, 2007.

Diaz, J. B. and Metcalf, F. T. On iteration procedures for equations of the first kind, ax = y, and picard's criterion for the existence of a solution. *Mathematics of Computation*, 24(112):923–935, 1970.

Fan, J. and Yao, Q. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660, 1998.

Fridman, V. A method of successive approximations for fredholm integral equations of the first kind. *Uspeki mat Nauk*, 11:233–234, 1965.

Fuller, W. A. *Measurement error models*. John Wiley & Sons, 2009.

Graham, B. S. and Powell, J. L. Identification and estimation of average partial effects in "irregular" correlated random coefficient panel data models. *Econometrica*, 80(5):2105–2152, 2012. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/23271443`.

Greene, W. *Econometric Analysis*. Prentice-Hall international editions. Prentice Hall, 1997.

Griliches, Z. Estimating the returns to schooling: Some econometric problems. *Econometrica*, 45(1):1–22, 1977. ISSN 00129682, 14680262.

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054, 1982.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep IV: A flexible approach for counterfactual prediction. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1414–1423. PMLR, 06–11 Aug 2017.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.

Hausman, J. A. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

Hilt, D. E., Seegrist, D. W., Service., U. S. F., and Northeastern Forest Experiment Station (Radnor, P. *Ridge, a computer program for calculating ridge regression estimates*, volume no.236. Upper Darby, Pa, Dept. of Agriculture, Forest Service, Northeastern Forest Experiment Station, 1977.

Hirano, K. and Imbens, G. *The Propensity Score with Continuous Treatments*, pp. 73–84. Wiley-Blackwell, July 2005. ISBN 047009043X. doi: 10.1002/0470090456.ch7.

Imbens, G. W. Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358, 2014. ISSN 08834237, 21688745. URL `http://www.jstor.org/stable/43288511`.

Imbens, G. W. and Angrist, J. D. Identification and estimation of local average treatment effects. *Econometrica*, 62 (2):467–475, 1994.

Imbens, G. W. and Newey, W. K. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009.

Jaeger, D. and Page, M. Degrees matter: New evidence on sheepskin effects in the returns to education. *The Review of Economics and Statistics*, 78(4):733–40, 1996.

Jeffreys, H. and Jeffreys, B. Lagrange's interpolation formula. *Methods of mathematical physics*, 3:260, 1988.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79 (4):1229–1245, 2017.

Mandel, J. Use of the singular value decomposition in regression analysis. *The American Statistician*, 36(1): 15–24, 1982.

Muscat, J. *Functional Analysis: An Introduction to Metric Spaces, Hilbert Spaces, and Banach Algebras*. Springer International Publishing, 2014.

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

Olive, D. J. Multiple linear regression. In *Linear regression*, pp. 17–83. Springer, 2017.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Pearl, J. An introduction to causal inference. *Int J Biostat*, 6(2):Article 7, Feb 2010.

Picard, É. Sur un théorème général relatif aux équations intégrales de première espèce et sur quelques problèmes de physique mathématique. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 29:79–97, 1910.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Söderström, T. Errors-in-variables methods in system identification. *Automatica*, 43(6):939–958, 2007.

Spence, M. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973. ISSN 00335533, 15314650. URL `http://www.jstor.org/stable/1882010`.

Stock, J. Instrumental variables in statistics and econometrics. In Smelser, N. J. and Baltes, P. B. (eds.), *International Encyclopedia of the Social Behavioral Sciences*, pp. 7577–7582. Pergamon, Oxford, 2001.

Su, L., Tu, Y., and Ullah, A. Testing additive separability of error term in nonparametric structural models. *Econometric Reviews*, 34(6-10):1057–1088, 2015.

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. *Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Terza, J. V., Basu, A., and Rathouz, P. J. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27 (3):531–543, 2008.

Tikhonov, A. N., Goncharsky, A., Stepanov, V., and Yagola, A. G. *Numerical methods for the solution of ill-posed problems*, volume 328. Springer Science & Business Media, 1995.

Torgovitsky, A. Identification of nonseparable models using instruments with small support. *Econometrica*, 83(3): 1185–1197, 2015.

Wang, L. and Tchetgen Tchetgen, E. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *J R Stat Soc Series B Stat Methodol*, 80(3):531–550, Jun 2018.

Wong, W. H. An equation for the identification of average causal effect in nonlinear models. *Statistica Sinica*, 32: 539–545, 2022.

Wooldridge, J. M. Unobserved heterogeneity and estimation of average partial effects. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, pp. 27–55, 2005.

Wooldridge, J. M. *Econometric analysis of cross section and panel data*. MIT press, 2010.

Appendix

A. Proofs

A.1. Proof of Proposition 3.1

Proposition 3.1. *Under Assumptions 1, 2, and 3, solving the function $\mathbb{E}[\partial_x Y_x]$ via the integral equation (2) is a well-posed problem, that is, there exists a unique solution and the solution changes continuously with changes in the input functions.*

Proof. First, note that a problem $\mathcal{K}f = g$ is a well-posed problem if

1. a solution exists,
2. the solution is unique, and
3. the solution changes continuously with changes in the input operator $\mathcal{K}$ and function g.

Problems in which one or more of the conditions fails to hold are called ill-posed problems (Tikhonov et al., 1995). First, since this integral equation has a unique solution from Wong (2022), it satisfies the first and second conditions. From Assumptions 1, 2, and 3, the operator $\mathcal{K}$ is a bounded operator. Next, we show that a bounded operator implies continuous. For any function $f, f^* \in \mathcal{H}$ ($f \neq f^*$),

$$\|\mathcal{K}(f) - \mathcal{K}(f^*)\|^2 = \|\mathcal{K}(f - f^*)\|^2 = \int_{\Omega_Z} \left(\int_{\Omega_X} k(x', x)\{f(x') - f^*(x')\}dx' \right)^2 dx, \tag{21}$$

from the Cauchy-Schwarz inequality

$$\leq \int_{\Omega_Z} \left(\int_{\Omega_X} k(x', x)^2 dx' \right) \left(\int_{\Omega_X} \{f(x') - f^*(x')\}^2 dx' \right) dx \tag{22}$$

$$= \left(\int_{\Omega_Z} \int_{\Omega_X} k(x', x)^2 dx' dx \right) \left(\int_{\Omega_X} \{f(x') - f^*(x')\}^2 dx' \right) \tag{23}$$

$$\leq \left(\int_{\Omega_Z} \int_{\Omega_X} k(x', x)^2 dx' dx \right) \|f - f^*\|^2 < \infty \tag{24}$$

Because

$$\|\mathcal{K}(f) - \mathcal{K}(f^*)\| \leq \sqrt{\left(\int_{\Omega_Z} \int_{\Omega_X} k(x', x)^2 dx' dx \right)} \|f - f^*\|, \tag{25}$$

the operator $\mathcal{K}$ is a continuous operator. Finally, from the open mapping theorem, the inverse operator $\mathcal{K}^{-1}$ is also continuous. Here, $\mathcal{K}^{-1} = \alpha \sum_{t=1}^{\infty} (\mathcal{I} - \alpha\mathcal{K})^t$ where $\mathcal{I}$ is an identity operator and $0 < \alpha < 2/\|\mathcal{K}\|$. Furthermore, for any $\mathcal{K}, \mathcal{K}^*$ ($\mathcal{K} \neq \mathcal{K}^*$) and g, g^* ($g \neq g^*$),

$$\begin{aligned} \|\mathcal{K}^{-1}(g) - \mathcal{K}^{*\,-1}(g^*)\| &\leq \|\mathcal{K}^{-1}(g) - \mathcal{K}^{*\,-1}(g)\| + \|\mathcal{K}^{*\,-1}(g) - \mathcal{K}^{*\,-1}(g^*)\| && (26) \\ &\leq \|\mathcal{K}^{-1}(g) - \mathcal{K}^{*\,-1}(g)\| + \|\mathcal{K}^{*\,-1}\| \|g - g^*\| && (27) \end{aligned}$$

holds. Since the function k is continuous, the solution changes continuously with changes in the input function $\mathcal{K}$ and g. $\square$

A.2. Proof of Lemmma 3.2

Lemma 3.2. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, and 3, the operator $\mathcal{K}$ satisfies the following three properties:*

1. *$\mathcal{K}$ is a compact operator: $\mathcal{K}$ maps a bounded set into a compact set in the sense of strong convergence.*
2. *$\mathcal{K}$ is self-adjoint: $\langle \mathcal{K}(a), b \rangle = \langle a, \mathcal{K}(b) \rangle$ for $a, b \in \mathcal{H}$.*
3. *$\mathcal{K}$ is positive semi-definite: $\langle \mathcal{K}(a), a \rangle \geq 0$ for $a \in \mathcal{H}$.*

Proof. First, this is an integral equation (2)

$$\mathbb{E}[Y|Z=z_0]-\mathbb{E}[Y|Z=z]=\int_{\Omega_X}\{\mathbb{P}_X[x|Z=z]-\mathbb{P}_X[x|Z=z_0]\}\mathbb{E}[\partial_x Y_x]dx, \tag{28}$$

and the operator $\mathcal{K}$ is

$$\mathcal{K}(f)=\int_{\Omega_X}\{\mathbb{P}_X[x|Z=z]-\mathbb{P}_X[x|Z=z_0]\}f(x)dx \tag{29}$$

for any $f\in\mathcal{H}$. The function $\mathcal{K}$ is satisfies

$$\int_{\Omega_Z}\int_{\Omega_X}|\{\mathbb{P}_X[x|Z=z]-\mathbb{P}_X[x|Z=z_0]\}|^2dxdz<\infty, \tag{30}$$

thus $\mathcal{K}$ is a compact integral kernel (Alexanderian, 2013). Second, since

$$\langle\mathcal{K}(f),g\rangle=\int_{\Omega_Z}\Big(\int_{\Omega_X}\{\mathbb{P}_X[x|Z=z]-\mathbb{P}_X[x|Z=z_0]\}f(x)dx\Big)g(z)dz \tag{31}$$

$$=\int_{\Omega_Z}f(x)\Big(\int_{\Omega_X}\{\mathbb{P}_X[x|Z=z]-\mathbb{P}_X[x|Z=z_0]\}g(z)dz\Big)dx=\langle f,\mathcal{K}(g)\rangle\,, \tag{32}$$

the operator $\mathcal{K}$ is selfadjoint. Third, since $\{\mathbb{P}_X[x|Z=z]-\mathbb{P}_X[x|Z=z_0]\}>0$ for all $x\in X$ and

$$<\mathcal{K}(f),f>=\int_{\Omega_Z}\Big(\int_{\Omega_X}\{\mathbb{P}_X[x|Z=z]-\mathbb{P}_X[x|Z=z_0]\}f(x)dx\Big)f(z)dz \tag{33}$$

holds. The integral operator $\mathcal{K}$ satisfies the three properties in lemma 3.2. □

A.3. Proof of Lemma 3.3

Lemma 3.3. *Under SCM $\mathcal{M}_{IV}$ and Assumption 1, $\mathcal{K}$ satisfies the Picard's condition.*

Proof. From Assumption 1, there exists the APCE, which is the solution of the integral equation (2). Since Picard's condition is the necessary condition for the existence of the solution, $\mathcal{K}$ satisfies the Picard condition. □

A.4. Proof of Theorem 3.4

Theorem 3.4. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, and 4, the Picard iteration scheme (5) converges strongly to the APCE, that is,* $\lim_{t\to\infty}\theta_t(x)=\mathbb{E}[\partial_x Y_x]$.

Proof. We use the following result (Theorem (c_1) in Diaz & Metcalf (1970)):

The operator $\mathcal{L}$ is assumed to be compact, selfadjoint, and positive semidefinite. Let $b\in\mathcal{H}$ be such that $\langle b,v\rangle$ for every u such that $\mathcal{L}(v)=0$. Then, the sequence of the Picard's iteration $\{\theta_t\}_{t=1}^{\infty}$ converge strongly, for every $\theta_0\in\mathcal{H}$ if and only if $\mathcal{L}$ satisfies Picard condtion.

The operator $\mathcal{K}$ is compact, selfadjoint, and positive semidefinite from the Lemma 3.2 and $\mathcal{K}$ satisfies the Picard condition from Lemma 3.3. Thus, the Picard iteration $\{\theta_t\}_{t=0}^{\infty}$ converge strongly. From the uniqueness of the solution, the convergence point is the APCE.

□

A.5. Proof of Theorem 3.5

Theorem 3.5. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 5, taking limits $N\to\infty$, $R\to\infty$, and $t\to\infty$, $\hat{\theta}_t(x)$ is a pointwise consistent estimator of the APCE $\mathbb{E}[\partial_x Y_x]$ for $x\in\Omega_X$ almost everywhere.*

Proof. The functions $\hat{\mu}(x)$ and $\hat{k}(x', x)$ can be written as

$$\begin{cases} \hat{\mu}(x) = \mu(x) + e(x) \\ \hat{k}(x', x) = k(x', x) + \epsilon(x', x) \end{cases}, \tag{34}$$

where functions $e(x)$ and $\epsilon(x', x)$ are error terms. Concisely, we represent the above relationships as below;

$$\begin{cases} \hat{\mu} = \mu + e \\ \hat{k} = k + \epsilon \end{cases}. \tag{35}$$

For the numerical integration, we choose the subinterval $[x_q, x_{q+1}]$ satisfies that $[x_0, x_Q] = [\min(\Omega_X), \max(\Omega_X)]$ and $\lim_{Q\to\infty} |x_{q+1} - x_q| = 0$ for $q = 1, \ldots, Q$. In addition, for the numerical interpolation, we choose the subinterval $[x_r, x_{r+1}]$ satisfies that $[x_0, x_R] = [\min(\Omega_X), \max(\Omega_X)]$ and $\lim_{R\to\infty} |x_{r+1} - x_r| = 0$ for $r = 1, \ldots, R$.

Then, we write down the Picard iteration with the ML error. Update the function $\hat{\theta}$ at $X = x_r$ by

$$\hat{\theta}_{t+1}(x_r) = \hat{\theta}_t(x_r) + \alpha\Big(\hat{\mu}(x_r) - \mathcal{I}[\hat{k}(x', x)\hat{\theta}_t(x'); \mathscr{X}]\Big) \tag{36}$$

$$\hat{\theta}_{t+1}(x_r) = \hat{\theta}_t(x_r) + \alpha\Big(\mu(x_r) + e(x_r) - \mathcal{I}[\{k(x', x_r) + \epsilon(x', x_r)\}\hat{\theta}_t(x'); \mathscr{X}]\Big) \tag{37}$$

$$\hat{\theta}_{t+1}(x_r) = \hat{\theta}_t(x_r) + \alpha\Big(\mu(x_r) + e(x_r) - \mathcal{I}[k(x', x_r)\hat{\theta}_t(x'); \mathscr{X}] - \mathcal{I}[\epsilon(x', x_r)\hat{\theta}_t(x'); \mathscr{X}]\Big), \tag{38}$$

where $\mathcal{I}$ means the numerical integration, which is represented as below, concretely;

$$\hat{\theta}_{t+1}(x_r) = \hat{\theta}_t(x_r) + \alpha\Big(\mu(x_r) + e(x_r) - \sum_{q=0}^{Q} A(k(x_q, x_r)\hat{\theta}_t(x_q), k(x_{q+1}, x_r)\hat{\theta}_t(x_{q+1}))(x_{q+1} - x_q) \tag{39}$$

$$- \sum_{q=0}^{Q} A(\epsilon(x_q, x_r)\hat{\theta}_t(x_q), \epsilon(x_{q+1}, x_r)\hat{\theta}_t(x_{q+1}))(x_{q+1} - x_q)\Big). \tag{40}$$

Then, we take limit $Q \to \infty$, and the numerical integration converge to integration. Thus,

$$\lim_{Q\to\infty} \hat{\theta}_{t+1}(x_r) = \lim_{Q\to\infty} \hat{\theta}_t(x_r) + \alpha\Big(\mu(x_r) + e(x_r) - \int_{\Omega_X} k(x', x_r) \lim_{Q\to\infty} \hat{\theta}_t(x')dx' - \int_{\Omega_X} \epsilon(x', x_r) \lim_{Q\to\infty} \hat{\theta}_t(x')dx'\Big) \tag{41}$$

holds.

Since $e(x_r)$ converges in probability to 0 taking limit $N \to \infty$, and $\epsilon(x', x_r)$ converges in probability to zero function for $r = 1, \ldots, R$,

$$\lim_{Q\to\infty, N\to\infty} \hat{\theta}_{t+1}(x_r) = \lim_{Q\to\infty, N\to\infty} \hat{\theta}_t(x_r) + \alpha\Big(\mu(x_r) - \int_{\Omega_X} k(x', x_r)\Big\{\lim_{Q\to\infty, N\to\infty} \hat{\theta}_t(x')\Big\}dx'\Big) \tag{42}$$

holds, which is the same as the Picard iteration of $\theta_t(x_r)$ for $t = 1, 2, 3, \ldots$, the estimator $\hat{\theta}_t(x_r)$ is a consistent estimator of $\theta_t(x_r)$ for $r = 1, \ldots, R$. Furthermore, taking the limit $t \to \infty$, $\lim_{t\to\infty} \hat{\theta}_t(x_r)$ is a consist estimator of APCE at $X = x_r$ since θ_t converge to APCE at $X = x_r$. From the property of the interpolation, taking the limit that $R \to \infty$, the function $\hat{\theta}(x)$ is a consistent estimator of the APCE for $x \in \Omega_X$. □

A.6. Proof of Theorem 3.6

Theorem 3.6. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 6, taking the limit $R \to \infty$, Algorithm 1 stops after a finite number of iterations for any $\epsilon > 0$.*

Proof. As Theorem 3.4, the Picard iteration scheme (10) also converges strongly to the solution under Assumption 6. Thus, Algorithm 1 stops after a finite number of iterations for any $\epsilon > 0$. □

A.7. Proof of corollary 3.7

Corollary 3.7. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 6, the sequence $\hat{\theta}_t$ converges linearly to $\hat{\theta}_\infty$.*

Proof. From the triangle inequality,

$$\frac{\|\hat{\theta}_{t+1}-\hat{\theta}_\infty\|}{\|\hat{\theta}_t-\hat{\theta}_\infty\|} = \frac{\|\sum_{t'=t+1}^{\infty}(I-\alpha\hat{\mathcal{K}})^{t'}(\alpha\hat{\mu})\|}{\|\sum_{t'=t}^{\infty}(I-\alpha\hat{\mathcal{K}})^{t'}(\alpha\hat{\mu})\|} = \frac{\|\sum_{t'=t}^{\infty}(I-\alpha\hat{\mathcal{K}})^{t'}(\alpha\hat{\mu})+(I-\alpha\hat{\mathcal{K}})^{t+1}(\alpha\hat{\mu})\|}{\|\sum_{t'=t}^{\infty}(I-\alpha\mathcal{K})^{t'}(\alpha\hat{\mu})\|} \tag{43}$$

$$\leq \frac{\|\sum_{t'=t}^{\infty}(I-\alpha\hat{\mathcal{K}})^{t'}(\alpha\hat{\mu})\|+\|(I-\alpha\hat{\mathcal{K}})^{t+1}(\alpha\hat{\mu})\|}{\|\sum_{t'=t}^{\infty}(I-\alpha\hat{\mathcal{K}})^{t'}(\alpha\hat{\mu})\|} \tag{44}$$

holds. Since $\|(I-\alpha\hat{\mathcal{K}})^{t+1}(\alpha\hat{\mu})\|$ is bounded for all t, $\|\hat{\theta}_{t+1}-\hat{\theta}_\infty\|/\|\hat{\theta}_t-\hat{\theta}_\infty\|$ is also bounded. Thus, the sequence $\hat{\theta}_t$ converges linearly. □

A.8. Proof of Theorem 3.8

Theorem 3.8. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, 6, and 7, letting $\hat{\mathcal{K}}^{-1}=\alpha\sum_{t=0}^{\infty}(I-\alpha\hat{\mathcal{K}})^t$, if $\|\hat{\mathcal{K}}^{-1}\|$ is bounded by M, then the expected absolute bias $\mathbb{E}[\|\hat{\theta}_\infty-\theta_\infty\|]$ is bounded by $M(A+\|\theta_\infty\|B)$, where*

$$A=\sqrt{\int_{\Omega_X}\mathbb{V}(\hat{\mu}(x))dx},\ B=\sqrt{\int_{\Omega_X}\mathbb{V}(\hat{k}(x,x))dx}. \tag{45}$$

Proof. The Picard iteration, both with the ML error and without the ML error

$$\begin{cases} \hat{\theta}_{t+1}(x)=\hat{\theta}_t(x)+\alpha\Big(\mu(x)-\int_{\Omega_X}k(x',x)\hat{\theta}_t(x')dx'\Big)+\alpha\Big(e(x)-\int_{\Omega_X}\epsilon(x',x)\hat{\theta}_t(x')dx'\Big) \\ \theta_{t+1}(x)=\theta_t(x)+\alpha\Big(\mu(x)-\int_{\Omega_X}k(x',x)\theta_t(x')dx'\Big) \end{cases}. \tag{46}$$

Thus, the error of the estimator, $\hat{\theta}_{t+1}(x)-\theta_{t+1}(x)$, becomes

$$\begin{aligned} &\hat{\theta}_{t+1}(x)-\theta_{t+1}(x) \\ &=\hat{\theta}_t(x)-\theta_t(x)-\alpha\int_{\Omega_X}k(x',x)\{\hat{\theta}_t(x')-\theta_t(x')\}dx'+\alpha\tilde{e}(x)-\alpha\int_{\Omega_X}\epsilon(x',x)\hat{\theta}_t(x')dx' &(47)\\ &=\hat{\theta}_t(x)-\theta_t(x)-\alpha\int_{\Omega_X}k(x',x)\{\hat{\theta}_t(x')-\theta_t(x')\}dx'+\alpha\tilde{e}(x)-\alpha\int_{\Omega_X}\epsilon(x',x)\{\hat{\theta}_t(x')-\theta_t(x')+\theta_t(x')\}dx', &(48)\end{aligned}$$

then

$$\begin{aligned} &\hat{\theta}_{t+1}(x)-\theta_{t+1}(x) \\ &=\hat{\theta}_t(x)-\theta_t(x)-\alpha\int_{\Omega_X}k(x',x)\{\hat{\theta}_t(x')-\theta_t(x')\}dx' &(49)\\ &+\alpha e(x)-\alpha\int_{\Omega_X}\epsilon(x',x)\{\hat{\theta}_t(x')-\theta_t(x')\}dx'+\alpha\int_{\Omega_X}\epsilon(x',x)\theta_t(x')dx'. &(50)\end{aligned}$$

Concisely, we denote the operator $I-\alpha\mathcal{K}-\alpha\mathcal{E}$

$$\begin{aligned} &(I-\alpha\mathcal{K}-\alpha\mathcal{E})(\hat{\theta}_t-\theta_t)(x) &(51)\\ &:=\hat{\theta}_t(x)-\theta_t(x)-\alpha\int_{\Omega_X}k(x',x)\{\hat{\theta}_t(x')-\theta_t(x')\}dx'-\alpha\int_{\Omega_X}\epsilon(x',x)\{\hat{\theta}_t(x')-\theta_t(x')\}dx', &(52)\end{aligned}$$

then

$$\hat{\theta}_{t+1}-\theta_{t+1}=(I-\alpha\mathcal{K}-\alpha\mathcal{E})(\hat{\theta}_t-\theta_t)+\alpha e+\alpha\mathcal{E}\theta_t. \tag{53}$$

We write down the Picard iteration with the ML error, since $\theta_1 = \hat{\theta}_1$,

$$\hat{\theta}_2 - \theta_2 = \alpha e + \alpha\mathcal{E}\theta_2 \tag{54}$$

$$\hat{\theta}_3 - \theta_3 = (I - \alpha\mathcal{K} - \alpha\mathcal{E})(\alpha e + \alpha\mathcal{E}\theta_2) + \alpha e + \alpha\mathcal{E}\theta_3 \tag{55}$$

$$\hat{\theta}_4 - \theta_4 = (I - \alpha\mathcal{K} - \alpha\mathcal{E})^2(\alpha e + \alpha\mathcal{E}\theta_2) + (I - \alpha\mathcal{K} - \alpha\mathcal{E})(\alpha e + \alpha\mathcal{E}\theta_3) + \alpha e + \alpha\mathcal{E}\theta_4 \tag{56}$$

$$\hat{\theta}_5 - \theta_5 = (I - \alpha\mathcal{K} - \alpha\mathcal{E})^4(\alpha e + \alpha\mathcal{E}\theta_2) + (I - \alpha\mathcal{K} - \alpha\mathcal{E})^2(\alpha e + \alpha\mathcal{E}\theta_3) \tag{57}$$

$$+(I - \alpha\mathcal{K} - \alpha\mathcal{E})(\alpha e + \alpha\mathcal{E}\theta_4) + \alpha e + \alpha\mathcal{E}\theta_4 \tag{58}$$

$$\vdots \tag{59}$$

holds; thus the error after t times iterations becomes

$$\hat{\theta}_t - \theta_t = \sum_{t'=0}^{t-2}(I - \alpha\mathcal{K} - \alpha\mathcal{E})^{t'}(\alpha e + \alpha\mathcal{E}\theta_{t'}) = \sum_{t'=0}^{t-2}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_{t'}). \tag{60}$$

Here, since the operator $(\mathcal{K} + \mathcal{E})$ is continuous (bounded) and

$$\| \sum_{t'=0}^{\infty}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_\infty) - \sum_{t'=0}^{t-2}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_{t'})\| \tag{61}$$

$$\leq \| \sum_{t'=0}^{\infty}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_\infty) - \sum_{t'=0}^{t-2}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_\infty)\| \tag{62}$$

$$+\| \sum_{t'=0}^{t-2}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_\infty) - \sum_{t'=0}^{t-2}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_t)\|, \tag{63}$$

$\sum_{t'=0}^{t-2}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_{t'})$ converges strongly to $\sum_{t'=0}^{\infty}(I - \alpha(\mathcal{K} + \mathcal{E}))^{t'}(\alpha e + \alpha\mathcal{E}\theta_\infty)$, if the operator $(\mathcal{K} + \mathcal{E})$ satisfies the Picard condition. It converges strongly and $\hat{\theta}_t - \theta_t$ converges strongly to the solution of the integral equation σ

$$\alpha e + \alpha\mathcal{E}\theta_\infty = \alpha(\mathcal{K} + \mathcal{E})\sigma \Leftrightarrow \sigma = (\mathcal{K} + \mathcal{E})^{-1}(e + \mathcal{E}\theta_\infty) \tag{64}$$

where $(\mathcal{K} + \mathcal{E})^{-1}(e + \mathcal{E}\theta_\infty) = \alpha\sum_{t=0}^{\infty}(I - \alpha(\mathcal{K} + \mathcal{E}))^t(e + \mathcal{E}\theta_\infty)$ (Diaz & Metcalf, 1970). The norm of the error is bounded by $\|\sigma\| \leq \|(\mathcal{K} + \mathcal{E})^{-1}\|\|e + \mathcal{E}\theta_\infty\| \leq \|(\mathcal{K} + \mathcal{E})^{-1}\|\{\|e\| + \|\mathcal{E}\|\|\theta_\infty\|\}$. This means

$$\|\hat{\theta}_\infty - \theta_\infty\| \leq \|\hat{\mathcal{K}}^{-1}\|\{\|\hat{\mu} - \mu\| + \|\hat{\mathcal{K}} - \mathcal{K}\|\|\theta_\infty\|\}. \tag{65}$$

If the operator $\|\hat{\mathcal{K}}^{-1}\|$ is bounded by M,

$$\|\hat{\theta}_\infty - \theta_\infty\| \leq M\{\|\hat{\mu} - \mu\| + \|\hat{\mathcal{K}} - \mathcal{K}\|\|\theta_\infty\|\} \tag{66}$$

holds. If $\hat{\mu}$ is equal to μ and $\hat{\mathcal{K}}$ is equal to $\mathcal{K}$, $\hat{\theta}_\infty$ is equal to θ_∞.

$\square$

A.9. Proof of Theorem 3.9

Theorem 3.9. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 7, when Algorithm 1 stops at $t = T$, the upper bound of the variance of $\hat{\theta}_T(x)$ is $\alpha^2(T-1)^2\nu(x) + \mathcal{O}(\alpha^3)$ as $\alpha \to 0$ for $x \in \Omega_X$, where $\nu(x)$ is*

$$\mathbb{V}(\hat{\mu}(x)) + \Big(\int_{\Omega_X}\sqrt{\mathbb{V}(\hat{k}(x', x))}|\theta_T(x')|dx'\Big)^2 + 2\sqrt{\mathbb{V}(\hat{\mu}(x))}\Big(\int_{\Omega_X}\sqrt{\mathbb{V}(\hat{k}(x', x))}|\theta_T(x')|dx'\Big).$$

Proof. As for the error at each x, the error of the estimator at x after t times iterations become

$$\hat{\theta}_t(x) - \theta_t(x) = \sum_{t'=0}^{t-2}(I - \alpha\mathcal{K} - \alpha\mathcal{E})^{t'}(\alpha e + \alpha\mathcal{E}\theta_t)(x) \tag{67}$$

$$= \alpha(t-1)(e+\mathcal{E}\theta_t)(x) + \mathcal{O}(\alpha^2) \tag{68}$$

Then, the absolute error becomes

$$|\hat{\theta}_t(x) - \theta_t(x)| = |(t-1)\alpha(e+\mathcal{E}\theta_t)(x)| + \mathcal{O}(\alpha^2) \tag{69}$$

and the squared error becomes

$$(\hat{\theta}_t(x) - \theta_t(x))^2 = (t-1)^2\alpha^2(e+\mathcal{E}\theta_t)(x)^2 + \mathcal{O}(\alpha^3) \tag{70}$$

Thus, the variance becomes

$$\mathbb{V}(\hat{\theta}_t(x)) = \mathbb{E}[\{\hat{\theta}_t(x) - \theta_t(x)\}^2] = (t-1)^2\alpha^2\mathbb{E}[(e+\mathcal{E}\theta_t)(x)^2] + \mathcal{O}(\alpha^3). \tag{71}$$

Here,

$$\mathbb{E}[(e+\mathcal{E}\theta_t)(x)^2] \leq \mathbb{E}[e(x)^2] + 2\mathbb{E}[e(x)\mathcal{E}(\theta_t)(x)] + \mathbb{E}[\mathcal{E}(\theta_t)(x)^2] \tag{72}$$

$$= \mathbb{V}(\hat{\mu}(x)) + 2\mathbb{E}[\{\hat{\mu}(x) - \mu(x)\}\{\hat{\mathcal{K}}(\theta_t)(x) - \mathcal{K}(\theta_t)(x)\}] + \mathbb{V}(\hat{\mathcal{K}}(\theta_t)(x)) \tag{73}$$

holds. From the unbiasedness of the ML, since $\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$ from the Cauchy–Schwarz inequality,

$$\mathbb{E}[(e+\mathcal{E}\theta_t)(x)^2] \leq \mathbb{V}(\hat{\mu}(x)) + 2\sqrt{\mathbb{V}(\hat{\mu}(x))\mathbb{V}(\hat{\mathcal{K}}(\theta_t)(x))} + \mathbb{V}(\hat{\mathcal{K}}(\theta_t)(x)) \tag{74}$$

holds. Furthermore, $\mathbb{V}(\hat{\mathcal{K}}(\theta_t)(x))$ is bounded by

$$\mathbb{V}(\hat{\mathcal{K}}(\theta_t)(x)) = E\Big[\Big(\int_{\Omega_X}\{\hat{k}(x',x) - k(x',x)\}\theta_t(x')dx'\Big)^2\Big] \tag{75}$$

$$\leq E\Big[\Big|\int_{\Omega_X}\{\hat{k}(x',x) - k(x',x)\}\theta_t(x')dx'\Big|\Big]^2 \tag{76}$$

$$\leq \Big(\int_{\Omega_X}\mathbb{E}[|\hat{k}(x',x) - k(x',x)|]\theta_t(x')|dx'\Big)^2 \tag{77}$$

$$\leq \Big(\int_{\Omega_X}\sqrt{\mathbb{V}(\hat{k}(x',x))}|\theta_t(x')|dx'\Big)^2 \tag{78}$$

$$\tag{79}$$

from the Hölder's inequality. Finally,

$$\mathbb{V}(\hat{\theta}_t(x)) \leq (t-1)^2\alpha^2\nu(x) + \mathcal{O}(\alpha^3) \tag{80}$$

holds, where

$$\nu(x) = \mathbb{V}(\hat{\mu}(x)) + 2\sqrt{\mathbb{V}(\hat{\mu}(x))\Big(\int_{\Omega_X}\sqrt{\mathbb{V}(\hat{k}(x',x))}|\theta_t(x')|dx'\Big)^2} + \Big(\int_{\Omega_X}\sqrt{\mathbb{V}(\hat{k}(x',x))}|\theta_t(x')|dx'\Big)^2. \tag{81}$$

□

A.10. Proof of Corollary 3.10

Corollary 3.10. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, 4, and 7, when Algorithm 1 stops at $t = T$, the variance of $\hat{\theta}_T(x)$ is*

$$\alpha^2(T-1)^2\mathbb{V}(\hat{\mu}(x)) + \mathcal{O}(\alpha^3) + \mathcal{O}(\{\max_{x'}\{\mathbb{V}(\hat{k}(x',x))\}\}^{1/2}) \tag{82}$$

as $\alpha \to 0, \{\max_{x'}\{\mathbb{V}(\hat{k}(x',x))\}\}^{1/2} \to 0$ for $x \in \Omega_X$.

Proof. The inequality

$$\Big(\int_{\Omega_X}\sqrt{\mathbb{V}(\hat{k}(x',x))}|\theta_t(x')|dx'\Big)^2 \le \max_{x'}\{\mathbb{V}(\hat{k}(x',x))\}\Big(\int_{\Omega_X}|\theta_t(x')|dx'\Big)^2 \tag{83}$$

holds; thus, from the Theorem 3.9

$$\mathbb{E}[\{\hat{\theta}_t(x)-\theta_t(x)\}^2] = (t-1)^2\alpha^2\mathbb{V}(\hat{\mu}(x))(\alpha^3)+\mathcal{O}(\alpha^3)+\mathcal{O}(\sqrt{\max_{x'}\{\mathbb{V}(\hat{k}(x',x))\}}) \tag{84}$$

holds. □

A.11. Proof of Theorem 4.1

Theorem 4.1. *Under SCM $\mathcal{M}_{IV}$ and Assumptions 1, 2, 3, and 8, taking a limit $N\to\infty$, the estimator $\hat{\boldsymbol{\theta}}$ given by Algorithm 2 is a pointwise consistent estimator of $\boldsymbol{\theta}$ in Eq. (15).*

Proof. $\hat{\boldsymbol{u}}$ and $\hat{\mathbf{D}}$ can be written as

$$\begin{cases} u_r = \theta_P d_r^P + \ldots + \theta_1 d_r^1 \\ \hat{u}_r = u_r + e_r, \text{ for } r = 1,\ldots,R \\ \hat{d}_r^p = d_r^p + \epsilon_r^p \text{ for } r=1,\ldots,R \text{ and } p=1,\ldots,P \end{cases}, \tag{85}$$

where $e_r,$ for $r=1,\ldots,R$ and ϵ_r^p for $r=1,\ldots,R$ and $p=1,\ldots,P$ are error terms. We denote $\boldsymbol{e}=(e_1,\ldots,e_R)$ and $\boldsymbol{\epsilon}=(\epsilon_1,\ldots,\epsilon_R)$. Then, the estimator $\hat{\boldsymbol{\theta}}$ becomes

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{D}}^T\hat{\boldsymbol{D}})^{-1}\hat{\boldsymbol{D}}^T\hat{\boldsymbol{u}} = \{(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{D}+\boldsymbol{\epsilon})\}^{-1}(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e}). \tag{86}$$

Since $\boldsymbol{D}\perp\!\!\!\perp\boldsymbol{\epsilon}$,

$$= (\boldsymbol{D}^T\boldsymbol{D}+\boldsymbol{\epsilon}^T\boldsymbol{\epsilon})^{-1}(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e}) \tag{87}$$

holds. From the Woodbury formula,

$$\begin{aligned}
&= (\boldsymbol{D}^T\boldsymbol{D}+\boldsymbol{\epsilon}^T\boldsymbol{\epsilon})^{-1}(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e}) && (88)\\
&= [(\boldsymbol{D}^T\boldsymbol{D})^{-1}-(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T(I+\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T)^{-1}\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}](\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e}) && (89)\\
&= (\boldsymbol{D}^T\boldsymbol{D})^{-1}(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e})-(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T(I+\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T)^{-1}\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e}) && (90)\\
&= \boldsymbol{\theta}+(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T\boldsymbol{u}+(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{e}+(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T\boldsymbol{e}\\
&\qquad -(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T(I+\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T)^{-1}\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e}). && (91)
\end{aligned}$$

Then,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}-\boldsymbol{\theta} &= (\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T\boldsymbol{u}+(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{e}+(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T\boldsymbol{e}\\
&\quad -(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T(I+\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{\epsilon}^T)^{-1}\boldsymbol{\epsilon}(\boldsymbol{D}^T\boldsymbol{D})^{-1}(\boldsymbol{D}+\boldsymbol{\epsilon})^T(\boldsymbol{u}+\boldsymbol{e}). && (92)
\end{aligned}$$

Taking the limit $N\to\infty$, $\boldsymbol{\epsilon}$ converges in probability to a zero matrix and $\boldsymbol{e}$ converges in probability to a zero vector from the consistency of the ML, thus $\hat{\boldsymbol{\theta}}$ is a consistent estimator. □

B. Details of numerical integration and interpolation

In this section, we explain numerical integration and interpolation.

B.1. Details of numerical integration

First, denote $\mathcal{I}[a(x); \mathscr{X}]$ be a numerical integration of the integration $\int_{\Omega_X} a(x)dx$ for a function $a \in \mathcal{H}$ given $\mathscr{X}$; and, it takes form as

$$\mathcal{I}[a(x); \mathscr{X}] = \sum_{q=1}^{R} I(x_q, x_{q+1})(x_{q+1} - x_q), \tag{93}$$

e.g., the left-hand rule

$$\mathcal{I}[a(x); \mathscr{X}] = \sum_{q=1}^{R} a(x_q)(x_{q+1} - x_q), \tag{94}$$

the mid-point rule

$$\mathcal{I}[a(x); \mathscr{X}] = \sum_{q=1}^{R} a(\frac{x_q + x_{q+1}}{2})(x_{q+1} - x_q), \tag{95}$$

and the trapezoidal rule

$$\mathcal{I}[a(x); \mathscr{X}] = \sum_{q=1}^{R} \frac{a(x_q) + a(x_{q+1})}{2}(x_{q+1} - x_q). \tag{96}$$

A total error of, at most, the mid-point rule is $(x_R - x_0)^3 B/24R^2$, where B is an upper bound for the second derivative of $a(x)$ and $x_{r+1} - x_r$ are equal for all $r = 1, \dots, R$. The total error converges to zero when $R \to \infty$, and the total error converges with the order $\mathcal{O}(R^{-2})$.

B.2. Details of numerical interpolation

Next, we explain the numerical interpolation. Given, the set $\mathscr{X}$ and their values $\hat{\theta}(x_r)$ for $r = 1, \dots, R$. We interpolate the function $\hat{\theta}$ by the linear combination

$$\hat{\theta}(x) = \sum_{r=1}^{R} w_r l_r(x), \tag{97}$$

e.g., the Lagrange interpolating polynomial

$$l_r(x) = \frac{x - x_0}{x_r - x_0} \cdots \frac{x - x_{r-1}}{x_r - x_{r-1}} \frac{x - x_{r+1}}{x_r - x_{r+1}} \cdots \frac{x - x_R}{x_r - x_R} = \prod_{0 \le m \le R, m \ne r} \frac{x - x_m}{x_r - x_m} \tag{98}$$

The coefficients $w_1, \dots, w_R$ are determined by solving the system of equations

$$\hat{\theta}(x_r) = \sum_{r=1}^{R} w_r l_r(x_r), \quad \text{for } r = 1, \dots, R. \tag{99}$$

When we interpolation the function $\theta(x)$ by the Lagrange interpolating polynomial $\hat{\theta}(x)$, whose degree is R, the errors are bounded by

$$|\theta(x) - \hat{\theta}(x)| \le \frac{Ch^R}{4R}, \tag{100}$$

where $C = \max_{x \in [x_0, x_R]} |\frac{d^R}{d^R x}\theta(x)|$ and $h = \max_{r=0,\dots,R-1} |x_{r+1} - x_r|$. The errors converge to zero when $R \to \infty$ and $\lim_{R \to \infty} |x_{r+1} - x_r| = 0$ for all $r = 1, \dots, R$. Since h can be represented as w/R, and error converge with the order $\mathcal{O}(R^{-R})$.

C. Additional Information on Numerical Experiments

C.1. Additional Information on Numerical Experiments

We give additional information on the numerical experiment based on the following SCM (Model 1);

$$\begin{cases} X = 5^{-2}Z^2 + 5^{-1}Z + 0.5 + (\frac{Z}{3} + 0.1)U \\ Y = X^3 + X^2 + X + U + E \end{cases} . \tag{101}$$

Table 6 and Table 9 show basic statistics of the parametric and N-APCE estimator when $N = 100$ and $N = 1000$. The approximate upper bound of the SD (Approx USD) by the equation (12) and approximate bound of the SD (Approx SD) by the equation (13) are also shown in Table 9. In addition, we compare the estimator with the TSPS in Table 7 and the NPTSLS in Table 8.

Table 5: Basic statistics of the test error of P-APCE over 100 runs for each degree; the bold number is the smallest.

$N = 100$	$D = 2$	3	4	5	6
Min.	0.022	0.019	0.173	0.081	0.091
1st Qu.	0.131	0.111	0.352	0.281	0.222
Median	0.173	0.157	0.439	0.352	0.283
Mean	0.183	**0.168**	0.443	0.352	0.300
3rd Qu.	0.235	0.211	0.520	0.410	0.362
Max.	0.416	0.442	0.747	0.668	0.604

$N = 1000$	$D = 2$	3	4	5	6
Min.	0.039	0.027	0.030	0.068	0.059
1st Qu.	0.085	0.073	0.097	0.165	0.237
Median	0.122	0.102	0.133	0.226	0.307
Mean	0.124	**0.113**	0.151	0.227	0.315
3rd Qu.	0.151	0.134	0.195	0.272	0.386
Max.	0.303	0.349	0.332	0.509	0.595

Table 6: Basic statistics of the P-APCE estimators when $N = 100$ and $N = 1000$; ”Degree=m” means ”the estimated coefficient of m-th degree term.”

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-22.384	-57.235	-21.697
1st Qu.	1.460	-24.815	2.947
Median	7.017	-8.754	7.463
3rdQu.	15.728	2.386	13.781
Max.	33.222	52.881	29.850
Mean	7.879	-11.128	8.525
SD	10.516	20.884	9.214

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-12.438	-22.410	-6.869
1st Qu.	-2.074	-4.150	0.615
Median	1.648	0.882	3.286
3rd Qu.	4.347	7.390	5.498
Max.	13.550	26.577	13.381
Mean	0.995	2.132	2.878
SD	4.951	9.477	3.997

Table 7: Basic statistics of the TSPS estimators when $N = 100$ and $N = 1000$; ”Degree=m” means ”the estimated coefficient of m-th degree term.”

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-124.484	-272.723	-222.915
1st Qu.	-40.675	-48.223	-55.086
Median	-11.869	35.187	-13.301
3rd Qu.	22.194	105.129	33.460
Max.	107.563	346.930	173.959
Mean	-9.749	29.489	-11.571
SD	48.039	118.938	69.845

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-36.485	-122.498	-49.695
1st Qu.	-10.755	-20.670	-11.550
Median	-4.443	14.456	-1.834
3rd Qu.	9.840	31.107	18.230
Max.	52.377	95.724	75.675
Mean	-1.590	8.403	1.531
SD	15.110	37.188	21.470

Table 8: Basic statistics of the NPTSLS estimators when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th basis function."

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-47.459	-294.671	-111.854
1st Qu.	-4.447	-52.320	-4.138
Median	8.343	-20.611	23.730
3rd Qu.	17.487	25.362	41.668
Max.	84.201	189.642	191.154
Mean	8.043	-18.671	21.677
SD	18.795	68.360	42.250

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-18.946	-103.199	-36.229
1st Qu.	-1.671	-25.216	2.625
Median	3.502	-4.251	13.997
3rd Qu.	9.191	13.436	25.771
Max.	31.341	77.682	71.694
Mean	3.453	-3.980	13.324
SD	9.310	32.709	19.411

Table 9: Basic statistics of the N-APCE estimator at $x = (0, 0.3, \ldots, 2.7, 3.0)$ when $N = 100$ and $N = 1000$.

$N = 100$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	1.87	3.28	5.23	7.72	10.75	14.32	18.43	23.08	28.27	34
Min.	-9.702	-5.306	-6.728	-4.243	-5.898	-1.685	6.280	5.449	7.099	10.360
1st Qu.	-0.486	0.303	4.325	6.304	7.545	13.049	15.140	20.630	25.599	33.482
Median	2.420	4.246	6.621	9.349	12.814	18.146	24.597	29.292	33.197	42.879
3rd Qu.	5.244	7.758	9.515	12.908	17.845	22.470	33.341	38.564	45.096	54.460
Max.	16.420	17.679	35.452	40.516	50.867	44.323	56.283	62.652	75.490	106.544
Mean	2.519	4.416	6.907	10.001	13.588	18.721	24.918	30.223	35.955	44.400
SD	4.386	4.577	5.387	6.335	8.269	8.145	11.114	12.389	14.135	16.371
Approx USD	4.578	51.623	79.003	82.493	79.648	72.004	34.580	29.335	19.716	24.110
Approx SD	4.578	5.481	5.408	5.478	7.472	11.881	15.197	14.683	19.716	24.110

$N = 1000$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	1.87	3.28	5.23	7.72	10.75	14.32	18.43	23.08	28.27	34
Min.	-1.462	0.443	2.734	4.350	6.826	9.510	12.423	15.366	21.334	27.282
1st Qu.	0.901	3.042	4.594	6.577	8.951	12.596	16.446	21.032	27.702	33.985
Median	2.057	3.765	5.432	7.454	10.386	14.192	18.306	23.622	30.236	37.913
3rd Qu.	2.846	4.308	5.915	8.554	12.135	16.151	20.624	26.410	33.065	40.830
Max.	4.685	6.464	9.256	11.474	15.248	21.648	26.449	34.193	42.474	50.738
Mean	1.822	3.647	5.397	7.514	10.561	14.394	18.601	24.017	30.521	37.536
SD	1.365	1.180	1.161	1.415	1.971	2.373	3.094	3.813	4.472	5.004
Approx USD	2.797	17.331	20.128	20.300	17.941	15.246	10.837	7.320	5.843	6.824
Approx SD	1.246	1.414	1.655	2.010	2.468	2.987	3.817	4.789	5.843	6.824

C.2. Additional Numerical Experiments: Non-Monotone Situation

Settings. We consider the following SCM (Model 2):

$$\begin{cases} X = \frac{1}{25}Z^2 + \frac{1}{5}Z + 0.5 + (\frac{Z}{3} + 0.1)U \\ Y = X^3 - 5X^2 + X + U + E \end{cases}. \tag{102}$$

This model has the properties of the non-separability I, and non-linearity for both functions Y_x and X_z. Furthermore, the function f_Y is not monotone function. Each realized value of U and E are generated by i.i.d. uniform distributions that $U[-1, 1]$, and R values of the IV are $(0, 0.3, 0.6, \ldots, 2.7, 3)$. Let the total sample size be 100 and 1000, which means that the sample size of each value of the IV is 10 and 100, respectively. We compute the numerical integration using the left hand rule. Let the initial function be $\hat{\theta}_1(x) = 0$ for $x \in \Omega_X$, and the stop condition ϵ be 0.1. We determined the smallest step size from $(1, 0.75, 0.5.0.25, \ldots)$ where the Algorithm 1 stops before 100 iterations; and the chosen step size is 0.25. By splitting the dataset into training data and test sets, we choose the degree of the candidate models.

Results. The basic statistics of the estimators of the P-APCE estimator are shown in Table 11, the basic statistics of the estimators of the N-APCE estimator are shown in Table 14. The approximate upper bound of the SD (Approx USD) by the equation (12) and approximate bound of the SD (Approx SD) by the equation (13) are also shown in Table 14. In addition, we compare the estimator with the TSPS shown in Table 12 and the NPTSLS in Table 13. The basic statistics of the test error (19) shown in Table 10. The boxplots of the prediction values or the estimators at each point for $(0, 0.3, 0.6, \ldots, 2.7, 3)$ are shown in Figure 3. Our P-APCE and N-APCE estimators are also work well in this situation.

Table 10: Basic statistics of the test error of P-APCE estimator over 100 runs for each degree; the bold number is the smallest.

$N = 100$	2	3	4	5	6
Min.	0.032	0.052	0.108	0.084	0.094
1st Qu.	0.132	0.117	0.214	0.202	0.260
Median	0.170	0.156	0.283	0.251	0.316
Mean	0.176	**0.170**	0.291	0.269	0.321
3rd Qu.	0.220	0.200	0.350	0.329	0.378
Max.	0.360	0.506	0.658	0.640	0.593

$N = 1000$	2	3	4	5	6
Min.	0.023	0.024	0.032	0.068	0.072
1st Qu.	0.082	0.082	0.085	0.163	0.259
Median	0.119	0.111	0.120	0.203	0.319
Mean	0.124	**0.118**	0.134	0.224	0.338
3rd Qu.	0.154	0.152	0.166	0.272	0.427
Max.	0.307	0.280	0.431	0.522	0.626

Table 11: Basic statistics of the P-APCE estimator over 100 runs when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th degree term." The true coefficients are $1, -10, 3$ for $D = 0, 1, 2$.

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-21.455	-112.524	-12.706
1st Qu.	-1.453	-35.753	0.845
Median	6.591	-19.507	6.357
3rd Qu.	14.245	-3.776	13.895
Max.	45.587	30.500	56.341
Mean	6.884	-21.293	7.836
SD	11.597	23.523	10.776

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-7.699	-35.297	-4.303
1st Qu.	-2.114	-15.779	0.479
Median	0.181	-8.598	2.429
3rd Qu.	3.940	-4.209	5.251
Max.	13.934	7.164	13.788
Mean	0.965	-9.980	3.011
SD	4.074	7.788	3.285

Table 12: Basic statistics of the TSPS estimators over 100 runs when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th degree term." The true coefficients are $1, -10, 3$ for $D = 0, 1, 2$.

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-79.369	-242.889	-97.491
1st Qu.	-17.842	-64.638	-19.566
Median	-2.105	-5.722	0.531
3rd Qu.	23.754	31.387	35.343
Max.	91.758	174.522	143.904
Mean	1.724	-14.264	6.211
SD	28.991	70.436	40.554

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-17.786	-79.012	-19.795
1st Qu.	-3.297	-27.551	-0.608
Median	2.531	-16.259	7.508
3rd Qu.	6.751	-2.533	14.251
Max.	27.773	32.155	44.101
Mean	2.310	-15.873	7.160
SD	0.705	21.219	12.169

Table 13: Basic statistics of the NPTSLS estimators when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th basis function."

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-54.375	-166.638	-99.904
1st Qu.	1.363	-81.848	6.876
Median	10.593	-48.348	25.811
3rd Qu.	18.377	-19.390	44.935
Max.	42.033	169.779	98.555
Mean	10.280	-51.501	27.218
SD	14.306	51.337	31.388

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-16.996	-83.916	-32.204
1st Qu.	-2.929	-41.881	0.029
Median	2.927	-25.112	11.778
3rd Qu.	7.329	-6.241	20.938
Max.	19.271	46.499	46.130
Mean	1.947	-22.755	10.307
SD	7.730	26.933	15.843

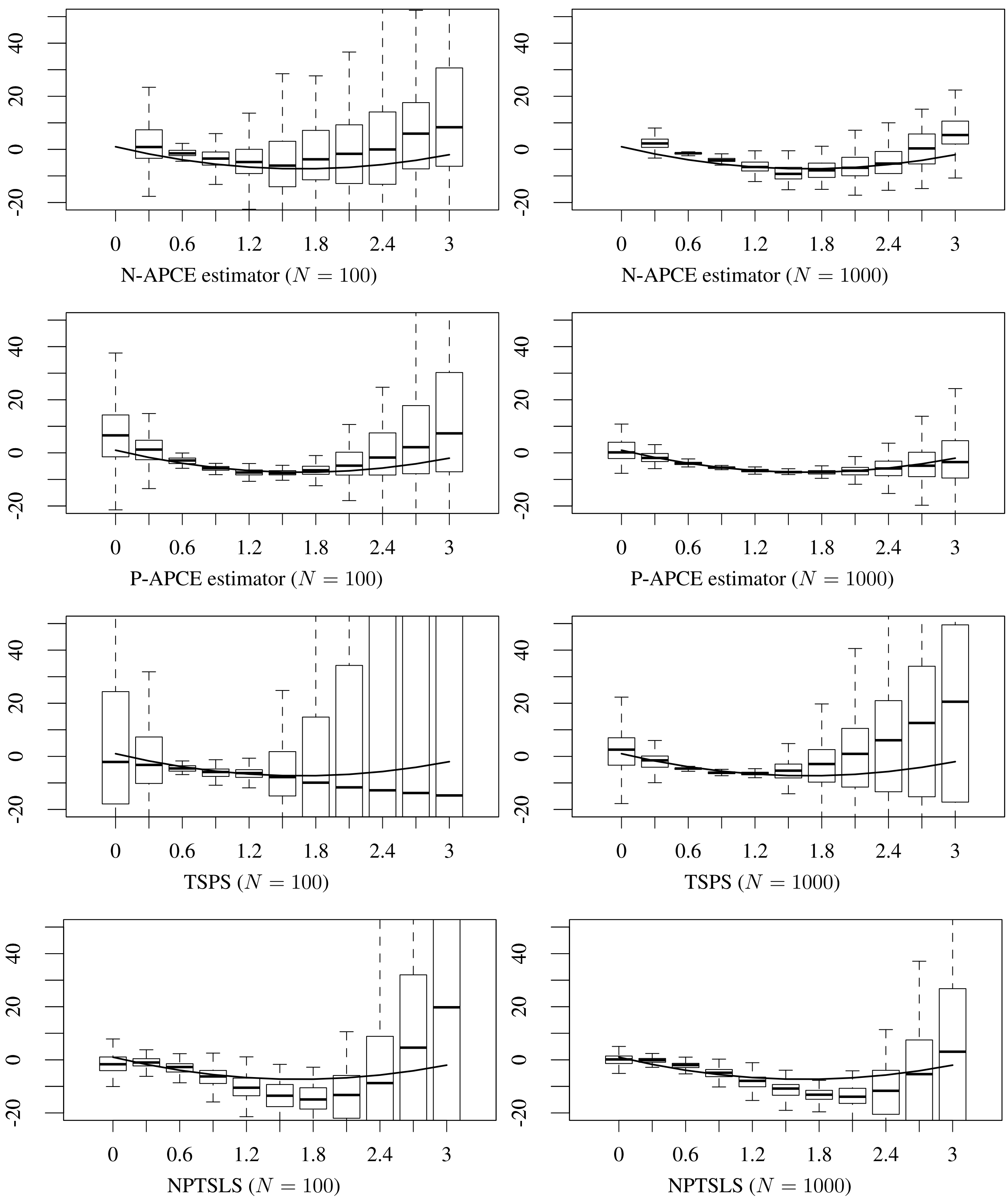

Figure 3: Boxplots of the estimated APCE by the P-APCE, TSPS, NPTSLS, and N-APCE estimators at $(0, 0.3, \dots, 2.7, 3.0)$. The black curve is the true APCE. The X-axis is the value of the treatment variable, and Y-axis is the value of APCE $\mathbb{E}[\partial_x Y_x]$ at x. In the N-APCE estimator, we can not identify the values at $x = 0$.

Table 14: Basic statistics of the N-APCE estimator over 100 runs when $N = 100$ and $N = 1000$.

$N = 100$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	-1.73	-3.92	-5.57	-6.68	-7.25	-7.28	-6.77	-5.72	-4.13	-2
Min.	-19.659	-5.338	-17.493	-22.555	-29.596	-31.164	-35.826	-32.640	-36.895	-46.024
1st Qu.	-3.301	-2.320	-5.896	-8.994	-14.030	-11.347	-12.774	-13.120	-6.930	-6.334
Median	0.889	-1.492	-3.454	-4.758	-6.105	-3.734	-1.670	-0.024	5.929	8.340
Mean	1.710	-0.674	-3.403	-3.510	-4.706	-1.540	-1.186	1.582	5.377	12.435
3rd Qu.	7.335	-0.366	-1.010	0.000	2.954	7.091	8.802	13.361	17.543	30.361
Max.	23.381	19.330	18.537	34.306	33.064	40.473	48.221	55.803	52.369	100.174
Mean	1.710	-0.674	-3.403	-3.510	-4.706	-1.540	-1.186	1.582	5.377	12.435
SD	8.078	3.598	5.058	9.526	12.692	13.507	16.213	20.295	17.986	26.968
Approx USD	7.710	38.040	40.960	48.700	38.167	44.404	33.996	26.703	24.922	33.460
Approx SD	7.710	7.360	7.924	11.130	12.215	20.588	17.775	26.703	24.922	33.460

$N = 1000$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	-1.73	-3.92	-5.57	-6.68	-7.25	-7.28	-6.77	-5.72	-4.13	-2
Min.	-4.375	-2.390	-7.031	-12.117	-15.204	-15.057	-17.252	-15.415	-14.759	-10.784
1st Qu.	0.754	-1.725	-4.548	-8.119	-11.112	-10.505	-9.901	-9.024	-5.432	2.184
Median	2.247	-1.461	-4.082	-6.637	-9.221	-7.893	-6.935	-5.334	0.359	5.412
3rd Qu.	3.865	-1.273	-3.391	-4.914	-6.808	-5.175	-2.996	-0.822	5.779	10.572
Max.	12.218	-0.920	-1.136	-0.534	1.298	1.175	7.217	10.010	23.768	27.480
Mean	2.256	-1.500	-3.987	-6.287	-8.858	-7.759	-6.373	-4.819	0.075	6.831
SD	2.698	0.334	1.001	2.468	3.281	3.674	4.901	5.854	7.649	7.570
Approx USD	2.747	17.956	19.587	17.226	12.704	9.871	8.979	8.874	8.469	9.297
Approx SD	2.227	2.195	2.619	3.102	3.894	4.557	6.512	7.344	8.469	9.297

C.3. Additional Numerical Experiments: Non-Polynomial Situation

Settings. We consider the following SCM (Model 3):

$$\begin{cases} X = \frac{1}{25}Z^2 + \frac{1}{5}Z + 0.5 + (\frac{Z}{3} + 0.1)U \\ Y = 0.05 * exp(X)^2 + U + E \end{cases} . \tag{103}$$

This model has the properties of the non-separability I, and non-linearity for both functions Y_x and X_z. Each realized value of U and E are generated by i.i.d. uniform distributions that $U[-1, 1]$, and R values of the IV are $(0, 0.3, 0.6, \dots, 2.7, 3)$. Let the total sample size be 100 and 1000, which means that the sample size of each value of the IV is 10 and 100, respectively. We compute the numerical integration using the left hand rule. Let the initial function be $\hat{\theta}_1(x) = 0$ for $x \in \Omega_X$, and the stop condition ϵ be 0.5. We determined the smallest step size from $(1, 0.75, 0.5 0.25, \dots)$ where the Algorithm 1 stops before 100 iterations; and the chosen step size is 0.25. By splitting the dataset into training data and test sets, we choose the degree of the candidate models.

Results. The basic statistics of the estimators of the P-APCE estimator are shown in Table 16, the basic statistics of the estimators of the N-APCE estimator are shown in Table 19. The approximate upper bound of the SD (Approx USD) by the equation (12) and approximate bound of the SD (Approx SD) by the equation (13) are also shown in Table 19. In addition, we compare the estimator with the TSPS shown in Table 17 and the NPTSLS in Table 18. The basic statistics of the test error (19) shown in Table 15. The boxplots of the prediction values or the estimators at each point for $(0, 0.3, 0.6, \dots, 2.7, 3)$ are shown in Figure 4. In this setting, both the P-APCE estimator and the TSPS are biased; therefore, the N-APCE estimator is superior to them. Even for the N-APCE estimator, the estimators have large bias around $x = 3$ due to the error of the numerical integration.

Table 15: Basic statistics of the test error of P-APCE estimator over 100 runs for each degree; the bold number is the smallest.

$N = 100$	2	3	4	5	6
Min.	0.124	0.067	0.127	0.084	0.082
1st Qu.	0.303	0.168	0.238	0.209	0.235
Median	0.372	0.231	0.320	0.288	0.290
Mean	0.403	**0.242**	0.336	0.291	0.311
3rd Qu.	0.507	0.291	0.409	0.357	0.371
Max.	0.726	0.551	0.678	0.617	0.679

$N = 1000$	2	3	4	5	6
Min.	0.244	0.027	0.016	0.051	0.127
1st Qu.	0.395	0.073	0.091	0.199	0.260
Median	0.471	0.102	0.123	0.261	0.336
Mean	0.490	**0.113**	0.136	0.275	0.340
3rd Qu.	0.575	0.134	0.167	0.339	0.398
Max.	0.956	0.349	0.325	0.643	0.648

Table 16: Basic statistics of the P-APCE estimator over 100 runs over 100 runs when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th degree term."

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-23.378	-88.548	-19.309
1st Qu.	1.852165	-24.751	2.342
Median	6.953	-13.343	6.144
3rd Qu.	12.079242	-3.926	11.402
Max.	43.864	46.541	40.170
Mean	6.718	-13.227	6.647
SD	10.726	21.301	9.403

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-8.518	-30.722	-5.328
1st Qu.	0.010	-10.270581	1.278
Median	3.387	-6.747	3.928
3rd Qu.	5.419	-0.2028813	5.421
Max.	15.775	15.547	13.924
Mean	2.781	-5.759	3.528
SD	4.223	8.097	3.412

Table 17: Basic statistics of the TSPS estimator over 100 runs when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th degree term."

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-46.126	-116.264	-74.016
1st Qu.	-11.111	-21.324	-12.864
Median	-1.101	2.165	0.348
3rd Qu.	8.514	27.131	14.994
Max.	47.440	120.329	69.432
Mean	-0.538	0.523	1.423
SD	18.106	44.650	26.131

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-16.857	-41.870	-23.039
1st Qu.	-2.651	-10.612	-1.069
Median	-0.012	-0.570	2.059
3rd Qu.	4.138	5.425	7.886
Max.	16.973	41.654	25.948
Mean	0.830	-2.717	3.365
SD	5.659	13.852	8.016

Table 18: Basic statistics of the NPTSLS estimators when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th basis function."

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-55.585	-240.341	-153.915
1st Qu.	-1.316	-43.386	-1.137
Median	6.079	-20.862	17.487
3rd Qu.	12.074	6.790	31.908
Max.	64.567	228.743	155.873
Mean	6.289	-22.454	16.472
SD	16.103	58.396	36.358

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-16.711	-79.781	-30.673
1st Qu.	-1.236	-34.321	3.256
Median	3.350	-14.028	11.031
3rd Qu.	9.626	2.062	23.735
Max.	23.115	57.069	48.874
Mean	3.767	-14.165	11.788
SD	7.511	25.951	15.163

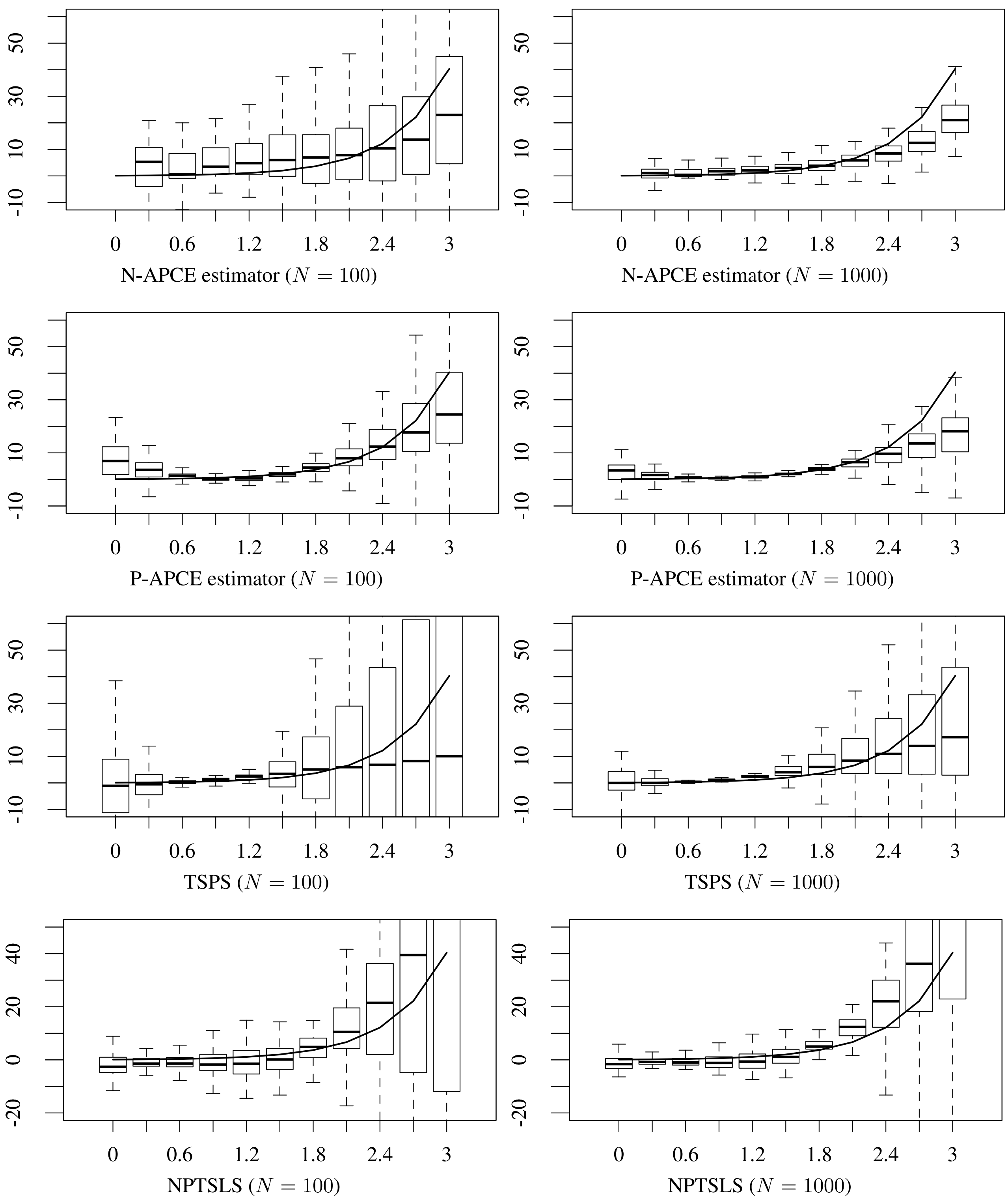

Figure 4: Boxplots of the estimated APCE by the P-APCE, TSPS, NPTSLS, and N-APCE estimators at $(0, 0.3, \ldots, 2.7, 3.0)$. The black curve is the true APCE. The X-axis is the value of the treatment variable, and Y-axis is the value of the APCE $\mathbb{E}[\partial_x Y_x]$ at x. In the N-APCE estimator, we can not identify the values at $x = 0$.

Table 19: Basic statistics of the N-APCE estimator over 100 runs when $N = 100$ and $N = 1000$.

$N = 100$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	0.182	0.332	0.605	1.102	2.009	3.660	6.669	12.151	22.141	40.343
Min.	-22.329	-12.684	-21.421	-8.009	-13.586	-16.736	-56.621	-73.827	-28.629	-81.148
1st Qu.	-3.919	-0.875	0.783	0.489	-0.063	-2.599	-1.035	-1.807	0.708	4.686
Median	5.324	0.660	3.469	4.827	5.958	6.927	7.838	10.374	13.687	22.997
3rd Qu.	10.705	8.456	10.612	12.139	15.306	15.262	17.582	25.986	29.788	45.007
Max.	53.618	29.044	83.069	120.382	161.105	60.605	68.022	65.310	84.309	137.687
Mean	4.148	4.020	6.738	8.787	9.517	8.890	9.479	11.404	17.209	25.526
SD	11.134	7.091	11.756	16.375	19.780	14.945	18.231	24.324	24.632	31.576
Approx USD	8.921	96.077	116.914	93.275	85.556	74.409	40.738	35.330	29.188	29.285
Approx SD	8.921	11.623	10.657	9.974	8.094	12.803	9.150	14.271	29.188	29.285

$N = 1000$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	0.182	0.332	0.605	1.102	2.009	3.660	6.669	12.151	22.141	40.343
Min.	-5.477	-0.822	-1.358	-3.381	-4.148	-4.708	-3.454	-7.389	1.450	7.301
1st Qu.	-0.749	-0.144	0.318	1.022	1.045	2.106	3.721	5.595	9.176	16.389
Median	1.081	0.431	1.749	2.109	2.986	3.847	5.869	8.484	12.474	21.056
3rd Qu.	2.539	2.462	2.867	3.627	4.340	5.791	7.750	11.273	16.639	26.486
Max.	8.303	8.744	12.767	13.220	8.764	19.982	20.475	23.981	34.558	41.229
Mean	1.001	1.396	1.996	2.636	2.740	4.205	6.098	8.429	13.254	21.470
SD	2.651	2.106	2.336	2.691	2.618	3.907	4.157	5.100	5.791	7.288
Approx USD	3.608	21.350	25.677	25.233	23.177	20.001	14.176	8.433	5.476	7.424
Approx SD	3.265	3.307	3.471	3.359	3.409	3.751	4.036	4.563	5.476	7.424

C.4. Additional Numerical Experiments: Separable Model

Settings. We consider the following SCM (Model 4):

$$\begin{cases} X = \frac{1}{25}Z^2 + \frac{1}{5}Z + 0.5 + 0.5U \\ Y = X^3 + X^2 + X + U + E \end{cases}. \tag{104}$$

This model has the properties of the non-linearity for both functions Y_x and X_z, and is a separable model. Each realized value of U and E are generated by i.i.d. uniform distributions that $U[-1, 1]$, and R values of the IV are $(0, 0.3, 0.6, \dots, 2.7, 3)$. Let the total sample size be 100 and 1000, which means that the sample size of each value of the IV is 10 and 100, respectively. We compute the numerical integration by the left-hand rule. Let the initial function be $\hat{\theta}_1(x) = 0$ for $x \in \Omega_X$, and the stop condition ϵ be 10. We determined the smallest step size from $(1, 0.5, 0.1.0.05, \dots)$ where the Algorithm 1 stops before 100 iterations; and the chosen step size is 0.1. By splitting the dataset into training data and test sets, we choose the degree of the candidate models.

Results. The basic statistics of the estimators of the P-APCE estimator are shown in Table 21, and the basic statistics of the estimators of the N-APCE estimator are shown in Table 24. The approximate upper bound of the SD (Approx USD) by the equation (12) and approximate bound of the SD (Approx SD) by the equation (13) are also shown in Table 24. In addition, we compare the estimator with the TSPS shown in Table 22 and the NPTSLS in Table 23. The basic statistics of the test error (19) are shown in Table 20. The boxplots of the prediction values or the estimators at each point for $(0, 0.3, 0.6, \dots, 2.7, 3)$ are shown in Figure 5. In this setting, all methods are unbiased; however, our estimators are superior to the TSPS because our estimators have small SD, especially for the N-APCE estimator. These results imply that our two P-APCE and N-APCE estimators are highly recommended even if the SCM satisfies two separabilities.

Table 20: Basic statistics of the test error of P-APCE estimator over 100 runs for each degree; the bold number is the smallest.

$N = 100$	2	3	4	5	6
Min.	0.032	0.021	0.096	0.043	0.490
1st Qu.	0.106	0.095	0.181	0.115	0.782
Median	0.150	0.120	0.242	0.168	0.908
Mean	0.168	**0.139**	0.264	0.181	0.917
3rd Qu.	0.212	0.175	0.335	0.227	1.056
Max.	0.400	0.389	0.601	0.456	1.430

$N = 1000$	2	3	4	5	6
Min.	0.052	0.053	0.088	0.053	0.185
1st Qu.	0.190	0.106	0.189	0.136	0.322
Median	0.251	0.143	0.251	0.200	0.383
Mean	0.267	**0.161**	0.252	0.204	0.394
3rd Qu.	0.332	0.200	0.302	0.258	0.466
Max.	0.599	0.405	0.561	0.567	0.799

Table 21: Basic statistics of the P-APCE estimator over 100 runs when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th degree term." The true coefficients are $1, 2, 3$ for $D = 0, 1, 2$.

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-17.688	-42.620	-31.250
1st Qu.	-1.185	-14.817	-2.916
Median	2.624	0.765	3.218
3rd Qu.	7.830	10.437	11.641
Max.	17.276	59.196	29.240
Mean	2.855	-0.955	3.764
SD	6.559	18.412	10.879

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-7.029	-22.765	-10.234
1st Qu.	-1.262	-4.727	-0.614
Median	0.620	3.008	2.357
3rd Qu.	3.419	8.567	7.179
Max.	10.107	24.833	16.936
Mean	1.004	2.067	2.927
SD	3.765	10.267	5.833

Table 22: Basic statistics of the TSPS estimator over 100 runs when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th degree term." The true coefficients are $1, 2, 3$ for $D = 0, 1, 2$.

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-122.595	-304.505	-213.737
1st Qu.	-20.949	-62.730	-27.802
Median	5.848	-9.143	9.009
3rd Qu.	28.291	55.801	40.003
Max.	120.738	336.857	190.197
Mean	3.347	-3.349	6.102
SD	36.736	91.202	53.758

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-25.805	-79.778	-34.393
1st Qu.	-5.239	-16.156	-7.000
Median	1.088	1.925	2.919
3rd Qu.	8.358	18.782	13.993
Max.	34.760	67.328	49.782
Mean	2.044	-0.023	4.204
SD	2.044	-0.023	4.204

Table 23: Basic statistics of the NPTSLS estimators when $N = 100$ and $N = 1000$; "Degree=m" means "the estimated coefficient of m-th basis function."

$N = 100$	Degree=0	Degree=1	Degree=2
Min.	-160.130	-382.764	-568.256
1st Qu.	-30.102	-43.643	-96.356
Median	-3.183	31.600	-13.400
3rd Qu.	9.766	147.958	45.862
Max.	65.174	750.735	367.791
Mean	-9.762	56.589	-33.426
SD	32.979	158.711	128.132

$N = 1000$	Degree=0	Degree=1	Degree=2
Min.	-35.144	-144.924	-118.251
1st Qu.	-9.750	-31.342	-35.186
Median	0.042	8.276	6.312
3rd Qu.	8.228	56.653	36.604
Max.	30.889	163.378	129.049
Mean	-0.989	14.404	0.233
SD	12.875	62.087	49.119

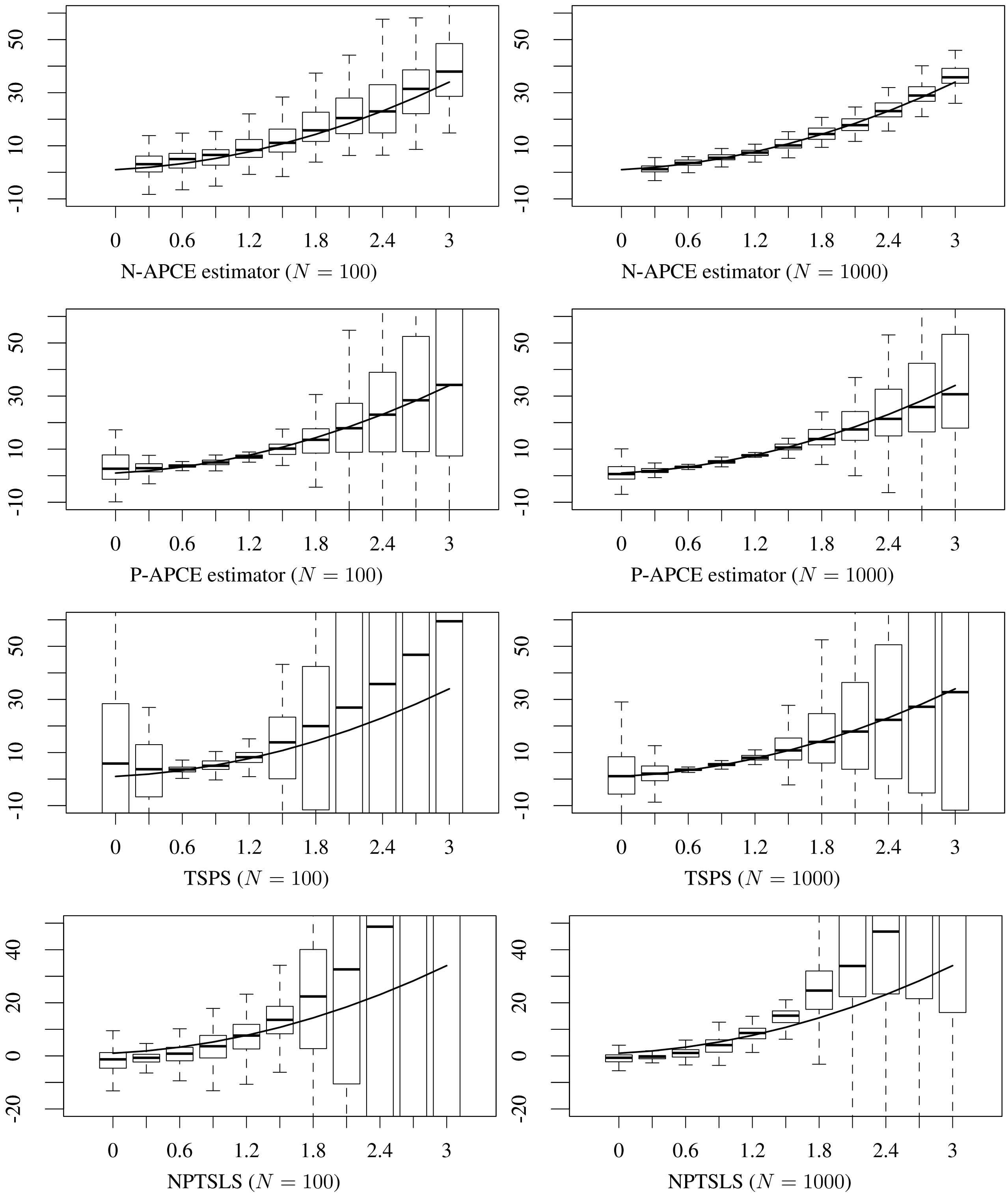

Figure 5: Boxplots of the estimated APCE by the P-APCE, TSPS, NPTSLS and N-APCE estimators at $(0, 0.3, \ldots, 2.7, 3.0)$. The black curve is the true APCE. The X-axis is the value of the treatment variable, and Y-axis is the value of the APCE $\mathbb{E}[\partial_x Y_x]$ at x. In the N-APCE estimator, we can not identify the values at $x = 0$.

Table 24: Basic statistics of the N-APCE estimator over 100 runs when $N = 100$ and $N = 1000$.

$N = 100$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	1.87	3.28	5.23	7.72	10.75	14.32	18.43	23.08	28.27	34
Min.	-11.605	-9.146	-9.557	-9.620	-1.594	3.851	6.348	6.452	8.634	14.821
1st Qu.	0.150	1.681	2.710	5.635	7.688	11.684	14.616	14.852	22.228	28.627
Median	3.063	4.997	6.548	8.422	11.102	15.816	20.449	22.942	31.435	37.952
3rd Qu.	6.100	7.135	8.491	12.347	16.158	22.444	27.947	32.868	38.507	48.481
Max.	13.837	14.769	23.106	25.179	41.409	42.040	61.082	62.592	86.438	95.344
Mean	2.681	4.076	5.827	8.926	12.499	17.696	22.271	25.203	32.722	39.305
SD	4.844	4.859	5.837	5.495	7.596	8.531	10.687	12.770	14.593	15.073
Approx USD	88.652	67.021	63.676	83.695	61.763	9.877	10.481	11.446	10.812	9.959
Approx SD	4.931	6.472	5.597	8.266	6.705	9.877	10.481	11.446	10.812	9.959

$N = 1000$	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3
True Value	1.87	3.28	5.23	7.72	10.75	14.32	18.43	23.08	28.27	34
Min.	-3.126	-0.782	2.010	3.847	5.457	9.405	11.638	15.549	20.959	24.410
1st Qu.	0.187	2.709	4.736	6.435	9.183	12.496	15.731	20.898	26.787	33.617
Median	1.184	3.561	5.479	7.425	10.142	14.422	17.763	23.054	28.922	35.793
3rd Qu.	2.381	4.608	6.598	8.266	12.331	16.625	20.121	26.083	32.220	39.129
Max.	5.562	5.923	8.978	10.595	15.305	23.484	26.917	31.931	40.139	49.402
Mean	1.218	3.365	5.609	7.343	10.600	14.653	18.040	23.364	29.256	36.121
SD	1.619	1.588	1.446	1.435	2.191	2.827	2.982	3.652	3.794	4.499
Approx USD	20.555	27.904	27.373	22.563	17.857	9.501	3.409	3.482	3.670	4.400
Approx SD	2.374	2.601	2.590	2.653	2.775	3.172	3.409	3.482	3.670	4.400