KIER DISCUSSION PAPER SERIES

KYOTO INSTITUTE
OF
ECONOMIC RESEARCH

Discussion Paper No.931

"DOUBLY ROBUST UNIFORM CONFIDENCE BAND FOR THE CONDITIONAL AVERAGE TREATMENT EFFECT FUNCTION"

SOKBAE LEE, RYO OKUI, AND YOON-JAE WHANG

January 2016

KYOTO UNIVERSITY

KYOTO, JAPAN

DOUBLY ROBUST UNIFORM CONFIDENCE BAND FOR THE CONDITIONAL AVERAGE TREATMENT EFFECT FUNCTION

SOKBAE LEE[1,2], RYO OKUI[3,4], AND YOON-JAE WHANG[1]

Abstract. In this paper, we propose a doubly robust method to present the heterogeneity of the average treatment effect with respect to observed covariates of interest. We consider a situation where a large number of covariates are needed for identifying the average treatment effect but the covariates of interest for analyzing heterogeneity are of much lower dimension. Our proposed estimator is doubly robust and avoids the curse of dimensionality. We propose a uniform confidence band that is easy to compute, and we illustrate its usefulness via Monte Carlo experiments and an application to the effects of smoking on birth weights.

Keywords: average treatment effect conditional on covariates, uniform confidence band, double robustness, Gaussian approximation.

JEL classification codes: C14, C21

[1]Department of Economics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 151-742, Republic of Korea.
[2]Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, UK.
[3]Institute of Economic Research, Kyoto University, Yoshida-Honmachi Sakyo, Kyoto, 606-8501, Japan.
[4]Department of Econometrics & OR, VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
E-mail addresses: sokbae@gmail.com, okui.ryo.3@gmail.com, whang@snu.ac.kr.
Date: This version: January 2016; the first draft: November 2014.
We would like to thank Yanchun Jin for capable research assistance and Takahiro Hoshino, Yu-Chin Hsu, Kengo Kato, Edward Kennedy, Taisuke Otsu, and Dylan Small for helpful comments. Lee's work was supported by the European Research Council (ERC-2009-StG-240910-ROMETA). Okui's work was supported by the Japan Society of the Promotion of Science (KAKENHI 25285067, 25780151, 15H03329). Whang's work was supported by the SNU College of Social Science Research Grant.

1. Introduction

In this paper, we propose a doubly robust method to present the heterogeneity of the average treatment effect with respect to observed covariates of interest. To describe our methodology, we consider the potential outcome framework. Let Y_1 and Y_0 be potential individual outcomes in two states, with treatment and without treatment, respectively. For each individual, the observed outcome Y is $Y = DY_1 + (1 - D)Y_0$, where D denotes an indicator variable for the treatment, with $D = 0$ if an individual is not treated and $D = 1$ if an individual is treated. We assume that independent and identically distributed observations $\{(Y_i, D_i, \mathbf{Z}_i) : i = 1, \ldots, n\}$ of $(Y, D, \mathbf{Z})$ are available, where $\mathbf{Z} \in \mathbb{R}^p$ denotes a p-dimensional vector of covariates.

Suppose that a researcher is interested in evaluating the average treatment effect conditional on only a subset of covariates $\mathbf{X}$, which is of a substantially lower dimension than $\mathbf{Z}$, where $\mathbf{Z} \equiv (\mathbf{X}^\top, \mathbf{V}^\top)^\top \in \mathbb{R}^d \times \mathbb{R}^m$, $p \equiv d + m$. That is, we are interested in a case where $d \ll p$.

The main object of interest in this paper is the conditional average treatment effect function (CATEF); namely:

$$g(\mathbf{x}) \equiv \mathbb{E}[Y_1 - Y_0 | \mathbf{X} = \mathbf{x}]. \tag{1.1}$$

When $d \geq 3$, it is difficult to plot $g(\mathbf{x})$, not to mention low precision due to the curse of dimensionality. Hence, for practical reasons, we focus on the case that $d = 1$ or $d = 2$, while p is often of a much higher dimension.

To achieve identification of the CATEF, we assume that Y_1 and Y_0 are independent of D conditional on $\mathbf{Z}$ (known as the unconfoundedness assumption):

$$(Y_1, Y_0) \perp D | \mathbf{Z}, \tag{1.2}$$

where $\perp$ denotes the independence. For (1.2) to be plausible in applications, applied researchers tend to consider a large number of covariates $\mathbf{Z}$. Note that in our setup, the treatment may be confounded in the sense that the treatment assignment may not be independent of the potential outcome variables given $\mathbf{X}$ only. To satisfy the unconfoundedness condition, a much larger set of conditioning variables $\mathbf{Z}$ needs to be employed.

Different roles of covariates between $\mathbf{X}$ and $\mathbf{V}$ are noted in the recent literature. For example, Ogburn, Rotnitzky, and Robins (2015) consider a similar issue in the context of the local average treatment effect (LATE) of Imbens and Angrist (1994). Ogburn, Rotnitzky, and Robins (2015) emphasize that conditioning on a large number of covariates $\mathbf{Z}$ may be required to make it plausible that the binary instrument is valid. In their empirical example, Ogburn, Rotnitzky, and Robins (2015) revisit the analyses of Poterba, Venti, and Wise (1995) and Abadie (2003) to examine whether participation in 401(k) pension plans increases household savings. In their example, the vector of covariates $\mathbf{Z}$ for the identifying assumption consists of income, age, marital status, and family size, whereas the variable of interest $\mathbf{X}$ is income. Abrevaya, Hsu, and Lieli (2015) also consider the case of investigating the effect of smoking during pregnancy on birth weights. They are interested in estimating (1.1) with $\mathbf{X}$ being the age of mother; however, as noted in Abrevaya, Hsu, and Lieli (2015), it is unlikely that conditioning only on the age of mother would achieve the unconfoundedness assumption with nonexperimental data. As a result, it is necessary to consider a high-dimensional $\mathbf{Z}$, including the age of the mother.

The fact that a high-dimensional $\mathbf{Z}$ needs to be employed for (1.2) to be plausible in an application makes a fully nonparametric estimation approach impractical because of the curse of dimensionality. For example, the propensity score is not nonparametrically estimable in moderately sized samples, if the dimension of $\mathbf{Z}$ is high. One

obvious alternative is to use a parametric model for the propensity score; however, it may lead to misleading results if the parametric model is misspecified.

With the aim of providing a practical method and, at the same time, reducing sensitivity to model misspecification, we propose to use a doubly robust method based on parametric regression and propensity score models. Our estimator of the CATEF is doubly robust in the sense that it is consistent when at least one of the regression model and the propensity score model is correctly specified. Specifically, we first estimate CATEF($\mathbf{Z}$) using a doubly robust procedure: we estimate a parametric regression model of the outcome on $\mathbf{Z}$ for each treatment status and a parametric model for the probability of selecting into the treatment given $\mathbf{Z}$; we then combine the parametric estimation results in a doubly robust fashion to construct an estimate of CATEF($\mathbf{Z}$). We then obtain an estimate of CATEF($\mathbf{X}$) by adopting the local linear smoothing of CATEF($\mathbf{Z}$). As a result, we avoid high-dimensional smoothing with respect to $\mathbf{Z}$ but mitigate the problem of misspecification by both the doubly robust estimation and low-dimensional smoothing with respect to $\mathbf{X}$.

We emphasize that we are willing to assume parametric specifications for the propensity score and regression models as functions of $\mathbf{Z}$ to avoid the curse of dimensionality, but not for CATEF($\mathbf{X}$). One may consider parametric estimation of CATEF($\mathbf{X}$), as Ogburn, Rotnitzky, and Robins (2015) estimate their LATE parameter using least squares approximations. However, note that even if the parametric specification of CATEF($\mathbf{Z}$) is correct, the resulting specification of CATEF($\mathbf{X}$) may not be correctly specified since $\mathbb{E}[\mathbf{Z}|\mathbf{X}]$ is possibly highly nonlinear. To avoid this misspecification, we estimate CATEF($\mathbf{X}$) nonparametrically.

Because the CATEF is a functional parameter, as a tool of inference, we propose to use a uniform confidence band for the CATEF. Our construction of the uniform confidence band is based on some analytic approximation of the supremum of a Gaussian

process using arguments built on Piterbarg (1996), combined with a Gaussian approximation result of Chernozhukov, Chetverikov, and Kato (2014) and an empirical process result of Ghosal, Sen, and van der Vaart (2000). Our method is simple to implement and does not rely on resampling techniques.

This paper contributes to the literature on doubly robust estimation by demonstrating that the doubly robust procedures are useful for estimating the CATEF. In this paper, we focus on the so-called augmented inverse probability weighting estimator that was originally proposed by Robins, Rotnitzky, and Zhao (1994) for the estimation of the mean (see also Robins and Rotnitzky, 1995; Scharfstein, Rotnitzky, and Robins, 1999). Their estimator appears to be the first estimator to be recognized as being doubly robust. Since then, many other alternative doubly robust estimators have been proposed in the literature. For example, the inverse probability weighting regression adjustment estimator (Kang and Schafer, 2007; Wooldridge, 2007, 2010) is widely known and has been implemented in statistical software packages. See the introduction of Tan (2010) for a comprehensive summary of other doubly robust estimators. Doubly robust estimators have been advocated for use in many different areas of application: See, for example, Lunceford and Davidian (2004) for medicine, Glynn and Quinn (2010) for political science, Wooldridge (2010) for economics, and Schafer and Kang (2008) for psychology. There are also doubly robust estimators available for different settings including instrumental variables estimation (Tan, 2006; Okui, Small, Tan, and Robins, 2012) and estimation under multivalued treatments (Derya Uysal, 2015). It would not be difficult to extend our method to allow these other doubly robust estimators and to consider different settings. However, to keep the analysis simple, in this paper, we focus on the augmented inverse probability weighting estimator of the CATEF.

The CATEF is mathematically equivalent to "V-adjusted variable importance" of van der Laan (2006), who proposes it as a measure of variable importance in prediction. van der Laan (2006) proposes a doubly robust estimator of V-adjusted variable importance. Contrary to ours, he considers the projection of the V-adjusted variable importance on a parametric working model and does not consider a nonparametric estimation. Moreover, a uniform confidence band is not examined in van der Laan (2006).

In a recent paper, Abrevaya, Hsu, and Lieli (2015) consider the estimation of the CATEF[1]; however, there are two main differences of this paper relative to Abrevaya, Hsu, and Lieli (2015). First, we propose the doubly robust procedure to estimate the CATEF. Abrevaya, Hsu, and Lieli (2015) consider the inverse probability weighting estimator. The inverse probability weighting estimator suffers from model misspecification when the propensity score model is misspecified and from the curse of dimensionality when it is estimated nonparametrically. Second, we present a method to construct a uniform confidence band, whereas Abrevaya, Hsu, and Lieli (2015) only provide a pointwise confidence interval.

The remainder of the paper is organized as follows. Section 2 presents the doubly robust estimation method, Section 3 gives an informal description of how to construct a two-sided, symmetric uniform confidence band when the dimension of $\mathbf{X}$ is one, and Section 4 deals with a general case and provides formal theoretical results. In Section 5, the results of Monte Carlo simulations demonstrate that in finite samples, our doubly robust estimator works well, and the proposed confidence band has desirable coverage properties. Section 6 gives an empirical application, and Section 7 concludes. The proofs are contained in Appendix A.

[1]Our paper is independent of Abrevaya, Hsu, and Lieli (2015) and it is started without knowing their work.

2. Doubly Robust Estimation of the Average Treatment Effect Conditional on Covariates of Interest

In this section, a doubly robust method for estimating the CATEF is proposed. We first estimate the CATEF for all the covariates using a doubly robust method. We then obtain the CATEF for the covariates of interest using a nonparametric approach.

Define:

$$\pi(\mathbf{z}) \equiv \mathbb{E}\left[D|\mathbf{Z}=\mathbf{z}\right],$$
$$\mu_j(\mathbf{z}) \equiv \mathbb{E}\left[Y|\mathbf{Z}=\mathbf{z},\ D=j\right] \text{ for } j=0,1,$$

where $\pi(\mathbf{z})$ is the propensity score and $\mu_j(\mathbf{z})$ for $j=0,1$ are called regression functions. Note that $\mu_j(\mathbf{z}) = E(Y_j|\mathbf{Z}=\mathbf{z})$ for $j=0,1$ under unconfoundedness. Let $\pi(\mathbf{z},\beta)$ and $\mu_j(\mathbf{z},\alpha_j)$ for $j=0,1$ denote parametric models of $\pi(\mathbf{z})$ and $\mu_j(\mathbf{z})$, respectively. ($\mu_j(\mathbf{z},\alpha_j)$ may also be called "marginal structural models" of Robins (2000).) A doubly robust procedure requires that either $\pi(\mathbf{z})$ or $\mu_j(\mathbf{z})$ for $j=0,1$ should be correctly specified, thereby allowing for misspecification in $\pi(\mathbf{z})$ or in $\mu_j(\mathbf{z})$. Let $\theta_0 \equiv (\alpha_{10}^\top, \alpha_{00}^\top, \beta_0^\top)^\top$ denote the vector of true or pseudo-true parameter values that optimize some criterion functions.

We consider the augmented inverse probability weighting approach. Let:

$$\begin{aligned}
\psi_1(\mathbf{W},\alpha_1,\beta) &\equiv \frac{DY}{\pi(\mathbf{Z},\beta)} - \frac{D-\pi(\mathbf{Z},\beta)}{\pi(\mathbf{Z},\beta)}\mu_1(\mathbf{Z},\alpha_1),\\
\psi_0(\mathbf{W},\alpha_0,\beta) &\equiv \frac{(1-D)\,Y}{1-\pi(\mathbf{Z},\beta)} + \frac{D-\pi(\mathbf{Z},\beta)}{1-\pi(\mathbf{Z},\beta)}\mu_0(\mathbf{Z},\alpha_0),\\
\psi(\mathbf{W},\theta) &\equiv \psi_1(\mathbf{W},\alpha_1,\beta) - \psi_0(\mathbf{W},\alpha_0,\beta),
\end{aligned}$$

where $\mathbf{W} \equiv (Y, \mathbf{Z}^\top)^\top$ and $\theta \equiv (\alpha_1^\top, \alpha_0^\top, \beta^\top)^\top$. The first terms in $\psi_1(\mathbf{W},\alpha_1,\beta)$ and $\psi_0(\mathbf{W},\alpha_0,\beta)$ correspond to inverse probability weighting. The second terms are augmented terms that make the procedure doubly robust.

The following lemma gives regularity conditions under which $g(\mathbf{x})$ is identified.

Lemma 1 (Identification of the CATEF)**.** *Assume that* (1.2) *holds and* $0 < \pi(\mathbf{Z}, \beta_0) < 1$ *almost surely. Suppose that either there exists* β_0 *such that* $\mathbb{E}[D|\mathbf{Z}] = \pi(\mathbf{Z}, \beta_0)$ *almost surely or there exist* α_{10} *and* α_{00} *such that* $\mathbb{E}[Y_1|\mathbf{Z}] = \mu_1(\mathbf{Z}, \alpha_{10})$ *and* $\mathbb{E}[Y_0|\mathbf{Z}] = \mu_0(\mathbf{Z}, \alpha_{00})$ *almost surely. Then:*

$$g(\mathbf{x}) = \mathbb{E}[\psi(\mathbf{W}, \theta_0)|\mathbf{X} = \mathbf{x}].$$

Lemma 1 suggests that one may estimate $g(\mathbf{x})$ by running the nonparametric regression of $\psi(\mathbf{W}, \hat{\theta})$ on $\mathbf{X}_i$, where $\hat{\theta}$ is a consistent parametric estimator of θ_0. Moreover, this lemma implies that the CATEF can be identified through $\psi(\mathbf{W}, \theta_0)$ if either the regression models ($\mu_1(\mathbf{z}, \alpha_1)$ and $\mu_0(\mathbf{z}, \alpha_0)$) or the propensity score model ($\pi(\mathbf{z}, \beta)$) is correctly specified (or both). That is, even if $\mu_1(\mathbf{z}, \alpha_1)$ and $\mu_0(\mathbf{z}, \alpha_0)$ do not represent the true conditional expectation functions, provided that $\pi(\mathbf{z}, \beta)$ is correct, the CATEF is identified. Similarly, even if $\pi(\mathbf{z}, \beta)$ is misspecified, provided that $\mu_1(\mathbf{z}, \alpha_1)$ and $\mu_0(\mathbf{z}, \alpha_0)$ are correct, the CATEF is identified.

2.1. **Parametric Estimation of θ.** For concreteness, we consider the following estimation procedure for θ_0. However, how θ_0 is estimated does not alter our results provided that the rate of convergence is sufficiently fast so that Assumption 1(7) given below is satisfied. For each $j = 0, 1$, we estimate α_j by least squares:

$$\hat{\alpha}_j \equiv \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} D_i^j (1 - D_i)^{1-j} (Y_i - \mu_j(\mathbf{Z}_i, \alpha))^2. \tag{2.1}$$

We estimate β by maximum likelihood (e.g., probit or logit):

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n} \left(D_i \log \pi(\mathbf{Z}_i, \beta) + (1 - D_i) \log(1 - \pi(\mathbf{Z}_i, \beta)) \right). \tag{2.2}$$

Remark 1. When the dimension of $\mathbf{Z}$ is not too high, an alternative to parametric estimation of $\psi(\mathbf{W}, \theta_0)$ is to estimate its nonparametric counterpart via local polynomial estimators as in Rothe and Firpo (2014). However, this would not work when the dimension of $\mathbf{Z}$ is sufficiently high (see related remarks in Rothe and Firpo (2014)). The latter is the case we focus on in the paper.

2.2. **Local Linear Estimation of g.** We consider a local linear estimator of $g(\mathbf{x})$.[2] Assume that $g(\mathbf{x})$ is twice continuously differentiable. For each $\mathbf{x} = (x_1, \ldots, x_d)$, the local linear estimator of $g(\mathbf{x})$ can be obtained by minimizing:

$$S_n(\gamma) \equiv \sum_{i=1}^{n} \left[\psi(\mathbf{W}_i, \hat{\theta}) - \gamma_0 - \gamma_1^\top (\mathbf{X}_i - \mathbf{x})\right]^2 \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right)$$

with respect to $\gamma \equiv (\gamma_0, \gamma_1^\top)^\top \in \mathbb{R}^{d+1}$, where $\mathbf{K}(\cdot)$ is a kernel function on $\mathbb{R}^d$ and h_n is a sequence of bandwidths. More specifically, let $\hat{g}(\mathbf{x}) = \mathbf{e}_1^\top \hat{\gamma}(\mathbf{x})$, where $\hat{\gamma}(\mathbf{x}) \equiv \arg\min_{\gamma \in \mathbb{R}^{d+1}} S_n(\gamma)$ and $\mathbf{e}_1$ is a column vector whose first entry is one, and the rest are zero. Because $\hat{\theta}$ can be estimated at a rate of $n^{-1/2}$, there is no estimation effect from the first stage, implying that we can carry out inference as if θ_0 were known.

3. An Informal Description of a Uniform Confidence Band

In this section, we provide an informal description of how to construct a two-sided, symmetric uniform confidence band. For simplicity, we focus on the leading case where $d = 1$. Let $\mathcal{I} \equiv [a, b]$ denote an interval of interest for which we build a uniform confidence band. Assume that $\mathcal{I}$ is a subset of the support of X. We use nonbold x to mean that x is one-dimensional.

Algorithm. Carry out the following steps to construct a $(1 - \alpha)$ uniform confidence band.

[2]In this paper, we consider cases in which $\mathbf{x}$ is continuous. When $\mathbf{x}$ is discreet, the CATEF can be estimated by the sample average of $\psi(\mathbf{W}, \hat{\theta})$ using the sub-sample for each value of $\mathbf{x}$. Moreover, constructing a confidence band is standard when $\mathbf{x}$ takes a finite number of values.

(1) Obtain $\hat{g}(x)$ using a local linear estimator with a bandwidth h_n such that:

$$h_n = \widehat{h} \times n^{1/5} \times n^{-2/7},$$

where $\widehat{h}$ is a commonly used optimal bandwidth in the literature (for example, the plug-in method of Ruppert, Sheather, and Wand (1995)).

(2) Obtain the pointwise standard error $\hat{s}(x)/(nh_n)^{1/2}$ of $\hat{g}(x)$ by constructing a feasible version of the asymptotic standard error formula:

$$\frac{\hat{s}(x)}{(nh_n)^{1/2}} \equiv \left\{ [nh_n f_X(x)]^{-1} \int K^2(u)du \, \sigma^2(x) \right\}^{1/2},$$

where f_X is the density of X and $\sigma^2(x)$ is the conditional variance function.

(3) To compute a critical value $c(1-\alpha)$, define:

$$\lambda \equiv \frac{-\int K(u)K''(u)du}{\int K^2(u)du}.$$

Let:

$$a_n \equiv a_n(\mathcal{I}) = \left(2\log(h_n^{-1}(b-a)) + 2\log\frac{\lambda^{1/2}}{2\pi} \right)^{1/2}.$$

Now set the critical value by:

$$c(1-\alpha) \equiv \left(a_n^2 - 2\log\{\log[(1-\alpha)^{-1/2}]\} \right)^{1/2}.$$

(4) For each $x \in \mathcal{I}$, we set the two-sided symmetric confidence band:

$$\hat{g}(x) - c(1-\alpha)\frac{\hat{s}(x)}{\sqrt{nh_n}} \leq g(x) \leq \hat{g}(x) + c(1-\alpha)\frac{\hat{s}(x)}{\sqrt{nh_n}}.$$

We make some remarks on the proposed algorithm. In step (1), the factor $n^{1/5} \times n^{-2/7}$ is multiplied in the definition of h_n to ensure that the bias is asymptotically negligible by undersmoothing. In step (2), one can estimate f_X and $\sigma^2(x) \equiv \mathrm{Var}\left[\psi(\mathbf{W}, \theta_0)|X = x\right]$ using the standard kernel density and regression estimators

with the same kernel function $K(\cdot)$ and the same bandwidth h and also with an estimator of θ_0. In step (3), we may restrict the bandwidth such that $h_n \ll (b-a)$ (which is satisfied asymptotically), thereby imposing the condition that $\log(h_n^{-1}(b-a))$ is positive. The critical value proposed in step (3) is strictly positive if α is not too close to one or if n is large enough.

Remark 2. We may compare our proposal with the critical value based on the $(1-\alpha)$ quantile of the Gumbel distribution, which is given by:

$$c_\infty(1-\alpha) \equiv a_n + \frac{-\log\{\log[(1-\alpha)^{-1/2}]\}}{a_n}.$$

Note that:

$$c_\infty(1-\alpha) - c(1-\alpha) = \left[\frac{-\log\{\log[(1-\alpha)^{-1/2}]\}}{a_n}\right]^2,$$

which is strictly positive for small α but converges to zero as a_n diverges. Hence, we expect that in finite samples, the confidence band based on $c_\infty(1-\alpha)$ is too wide and has higher coverage probability than the nominal level. It is shown in the next section that the critical value based on the Gumbel distribution is accurate only up to the logarithmic rate, where our proposed critical value is precise in a polynomial rate. This is because our proposal uses a higher-order expansion of Piterbarg (1996), whose approximation error is of a polynomial rate. See Theorem 2 in Section 4 for details.

Remark 3. Our construction of critical values is based on a simple analytic method that is easy to compute. Alternatively, one may rely on bootstrap methods to compute critical values for the uniform confidence band. For example, see Claeskens and Keilegom (2003) for smoothed bootstrap confidence bands and Chernozhukov, Lee, and Rosen (2013) for multiplier bootstrap confidence bands. Chernozhukov, Chetverikov, and Kato (2013) show that in general settings including high dimensional models,

Gaussian multiplier bootstrap methods yield critical values for which the approximation error decreases polynomially in the sample size. Roughly speaking, both our simple analytic correction and multiplier bootstrap methods yield critical values that are accurate at polynomial rates. A refined theoretical analysis is necessary to determine which type of the critical value is better asymptotically.

Remark 4. The proposed confidence band can be used to test whether the CATEF is constant. Suppose that our null hypothesis is that $g(\mathbf{x})$ is constant in $\mathcal{I}$. This null hypothesis can be written as $g(\mathbf{x}) = g_{\mathcal{I}}$, where $g_{\mathcal{I}} = \mathbb{E}[g(\mathbf{x})|\mathbf{x} \in \mathcal{I}]$. Since $g_{\mathcal{I}}$ can be estimated at the parametric $(\sqrt{n}\,)$ rate and the estimator thus converges faster than $\hat{g}(\mathbf{x})$, we can ignore the estimation error in $\hat{g}(\mathbf{x})$. We reject the constancy of $g(\mathbf{x})$, if the confidence band does not include the estimate of $g_{\mathcal{I}}$ for some $\mathbf{x} \in \mathcal{I}$.

4. Asymptotic Theory

In this section, we establish asymptotic theory. Let $U \equiv \psi(\mathbf{W}, \theta_0) - g(\mathbf{X})$ and let $U_i \equiv \psi(\mathbf{W}_i, \theta_0) - g(\mathbf{X}_i)$ for $i = 1, \ldots, n$. Let $\hat{s}^2(\mathbf{x})$ be the estimator of the asymptotic variance of $\hat{g}(\mathbf{x})$. Let $s_n^2(\mathbf{x})$ denote the population version of the asymptotic variance of the estimator:

$$s_n^2(\mathbf{x}) \equiv \frac{1}{h_n^d} \mathbb{E}\left[\frac{U^2}{f_{\mathbf{X}}^2(\mathbf{x})} \mathbf{K}^2 \left(\frac{\mathbf{X} - \mathbf{x}}{h_n} \right) \right].$$

Assume that the d-dimensional kernel function is the product of d univariate kernel functions. That is, $\mathbf{K}(\mathbf{s}) = \prod_{j=1}^d K(s_j)$, where $\mathbf{s} \equiv (s_1, \ldots, s_d)$ is a d-dimensional vector and $K(\cdot)$ is a kernel function on $\mathbb{R}$. Let $\rho_d(\mathbf{s}) = \prod_{j=1}^d \rho(s_j)$, where:

$$\rho(s_j) \equiv \frac{\int K(u) K(u - s_j) du}{\int K^2(u) du}, \tag{4.1}$$

for each j. We make the following assumptions.

Assumption 1. *Let $d < 4$.*

(1) $\mathcal{I} \equiv \prod_{j=1}^{d}[a_j, b_j]$, *where* $a_j < b_j, j = 1, \ldots, d$, *and* $\mathcal{I}$ *is a strict subset of the support of* $\boldsymbol{X}$.

(2) *The distribution of* $\mathbf{X}$ *has a bounded Lebesgue density* $f_{\mathbf{X}}(\cdot)$ *on* $\mathbb{R}^d$. *Furthermore,* $f_{\mathbf{X}}(\cdot)$ *is bounded below from zero with continuous derivatives on* $\mathcal{I}$.

(3) *The density of* U *is bounded,* $\mathbb{E}[U^2|\mathbf{X} = \mathbf{x}]$ *is continuous on* $\mathcal{I}$, *and* $\sup_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}[U^4|\mathbf{X} = \mathbf{x}] < \infty$.

(4) $g(\cdot)$ *is twice continuously differentiable on* $\mathcal{I}$.

(5) $\mathbf{K}(\mathbf{s}) = \prod_{j=1}^{d} K(s_j)$, *where* $K(\cdot)$ *is a kernel function on* $\mathbb{R}$ *that has finite support on* $[-1, 1]$, $\int_{-1}^{1} uK(u)du = 0$, $\int_{-1}^{1} K(u)du = 1$, *symmetric around zero, and six times differentiable.*

(6) $h_n = Cn^{-\eta}$, *where* C *and* η *are positive constants such that* $\eta < 1/(2d)$ *and* $\eta > 1/(d + 4)$.

(7) $\inf_{n\geq 1} \inf_{\mathbf{x}\in\mathcal{I}} s_n(\mathbf{x}) > 0$ *and* $s_n(\mathbf{x})$ *is continuous for each* $n \geq 1$. *Furthermore,* $\mathbf{x} \mapsto \mathbb{E}[U^2|\mathbf{X} = \mathbf{x}] f_{\mathbf{X}}(\mathbf{x})$ *is Lipschitz continuous.*

(8) *There exists an estimator* $\hat{s}^2(\mathbf{x})$ *such that*

$$\sup_{\mathbf{x}\in\mathcal{I}} \left|\hat{s}^2(\mathbf{x}) - s_n^2(\mathbf{x})\right| = O_p(n^{-c})$$

for some constant $c > 0$.

(9) $\max\left\{(nh_n^d)^{1/2}|\psi(\mathbf{W}_i, \hat{\theta}) - \psi(\mathbf{W}_i, \theta_0)| : i = 1, \ldots, n\right\} = O_p(n^{-c})$ *for some constant* $c > 0$.

Most of the assumptions are standard. One of the bandwidth conditions in h_n (that is, $\eta > 1/(d + 4)$) imposes undersmoothing, so that we can ignore the bias asymptotically. The rule-of-thumb bandwidth proposed in Section 3 satisfies the required rate conditions.

Remark 5. Note that $d < 4$ is necessary to ensure that $\eta < 1/(2d)$ and $\eta > 1/(d+4)$ can hold jointly. It is possible to extend our asymptotic theory to the case that

$d \geq 4$ using a higher-order local polynomial estimator under stronger smoothness conditions. In this paper, we limit our attention to the local linear estimator since we are mainly interested in low dimensional $\boldsymbol{x}$.

Remark 6. An estimator of $\hat{s}^2(\mathbf{x})$ is readily available. For example, we may set

$$\hat{s}^2(\mathbf{x}) = \frac{1}{h_n^d} \sum_{i=1}^{n} \frac{\hat{U}_i^2}{\hat{f}_{\mathbf{X}}^2(\mathbf{x})} \mathbf{K}^2 \left(\frac{\mathbf{X} - \mathbf{x}}{h_n} \right),$$

where $\hat{U}_i \equiv \psi(\mathbf{W}_i, \hat{\theta}) - \hat{g}(\mathbf{X}_i)$ and $\hat{f}_{\mathbf{X}}(\cdot)$ is the kernel density estimator. Alternatively, since

$$s_n^2(\mathbf{x}) \to \frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \int \mathbf{K}^2(\mathbf{u})\, d\mathbf{u}, \text{ as } n \to \infty,$$

where $\sigma^2(\mathbf{x}) \equiv \mathbb{E}[U^2 | \mathbf{X} = \mathbf{x}]$, we may use

$$\hat{s}^2(\mathbf{x}) = \frac{\hat{\sigma}^2(\mathbf{x})}{\hat{f}_{\mathbf{X}}(\mathbf{x})} \int \mathbf{K}^2(\mathbf{u})\, d\mathbf{u},$$

with a suitable nonparametric estimator $\hat{\sigma}^2(\mathbf{x})$ of $\sigma^2(\mathbf{x})$ such as the kernel regression estimator using $\{(\hat{U}_i^2, \mathbf{X}_i) : i = 1, \ldots, n\}$. For either estimator, it is straightforward to verify condition (8) of Assumption 1 using the standard results in kernel estimation.

Remark 7. Condition (9) of Assumption 1 is satisfied, for example, if $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$, functions $\beta \mapsto \pi(\mathbf{Z}, \beta)$ and $\alpha_j \mapsto \mu_j(\mathbf{Z}, \alpha_j), j = 0, 1$, are Lipschitz continuous, $\pi(\mathbf{Z}, \beta_0)$ is bounded between ϵ and $1 - \epsilon$ for some constant $\epsilon > 0$, provided that some weak moment conditions on $(Y, \mathbf{Z})$ hold.

Let $a_n \equiv a_n(\mathcal{I})$ be the largest solution to the following equation:

$$\text{mes}(\mathcal{I}) h_n{}^{-d} \lambda^{d/2} (2\pi)^{-(d+1)/2} a_n^{d-1} \exp(-a_n^2/2) = 1, \tag{4.2}$$

where $\text{mes}(\mathcal{I})$ is the Lebesgue measure of $\mathcal{I}$; that is, $\text{mes}(\mathcal{I}) = \prod_{j=1}^{d}(b_j - a_j)$ and:

$$\lambda = \frac{-\int K(u)K''(u)du}{\int K^2(u)du}. \tag{4.3}$$

The following is the main theoretical result of our paper.

Theorem 2. *Let Assumption 1 hold. Then there exists $\kappa > 0$ such that, uniformly in t, on any finite interval:*

$$\begin{aligned}&\mathbb{P}\left(a_n\left[\max_{\mathbf{x}\in\mathcal{I}}\left|\frac{\hat{g}(\mathbf{x}) - g(\mathbf{x})}{\hat{s}(\mathbf{x})}\right| - a_n\right] < t\right) = \\ &\exp\left(-2e^{-t-t^2/2a_n^2}\right)\sum_{m=0}^{[(d-1)/2]} h_{m,d-1}a_n^{-2m}\left(1 + \frac{t}{a_n^2}\right)^{d-2m-1} + O(n^{-\kappa}),\end{aligned} \tag{4.4}$$

as $n \to \infty$, where $h_{m,d-1} \equiv \frac{(-1)^m(d-1)!}{m!2^m(d-2m-1)!}$ and and $[\cdot]$ is the integer part of a number.

Notice that the approximation error is of a polynomial rate. As a result, a critical value based on the leading term of the right-hand side of (4.4) provides a better approximation than one based on the Gumbel approximation. The result in Theorem 2 may be of independent interest for constructing the uniform confidence band in nonparametric regression beyond the scope of estimating the CATEF in our context.

Remark 8. In a setting different from here, Lee, Linton, and Whang (2009) propose analytic critical values based on Piterbarg (1996) in order to test for stochastic monotonicity, compare its performance with the bootstrap critical values in their Monte Carlo experiments, and find that both perform well in finite samples. However, the discussions in Lee, Linton, and Whang (2009) are informal and rely on the results of Monte Carlo experiments without the formal proof of establishing the polynomial approximation error.

The conclusion of the theorem can be simplified for special cases. In particular, if $d = 1$, then:

$$\mathbb{P}\left(a_n\left[\max_{\mathbf{x}\in\mathcal{I}}\left|\frac{\hat{g}(\mathbf{x})-g(\mathbf{x})}{\hat{s}(\mathbf{x})}\right|-a_n\right]<t\right)=\exp\left(-2e^{-t-t^2/2a_n^2}\right)+O(n^{-\kappa}),$$

where a_n is the largest solution to $\text{mes}(\mathcal{I}){h_n}^{-1}\lambda^{1/2}(2\pi)^{-1}\exp(-a_n^2/2)=1$. Also, if $d = 2$, then:

$$\mathbb{P}\left(a_n\left[\max_{\mathbf{x}\in\mathcal{I}}\left|\frac{\hat{g}(\mathbf{x})-g(\mathbf{x})}{\hat{s}(\mathbf{x})}\right|-a_n\right]<t\right)=\exp\left(-2e^{-t-t^2/2a_n^2}\right)\left(1+\frac{t}{a_n^2}\right)+O(n^{-\kappa}),$$

where a_n is the largest solution to $\text{mes}(\mathcal{I}){h_n}^{-2}\lambda^2(2\pi)^{-3/2}a_n\exp(-a_n^2/2)=1$.

4.1. **Construction of critical values.** We use the leading term on the right-hand side of (4.4) as a distribution-like function to construct a uniform confidence band. For example, if $d = 1$, we may construct a critical value $c(1-\alpha)$ that satisfies:

$$F_{n,1}(c)\geq 1-\alpha,$$

where $F_{n,1}(t)\equiv\exp\left(-2e^{-t-t^2/2a_n^2}\right)$. This yields the critical value presented in the Algorithm of Section 3. Similarly, if $d = 2$, we can use:

$$F_{n,2}(c)\geq 1-\alpha,$$

where $F_{n,2}(t)\equiv\exp\left(-2e^{-t-t^2/2a_n^2}\right)\left(1+\frac{t}{a_n^2}\right)$. In finite samples, it might be useful to impose monotonicity of $F_{n,j}(\cdot)$ by rearrangement (see, e.g., Chernozhukov, Ferndádez-Val, and Galichon (2009)).

Remark 9. Theorem 2 implies that:

$$\lim_{n\to\infty}\mathbb{P}\left(a_n\left[\max_{\mathbf{x}\in\mathcal{I}}\left|\frac{\hat{g}(\mathbf{x})-g(\mathbf{x})}{\hat{s}(\mathbf{x})}\right|-a_n\right]<t\right)=\exp\left(-2e^{-t}\right).$$

Thus, one may construct analytical critical values based on the Gumbel distribution. However, this approximation is accurate only up to the logarithmic rate in view of Theorem 2.

5. Monte Carlo Experiments

In this section, we present the results of Monte Carlo experiments. These experiments are conducted to see the finite sample performances of the proposed doubly robust estimator and the proposed uniform confidence band. The simulations are conducted by R 3.1.2 with Windows 7. The number of replications is 5000.

5.1. **Data generating process.** The data generating process follows the potential outcome framework, and we borrow the design used by Abrevaya, Hsu, and Lieli (2015). The notations for the variables are the same as those used in the theoretical part of the paper. Recall that p is the number of covariates (i.e., the dimension of $\mathbf{Z}$). We consider two cases: $p = 2$ and $p = 4$. We explain the data generating process for each value of p.

The data generating process for $p = 2$ is the following. The vector of covariates $\mathbf{Z} = (X_1, X_2)^\top$ is generated by:

$$X_1 = \epsilon_1, \quad X_2 = (1 + 2X_1)^2(-1 + X_1)^2 + \epsilon_2,$$

where $\epsilon_i \sim U[-0.5, 0.5]$ for $i = 1, 2$ and ϵ_1 and ϵ_2 are independent. The potential outcomes are generated by:

$$Y_1 = X_1 X_2 + v, \quad Y_0 = 0,$$

where $v \sim N(0, 0.25^2)$ and v is independent of (ϵ_1, ϵ_2). The treatment status D is generated by:

$$D = \mathbf{1}\{\Lambda(X_1 + X_2) > U\},$$

where $U \sim U[0,1]$, U is independent of $(\epsilon_1, \epsilon_2, v)$ and Λ is the logistic function. Thus, the propensity score is $\pi(\mathbf{Z}) = \Lambda(X_1 + X_2)$. The observed outcome is $Y = DY_1$.

Next, we present the data generating process for $p = 4$. The covariates $\mathbf{Z} = (X_1, X_2, X_3, X_4)^\top$ are generated by:

$$X_1 = \epsilon_1, \quad X_2 = 1 + 2X_1 + \epsilon_2, \quad X_3 = 1 + 2X_1 + \epsilon_3, \quad X_4 = (-1 + X_1)^2 + \epsilon_4,$$

where $\epsilon_i \sim U[-0.5, 0.5]$ for $i = 1, 2, 3, 4$ and ϵ_1, ϵ_2, ϵ_3 and ϵ_4 are independent. The potential outcomes are generated by:

$$Y_1 = X_1 X_2 X_3 X_4 + v, \quad Y_0 = 0,$$

where $v \sim N(0, 0.25^2)$ and v is independent of $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$. The treatment status D is generated by:

$$D = \mathbf{1}\{\Lambda(0.5(X_1 + X_2 + X_3 + X_4)) > U\},$$

where $U \sim U[0,1]$ and U is independent of $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 v)$. The propensity score is $\pi(\mathbf{Z}) = \Lambda(0.5(X_1 + X_2 + X_3 + X_4))$. The observed outcome is $Y = DY_1$.

The parameter of interest is the CATEF for $\mathbf{X} = X_1$. In our specification, the CATEF is the same in both $p = 2$ and $p = 4$ cases. The CATEF can be written as:

$$CATEF(x_1) = x_1(1 + 2x_1)^2(-1 + x_1)^2.$$

We examine the performance of various statistical procedures regarding this CATEF.

5.2. **Model specification.** To estimate and conduct statistical inferences on $CATEF(x_1)$ using our doubly robust procedure, we need to specify a model for the regression $\mu_j(\mathbf{z})$ for $j = 0, 1$ and a model for the propensity score $\pi(\mathbf{z})$. We consider two regression models and two propensity score models. One of two models is correctly specified, but the other model is misspecified. We note that our doubly robust procedure is

predicted to work well provided that at least one of the regression models and the propensity score models is correctly specified.

We first discuss the model specifications for $p = 2$. The first regression model is:

$$\mu_1(\mathbf{z}, \alpha_1) = \alpha_{10} + \alpha_{11} X_1 X_2, \quad \mu_0(\mathbf{z}, \alpha_0) = \alpha_{00} + \alpha_{01} X_1 X_2.$$

This model is correctly specified. The coefficients are estimated by OLS using $X_1 X_2$ as the explanatory variable. The second regression model, which is misspecified, is:

$$\mu_1(\mathbf{z}, \alpha_1) = \alpha_{10} + \alpha_{11} X_1 + \alpha_{12} X_2, \quad \mu_0(\mathbf{z}, \alpha_0) = \alpha_{00} + \alpha_{01} X_1 + \alpha_{02} X_2.$$

The model is estimated by OLS using X_1 and X_2 as explanatory variables.

We also consider two models for propensity score. When $p = 2$, the first model is correctly specified and it is:

$$\pi(\mathbf{z}, \beta) = \Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2).$$

The model is estimated by maximum likelihood. The misspecified model is:

$$\pi(\mathbf{z}, \beta) = \left(\Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2)\right)^2.$$

This misspecified propensity score is computed by taking the square of the estimated correctly specified propensity score.

The model specifications for $p = 4$ are similar to those for $p = 2$ but include more covariates. The first regression model (correctly specified) is:

$$\mu_1(\mathbf{z}, \alpha_1) = \alpha_{10} + \alpha_{11} X_1 X_2 X_3 X_4, \quad \mu_0(\mathbf{z}, \alpha_0) = \alpha_{00} + \alpha_{01} X_1 X_2 X_3 X_4.$$

The model is estimated by OLS using $X_1X_2X_3X_4$ as the explanatory variable. The misspecified version is:

$$\mu_1(\mathbf{z}, \alpha_1) = \alpha_{10} + \alpha_{11}X_1 + \alpha_{12}X_2 + \alpha_{13}X_3 + \alpha_{14}X_4,$$

$$\mu_0(\mathbf{z}, \alpha_0) = \alpha_{00} + \alpha_{01}X_1 + \alpha_{02}X_2 + \alpha_{03}X_3 + \alpha_{04}X_4.$$

The model is estimated by OLS using X_1, X_2, X_3 and X_4 as explanatory variables. For the propensity score, we first estimate the correctly specified logit model:

$$\pi(\mathbf{z}, \beta) = \Lambda(\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4).$$

We then compute a misspecified propensity score by taking the square of the estimated propensity score based on the correctly specified model.

We estimate $CATEF(x_1)$ for $x_1 \in \{-0.4, -0.2, 0, 0.2, 0.4\}$ and compute the mean bias ("MEAN"), standard deviation ("SD"), the average of standard error for $\widehat{CATEF}(x_1)$ ("ASE"), and the mean squared error times 100 ("MSE×100"). The local linear regression is conducted with the Gaussian kernel, and the preliminary bandwidth ($\hat{h}$ in Algorithm (1)) is chosen by the method of Ruppert, Sheather, and Wand (1995). We also compute the "BIAS", "SE" and "MSE×100" of the corresponding inverse probability weighting estimators and the regression adjustment estimators. Note that the difference between the proposed method and those alternative methods arises only in the estimation of $\psi(\mathbf{W}, \theta_0)$ and the other steps are the same.

We examine the coverage probability of the uniform confidence band for $CATEF(x_1)$ for the range $-0.4 \leq x_1 \leq 0.4$. The nominal coverage probabilities that we consider are 99%, 95% and 90%. We compute the empirical coverage ("CP"), the mean critical value ("Mcri"), and the standard deviation of critical value ("Sdcri"). We also compute the coverage probabilities of the confidence band when the critical values are computed by the Gumbel approximation ("GCP").

5.3. **Results.** Tables 1 and 2 summarize the results on the properties of the estimators. The proposed doubly robust estimator of the CATEF works well in finite samples. As the theory indicates, the proposed estimator exhibits small bias provided that at least one of the regression models and the propensity score models is correctly specified. Its standard deviation is typically smaller than that of the inverse probability weighting estimator, and when it is not, the difference is small. We find that the regression adjustment estimator is very precise when the regression model is correctly specified. However, it is markedly vulnerable against model misspecification. The inverse probability weighting estimator also suffers from model misspecification, although it is arguably more robust than the regression adjustment estimator. When both models are misspecified, the doubly robust estimator is biased; however, its mean squared error is typically smaller than those of the alternatives, although we notice that at some points ($x = 0$ for the inverse probability weighting estimator and $x = 0.2$ for the regression adjustment estimator), the alternative estimators happen to have small biases. All the estimators have large mean squared errors when $x = -0.4$. This result is a consequence of the well-known result in nonparametric statistics that it is difficult to estimate the value of a function near the end of the support. The standard error for the proposed doubly robust estimator is, on average, close to the standard deviation.

Tables 3 and 4 summarize the finite sample properties of the proposed uniform confidence band. The results show that our uniform confidence band has a good coverage property provided that one of the models is correctly specified. When both models are misspecified, $p = 4$ and $N = 1000$, the size distortion is heavy. The average values of the 95% critical values are around 3. Because the pointwise critical value is 1.96 and is much smaller than the uniformly valid critical value, it demonstrates the importance of the uniform property of confidence band. The standard deviations of

the critical values are small because they change only if the bandwidth changes. The confidence band based on the Gumbel approximation is very conservative.

The results of the Monte Carlo simulation confirm that the proposed doubly robust estimator indeed works well in finite samples provided that one of the regression and propensity score models is correctly specified. The proposed uniform confidence band also has good coverage properties.

6. An Empirical Application

We apply our uniformly valid confidence band for the CATEF for the effect of maternal smoking on birth weight where the argument of the CATEF is the mother's age. Our aim here is to illustrate our confidence band in comparison with alternative confidence bands. We first discuss the background of this application and the dataset used. We then compute various confidence bands for the CATEF and discuss the results.

While the purpose of this application is to illustrate our uniformly valid confidence band and not to present new insights on the effect of smoking, it is still informative to discuss the background of this application. Many studies document that low birth weight is associated with prolonged negative effects on health and educational or labor market outcomes throughout life, although there has been a debate over its magnitude. See, e.g., Almond and Currie (2011) for a review. Maternal smoking is considered to be the most important preventable negative cause of low birth weight (Kramer, 1987). There are many studies that evaluate the effect of maternal smoking on low birth weight (Almond and Currie, 2011). The program evaluation approach is employed by, for example, Almond, Chay, and Lee (2005), da Veiga and Wilder (2008) and Walker, Tekin, and Wallace (2009), and panel data analysis is carried out by Abrevaya (2006) and Abrevaya and Dahl (2008). Here, we are interested in how the effect of smoking changes across different age groups of mothers. Walker, Tekin,

and Wallace (2009) examine whether the effect of smoking is larger for teen mothers than for adult mothers and find mixed evidence. Abrevaya, Hsu, and Lieli (2015) also consider this problem in their application.

Our dataset consists of observations from white mothers in Pennsylvania in the USA. The dataset is an excerpt from Cattaneo (2010) and is obtained from the STATA website ("`http://www.stata-press.com/data/r13/cattaneo2.dta`"). Note that the dataset was originally used in Almond, Chay, and Lee (2005). We restrict our sample to white and non-Hispanic mothers, and the sample size is 3754. The outcome of interest (Y) is infant birth weight measured in grams. The treatment variable (D) is a binary variable that is equal to 1 if the mother smokes and 0 otherwise. The set of covariates $\mathbf{Z}$ includes the mother's age, an indicator variable for alcohol consumption during pregnancy, an indicator for the first baby, the mother's educational attainment, an indicator for the first prenatal visit in the first trimester, the number of prenatal care visits, and an indicator for whether there was a previous birth where the newborn died. We are interested in how the effect of smoking varies across different values of the mother's age. Therefore, $\mathbf{X}$ is mother's age in this application.

To estimate the CATEF, we use linear regression models for the regression part and a logit model for propensity score. The explanatory variables used in the regression models and the logit models consist of all the elements of $\mathbf{Z}$, the square of the mother's age, and the interaction terms between the mother's age and all other elements of $\mathbf{Z}$. We estimate the CATEF in the interval between ages 15 and 35.

We compute the following three 95% confidence bands for the CATEF. "Our CB" is the confidence band proposed in this paper. Because $\mathbf{X}$ is univariate in this application, we follow the algorithm in Section 3. We use the Gaussian kernel. The preliminary bandwidth ($\hat{h}$) is chosen by the method of Ruppert, Sheather, and Wand (1995). "Gumbel CB" is the confidence band in which $c(1-\alpha)$ in the algorithm is replaced by that based on the Gumbel approximation (see Remark 1). "PW CB"

is a pointwise valid confidence band where we replace $c(1-\alpha)$ in the algorithm by the corresponding value from the standard normal distribution (i.e., 1.96). This provides a valid confidence interval for each point of the CATEF. However, its uniform coverage rate would be smaller than 95%.

Figure 1 plots the estimated CATEF and the three 95% confidence bands for the range between 15 and 35 years of age. The figure also contains the average treatment effect estimate (AIPW estimate) for a reference.

The widths of the three confidence bands are substantially different. The confidence band based on the Gumbel approximation provides the widest band and may not be very informative. The confidence band that is valid only in a pointwise sense gives the narrowest band. This band is not uniformly valid and so may provide misleading information about the CATEF. On the other hand, this provides valuable information if we are interested at a particular point of the CATEF. The confidence band we propose lay between "Gumbel CB" and "PW CB". While this band is wider than "PW CB", it is much narrower than "Gumbel CB" and is uniformly asymptotically valid. We see from this figure that our confidence band is informative while being uniformly valid.

The estimated CATEF is decreasing from 15 to around 25 years of age. It is rather stable for the range above 25 years of age. All confidence bands indicate that the CATEF is estimated imprecisely near the ends of the range. Nonetheless, the estimated CATEF indicates that smoking may not have a strong impact when the mother is young. The CATEF is estimated relatively precisely in the middle of the range. For the range between 20 and 30 years of age, even the band based on the Gumbel approximation, which is the widest, does not contain 0. This result provides robust evidence that smoking has a negative impact on birth weight at least for mothers who are 20 to 30 years old. In this particular dataset, the statistical evidence against a constant smoking effect is somewhat weak. The confidence band that is valid

only in a pointwise sense may provide an impression that the smoking effect depends on the mother's age. However, the uniformly valid confidence band that we propose marginally contains the straight line that is equal to the ATE estimate. This result illustrates that there is a caveat when we use pointwise confidence intervals, as well as the importance of using uniformly valid confidence bands.

7. Conclusion

In this paper, we propose a doubly robust method for estimating the CATEF. We consider the situation where a high-dimensional vector of covariates is needed for identifying the average treatment effect but the covariates of interest are of much lower dimension. Our proposed estimator is doubly robust and does not suffer from the curse of dimensionality. We propose a uniform confidence band that is easy to compute, and we illustrate its usefulness via Monte Carlo experiments and an application to the effects of smoking on birth weights.

There are a few topics to be explored in the future. First, it would be useful to consider the issue of asymptotic biases of the proposed estimator without relying on undersmoothing. For example, it might be possible to extend the approach of Hall and Horowitz (2013) that avoids undersmoothing for our purposes. Second, it would be an interesting exercise to develop a method for estimating the quantile treatment effects conditional on covariates. Third, it is possible to extend our approach to the local average treatment effect. As mentioned in the Introduction, Ogburn, Rotnitzky, and Robins (2015) consider conditioning on $\mathbf{Z}$ to achieve identification, but they estimate the local average treatment effect, say LATE($\mathbf{X}$), as a function of $\mathbf{X}$. However, their specification of LATE($\mathbf{X}$) is parametric. Our approach can be adapted to specify LATE($\mathbf{X}$) nonparametrically and to develop a corresponding uniform confidence band. Fourth, this paper does not cover marginal treatment effects that can be identified using the method of local instrumental variables developed by Heckman and

Vytlacil (1999, 2005). It would be interesting to develop a uniform confidence band for the marginal treatment effects.

Appendix A. Proofs

Proof of Lemma 1. Because $DY = DY_1$ and Y_1 and D are independent of each other conditional on $\mathbf{Z}$, write:

(A.1)
$$\mathbb{E}\left[\psi_1(\mathbf{W}, \alpha_{10}, \beta_0)|\mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[\frac{\mathbb{E}\left[D|\mathbf{Z}\right]\mathbb{E}\left[Y_1|\mathbf{Z}\right]}{\pi(\mathbf{Z}, \beta_0)} - \frac{\mathbb{E}\left[D|\mathbf{Z}\right] - \pi(\mathbf{Z}, \beta_0)}{\pi(\mathbf{Z}, \beta_0)}\mu_1(\mathbf{Z}, \alpha_{10})\middle|\mathbf{X} = \mathbf{x}\right].$$

Suppose that there exists β_0 such that $\mathbb{E}\left[D|\mathbf{Z}\right] = \pi(\mathbf{Z}, \beta_0)$ almost surely. Then the right-hand side of (A.1) reduces to:

$$\mathbb{E}\left[\mathbb{E}\left[Y_1|\mathbf{Z}\right]|\mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[Y_1|\mathbf{X} = \mathbf{x}\right].$$

Suppose now that there exists α_{10} such that $\mathbb{E}\left[Y_1|\mathbf{Z}\right] = \mu_1(\mathbf{Z}, \alpha_{10})$ almost surely. Then the right-hand side of (A.1) again reduces to:

$$\mathbb{E}\left[\mu_1(\mathbf{Z}, \alpha_{10})|\mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[Y_1|\mathbf{X} = \mathbf{x}\right].$$

Analogously, we have $\mathbb{E}\left[\psi_0(\mathbf{W}, \alpha_{00}, \beta_0)|\mathbf{X} = \mathbf{x}\right] = \mathbb{E}\left[Y_0|\mathbf{X} = \mathbf{x}\right]$. □

The remainder of the appendix gives the proof of Theorem 2. We first establish the linear expansion of the local linear estimator.

Lemma 3.

$$\sup_{\mathbf{x}\in\mathcal{I}}\sqrt{nh_n^d}\left|\frac{\hat{g}(\mathbf{x}) - g(\mathbf{x})}{\hat{s}(\mathbf{x})} - \frac{1}{nh_n^d s_n(\mathbf{x})}\sum_{i=1}^n \frac{U_i}{f_{\mathbf{X}}(\mathbf{x})}\mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right)\right| = O_p\left(n^{-c}\right)$$

for some positive constant $c > 0$.

Proof of Lemma 3. Let $\hat{\mathbf{\Psi}}$ and $\mathbf{\Psi_0}$ denote the n-dimensional vectors such that $\hat{\mathbf{\Psi}} = \{\psi(\mathbf{W}_i, \hat{\theta})\}_{i=1}^n$ and $\mathbf{\Psi_0} = \{\psi(\mathbf{W}_i, \theta_0)\}_{i=1}^n$, respectively. Let $\mathbf{\Gamma}(\boldsymbol{x})$ be the $n \times (d+1)$ matrix whose i-th row is $[1, (\mathbf{X}_i - \mathbf{x})^\top]$, $\mathbf{\Omega}(\boldsymbol{x})$ the n-dimensional diagonal matrix whose i-th element is $h_n^{-1}\mathbf{K}\left[(\mathbf{X}_i - \mathbf{x})/h_n\right]$, $\boldsymbol{G} := [g(\boldsymbol{X}_i)]_{i=1}^n$ the n-dimensional vector of regression functions evaluated at data points, and $\boldsymbol{U} := (U_i)_{i=1}^n$ the n-dimensional vector of regression errors. Let $\mathbf{e}_1$ denote the $(d+1) \times 1$ vector whose first element is one and all others are zeros. Write

$$\hat{g}(\boldsymbol{x}) - g(\boldsymbol{x}) = T_{n1}(\boldsymbol{x}) + T_{n2}(\boldsymbol{x}) + R_{n1}(\boldsymbol{x}),$$

where

$$\begin{aligned}
T_{n1}(\boldsymbol{x}) &= \mathbf{e}_1^\top \left[\mathbf{\Gamma}(\boldsymbol{x})^\top \mathbf{\Omega}(\boldsymbol{x}) \mathbf{\Gamma}(\boldsymbol{x})\right]^{-1} \mathbf{\Gamma}(\boldsymbol{x})^\top \mathbf{\Omega}(\boldsymbol{x}) \boldsymbol{U}, \\
T_{n2}(\boldsymbol{x}) &= \mathbf{e}_1^\top \left[\mathbf{\Gamma}(\boldsymbol{x})^\top \mathbf{\Omega}(\boldsymbol{x}) \mathbf{\Gamma}(\boldsymbol{x})\right]^{-1} \mathbf{\Gamma}(\boldsymbol{x})^\top \mathbf{\Omega}(\boldsymbol{x}) \boldsymbol{G}, \\
R_n(\boldsymbol{x}) &= \mathbf{e}_1^\top \left[\mathbf{\Gamma}(\boldsymbol{x})^\top \mathbf{\Omega}(\boldsymbol{x}) \mathbf{\Gamma}(\boldsymbol{x})\right]^{-1} \mathbf{\Gamma}(\boldsymbol{x})^\top \mathbf{\Omega}(\boldsymbol{x}) \left(\hat{\mathbf{\Psi}} - \mathbf{\Psi_0}\right).
\end{aligned}$$

We first consider the leading stochastic term $T_{n1}(\boldsymbol{x})$. As in equation (2.10) of Ruppert and Wand (1994), we have that

(A.2)

$$\begin{aligned}
&n^{-1}\mathbf{\Gamma}(\boldsymbol{x})^\top \mathbf{\Omega}(\boldsymbol{x}) \mathbf{\Gamma}(\boldsymbol{x}) \\
&= \begin{pmatrix}
\frac{1}{nh_n^d} \sum_{i=1}^n \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right) & \frac{1}{nh_n^d} \sum_{i=1}^n \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right) (\mathbf{X}_i - \mathbf{x})^\top \\
\frac{1}{nh_n^d} \sum_{i=1}^n \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right) (\mathbf{X}_i - \mathbf{x}) & \frac{1}{nh_n^d} \sum_{i=1}^n \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right) (\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^\top
\end{pmatrix}.
\end{aligned}$$

By the standard empirical process results (see e.g., Pollard (1984) or van der Vaart and Wellner (1996)) combined with the usual change of variables, we have that uniformly

in $\mathbf{x} \in \mathcal{I}$,

$$\frac{1}{nh_n^d}\sum_{i=1}^n \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right) = f_{\mathbf{X}}(\mathbf{x}) + o(h_n) + O_p\left[\left(\frac{\log n}{nh_n^d}\right)^{1/2}\right],$$

$$\frac{1}{nh_n^d}\sum_{i=1}^n \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right)(\mathbf{X}_i - \mathbf{x}) = h_n^2 \frac{\partial f_{\mathbf{X}}(\mathbf{x})}{\partial \mathbf{x}}\mu_2(K) + o(h_n^2) + O_p\left[h_n\left(\frac{\log n}{nh_n^d}\right)^{1/2}\right],$$

$$\frac{1}{nh_n^d}\sum_{i=1}^n \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right)(\mathbf{X}_i - \mathbf{x})(\mathbf{X}_i - \mathbf{x})^\top = h_n^2 f_{\mathbf{X}}(\mathbf{x})\mu_2(K) + o(h_n^2) + O_p\left[h_n^2\left(\frac{\log n}{nh_n^d}\right)^{1/2}\right],$$

where $\mu_2(K) := \int u^2 K(u)du$.

Throughout the remainder of the proof, we let $r_n(\mathbf{x}) = O_p(n^{-c})$, uniformly in $\boldsymbol{x}$, be a sequence that can be different in different places for some constant $c > 0$. Then as in (2.11) of Ruppert and Wand (1994), we have that

(A.3)

$$\left[n^{-1}\boldsymbol{\Gamma}(\boldsymbol{x})^\top \boldsymbol{\Omega}(\boldsymbol{x})\boldsymbol{\Gamma}(\boldsymbol{x})\right]^{-1}$$

$$= \begin{pmatrix} f_{\mathbf{X}}(\mathbf{x})^{-1}[1 + r_n(\mathbf{x})] & -[\partial f_{\mathbf{X}}(\mathbf{x})/\partial \mathbf{x}]^\top f_{\mathbf{X}}(\mathbf{x})^{-2}[1 + r_n(\mathbf{x})] \\ -[\partial f_{\mathbf{X}}(\mathbf{x})/\partial \mathbf{x}] f_{\mathbf{X}}(\mathbf{x})^{-2}[1 + r_n(\mathbf{x})] & \left[\mu_2(K) f_{\mathbf{X}}(\mathbf{x}) h_n^2 \mathbf{I}_d\right]^{-1}[1 + r_n(\mathbf{x})] \end{pmatrix},$$

where $\mathbf{I}_d$ is the d-dimensional identity matrix. The little $o_p(\cdot)$ terms in equation (2.11) of Ruppert and Wand (1994) are pointwise; however, (A.3) holds uniformly in $\mathbf{x} \in \mathcal{I}$ with polynomially decaying terms $r_n(\mathbf{x})$ under our assumptions.

Let $\boldsymbol{\Gamma}_i(\boldsymbol{x}) := [1, (\mathbf{X}_i - \mathbf{x})^\top]^\top$. Since

$$n^{-1}\boldsymbol{\Gamma}(\boldsymbol{x})^\top \boldsymbol{\Omega}(\boldsymbol{x})\boldsymbol{U} = \frac{1}{nh_n^d}\sum_{i=1}^n U_i \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right)\boldsymbol{\Gamma}_i(\boldsymbol{x}),$$

we have by (A.3) that $T_{n1}(\boldsymbol{x}) = T_{n11}(\boldsymbol{x}) + T_{n12}(\boldsymbol{x})$, where

$$T_{n11}(\boldsymbol{x}) = \frac{1}{nh_n^d f_{\mathbf{X}}(\mathbf{x})}\sum_{i=1}^n U_i \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n}\right)[1 + r_n(\mathbf{x})],$$

$$T_{n12}(\boldsymbol{x}) = -\frac{1}{nh_n^d[f_{\mathbf{X}}(\mathbf{x})]^2}\sum_{i=1}^n U_i\mathbf{K}\left(\frac{\mathbf{X}_i-\mathbf{x}}{h_n}\right)\left[\frac{\partial f_{\mathbf{X}}(\mathbf{x})}{\partial \mathbf{x}}\right]^\top(\mathbf{X}_i-\mathbf{x})[1+r_n(\mathbf{x})].$$

Again using the standard empirical process result and the method of change of variables,

$$T_{n11}(\boldsymbol{x}) = O_p\left[\left(\frac{\log n}{nh_n^d}\right)^{1/2}\right] \text{ and } T_{n12}(\boldsymbol{x}) = O_p\left[h_n\left(\frac{\log n}{nh_n^d}\right)^{1/2}\right]$$

uniformly in $\mathbf{x}\in\mathcal{I}$. Therefore, we have shown that

$$T_{n1}(\boldsymbol{x}) = \frac{1}{nh_n^d f_{\mathbf{X}}(\mathbf{x})}\sum_{i=1}^n U_i\mathbf{K}\left(\frac{\mathbf{X}_i-\mathbf{x}}{h_n}\right)[1+r_n(\mathbf{x})]. \tag{A.4}$$

We now move on the other remainder terms. The proof of Theorem 2.1 (in particular, equation (2.3)) of Ruppert and Wand (1994) implies that $T_{n2}(\boldsymbol{x}) = O(h_n^2)$ uniformly in $\mathbf{x}\in\mathcal{I}$. The condition that $nh_n^{d+4}\to 0$ at a polynomial rate in n corresponds to the undersmoothing condition. It is straightforward to show that $(nh_n^d)^{1/2}R_n(\boldsymbol{x}) = O(n^{-c})$ uniformly in $\mathbf{x}$ for some constant $c>0$ due to Assumption 1(9) that

$$\max\left\{(nh_n^d)^{1/2}|\psi(\mathbf{W}_i,\hat{\theta})-\psi(\mathbf{W}_i,\theta_0)|i=1,\ldots,n\right\} = O_p(n^{-c})$$

for some constant $c>0$

Note that by conditions (7) and (8) of Assumption 1, we have that $\inf_{n\geq 1}\inf_{\mathbf{x}\in\mathcal{I}} s_n(\mathbf{x}) > 0$ and $\sup_{\mathbf{x}\in\mathcal{I}}|\hat{s}^2(\mathbf{x})-s_n^2(\mathbf{x})| = O_p(n^{-c})$. Hence, the lemma follows from (A.4) immediately. □

Define:

$$T_n(\mathbf{x}) \equiv \frac{1}{nh_n^d}\sum_{i=1}^n U_i\mathbf{K}\left(\frac{\mathbf{X}_i-\mathbf{x}}{h_n}\right) \text{ and } c_n(\mathbf{x}) \equiv \left\{\frac{1}{h_n^d}\mathbb{E}\left[U^2\mathbf{K}^2\left(\frac{\mathbf{X}-\mathbf{x}}{h_n}\right)\right]\right\}^{-1/2}.$$

Note that $c_n(\mathbf{x}) = [f_{\mathbf{X}}(\mathbf{x})s_n(\mathbf{x})]^{-1}$. By Lemma 3:

$$\max_{\mathbf{x}\in\mathcal{I}} \sqrt{nh_n^d} \left| \frac{\hat{g}(\mathbf{x}) - g(\mathbf{x})}{\hat{s}(\mathbf{x})} - c_n(\mathbf{x})T_n(\mathbf{x}) \right| = O_p\left(n^{-c}\right).$$

We now use the result of Chernozhukov, Chetverikov, and Kato (2014) to obtain Gaussian approximations. Define:

$$W_n \equiv \sup_{\mathbf{x}\in\mathcal{I}} c_n(\mathbf{x})\sqrt{nh_n^d}\left[T_n(\mathbf{x}) - \mathbb{E}T_n(\mathbf{x})\right]. \tag{A.5}$$

Chernozhukov, Chetverikov, and Kato (2014) established an approximation of W_n by a sequence of suprema of Gaussian processes. For each $n \geq 1$, let $\tilde{B}_{n,1}$ be a centered Gaussian process indexed by $\mathcal{I}$ with covariance function:

$$\begin{aligned} &\mathbb{E}[\tilde{B}_{n,1}(\mathbf{x})\tilde{B}_{n,1}(\mathbf{x}')] = \\ &h_n^{-d}c_n(\mathbf{x})c_n(\mathbf{x}')\mathrm{Cov}\left[U^2\mathbf{K}\left(\frac{\mathbf{X}-\mathbf{x}}{h_n}\right)\mathbf{K}\left(\frac{\mathbf{X}-\mathbf{x}'}{h_n}\right)\right]. \end{aligned} \tag{A.6}$$

Proposition 3.2 of Chernozhukov, Chetverikov, and Kato (2014) establishes the following approximation result.

Lemma 4. *Let Assumption 1 hold. Then for every $n \geq 1$, there is a tight Gaussian random variable $\tilde{B}_{n,1}$ in $\ell^\infty(\mathcal{I})$ with mean zero and covariance function* (A.6)*, and there is a sequence $\tilde{W}_{n,1}$ of random variables such that $\tilde{W}_{n,1} =_d \sup_{\mathbf{x}\in\mathcal{I}} \tilde{B}_{n,1}(\mathbf{x})$ and as $n \to \infty$:*

$$|W_n - \tilde{W}_{n,1}| = O_{\mathbb{P}}\left\{(nh_n^d)^{-1/6}\log n + (nh_n^d)^{-1/4}\log^{5/4} n + (n^{1/2}h_n^d)^{-1/2}\log^{3/2} n\right\}.$$

Proof of Lemma 4. To apply Proposition 3.2 of Chernozhukov, Chetverikov, and Kato (2014), we first note that Assumption 1 implies that all the regularity conditions for Proposition 3.2 of Chernozhukov, Chetverikov, and Kato (2014) are satisfied. They are:

(1) $\sup_{\mathbf{x}\in\mathbb{R}^d}\mathbb{E}[U^4|\mathbf{X}=\mathbf{x}]<\infty$.

(2) $\mathbf{K}(\cdot)$ is a bounded and continuous kernel function on $\mathbb{R}^d$, and such that the class of functions $\mathbf{K}\equiv\{\mathbf{t}\mapsto\mathbf{K}(h\mathbf{t}+\mathbf{x}):h>0,\mathbf{x}\in\mathbb{R}^d\}$ is a VC type with envelope $\|\mathbf{K}\|_\infty$.

(3) The distribution of $\mathbf{X}$ has a bounded Lebesgue density $p(\cdot)$ on $\mathbb{R}^d$.

(4) $h_n\to 0$ and $\log(1/h_n)=O(\log n)$ as $n\to\infty$.

(5) $C_{\mathcal{I}}\equiv\sup_{n\geq 1}\sup_{\mathbf{x}\in\mathcal{I}}|c_n(\mathbf{x})|<\infty$. Moreover, for every fixed $n\geq 1$ and for every $\mathbf{x}_m\in\mathcal{I}\to\mathbf{x}\in\mathcal{I}$ pointwise, $c_n(\mathbf{x}_m)\to c_n(\mathbf{x})$.

Then the desired result is an immediate consequence of Proposition 3.2 of Chernozhukov, Chetverikov, and Kato (2014) with a singleton set $\mathcal{G}=\{U\}$ and with $q=4$ (using their notation) in verifying condition (B1)' of Chernozhukov, Chetverikov, and Kato (2014). □

We now show that the Gaussian field obtained in Lemma 4 can be further approximated by a stationary Gaussian field.

Lemma 5. *Let Assumption 1 hold. Then for every $n\geq 1$, there is a tight Gaussian random variable $\tilde{B}_{n,2}$ in $\ell^\infty(\mathcal{I}_n)$ with mean zero and covariance function:*

$$\mathbb{E}[\tilde{B}_{n,2}(\mathbf{s})\tilde{B}_{n,2}(\mathbf{s}')]=\rho_d(\mathbf{s}-\mathbf{s}')$$

for $\mathbf{s},\mathbf{s}'\in\mathcal{I}_n\equiv h_n^{-1}\mathcal{I}$, and there is a sequence of random variables such that $\tilde{W}_{n,2}=_d\sup_{\mathbf{x}\in\mathcal{I}}\tilde{B}_{n,2}(h_n^{-1}\mathbf{x})$ and as $n\to\infty$:

$$|\tilde{W}_{n,1}-\tilde{W}_{n,2}|=O_{\mathbb{P}}\left(h_n\sqrt{\log h_n^{-d}}\right).$$

Proof of Lemma 5. This lemma can be proved as in the proof of Lemma 3.4 of Ghosal, Sen, and van der Vaart (2000). Let:

$$\phi_{n,\mathbf{x}}(U_i,\mathbf{X}_i):=\left\{\mathbb{E}\left[U^2\mathbf{K}^2\left(\frac{\mathbf{X}-\mathbf{x}}{h_n}\right)\right]\right\}^{-1/2}U_i\mathbf{K}\left(\frac{\mathbf{X}_i-\mathbf{x}}{h_n}\right),$$

$$\varphi_{n,\mathbf{x}}(U_i, \mathbf{X}_i) := \left\{ h_n^d \mathbb{E}\left[U^2 | \mathbf{X}_i\right] f_{\mathbf{X}}(\mathbf{X}_i) \int \mathbf{K}^2(\mathbf{u}) d\mathbf{u} \right\}^{-1/2} U_i \mathbf{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_n} \right).$$

As in Remark 8.3 of Ghosal, Sen, and van der Vaart (2000) and in the proof of Lemma 3.4 of Ghosal, Sen, and van der Vaart (2000), we can regard Gaussian processes $\tilde{B}_{n,1}$ and $\tilde{B}_{n,2}$ as Brownian bridges $B_n(\phi_{n,\mathbf{x}})$ and $B_n(\varphi_{n,\mathbf{x}})$, respectively, in the sense that $\mathbb{E}B_n(g) = 0$ and $\mathbb{E}[B_n(g)B_n(g')] = \text{cov}(g, g')$ for $g = \phi_{n,\mathbf{x}}, g' = \phi_{n,\mathbf{x}'}$ or $g = \varphi_{n,\mathbf{x}}, g = \varphi_{n,\mathbf{x}'}$.

Define $\delta_n(\mathbf{x}) := B_n(\phi_{n,\mathbf{x}}) - B_n(\varphi_{n,\mathbf{x}})$. Note that $\delta_n(\mathbf{x})$ is also a mean zero Gaussian process with:

$$\mathbb{E}\left[\delta_n(\mathbf{x})\delta_n(\mathbf{x}')\right] = \mathbb{E}\left[\{\phi_{n,\mathbf{x}}(U, \mathbf{X}) - \varphi_{n,\mathbf{x}}(U, \mathbf{X})\}\{\phi_{n,\mathbf{x}'}(U, \mathbf{X}) - \varphi_{n,\mathbf{x}'}(U, \mathbf{X})\}\right].$$

Note that:

$$\begin{aligned}
\mathbb{E}\left[\{\delta_n(\mathbf{x})\}^2\right] = \int \Bigg(& \left\{ \int \mathbb{E}\left[U^2 | \mathbf{X} = \mathbf{x} + h\mathbf{u}\right] \mathbf{K}^2(\mathbf{u}) f_{\mathbf{X}}(\mathbf{x} + h\mathbf{u}) d\mathbf{u} \right\}^{-1/2} \\
& - \left\{ \mathbb{E}\left[U^2 | \mathbf{X} = \mathbf{x} + h\mathbf{t}\right] f_{\mathbf{X}}(\mathbf{x} + h\mathbf{t}) \int \mathbf{K}^2(\mathbf{u}) d\mathbf{u} \right\}^{-1/2} \Bigg)^2 \\
& \times \mathbb{E}\left[U^2 | \mathbf{X} = \mathbf{x} + h\mathbf{t}\right] \mathbf{K}^2(\mathbf{t}) f_{\mathbf{X}}(\mathbf{x} + h\mathbf{u}) d\mathbf{t} \\
& = O(h_n^2),
\end{aligned}$$

because $\mathbf{x} \mapsto \mathbb{E}\left[U^2 | \mathbf{X} = \mathbf{x}\right] f_{\mathbf{X}}(\mathbf{x})$ is Lipschitz continuous. Thus, the L_2-diameter of $\delta_n(\cdot)$ is $O(h_n)$. In addition, we can show that there exists a constant $C > 0$ such that:

$$\mathbb{E}\left[\{\delta_n(\mathbf{x}) - \delta_n(\mathbf{x}')\}^2\right] \leq Ch^{-2} \left\|\mathbf{x} - \mathbf{x}'\right\|^2.$$

Then arguments similar to those used in the proof of Lemma 3.4 of Ghosal, Sen, and van der Vaart (2000) yield the desired result.

□

Proof of Theorem 2. First note that $a_n = O(\sqrt{\log n}\,)$ because $h_n = Cn^{-\eta}$. Lemmas 4 and 5 together imply that:

$$\max_{\mathbf{x}\in\mathcal{I}} \left| \frac{\hat{g}(\mathbf{x}) - g(\mathbf{x})}{\hat{s}(\mathbf{x})} - \tilde{B}_{n,2}(h_n^{-1}\mathbf{x}) \right| = o_p\left(a_n\right).$$

Note that $\tilde{B}_{n,2}$, defined in Theorem 5, is a homogeneous Gaussian field with zero mean and the covariance function $\rho_d(\mathbf{s})$. Because of the assumption on $K(\cdot)$, the covariance function $\rho_d(\mathbf{s})$ has finite support and is six times differentiable. The latter property implies that the Gaussian process $\tilde{B}_{n,2}$ is three times differentiable in the mean square sense (see, e.g., Chapter 4 of Rasmussen and Williams (2006)). Then by Theorem 14.3 of Piterbarg (1996) and also by Theorem 3.2 of Konakov and Piterbarg (1984), there exists $\kappa > 0$ such that uniformly in t, on any finite interval:

$$\mathbb{P}\left(a_n \left[\max_{\mathbf{x}\in\mathcal{I}} \left| \tilde{B}_{n,2}(h_n^{-1}\mathbf{x}) \right| - a_n \right] < t \right) =$$
$$\exp\left(-2e^{-t-t^2/2a_n^2}\right) \sum_{m=0}^{[(d-1)/2]} h_{m,d-1} a_n^{-2m} \left(1 + \frac{t}{a_n^2}\right)^{d-2m-1} + O(n^{-\kappa})$$

as $n \to \infty$, where a_n is obtained as the largest solution to the equation:

$$\frac{\mathrm{mes}(\mathcal{I}) {h_n}^{-d} \sqrt{\det\Lambda_2}}{(2\pi)^{(d+1)/2}} a_n^{d-1} e^{-a_n^2/2} = 1,$$

Λ_2 is the covariance matrix of the vector of the first derivative of the Gaussian field $\tilde{B}_{n,2}$:

$$\Lambda_2 \equiv \mathrm{cov\ grad}\ \tilde{B}_{n,2}(\mathbf{t}) = \left(-\frac{\partial^2 r(0)}{\partial t_i \partial t_j}, i, j = 1, \ldots, d \right),$$

and $[\cdot]$ is the integer part of a number. Simple calculation yields $\sqrt{\det\Lambda_2} = \lambda^{d/2}$ with λ defined in (4.3). □

References

Abadie, A. (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113(2), 231–263.

Abrevaya, J. (2006): "Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach," *Journal of Applied Econometrics*, 21, 489–519.

Abrevaya, J., and C. M. Dahl (2008): "The effects of birth inputs on birthweight: evidence from quantile estimation on panel data," *Journal of Business and Economic Statistics*, 26, 379–397.

Abrevaya, J., Y.-C. Hsu, and R. P. Lieli (2015): "Estimating Conditional Average Treatment Effects," *Journal of Business and Economic Statistics*, 33(4), 485–505.

Almond, D., K. Y. Chay, and D. S. Lee (2005): "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, 120, 1031–1083.

Almond, D., and J. Currie (2011): "Human Capital Development before Age Five," in *Handbook of Labor Economics*, ed. by D. Card, and O. Ashenfelter, vol. 4b, chap. 15, pp. 1315–1486. Elsevier.

Cattaneo, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155, 138–154.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *Annals of Statistics*, 41(6), 2786–2819.

——— (2014): "Gaussian approximation of suprema of empirical processes," *Annals of Statistics*, 42(4), 1564–1597.

Chernozhukov, V., I. Ferndádez-Val, and A. Galichon (2009): "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika*, 96(3), 559–575.

Chernozhukov, V., S. Lee, and A. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

Claeskens, G., and I. v. Keilegom (2003): "Bootstrap confidence bands for regression curves and their derivatives," *Annals of Statistics*, 31(6), 1852–1884.

da Veiga, P. V., and R. P. Wilder (2008): "Maternal Smoking During Pregnancy and Birthweight: A Propensity Score Matching Approach," *Maternal and Child Health Journal*, 12(2), 194–203.

Derya Uysal, S. (2015): "Doubly Robust Estimation of Causal Effects with Multivalued Treatments: An Application to the Returns to Schooling," *Journal of Applied Econometrics*, 30(5), 763–786.

Ghosal, S., A. Sen, and A. W. van der Vaart (2000): "Testing Monotonicity of Regression," *Annals of Statistics*, 28, 1054–1082.

Glynn, A. N., and K. M. Quinn (2010): "An Introduction to the Augmented Inverse Propensity Weighted Estimator," *Political Analysis*, 18, 36–56.

Hall, P., and J. Horowitz (2013): "A simple bootstrap method for constructing nonparametric confidence bands for functions," *Annals of Statistics*, 41(4), 1892–1921.

Heckman, J. J., and E. J. Vytlacil (1999): "Local instrumental variable and latent variable models for identifying and bounding treatment effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2005): "Structural equations, treatment, effects and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

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

Kang, J. D. Y., and J. L. Schafer (2007): "Demystifying Double Robustness: A comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22(4), 523–539.

KONAKOV, V., AND V. PITERBARG (1984): "On the convergence rate of maximal deviation distribution for kernel regression estimates," *Journal of Multivariate Analysis*, 15(3), 279–294.

KRAMER, M. S. (1987): "Intrauterine Growth and Gestational Duration Determinants," *Pediatrics*, 80, 502–511.

LEE, S., O. LINTON, AND Y.-J. WHANG (2009): "Testing for Stochastic Monotonicity," *Econometrica*, 77(2), 585–602.

LUNCEFORD, J. K., AND M. DAVIDIAN (2004): "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in Medicine*, 23, 2937–2960.

OGBURN, E. L., A. ROTNITZKY, AND J. M. ROBINS (2015): "Doubly robust estimation of the local average treatment effect curve," *Journal of the Royal Statistical Society, Series B*, 77(Part 2), 373–396.

OKUI, R., D. S. SMALL, Z. TAN, AND J. M. ROBINS (2012): "Doubly Robust Instrumental Variable Regression," *Statistica Sinica*, 22, 173–205.

PITERBARG, V. I. (1996): *Asymptotic Methods in the Theory of Gaussian Processes and Fields*. American Mathematical Society, Providence, RI.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer, New York, NY.

POTERBA, J. M., S. F. VENTI, AND D. A. WISE (1995): "Do 401(k) contributions crowd out other personal saving?," *Journal of Public Economics*, 58(1), 1–32.

RASMUSSEN, C. E., AND C. K. I. WILLIAMS (2006): *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

ROBINS, J. M. (2000): "Marginal structural models versus structural nested models as tools for causal inference," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, ed. by M. E. Halloran, and D. Berry, vol. 116 of *The IMA Volumes in Mathematics and its Applications*, pp. 95–133. Springer, New York.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90(429), 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of Regression Coefficients when Some Regressors are not always oberved," *Journal of the American Statistical Association*, 89, 846–866.

Rothe, C., and S. Firpo (2014): "Semiparametric Two-Step Estimation Using Doubly Robust Moment Conditions," working paper.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995): "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, 90(432), 1257–1270.

Ruppert, D., and M. P. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics*, 22(3), 1346–1370.

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

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999): "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the Americal Statistical Association*, 94(448), 1096–1120.

Tan, Z. (2006): "Regression and Weighting Methods for Causal Inference Using Instrumental Variables," *Journal of the American Statistical Association*, 101(476), 1607–1618.

——— (2010): "Bounded, efficient and doubly robust estimation with inverse weighting," *Biometrika*, 97(3), 661–682.

van der Laan, M. J. (2006): "Statistical Inference for Variable Importance," *The International Journal of Biostatistics*, 2(1), Article 2.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY.

Walker, M., E. Tekin, and S. Wallace (2009): "Teen Smoking and Birth Outcomes," *Southern Economic journal*, 75, 892–907.

Wooldridge, J. M. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141, 1281–1301.

——— (2010): *Econometric analysis of cross section and panel data*. The MIT press, second edn.

TABLE 1. Monte Carlo results: CATEF estimates, $p = 2$

	DR				IPW			RA		
At $\mathbf{x} =$	BIAS	SD	ASE	MSE×100	BIAS	SD	MSE×100	BIAS	SD	MSE×100
$N = 500$										
True propensity score model, False regression model										
-0.4	-0.001	0.06	0.057	0.36	-0.001	0.055	0.3	-0.158	0.036	2.626
-0.2	0.005	0.042	0.04	0.182	0.004	0.043	0.183	0.058	0.023	0.394
0	0.004	0.037	0.036	0.138	0.004	0.036	0.129	0.097	0.017	0.978
0.2	0	0.035	0.034	0.124	0	0.037	0.14	-0.002	0.018	0.034
0.4	-0.004	0.038	0.036	0.149	-0.004	0.044	0.192	-0.053	0.026	0.343
False propensity score model, True regression model										
-0.4	0.002	0.134	0.117	1.784	0.006	0.147	2.171	-0.001	0.024	0.059
-0.2	0.005	0.069	0.07	0.484	-0.06	0.077	0.955	0.002	0.022	0.047
0	0.005	0.048	0.048	0.23	0.004	0.05	0.254	0.001	0.016	0.026
0.2	0	0.041	0.041	0.169	0.056	0.047	0.543	-0.001	0.017	0.03
0.4	-0.005	0.042	0.042	0.176	0.092	0.056	1.155	-0.002	0.029	0.085
True propensity score model, True regression model										
-0.4	0	0.052	0.048	0.268	-0.001	0.055	0.3	-0.001	0.024	0.059
-0.2	0.004	0.042	0.039	0.175	0.004	0.043	0.183	0.002	0.022	0.047
0	0.004	0.036	0.035	0.132	0.004	0.036	0.129	0.001	0.016	0.026
0.2	0	0.035	0.034	0.119	0	0.037	0.14	-0.001	0.017	0.03
0.4	-0.004	0.038	0.035	0.145	-0.004	0.044	0.192	-0.002	0.029	0.085
False propensity score model, False regression model										
-0.4	0.282	0.178	0.166	11.125	0.006	0.147	2.171	-0.158	0.036	2.626
-0.2	-0.023	0.075	0.076	0.615	-0.06	0.077	0.955	0.058	0.023	0.394
0	-0.034	0.052	0.052	0.387	0.004	0.05	0.254	0.097	0.017	0.978
0.2	-0.005	0.044	0.045	0.195	0.056	0.047	0.543	-0.002	0.018	0.034
0.4	-0.002	0.045	0.046	0.203	0.092	0.056	1.155	-0.053	0.026	0.343
$N = 1000$										
True propensity score model, False regression model										
-0.4	0	0.045	0.043	0.199	0	0.04	0.163	-0.156	0.025	2.51
-0.2	0.004	0.032	0.03	0.103	0.004	0.032	0.103	0.06	0.016	0.382
0	0.002	0.027	0.027	0.076	0.002	0.026	0.071	0.098	0.012	0.974
0.2	-0.001	0.026	0.026	0.068	-0.001	0.027	0.075	-0.002	0.013	0.016
0.4	-0.003	0.028	0.027	0.082	-0.003	0.032	0.101	-0.053	0.018	0.313
False propensity score model, True regression model										
-0.4	0.002	0.094	0.087	0.881	0.005	0.105	1.106	0	0.018	0.034
-0.2	0.005	0.051	0.052	0.265	-0.06	0.058	0.689	0.001	0.016	0.025
0	0.003	0.035	0.036	0.124	0.002	0.037	0.138	0.001	0.012	0.014
0.2	-0.001	0.03	0.031	0.092	0.055	0.034	0.425	0	0.012	0.015
0.4	-0.004	0.03	0.031	0.093	0.091	0.04	0.994	-0.001	0.02	0.041
True propensity score model, True regression model										
-0.4	0	0.039	0.036	0.148	0	0.04	0.163	0	0.018	0.034
-0.2	0.004	0.031	0.03	0.1	0.004	0.032	0.103	0.001	0.016	0.025
0	0.002	0.027	0.026	0.072	0.002	0.026	0.071	0.001	0.012	0.014
0.2	-0.001	0.026	0.026	0.067	-0.001	0.027	0.075	0	0.012	0.015
0.4	-0.003	0.028	0.027	0.077	-0.003	0.032	0.101	-0.001	0.02	0.041
False propensity score model, False regression model										
-0.4	0.276	0.127	0.123	9.214	0.005	0.105	1.106	-0.156	0.025	2.51
-0.2	-0.026	0.057	0.057	0.396	-0.06	0.058	0.689	0.06	0.016	0.382
0	-0.037	0.039	0.039	0.288	0.002	0.037	0.138	0.098	0.012	0.974
0.2	-0.006	0.033	0.034	0.114	0.055	0.034	0.425	-0.002	0.013	0.016
0.4	0	0.033	0.035	0.112	0.091	0.04	0.994	-0.053	0.018	0.313

TABLE 2. Monte Carlo results: CATEF estimates, $p = 4$

	DR				IPW			RA		
At $\mathbf{x} =$	BIAS	SD	ASE	MSE×100	BIAS	SD	MSE×100	BIAS	SD	MSE×100
$N = 500$										
True propensity score model, False regression model										
-0.4	-0.001	0.042	0.04	0.175	-0.001	0.039	0.153	-0.098	0.03	1.055
-0.2	0.005	0.038	0.037	0.147	0.005	0.037	0.141	0.062	0.022	0.43
0	0.004	0.036	0.034	0.128	0.003	0.035	0.121	0.079	0.02	0.671
0.2	-0.002	0.038	0.037	0.144	-0.002	0.04	0.158	-0.017	0.025	0.089
0.4	-0.002	0.063	0.059	0.399	-0.001	0.065	0.429	-0.044	0.04	0.351
False propensity score model, True regression model										
-0.4	-0.001	0.052	0.05	0.271	-0.009	0.055	0.314	-0.001	0.019	0.037
-0.2	0.005	0.045	0.045	0.207	-0.02	0.049	0.28	0.003	0.019	0.037
0	0.003	0.041	0.04	0.17	0.003	0.043	0.185	0.003	0.014	0.02
0.2	-0.002	0.041	0.041	0.171	0.033	0.046	0.324	-0.001	0.024	0.056
0.4	-0.002	0.063	0.059	0.392	0.045	0.073	0.732	-0.001	0.06	0.365
True propensity score model, True regression model										
-0.4	-0.001	0.039	0.037	0.153	-0.001	0.039	0.153	-0.001	0.019	0.037
-0.2	0.005	0.037	0.035	0.138	0.005	0.037	0.141	0.003	0.019	0.037
0	0.003	0.035	0.033	0.122	0.003	0.035	0.121	0.003	0.014	0.02
0.2	-0.002	0.038	0.037	0.144	-0.002	0.04	0.158	-0.001	0.024	0.056
0.4	-0.002	0.062	0.058	0.387	-0.001	0.065	0.429	-0.001	0.06	0.365
False propensity score model, False regression model										
-0.4	0.051	0.057	0.057	0.582	-0.009	0.055	0.314	-0.098	0.03	1.055
-0.2	0	0.049	0.048	0.24	-0.02	0.049	0.28	0.062	0.022	0.43
0	-0.009	0.043	0.042	0.194	0.003	0.043	0.185	0.079	0.02	0.671
0.2	-0.001	0.042	0.042	0.178	0.033	0.046	0.324	-0.017	0.025	0.089
0.4	-0.003	0.066	0.064	0.432	0.045	0.073	0.732	-0.044	0.04	0.351
$N = 500$										
True propensity score model, False regression model										
-0.4	-0.002	0.03	0.03	0.092	-0.002	0.029	0.083	-0.098	0.021	1.015
-0.2	0.003	0.029	0.028	0.083	0.003	0.028	0.079	0.062	0.016	0.407
0	0.003	0.026	0.026	0.069	0.003	0.025	0.066	0.079	0.015	0.646
0.2	-0.001	0.028	0.028	0.078	-0.001	0.029	0.086	-0.017	0.018	0.062
0.4	-0.003	0.048	0.044	0.228	-0.002	0.05	0.248	-0.045	0.029	0.286
False propensity score model, True regression model										
-0.4	-0.002	0.038	0.037	0.143	-0.01	0.04	0.171	-0.001	0.014	0.019
-0.2	0.003	0.034	0.033	0.115	-0.023	0.037	0.186	0.002	0.014	0.02
0	0.003	0.03	0.03	0.093	0.003	0.031	0.1	0.002	0.01	0.01
0.2	-0.001	0.03	0.031	0.091	0.036	0.034	0.242	-0.001	0.017	0.03
0.4	-0.003	0.047	0.044	0.222	0.043	0.055	0.493	-0.002	0.046	0.21
True propensity score model, True regression model										
-0.4	-0.002	0.029	0.028	0.082	-0.002	0.029	0.083	-0.001	0.014	0.019
-0.2	0.003	0.027	0.026	0.077	0.003	0.028	0.079	0.002	0.014	0.02
0	0.003	0.026	0.025	0.066	0.003	0.025	0.066	0.002	0.01	0.01
0.2	-0.001	0.028	0.028	0.077	-0.001	0.029	0.086	-0.001	0.017	0.03
0.4	-0.003	0.047	0.043	0.222	-0.002	0.05	0.248	-0.002	0.046	0.21
False propensity score model, False regression model										
-0.4	0.049	0.041	0.043	0.412	-0.01	0.04	0.171	-0.098	0.021	1.015
-0.2	-0.003	0.036	0.036	0.134	-0.023	0.037	0.186	0.062	0.016	0.407
0	-0.01	0.032	0.031	0.11	0.003	0.031	0.1	0.079	0.015	0.646
0.2	0	0.031	0.032	0.095	0.036	0.034	0.242	-0.017	0.018	0.062
0.4	-0.004	0.05	0.048	0.248	0.043	0.055	0.493	-0.045	0.029	0.286

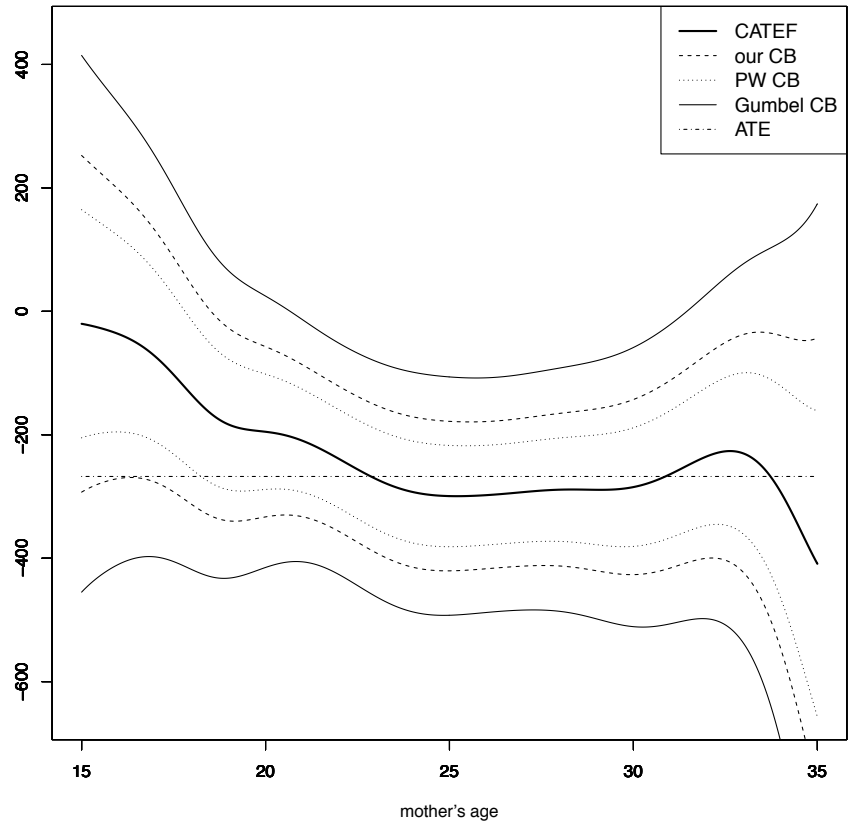
TABLE 3. Monte Carlo results, CATEF confidence band, $p = 2$

Confidence level	CP	Mcri	Sdcri	GCP
	$N = 500$			
True propensity score model, False regression model				
99%	0.978	3.485	0.078	0.999
95%	0.917	2.98	0.091	0.994
90%	0.855	2.728	0.099	0.977
False propensity score model, True regression model				
99%	0.99	3.468	0.106	1
95%	0.958	2.96	0.123	0.998
90%	0.912	2.705	0.134	0.99
True propensity score model, True regression model				
99%	0.981	3.483	0.072	1
95%	0.929	2.978	0.084	0.995
90%	0.864	2.726	0.092	0.98
False propensity score model, False regression model				
99%	0.983	3.505	0.098	1
95%	0.916	3.003	0.113	0.997
90%	0.823	2.753	0.123	0.982
	$N = 1000$			
True propensity score model, False regression model				
99%	0.982	3.525	0.067	0.999
95%	0.93	3.027	0.078	0.994
90%	0.875	2.779	0.085	0.979
False propensity score model, True regression model				
99%	0.993	3.5	0.093	1
95%	0.96	2.998	0.108	0.997
90%	0.919	2.747	0.117	0.994
True propensity score model, True regression model				
99%	0.986	3.524	0.061	1
95%	0.936	3.026	0.071	0.995
90%	0.881	2.778	0.077	0.983
False propensity score model, False regression model				
99%	0.948	3.545	0.088	1
95%	0.774	3.05	0.102	0.991
90%	0.614	2.804	0.111	0.947

TABLE 4. Monte Carlo results, CATEF confidence band, $p = 4$

Confidence level	CP	Mcri	Sdcri	GCP
	$N = 500$			
True propensity score model, False regression model				
99%	0.982	3.488	0.081	1
95%	0.926	2.984	0.095	0.994
90%	0.863	2.732	0.103	0.981
False propensity score model, True regression model				
99%	0.987	3.478	0.082	1
95%	0.941	2.972	0.095	0.996
90%	0.886	2.718	0.104	0.985
True propensity score model, True regression model				
99%	0.981	3.489	0.082	1
95%	0.924	2.985	0.095	0.996
90%	0.861	2.733	0.103	0.981
False propensity score model, False regression model				
99%	0.986	3.486	0.079	1
95%	0.938	2.982	0.092	0.996
90%	0.879	2.73	0.1	0.983
	$N = 1000$			
True propensity score model, False regression model				
99%	0.985	3.524	0.07	1
95%	0.931	3.026	0.081	0.996
90%	0.877	2.778	0.088	0.982
False propensity score model, True regression model				
99%	0.988	3.513	0.07	1
95%	0.945	3.013	0.081	0.997
90%	0.893	2.764	0.088	0.986
True propensity score model, True regression model				
99%	0.984	3.526	0.069	1
95%	0.933	3.029	0.08	0.995
90%	0.874	2.781	0.087	0.982
False propensity score model, False regression model				
99%	0.986	3.523	0.066	1
95%	0.923	3.025	0.077	0.996
90%	0.855	2.777	0.084	0.98

FIGURE 1. CATEF for the effect of smoking on birth weights, 95% confidence bands



Note: "CATEF" = the estimated CATEF; "our CB" = the uniformly valid confidence band proposed in this paper; "PW CB" = the confidence band that is valid only in a pointwise sense; "Gumbel CB" = the uniformly valid confidence band based on the Gumbel approximation; "ATE" = the estimated value of the average treatment effect.