Received: 6 January 2016 | Revised: 26 October 2016

DOI: 10.1002/jae.2574

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Doubly robust uniform confidence band for the conditional average treatment effect function

Sokbae Lee[1,2] | Ryo Okui[3,4] | Yoon-Jae Whang[5]

[1]Department of Economics, Columbia University, New York, NY, USA

[2]Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London, UK

[3]NYU Shanghai, 1555 Century Avenue, Pudong, Shanghai, 200122, China

[4]Department of Economics, University of Gothenburg, Gothenburg, Sweden

[5]Department of Economics, Seoul National University, Seoul, Republic of Korea

Correspondence
Sokbae Lee, Department of Economics, Columbia University, 1022 International Affairs Building, 420 West 118th Street, New York, NY 10027, USA.
Email: sl3841@columbia.edu

Funding information
European Research Council, Grant/Award Number: ERC-2009-StG-240910-ROMETA, ERC-2014-CoG-646917-ROMIA; Japan Society of the Promotion of Science, Grant/Award Number: KAKENHI 25285067, 25780151, 15H03329, 16K03598; SNU College of Social Science Research Grant

Summary

In this paper, we propose a doubly robust method to estimate 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.

1 | INTRODUCTION

In this paper, we propose a doubly robust method to estimate 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 $\{(Y_i, D_i, \mathbf{Z}_i) : i = 1, \dots, n\}$ 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:

$$g(\mathbf{x}) \equiv \mathbb{E}[Y_1 - Y_0 | \mathbf{X} = \mathbf{x}]. \tag{1}$$

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):

$$(Y_1, Y_0) \perp D | \mathbf{Z}, \tag{2}$$

where $\perp$ denotes the independence. For Equation 2 to be plausible in applications, applied researchers tend to consider a large number of covariates $\mathbf{Z}$. Note that in our set-up 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 et al. 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 et al. 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 Equation 1 with $\mathbf{X}$ being the age of mother; however, as noted in Abrevaya et al., it is unlikely that conditioning only on the age of the 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 Equation 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 models 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 nonparametric 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 et al. (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, for example, $\mathbb{E}[\mathbf{Z}|\mathbf{X}]$ 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 & Rotnitzky, 1995; Scharfstein, Rotnitzky, & Robins, 1999). Their estimator appears to be the first 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 & 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 (Okui, Small, Tan, & Robins, 2012; Tan, 2006) and estimation under multivalued treatments (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 the "*V*-adjusted variable importance" of van der Laan (2006), who proposes it as a measure of variable importance in prediction. van der Laan 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 by van der Laan.

In a recent paper, Abrevaya et al. (2015) consider the estimation of the CATEF;[1] however, there are two main differences of this paper relative to Abrevaya et al. First, we propose the doubly robust procedure to estimate the CATEF. Abrevaya et al. 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 et al. 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 proof of the main theorem is given in Section 8. Supporting Information appendices provide additional materials that are omitted from the main text.

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

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

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.[2] 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

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

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.

[1] Our paper is independent of Abrevaya et al. (2015) and it is started without knowing their work.
[2] $\mu_j(\mathbf{z},\alpha_j)$ may also be called "marginal structural models" of Robins (2000).

Lemma 1. *(Identification of the CATEF) Assume that Equation 2 holds and* $0 < \pi(\mathbf{Z}, \beta_0) < 1$ *almost surely. Suppose that either* β_0 *satisfies* $\mathbb{E}[D|\mathbf{Z}] = \pi(\mathbf{Z}, \beta_0)$ *almost surely or* α_{10} *and* α_{00} *satisfy* $\mathbb{E}[Y_1|\mathbf{Z}] = \mu_1(\mathbf{Z}, \alpha_{10})$ *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.

Remark 1. *In this paper, we focus on cases in which* $\boldsymbol{X}$ *is continuous. When* $\boldsymbol{X}$ *is discrete, the CATEF can be estimated by the sample average of* $\psi(\mathbf{W}, \hat{\theta})$ *using the subsample for each possible value of* $\boldsymbol{X}$ *and an estimator* $\hat{\theta}$ *of* θ_0*. Moreover, constructing a confidence band is standard when* $\boldsymbol{X}$ *takes a finite number of values.*

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:

$$\hat{\alpha}_j \equiv \arg\min_{\alpha_j} \sum_{i=1}^{n} D_i^j (1 - D_i)^{1-j} [Y_i - \mu_j(\mathbf{Z}_i, \alpha_j)]^2. \tag{3}$$

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

$$\hat{\beta} \equiv \arg\max_{\beta} \sum_{i=1}^{n} (D_i \log \pi(\mathbf{Z}_i, \beta) + (1 - D_i) \log(1 - \pi(\mathbf{Z}_i, \beta))). \tag{4}$$

Remark 2. *When the dimension of* $\boldsymbol{Z}$ *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 (2016). However, this would not work when the dimension of* $\boldsymbol{Z}$ *is sufficiently high (see related remarks in Rothe and Firpo). 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 $g(\mathbf{x})$. 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.

2.3 | Effect of first-stage estimation

In our setting, we can carry out inference as if θ_0 were known. This result would not be a surprise given that our first-stage estimation is parametric and our second-stage estimation is nonparametric: the rate of the convergence in the first-stage estimation is faster than that of the second stage. This feature of no first-order effect of the first-stage estimation in the second stage turns out to be more general than our set-up. It is indeed closely related to doubly robustness.

If we model $g(\mathbf{x})$ parametrically or more generally approximate $g(\mathbf{x})$ by linear projection, it can be estimated by running an ordinary least squares (OLS) of $\psi(\mathbf{W}, \hat{\theta})$ on $\mathbf{X}$. Because of the built-in feature of double robustness, it can be shown that the limiting distribution of the OLS estimator of $\psi(\mathbf{W}, \hat{\theta})$ on $\mathbf{X}$ is equivalent to that of the infeasible OLS estimator of $\psi(\mathbf{W}, \theta_0)$ on $\mathbf{X}$. Furthermore, even if we estimate $\pi(\cdot)$ and $\mu_j(\cdot)(j = 0, 1)$ nonparametrically when the dimension of $\mathbf{Z}$ is moderate, there will be no estimation effect from the first stage as well. For example, see Chen, Hong, and Tarozzi (2008), Rothe and Firpo (2016), and Chernozhukov, Escanciano, Ichimura, and Newey (2016), among others, for related results.

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.

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

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

where $\hat{h}$ is a commonly used optimal bandwidth in the literature (e.g., the plug-in method of Ruppert, Sheather, and Wand (1995), which is explained in Supporting Information Appendix A). We use the Gaussian kernel in our simulations and empirical application.

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:

$$\frac{\hat{s}(x)}{(nh_n)^{1/2}} \equiv \left\{ [nh_n \hat{f}_X(x)]^{-1} \int K^2(u)\mathrm{d}u \; \hat{\sigma}^2(x) \right\}^{1/2}, \tag{5}$$

where $\hat{f}_X$ is the kernel density estimator:

$$\hat{f}_X(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n} \right),$$

and $\hat{\sigma}^2(x)$ is the conditional variance function estimator:

$$\hat{\sigma}^2(x) = \frac{1}{(n - \dim(\theta))h_n} \sum_{i=1}^{N} \frac{\hat{\mathrm{U}}_i^2}{\hat{f}_X(x)} K\left(\frac{X_i - x}{h_n} \right). \tag{6}$$

Here, $\hat{\mathrm{U}}_i = \psi(\mathbf{W}_i, \hat{\theta}) - \hat{g}(X_i)$ and $\dim(\theta)$ is the dimension of θ.

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

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

Note that $\lambda = 0.5$ if $K(\cdot)$ is the Gaussian kernel.[3] 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 for the two-sided symmetric uniform confidence band 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:

[3]Note that $\lambda = 1.98$ for the biweight kernel and $\lambda = 2.5$ for the Epanechnikov kernel.

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

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 $\sigma^2(x) \equiv \text{var}[\psi(\mathbf{W}, \theta_0)|X = x]$ 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 3. *It is straightforward to modify the algorithm above to construct one-sided symmetric confidence bands. Define a new critical value by*

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

Then, for each $x \in \mathcal{I}$, we set the one-sided symmetric confidence bands:

$$\hat{g}(x) - c_{one\text{-}sided}(1-\alpha)\frac{\hat{s}(x)}{\sqrt{nh_n}} \leq g(x),$$

or

$$g(x) \leq \hat{g}(x) + c_{one\text{-}sided}(1-\alpha)\frac{\hat{s}(x)}{\sqrt{nh_n}}.$$

Remark 4. *When x is more than one dimension, the algorithm may be revised as follows. Obviously, we need to use multivariate kernels, and pointwise standard errors should be adjusted because the rate of convergence becomes nh_n^d. The value of λ stays the same when we use a product kernel. For example, if $\mathbf{K}$ is the product Gaussian kernel, then $\lambda = 0.5$. The formulas of a_n and $c(1-\alpha)$ need to be changed. a_n is the largest solution to the following equation:*

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

where $mes(\mathcal{I})$ is the Lebesgue measure of $\mathcal{I}$. When $d = 2$, the critical value has the form $c(1-\alpha) \equiv a_n + c/a_n$, where c is the smallest value that satisfies

$$\exp\left(-2e^{-c-c^2/2a_n^2}\right)\left(1 + \frac{c}{a_n^2}\right) \geq 1 - \alpha.$$

When $d = 3$, we have that $c(1-\alpha) \equiv a_n + c/a_n$, where c is the smallest value that satisfies

$$\exp\left(-2e^{-c-c^2/2a_n^2}\right)\left(\left(1 + \frac{c}{a_n^2}\right)^2 - 2\frac{1}{a_n^2}\right) \geq 1 - \alpha.$$

We note that, in this paper, we assume that $d < 4$ (see Assumption 1).

Remark 5. *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 a 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 6. *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 critical value is better asymptotically.*

Remark 7. *The proposed confidence band can be used to test whether the CATEF is constant. Suppose our null hypothesis is that $g(\boldsymbol{x})$ 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 for $g_{\mathcal{I}}$. We reject the constancy of $g(\boldsymbol{x})$, 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

$$\rho(s_j) \equiv \frac{\int K(u)K(u-s_j)du}{\int K^2(u)\mathrm{d}u}, \tag{7}$$

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 $\boldsymbol{X}$ has a bounded Lebesgue density $f_{\boldsymbol{X}}(\cdot)$ on $\mathbb{R}^d$. Furthermore, $f_{\boldsymbol{X}}(\cdot)$ 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 $[-1, 1]$, $\int_{-1}^1 uK(u)\mathrm{d}u = 0$, $\int_{-1}^1 K(u)\mathrm{d}u = 1$, 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 $s_n(\boldsymbol{x})$ is continuous for each $n \geq 1$. Furthermore, $\mathbf{x} \mapsto \mathbb{E}\left[U^2|\mathbf{X} = \mathbf{x}\right] f_{\mathbf{X}}(\mathbf{x})$ 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. Condition 2 of Assumption 1 rules out discrete covariates. If all regressors are discrete, then the estimation problem reduces to a parametric estimation problem. In this case, one may consider a multiple testing approach as in Lee and Shaikh (2014) by defining subpopulations with observed cells of discrete covariates. If some covariates are discrete and others are continuous, then one may use a smoothing approach proposed in Li and Racine (2004).

Condition 5 assumes that the kernel function has finite support. This assumption is for the simplicity of the paper and can be dropped at the expense of complicated proofs. It also assumes that the kernel function is differentiable. This assumption is

crucial and excludes, for example, the uniform kernel. One of the bandwidth conditions in h_n (i.e., $\eta > 1/(d+4)$ in condition 6) 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 8. *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 9. *An estimator of $\hat{s}^2(\mathbf{x})$ is readily available. For example, we may consider*

$$\hat{s}^2(\mathbf{x}) = \frac{\hat{\sigma}^2(\mathbf{x})}{\hat{f}_{\mathbf{X}}(\mathbf{x})} \int \mathbf{K}^2(\mathbf{u})\, d\mathbf{u}, \tag{8}$$

where $\hat{f}_{\mathbf{X}}(\cdot)$ is the kernel density estimator and $\hat{\sigma}^2(\mathbf{x})$ is a nonparametric estimator of $\sigma^2(\boldsymbol{x})$ using $\{(\hat{U}_i^2, \mathbf{X}_i) : i = 1, \ldots, n\}$ with $\hat{U}_i \equiv \psi(\mathbf{W}_i, \hat{\theta}) - \hat{g}(\mathbf{X}_i)$. Recall Equation 6 for its concrete form for the one-dimensional case. Alternatively, we may set

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

For either estimator, it is straightforward to verify condition 8 of Assumption 1 using the standard results in kernel estimation.

Remark 10. *Note that condition 9 of Assumption 1 is merely a sufficient (but not necessary) condition. This condition 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:

$$\text{mes}(\mathcal{I}) h_n{}^{-d} \lambda^{d/2} (2\pi)^{-(d+1)/2} a_n^{d-1} \exp(-a_n^2/2) = 1, \tag{9}$$

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

$$\lambda = \frac{-\int K(u)K''(u)du}{\int K^2(u)du}. \tag{10}$$

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:*

$$\mathbb{P}\left(a_n\left[\max_{\mathbf{x}\in\mathcal{I}}\left|\frac{\hat{g}(\mathbf{x}) - g(\mathbf{x})}{\hat{s}(\mathbf{x})}\right| - a_n\right] < t\right) = \exp\left(-2e^{-t-t^2/2a_n^2}\right) \sum_{m=0}^{\lfloor (d-1)/2 \rfloor} h_{m,d-1} a_n^{-2m} \left(1 + \frac{t}{a_n^2}\right)^{d-2m-1} + O(n^{-\kappa}), \tag{11}$$

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

Note 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 Equation 11 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 11. *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 et al. 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 $\text{mes}(\mathcal{I})h_n{}^{-2}\lambda^2(2\pi)^{-3/2}a_n\exp(-a_n^2/2)=1$.

Remark 12. *It is standard to obtain pointwise confidence intervals based on normal approximations. Recall that our two-sided symmetric uniform confidence band has the form*

$$\hat{g}(\mathbf{x})-c(1-\alpha)\frac{\hat{s}(\mathbf{x})}{\sqrt{nh_n^d}}\le g(x)\le\hat{g}(\mathbf{x})+c(1-\alpha)\frac{\hat{s}(\mathbf{x})}{\sqrt{nh_n^d}},\tag{12}$$

where $c(1-\alpha)$ is obtained from Theorem 2. To obtain two-sided symmetric pointwise confidence intervals, we just need to replace $c(1-\alpha)$ with the usual normal critical value $\Phi^{-1}(1-\alpha/2)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The pointwise confidence interval given in Equation 12 is different from the one resulting from Abrevaya et al. (2015, Theorem 2) in terms of the formula for $\hat{s}(\mathbf{x})$ in Equation 8. In their case, they need to estimate $\hat{\sigma}^2(\mathbf{x})$ using $\{(\tilde{U}_i^2,\mathbf{X}_i):i=1,\ldots,n\}$ with

$$\tilde{U}_i\equiv\frac{D_iY_i}{\pi(\mathbf{Z}_i,\hat{\beta})}-\frac{(1-D_i)Y_i}{1-\pi(\mathbf{Z}_i,\hat{\beta})}\hat{g}(\mathbf{X}_i).$$

Remark 13. *A one-sided version of the uniform confidence band is readily available. Combining Theorems 14.1 and 14.2 of Piterbarg (1996) with arguments identical to those used in the proof of Theorem 4.7 yields the following proposition. Under Assumption 1, there exists $\kappa>0$ such that, uniformly in t, on any finite interval:*

$$\mathbb{P}\left(a_n\left[\max_{\mathbf{x}\in\mathcal{I}}\frac{\hat{g}(\mathbf{x})-g(\mathbf{x})}{\hat{s}(\mathbf{x})}-a_n\right]<t\right)=\exp\left(-e^{-t-t^2/2a_n^2}\right)\sum_{m=0}^{\lfloor(d-1)/2\rfloor}h_{m,d-1}a_n^{-2m}\left(1+\frac{t}{a_n^2}\right)^{d-2m-1}+O(n^{-\kappa}),\tag{13}$$

as $n\to\infty$. Note that the only differences between Equations 11 and 13 are that (i) there is no absolute value on the left side of the equation in 13 and (ii) there is no factor 2 inside the exponential function in 13. Hence, for example, if $d = 1$, then

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

4.1 | Construction of critical values

We use the leading term on the right-hand side of Equation 11 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

$$F_{n,1}(c)\ge1-\alpha,$$

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

$$F_{n,2}(c)\ge1-\alpha,$$

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, Fernández-Val, & Galichon, 2009).

Remark 14. *Theorem 2 implies that*

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

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 performance of the proposed doubly robust estimator and the proposed uniform confidence band. The simulations are conducted using R 3.3.1 with Windows 10. The number of replications is 5,000.

5.1 | Data-generating process

The data-generating process follows the potential outcome framework. The notations for the variables are the same as those used in the theoretical part of the paper. We consider cases with $p = 10, 30$ and $N = 500, 2000$.

The data-generating process is the following. The vector of covariates $\mathbf{Z} = (X_1, \ldots, X_p)^\top$ is generated by $\mathbf{Z} \sim N(0, I_p)$, where I_p is the p-dimensional identity matrix. The potential outcomes are generated by

$$Y_1 = 10 + \sum_{k=1}^{p} \frac{1}{\sqrt{p}} X_k + v, \quad Y_0 = 0,$$

where $v \sim N(0, 1)$ and v is independent of $\mathbf{Z}$. The treatment status D is generated by

$$D = \mathbf{1}\left\{ \Lambda\left(\sum_{k=p/2}^{p} \frac{1}{\sqrt{p/2}} X_k \right) > U \right\},$$

where $U \sim U[0, 1]$, U is independent of $(\mathbf{Z}^\top, v)$ and Λ is the logistic function. Thus the propensity score is $\pi(\mathbf{Z}) = \Lambda\left(\sum_{k=p/2}^{p} X_k/\sqrt{p/2}\right)$. The observed outcome is $Y = DY_1$.

The parameter of interest is the CATEF for $\mathbf{X} = X_1$. In our specification, the CATEF can be written as

$$\text{CATEF}(x_1) = 10 + x_1/\sqrt{p}.$$

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 model and the propensity score model is correctly specified.

We first discuss the model specifications for the regression part. The first regression model is

$$\mu_1(\mathbf{z}, \alpha_1) = \alpha_{10} + \sum_{k=1}^{p} \alpha_{1k} X_k, \quad \mu_0(\mathbf{z}, \alpha_0) = \alpha_{00} + \sum_{k=1}^{p} \alpha_{0k} X_k.$$

This model is correctly specified. The coefficients are estimated by OLS using $(1, X_1, \ldots, X_p)$ as the explanatory variable. The second regression model, which is misspecified, is

$$\mu_1(\mathbf{z}, \alpha_1) = \alpha_{10} + \sum_{k=1}^{p/2} \alpha_{1k} X_k, \quad \mu_0(\mathbf{z}, \alpha_0) = \alpha_{00} + \sum_{k=1}^{p/2} \alpha_{0k} X_k.$$

The model is estimated by OLS using $(1, X_1, \ldots, X_{p/2})$ as explanatory variables. This model is misspecified because it suffers from sample selection bias introduced by omitting the second half of the regressors, which affects the treatment status.

We also consider two models for propensity score. The model for propensity score is

$$\pi(\mathbf{z}, \beta) = \Lambda\left(\beta_0 + \sum_{k=1}^{p} \beta_k X_k\right).$$

The misspecified model is

$$\pi(\mathbf{z}, \beta) = \Lambda\left(\beta_0 + \sum_{k=1}^{p/2} \beta_k X_k\right).$$

Similarly to the case of the regression part, misspecification is introduced by omitting the second half of the regressors. The models for propensity score are estimated by maximum likelihood.

We estimate CATEF(x_1) for $x_1 \in \{-1, -0.5, 0, 0.5, 1\}$ and compute the mean bias ("BIAS"), standard deviation ("SD"), the average of standard error for $\widehat{\text{CATEF}}(x_1)$ ("ASE"), and the root mean squared error ("RMSE"). 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 et al. (1995). We also compute the "BIAS", "SE" and "RMSE" 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 $-1 \leq x_1 \leq 1$. 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 based on the critical values computed by the Gumbel approximation ("GCP").

5.3 | Results

Tables 1 and 2 summarize the finite-sample properties of the proposed uniform confidence band. The results show that our uniform confidence band has a reasonably good coverage property provided that one of the models is correctly specified. When both models are misspecified, the size distortion is heavy. We observe that the size distortion is heavier when the regression model is misspecified than that in the case of propensity score misspecification. This result indicates that we should carefully model the regression part in order to obtain reliable confidence bands. The average values of the 95% critical values are around 2.75. Because the pointwise critical value is 1.96 and is much smaller than the uniformly valid critical value, this 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.

We now comment on other properties of the estimators. See Tables 3 and 4 in Supporting Information Appendix C for details. 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 model and the propensity score model is correctly specified. We find that the regression adjustment estimator is very precise when the regression model is correctly specified. However, it suffers from substantial bias when the regression model is misspecified. The inverse probability weighting estimator also suffers from model misspecification. Moreover, its standard deviation is much larger than those of the doubly robust and regression adjustment estimators. When both models are misspecified, all three estimators suffer from heavy bias. The inverse probability weighting estimator has the largest RMSE because its distribution is more diverse than those of the other two estimators. All the estimators have larger standard deviations when $x = 1$ and $x = -1$ compared to those in other points. This is because the number of observations around $x = 1$ or $x = -1$ is expected to be smaller than that around, for example, $x = 0$, which is the center of the distribution. On the other hand, the magnitude of the bias does not vary much across data points. The standard error for the proposed doubly robust estimator is slightly smaller than the standard deviation, but the difference is not large.

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.

TABLE 1 Monte Carlo results, CATEF confidence band, $p = 10$

Confidence level	CP	Mcri	Sdcri	GCP
	$N = 500$			
True propensity score model, true regression model				
99%	0.986	3.290	0.091	0.999
95%	0.939	2.750	0.107	0.999
90%	0.887	2.474	0.118	0.995
True propensity score model, false regression model				
99%	0.953	3.300	0.112	0.998
95%	0.881	2.762	0.131	0.991
90%	0.823	2.487	0.144	0.979
False propensity score model, true regression model				
99%	0.980	3.284	0.075	1.000
95%	0.926	2.743	0.088	0.998
90%	0.856	2.466	0.097	0.994
False propensity score model, false regression model				
99%	0.561	3.285	0.079	0.999
95%	0.299	2.744	0.093	0.988
90%	0.186	2.468	0.102	0.965
	$N = 3,000$			
True propensity score model, true regression model				
99%	0.988	3.299	0.090	1.000
95%	0.941	2.761	0.106	1.000
90%	0.880	2.486	0.117	0.997
True propensity score model, false regression model				
99%	0.966	3.303	0.099	1.000
95%	0.907	2.765	0.116	0.993
90%	0.848	2.491	0.128	0.985
False propensity score model, true regression model				
99%	0.987	3.296	0.082	1.000
95%	0.929	2.757	0.097	0.998
90%	0.863	2.482	0.106	0.995
False propensity score model, false regression model				
99%	0.014	3.296	0.081	0.986
95%	0.001	2.756	0.096	0.912
90%	0.000	2.481	0.105	0.843

Note. 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 based on the critical values computed by the Gumbel approximation ("GCP").

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 datasets used. We use two different datasets: the dataset from Pennsylvania and that from North Carolina. 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,

TABLE 2 Monte Carlo results, CATEF confidence band, $p = 30$

Confidence level	CP	Mcri	Sdcri	GCP
	$N = 500$			
True propensity score model, true regression model				
99%	0.993	3.294	0.096	1.000
95%	0.960	2.754	0.113	0.999
90%	0.915	2.478	0.124	0.997
True propensity score model, false regression model				
99%	0.969	3.302	0.113	0.999
95%	0.909	2.763	0.132	0.994
90%	0.853	2.489	0.145	0.985
False propensity score model, true regression model				
99%	0.989	3.284	0.078	1.000
95%	0.952	2.742	0.092	0.999
90%	0.904	2.466	0.101	0.996
False propensity score model, false regression model				
99%	0.683	3.284	0.077	0.999
95%	0.410	2.743	0.091	0.994
90%	0.264	2.466	0.099	0.977
	$N = 2,000$			
True propensity score model, true regression model				
99%	0.990	3.296	0.085	1.000
95%	0.950	2.757	0.100	1.000
90%	0.893	2.481	0.110	0.998
True propensity score model, false regression model				
99%	0.975	3.304	0.101	0.999
95%	0.920	2.766	0.119	0.996
90%	0.863	2.492	0.130	0.988
False propensity score model, true regression model				
99%	0.987	3.296	0.079	1.000
95%	0.935	2.756	0.093	0.998
90%	0.876	2.481	0.103	0.993
False propensity score model, false regression model				
99%	0.021	3.296	0.081	0.990
95%	0.003	2.757	0.096	0.921
90%	0.001	2.482	0.106	0.852

Note. See note to Table 1.

although there has been a debate over its magnitude. See, for example, 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 & Currie, 2011). The program evaluation approach is employed, for example, by 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 et al. (2009) examine whether the effect of smoking is larger for teen mothers than for adult mothers and find mixed evidence. Abrevaya et al. (2015) also consider this problem in their application.

6.1 | Pennsylvania data

The first 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 et al. (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 the 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 model 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 et al. (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 5). "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." Although 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 esti-

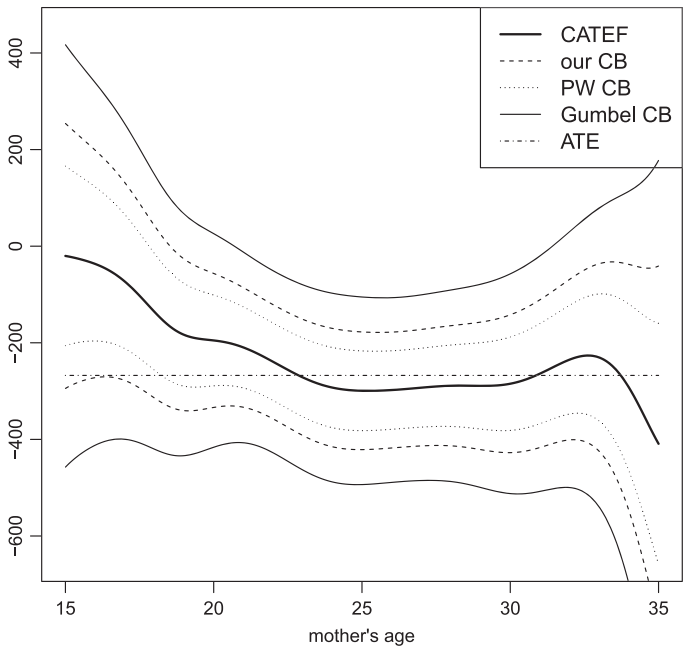

FIGURE 1 CATEF for the effect of smoking on birth weights, Pennsylvania data, 95% confidence bands. "CATEF" = the estimated CATEF; "our CB" = the uniformly valid confidence band proposed in this paper; "PW CB" = the confidence band that is valid only in a pointwise sense; "Gumbel CB" = the uniformly valid confidence band based on the Gumbel approximation; "ATE" = the estimated value of the average treatment effect

mated 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 20–30 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.

6.2 | North Carolina data

The second dataset is based on the records between 1988 and 2002 by the North Carolina State Center Health Services. This dataset is used in Abrevaya et al. (2015) and obtained from Robert Lieli's website (http://www.personal.ceu.hu/staff/Robert_Lieli/cate-birthdata.zip). We restrict our sample to white and first-time mothers, and the sample size is 433,558. As in the case of the Pennsylvania data, the outcome is infant birth weight measured in grams and the treatment variable is an indicator for smoking status. The set of covariates $\mathbf{Z}$ includes those used in the analysis of the Pennsylvania data, except an indicator for the first baby because we focus on first-time mothers, and in addition, it includes indicators for gestational diabetes, hypertension, amniocentesis, and ultrasound exams. Again, $\mathbf{X}$ is the mother's age in this application. The specification for the estimation of the CATEF is the same as before.

The purpose of using this much larger dataset is to examine the effect of the sample size. Our method involves nonparametric kernel regression and it might require a large sample size to yield a reliable result. For example, the result from the Pennsylvania data indicates that the effect of smoking is very small for very young mothers. One might argue that such a result is an artifact of small sample size. The other issue is that the confidence bands obtained using the Pennsylvania data are somewhat wide. We hope that using this larger dataset provides us with narrower confidence bands and more informative statistical results.

Figure 2 plots the estimated CATEF and the three 95% confidence bands for the range between 15 and 35 years of age. Note that the scale of the vertical axis is different from Figure 1. We now obtain much narrower confidence bands. The widths of the three (uniform, pointwise and Gumbel) confidence bands are still different. The estimated CATEF for young mothers is negative and statistically different from 0. The previous result that it is close to 0 may be considered as an artifact of small sample size. The estimated CATEF is decreasing from around 17 to around 29 years of age. For the range above 30 years of age, we obtain relatively wide confidence bands. We reject the null hypothesis of no effect of smoking on birth weights uniformly over 15–35 years of age. These confidence bands do not support the hypothesis that the CATEF is constant because the ATE line exceeds the confidence bands.

One might argue that the difference in the results may stem from the fact that the North Carolina data contain richer information and we use a larger set of covariates. We reexamine the North Carolina data based on the same set of covariates as that for the Pennsylvania data, except an indicator for the first baby. Figure 3 in the Supporting Information Appendix plots the estimated CATEF and confidence bands obtained using this set of covariates. The results in Figure 3 are qualitatively very similar to those in Figure 2. We thus believe that the difference between the results from the Pennsylvania data and the North Carolina data are not from the difference in the covariates but from the difference in the sizes of these two samples.

We thus interpret our findings to indicate that the different results come from the difference in sample size, yet our confidence bands reasonably quantify the uncertainty from small sample size. While two datasets yield different estimates of CATEF, the confidence bands from the Pennsylvania data include the estimated CATEF and the confidence bands from the North Carolina data.

While we use the same dataset as that used in Abrevaya et al. (2015), it is somewhat difficult to compare their results with ours because of differences in the implementations. In particular, the bandwidths are very different. Our choice of bandwidth is around 0.2, whereas theirs are between 1.4 and 11.2. Nonetheless, we make several remarks. Using small bandwidths (1.4 and 2.8), Abrevaya et al. observe almost no effect for young mothers and a large negative effect for 25- to 30-year-old mothers. We do not observe such a large difference in the effect across different age groups. Our confidence band is as tight as their confidence band obtained with bandwidth 11.2 even though we use a much smaller bandwidth and our confidence band is uniform. This is possibly because we use an AIPW method, which yields a more efficient estimate than an IPW method does.

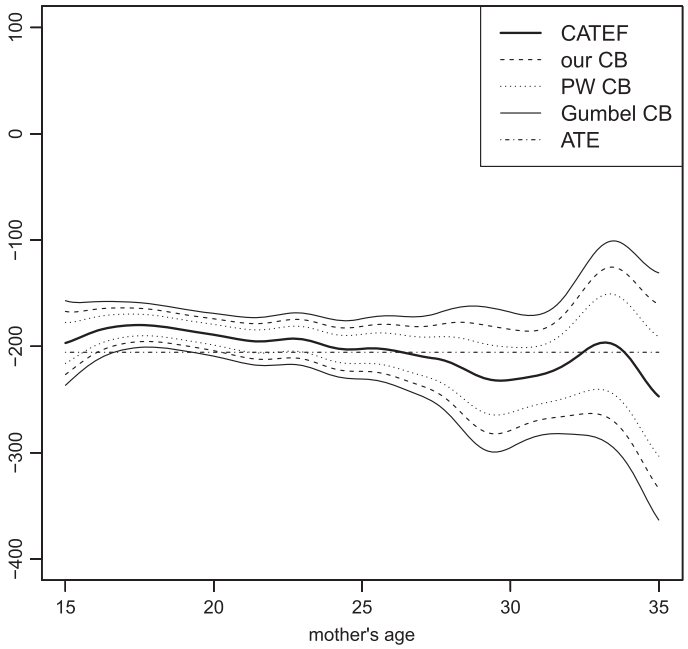

FIGURE 2 CATEF for the effect of smoking on birth weights, North Carolina data, with 95% confidence bands. See caption to Figure 1 for further details

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 et al. (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.

8 | PROOFS

This section provides the proof of Theorem 2. To save space, the proofs of Lemmas 1, 3, 4, and 5 are given in Supporting Information Appendix B. We first establish the linear expansion of the local linear estimator.

Lemma 3. *Let Assumption 1 hold. Then*

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

for some positive constant $c > 0$.

Define

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

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 et al. (2014) to obtain Gaussian approximations. Define

$$W_n \equiv \sup_{\mathbf{x}\in\mathcal{I}} c_n(\mathbf{x})\sqrt{nh_n^d}\left[T_n(\mathbf{x}) - \mathbb{E}T_n(\mathbf{x})\right]. \tag{14}$$

Chernozhukov et al. (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

$$\mathbb{E}[\tilde{B}_{n,1}(\mathbf{x})\tilde{B}_{n,1}(\mathbf{x}')] = h_n^{-d}c_n(\mathbf{x})c_n(\mathbf{x}')\text{cov}\left[U^2\mathbf{K}\left(\frac{\mathbf{X}-\mathbf{x}}{h_n}\right)\mathbf{K}\left(\frac{\mathbf{X}-\mathbf{x}'}{h_n}\right)\right]. \tag{15}$$

Proposition 3.2 of Chernozhukov et al. (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 (Equation 15), 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 $n \to \infty$:*

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

Next we 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 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 Lemma 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 & 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

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

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

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

and $[\cdot]$ is the integer part of a number. Simple calculation yields $\sqrt{\det\quad\Lambda_2}=\lambda^{d/2}$ with λ defined in Equation 10.

ACKNOWLEDGMENTS

We would like to thank Yanchun Jin for capable research assistance, Takahiro Hoshino, Yu-Chin Hsu, Artūras Juodis, Kengo Kato, Edward Kennedy, Taisuke Otsu, Dylan Small, seminar participants at various institutes, the co-editor Edward Vytlacil and three anonymous referees for helpful comments, and Robert Lieli for providing the data to us. Lee's work was supported by the European Research Council (ERC-2009-StG-240910-ROMETA and ERC-2014-CoG-646917-ROMIA). Okui's work was supported by the Japan Society of the Promotion of Science (KAKENHI 25285067, 25780151, 15H03329, 16K03598). Whang's work was supported by the SNU College of Social Science Research Grant. This work was conducted while Okui was at Kyoto University and Vrije Universiteit Amsterdam.

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., & Dahl, C. M. (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., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business and Economic Statistics*, *33*(4), 485–505.

Almond, D., Chay, K. Y., & Lee, D. S. (2005). The costs of low birth weight. *Quarterly Journal of Economics*, *120*, 1031–1083.

Almond, D., & Currie, J. (2011). Human capital development before age five, *Handbook of Labor Economics* (pp. 1315–1486), Vol. *4b*. Amsterdam, Netherlands: Elsevier.

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

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

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

Chernozhukov, V., Chetverikov, D., & Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics*, *42*(4), 1564–1597.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., & Newey, W. K. (2016). Locally robust semiparametric estimation. (*Cemmap Working Paper CWP31/16*).

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

Chernozhukov, V., Lee, S., & Rosen, A. (2013). Intersection bounds: Estimation and inference. *Econometrica*, *81*(2), 667–737.

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

da Veiga, P. V., & Wilder, R. P. (2008). Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Maternal and Child Health Journal*, *12*(2), 194–203.

Ghosal, S., Sen, A., & van der Vaart, A. W. (2000). Testing monotonicity of regression. *Annals of Statistics*, *28*, 1054–1082.

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, *18*, 36–56.

Hall, P., & Horowitz, J. (2013). A simple bootstrap method for constructing non-parametric confidence bands for functions. *Annals of Statistics*, *41*(4), 1892–1921.

Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variable and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, *96*(8), 4730–4734.

Heckman, J. J., & Vytlacil, E. J. (2005). Structural equations, treatment, effects and econometric policy evaluation. *Econometrica*, *73*(3), 669–738.

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

Kang, J. D. Y., & Schafer, J. L. (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., & Piterbarg, V. (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., Linton, O., & Whang, Y.-J. (2009). Testing for stochastic monotonicity. *Econometrica*, *77*(2), 585–602.

Lee, S., & Shaikh, A. M. (2014). Multiple testing and heterogeneous treatment effects: Re-evaluating the effect of PROGRESA on school enrollment. *Journal of Applied Econometrics*, *29*(4), 612–626.

Li, Q., & Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, *14*(2), 485–512.

Lunceford, J. K., & Davidian, M. (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., Rotnitzky, A., & Robins, J. M. (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., Small, D. S., Tan, Z., & Robins, J. M. (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*. Providence, RI: American Mathematical Society.

Poterba, J. M., Venti, S. F., & Wise, D. A. (1995). Do 401(k) contributions crowd out other personal saving?. *Journal of Public Economics*, *58*(1), 1–32.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference, *The IMA Volumes in Mathematics and its Applications* (pp. 95–133), Statistical models in epidemiology, the environment, and clinical trials, Vol. *116*. New York, NY: Springer.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122–129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846–866.

Rothe, C., & Firpo, S. (2016). Properties of doubly robust estimators when nuisance functions are estimated nonparametrically. (*Working paper*). http://www.christophrothe.net

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

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

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American 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.

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

Uysal, S. D. (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.

van der Laan, M. J. (2006). Statistical inference for variable importance. *International Journal of Biostatistics*, *2*(1), Article 2.

Walker, M., Tekin, E., & Wallace, S. (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.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). Cambridge, MA: MIT press.

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Lee S, Okui R, Whang Y-J. Doubly robust uniform confidence band for the conditional average treatment effect function. *J Appl Econ*. 2017;32:1207–1225. https://doi.org/10.1002/jae.2574