ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 4, *pp.* 1229–1245

Non-parametric methods for doubly robust estimation of continuous treatment effects

Edward H. Kennedy, Zongming Ma, Matthew D. McHugh and Dylan S. Small

University of Pennsylvania, Philadelphia, USA

[Received August 2015. Revised August 2016]

Summary. Continuous treatments (e.g. doses) arise often in practice, but many available causal effect estimators are limited by either requiring parametric models for the effect curve, or by not allowing doubly robust covariate adjustment. We develop a novel kernel smoothing approach that requires only mild smoothness assumptions on the effect curve and still allows for misspecification of either the treatment density or outcome regression. We derive asymptotic properties and give a procedure for data-driven bandwidth selection. The methods are illustrated via simulation and in a study of the effect of nurse staffing on hospital readmissions penalties.

Keywords: Causal inference; Dose–response; Efficient influence function; Kernel smoothing; Semiparametric estimation

1. Introduction

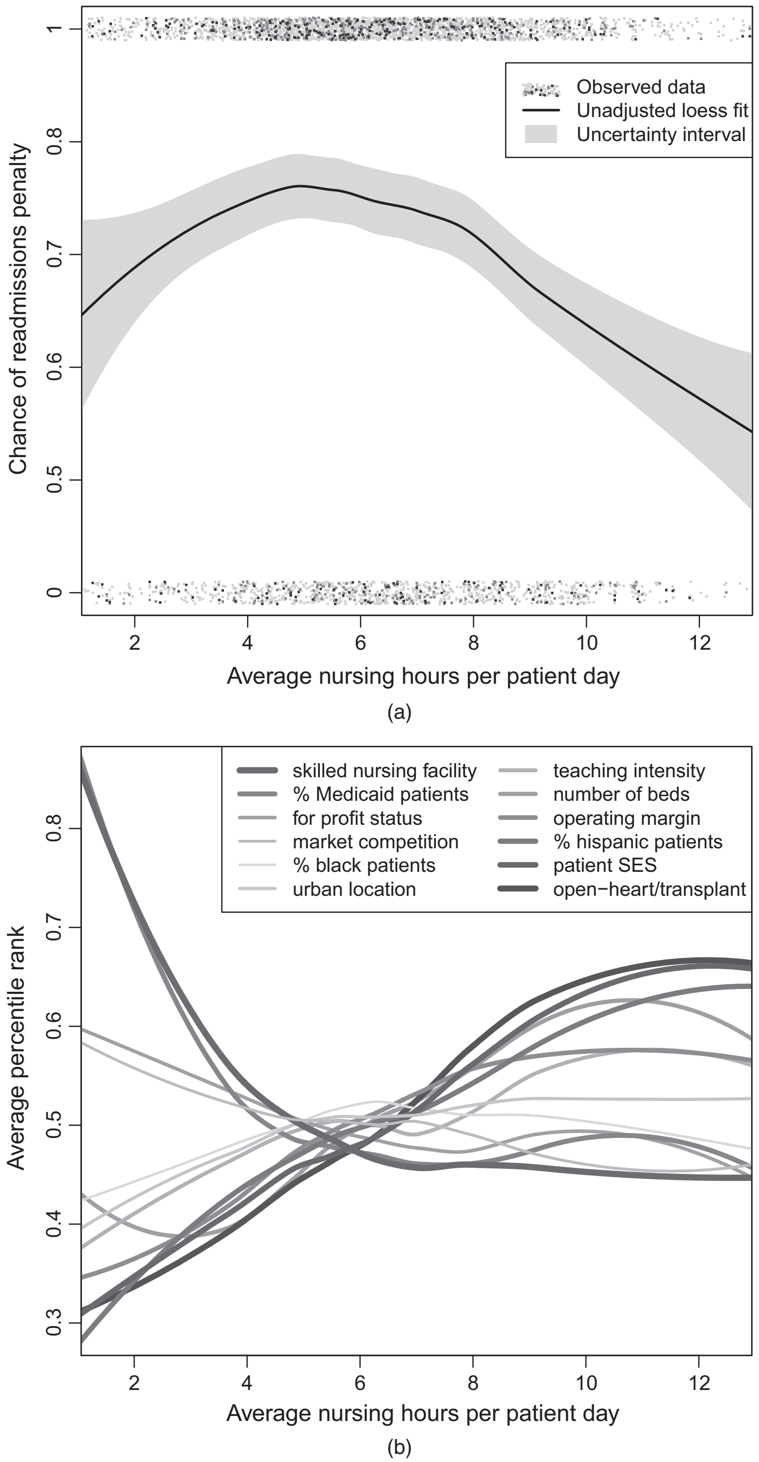
Continuous treatments or exposures (such as dose, duration and frequency) arise very often in practice, especially in observational studies. Importantly, such treatments lead to effects that are naturally described by curves (e.g. dose–response curves) rather than scalars, as might be the case for binary treatments. Two major methodological challenges in continuous treatment settings are

(a) to allow for flexible estimation of the dose–response curve (e.g. to discover underlying structure without imposing *a priori* shape restrictions) and
(b) to adjust properly for high dimensional confounders (i.e. pretreatment covariates related to treatment assignment and outcome).

Consider a recent example involving the ‘Hospital readmissions reduction program’, instituted by the Centers for Medicare and Medicaid Services in 2012, which aimed to reduce preventable hospital readmissions by penalizing hospitals with excess readmissions. McHugh *et al.* (2013) were interested in whether nurse staffing (measured in nurse hours per patient day) affected hospitals’ risk of excess readmissions penalty. Fig. 1(a) shows data for 2976 hospitals, with nurse staffing (the ‘treatment’) on the x-axis, whether each hospital was penalized (the outcome) on the y-axis and a LOESS curve fit to the data (without any adjustment). One way to characterize effects is to imagine setting all hospitals’ nurse staffing to the same level, and seeing whether changes in this level yield changes in excess readmissions risk. Such questions cannot be answered by simply comparing hospitals’ risk of penalty across levels of nurse staffing, since hospitals differ in many important ways that could be related to both nurse staffing and

Address for correspondence: Edward H. Kennedy, Department of Statistics, Carnegie Mellon University, Room 228A, Baker Hall, 1098 Morewood Avenue, Pittsburgh, PA 15213, USA.
E-mail: edward@stat.cmu.edu

1369–7412/17/791229



(a)

(b)

Fig. 1. (a) Observed treatment and outcome data with an unadjusted LOESS fit and (b) average covariate value as a function of exposure, after transforming to percentiles to display on a common scale excess readmissions (e.g. size, location and teaching status, among many other factors). Fig. 1(b) displays the extent of these hospital differences, showing for example that hospitals with more nurse staffing are also more likely to be high technology hospitals and to see patients with higher socio-economic status. To estimate the effect curve correctly, and to compare the risk of readmissions penalty at different nurse staffing levels fairly, we must adjust for hospital characteristics appropriately.

In practice, the most common approach for estimating continuous treatment effects is based on regression modelling of how the outcome relates to covariates and treatment (e.g. Imbens (2004) and Hill (2011)). However, this approach relies entirely on correct specification of the outcome model, does not incorporate available information about the treatment mechanism and is sensitive to the curse of dimensionality by inheriting the rate of convergence of the outcome regression estimator. Hirano and Imbens (2004), Imai and van Dyk (2004) and Galvao and Wang (2015) adapted propensity-score-based approaches to the continuous treatment setting, but these similarly rely on correct specification of at least a model for treatment (e.g. the conditional treatment density).

In contrast, semiparametric doubly robust estimators (Robins and Rotnitzky, 2001; van der Laan and Robins, 2003) are based on modelling both the treatment and the outcome processes and, remarkably, give consistent estimates of effects as long as one of these two nuisance processes is modelled sufficiently well (not necessarily both). Beyond giving two independent chances at consistent estimation, doubly robust methods can also attain faster rates of convergence than their nuisance (i.e. outcome and treatment process) estimators when both models are consistently estimated; this makes them less sensitive to the curse of dimensionality and can allow for inference even after using flexible machine-learning-based adjustment. However, standard semiparametric doubly robust methods for dose–response estimation rely on parametric models for the effect curve, either by explicitly assuming a parametric dose–response curve (Robins, 2000; van der Laan and Robins, 2003) or else by projecting the true curve onto a parametric working model (Neugebauer and van der Laan, 2007). Unfortunately, the first approach can lead to substantial bias under model misspecification, and the second can be of limited practical use if the working model is far from the truth.

Recent work has extended semiparametric doubly robust methods to more complicated non-parametric and high dimensional settings. In a foundational paper, van der Laan and Dudoit (2003) proposed a powerful cross-validation framework for estimator selection in general censored data and causal inference problems. Their empirical risk minimization approach allows global non-parametric modelling in general semiparametric settings involving complex nuisance parameters. For example, Díaz and van der Laan (2013) considered global modelling in the dose–response curve setting and developed a doubly robust substitution estimator of risk. In non-parameric problems it is also important to consider non-global learning methods, e.g. via local and penalized modelling (Györfi *et al.*, 2002). Rubin and van der Laan (2005, 2006a, b) proposed extensions to such paradigms in numerous important problems, but Rubin and van der Laan (2005, 2006a) considered weighted averages of dose–response curves and Rubin and van der Laan (2006b) did not consider doubly robust estimation.

In this paper we present a new approach for causal dose–response estimation that is doubly robust without requiring parametric assumptions and which can naturally incorporate general machine learning methods. The approach is motivated by semiparametric theory for a particular stochastic intervention effect and a corresponding doubly robust mapping. Our method has a simple two-stage implementation that is fast and easy to use with standard software: in the first stage a pseudo-outcome is constructed based on the doubly robust mapping, and in the second stage the pseudo-outcome is regressed on treatment via off-the-shelf non-parametric regression and machine learning tools. We provide asymptotic results for a kernel version of our approach under weak assumptions which require only mild smoothness conditions on the effect curve and allow flexible data-adaptive estimation of relevant nuisance functions. We also discuss a simple method for bandwidth selection based on cross-validation. The methods are illustrated via simulation, and in the study discussed earlier about the effect of hospital nurse staffing on excess readmission penalties.

2. Background

2.1. Data and notation

Suppose that we observe an independent and identically distributed sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ where $\mathbf{Z} = (\mathbf{L}, A, Y)$ has support $\mathcal{Z} = (\mathcal{L} \times \mathcal{A} \times \mathcal{Y})$. Here $\mathbf{L}$ denotes a vector of covariates, A a continuous treatment or exposure and Y some outcome of interest. We characterize causal effects by using potential outcome notation (Rubin, 1974), and so let Y^a denote the potential outcome that would have been observed under treatment level a.

We denote the distribution of $\mathbf{Z}$ by P, with density $p(\mathbf{z}) = p(y|\mathbf{l}, a)\, p(a|\mathbf{l})\, p(\mathbf{l})$ with respect to some dominating measure. We let $\mathbb{P}_n$ denote the empirical measure so that empirical averages $n^{-1}\Sigma_i f(\mathbf{Z}_i)$ can be written as $\mathbb{P}_n\{f(\mathbf{Z})\} = \int f(\mathbf{z})\mathrm{d}\mathbb{P}_n(\mathbf{z})$. To simplify the presentation we denote the mean outcome given covariates and treatment with $\mu(\mathbf{l}, a) = \mathbb{E}(Y|\mathbf{L} = \mathbf{l}, A = a)$, denote the conditional treatment density given covariates with $\pi(a|\mathbf{l}) = \partial P(A \leqslant a|\mathbf{L} = \mathbf{l})/\partial a$ and denote the marginal treatment density with $\varpi(a) = \partial P(A \leqslant a)/\partial a$. Finally, we use $\|f\| = \{\int f(\mathbf{z})^2 \mathrm{d}P(\mathbf{z})\}^{1/2}$ to denote the $L_2(P)$ norm, and we use $\|f\|_{\mathcal{X}} = \sup_{x \in \mathcal{X}} |f(x)|$ to denote the uniform norm of a generic function f over $x \in \mathcal{X}$.

2.2. Identification

In this paper our goal is to estimate the effect curve $\theta(a) = \mathbb{E}(Y^a)$. Since this quantity is defined in terms of potential outcomes that are not directly observed, we must consider assumptions under which it can be expressed in terms of observed data. A full treatment of identification in the presence of continuous random variables was given by Gill and Robins (2001); we refer the reader there for details. The assumptions that are most commonly employed for identification are as follows (the following assumptions must hold for any $a \in \mathcal{A}$ at which $\theta(a)$ is to be identified).

Assumption 1. Consistency: $A = a$ implies $Y = Y^a$.

Assumption 2. Positivity: $\pi(a|\mathbf{l}) \geqslant \pi_{\min} > 0$ for all $\mathbf{l} \in \mathcal{L}$.

Assumption 3. Ignorability: $\mathbb{E}(Y^a|\mathbf{L}, A) = \mathbb{E}(Y^a|\mathbf{L})$.

Assumptions 1–3 can all be satisfied by design in randomized trials, but in observational studies they may be violated and are generally untestable. The consistency assumption ensures that potential outcomes are defined uniquely by a subject's own treatment level and not others' levels (i.e. no interference), and also not by the way that treatment is administered (i.e. no different versions of treatment). Positivity says that treatment is not assigned deterministically, in the sense that every subject has some chance of receiving treatment level a, regardless of covariates; this can be a particularly strong assumption with continuous treatments. Ignorability says that the mean potential outcome under level a is the same across treatment levels once we condition on covariates (i.e. treatment assignment is unrelated to potential outcomes within strata of covariates) and requires sufficiently many relevant covariates to be collected. Using the same logic as with discrete treatments, it is straightforward to show that under assumptions 1–3 the effect curve $\theta(a)$ can be identified with observed data as

$$\theta(a) = \mathbb{E}\{\mu(\mathbf{L}, a)\} = \int_{\mathcal{L}} \mu(\mathbf{l}, a)\, \mathrm{d}P(\mathbf{l}). \tag{1}$$

Even if we are not willing to rely on assumptions 1 and 3, it may often still be of interest to estimate $\theta(a)$ as an adjusted measure of association, defined purely in terms of observed data.

3. Main results

In this section we develop doubly robust estimators of the effect curve $\theta(a)$ without relying on parametric models. First we describe the logic behind our proposed approach, which is based on finding a doubly robust mapping whose conditional expectation given treatment equals the effect curve of interest, as long as one of two nuisance parameters is correctly specified. To find this mapping, we derive a novel efficient influence function for a stochastic intervention parameter. Our proposed method is based on regressing this doubly robust mapping on treatment by using off-the-shelf non-parametric regression and machine learning methods. We derive asymptotic properties for a particular version of this approach based on local linear kernel smoothing. Specifically, we give conditions for consistency and asymptotic normality, and we describe how to use cross-validation to select the bandwidth parameter in practice.

3.1. Set-up and doubly robust mapping

If $\theta(a)$ is assumed known up to a finite dimensional parameter, e.g. $\theta(a) = \psi_0 + \psi_1 a$ for $(\psi_0, \psi_1) \in \mathbb{R}^2$, then standard semiparametric theory can be used to derive the efficient influence function, from which we can obtain the efficiency bound and an efficient estimator (Bickel *et al.*, 1993; van der Laan and Robins, 2003; Tsiatis, 2006). However, such theory is not directly available if we assume, for example, only mild smoothness conditions on $\theta(a)$ (e.g. differentiability). This is because without parametric assumptions $\theta(a)$ is not pathwise differentiable, and root n consistent estimators do not exist (Bickel *et al.*, 1993; Díaz and van der Laan, 2013). In this case there is no developed efficiency theory.

To derive doubly robust estimators for $\theta(a)$ without relying on parametric models, we adapt semiparametric theory in a novel way that is similar to the approach of Rubin and van der Laan (2005, 2006a). Our goal is to find a function $\xi(\mathbf{Z}; \pi, \mu)$ of the observed data $\mathbf{Z}$ and nuisance functions (π, μ) such that

$$\mathbb{E}\{\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu}) | A = a\} = \theta(a)$$

if either $\bar{\pi} = \pi$ or $\bar{\mu} = \mu$ (not necessarily both). Given such a mapping, off-the-shelf non-parametric regression and machine learning methods could be used to estimate $\theta(a)$ by regressing $\xi(\mathbf{Z}; \hat{\pi}, \hat{\mu})$ on treatment A, based on estimates $\hat{\pi}$ and $\hat{\mu}$.

This doubly robust mapping is intimately related to semiparametric theory and especially the efficient influence function for a particular parameter. Specifically, if $\mathbb{E}\{\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu}) | A = a\} = \theta(a)$ then it follows that $\mathbb{E}\{\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu})\} = \psi$ for

$$\psi = \int_{\mathcal{A}} \int_{\mathcal{L}} \mu(\mathbf{l}, a) \varpi(a)\, \mathrm{d}P(\mathbf{l})\, \mathrm{d}a. \tag{2}$$

This indicates that a natural candidate for the unknown mapping $\xi(\mathbf{Z}; \pi, \mu)$ would be a component of the efficient influence function for the parameter ψ, since, for regular parameters such as ψ in semiparametric or non-parametric models, the efficient influence function $\phi(\mathbf{Z}; \pi, \mu)$ will be doubly robust in the sense that $\mathbb{E}\{\phi(\mathbf{Z};\bar{\pi},\bar{\mu})\}=0$, if either $\bar{\pi}=\pi$ or $\bar{\mu}=\mu$ (Robins and Rotnitzky, 2001; van der Laan and Robins, 2003). This implies that $\mathbb{E}\{\phi(\mathbf{Z};\pi,\mu)\}=\mathbb{E}\{\xi(\mathbf{Z};\pi,\mu)-\psi\}=0$ so $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})\}=\psi$ if either $\bar{\pi}=\pi$ or $\bar{\mu}=\mu$. This kind of logic was first used by Rubin and van der Laan (2005, 2006a) for full data parameters that are functions of covariates rather than treatment (i.e. censoring) variables.

The parameter ψ is also of interest in its own right. In particular, it represents the average outcome under an intervention that randomly assigns treatment based on the density ϖ (i.e. a randomized trial). Thus comparing the value of this parameter with the average observed outcome provides a test of treatment effect; if the values differ significantly, then there is evidence that the observational treatment mechanism impacts outcomes for at least some units. Stochastic interventions were discussed by Díaz and van der Laan (2012), for example, but the efficient influence function for ψ has not been given before under a non-parametric model. Thus in theorem 1 below we give the efficient influence function for this parameter respecting the fact that the marginal density ϖ is unknown.

Theorem 1. Under a non-parametric model, the efficient influence function for ψ defined in equation (2) is $\xi(\mathbf{Z};\pi,\mu)-\psi+\int_{\mathcal{A}}\{\mu(\mathbf{L},a)-\int_{\mathcal{L}}\mu(\mathbf{l},a)\,\mathrm{d}P(\mathbf{l})\}\varpi(a)\,\mathrm{d}a$, where

$$\xi(\mathbf{Z};\pi,\mu)=\frac{Y-\mu(\mathbf{L},A)}{\pi(A|\mathbf{L})}\int_{\mathcal{L}}\pi(A|\mathbf{l})\mathrm{d}P(\mathbf{l})+\int_{\mathcal{L}}\mu(\mathbf{l},A)\,\mathrm{d}P(\mathbf{l}).$$

A proof of theorem 1 is given in the on-line appendix (section 2). Importantly, we also prove that the function $\xi(\mathbf{Z};\pi,\mu)$ satisfies its desired double-robustness property, i.e. that $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})|A=a\}=\theta(a)$ if either $\bar{\pi}=\pi$ or $\bar{\mu}=\mu$. As mentioned earlier, this motivates estimating the effect curve $\theta(a)$ by estimating the nuisance functions (π,μ), and then regressing the estimated pseudo-outcome

$$\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})=\frac{Y-\hat{\mu}(\mathbf{L},A)}{\hat{\pi}(A|\mathbf{L})}\int_{\mathcal{L}}\hat{\pi}(A|\mathbf{l})\mathrm{d}\mathbb{P}_n(\mathbf{l})+\int_{\mathcal{L}}\hat{\mu}(\mathbf{l},A)\mathrm{d}\mathbb{P}_n(\mathbf{l})$$

on treatment A by using off-the-shelf non-parametric regression or machine learning methods. In the next subsection we describe our proposed approach in more detail and analyse the properties of an estimator based on kernel estimation.

3.2. Approach proposed

In the previous subsection we derived a doubly robust mapping $\xi(\mathbf{Z};\pi,\mu)$ for which $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})|A=a\}=\theta(a)$ as long as either $\bar{\pi}=\pi$ or $\bar{\mu}=\mu$. This indicates that doubly robust non-parametric estimation of $\theta(a)$ can proceed with a simple two-step procedure, where both steps can be accomplished with flexible machine learning. To summarize, our proposed method is as follows.

Step 1: estimate nuisance functions (π,μ) and obtain predicted values.
Step 2: construct pseudo-outcome $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$ and regress on treatment variable A.

We give sample code implementing this method in the on-line appendix (section 9).

In what follows we present results for an estimator that uses kernel smoothing in step 2. Such an approach is related to kernel approximation of a full data parameter in censored data settings. Robins and Rotnitzky (2001) gave a general discussion and considered density estimation with missing data, whereas van der Laan and Robins (1998), van der Laan and Yu (2001) and van der Vaart and van der Laan (2006) used the approach for current status survival analysis; Wang *et al.* (2010) used it implicitly for non-parametric regression with missing outcomes.

As indicated above, however, a wide variety of flexible methods could be used in our step 2, including local partitioning or nearest neighbour estimation, global series or spline methods with complexity penalties, or cross-validation-based combinations of methods, e.g. Super Learner (van der Laan *et al.*, 2007). In general we expect that the results that we report in this paper will hold for many such methods. To see why, let $\hat{\theta}$ denote the proposed estimator described above (based on some initial nuisance estimators $(\hat{\pi},\hat{\mu})$ and a particular regression method in step 2), and let $\bar{\theta}$ denote an estimator based on an oracle version of the pseudo-outcome $\xi(\mathbf{Z};\bar{\pi},\bar{\mu})$ where $(\bar{\pi},\bar{\mu})$ are the unknown limits to which the estimators $(\hat{\pi},\hat{\mu})$ converge. Then $\|\hat{\theta}-\theta\| \leqslant \|\hat{\theta}-\bar{\theta}\| + \|\bar{\theta}-\theta\|$, where the second term on the right-hand side can be analysed with standard theory since $\bar{\theta}$ is a regression of a simple fixed function $\xi(\mathbf{Z};\bar{\pi},\bar{\mu})$ on A, and the first term will be small depending on the convergence rates of $\hat{\pi}$ and $\hat{\mu}$. A similar point was discussed by Rubin and van der Laan (2005, 2006a).

The local linear kernel version of our estimator is $\hat{\theta}_h(a)=\mathbf{g}_{ha}(a)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_h(a)$, where $\mathbf{g}_{ha}(t)=(1,(t-a)/h)^{\mathrm{T}}$ and

$$\hat{\boldsymbol{\beta}}_h(a)=\underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\,\mathbb{P}_n[K_{ha}(A)\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})-\mathbf{g}_{ha}(A)^{\mathrm{T}}\boldsymbol{\beta}\}^2] \tag{3}$$

for $K_{ha}(t)=h^{-1}K\{(t-a)/h\}$ with K a standard kernel function (e.g. a symmetric probability density) and h a scalar bandwidth parameter. This is a standard local linear kernel regression of $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$ on A. For overviews of kernel smoothing see, for example, Fan and Gijbels (1996), Wasserman (2006) and Li and Racine (2007). Under near violations of positivity, the above estimator could potentially lie outside the range of possible values for $\theta(a)$ (e.g. if Y is binary); thus we present a targeted minimum-loss-based estimator in the on-line appendix (section 4), which does not have this problem. Alternatively one could project onto a logistic model in equation (3).

3.3. Consistency of kernel estimator

In theorem 2 below we give conditions under which the proposed kernel estimator $\hat{\theta}_h(a)$ is consistent for $\theta(a)$, and we also give the corresponding rate of convergence. In general this result follows if the bandwidth decreases with sample size sufficiently slowly, and if either of the nuisance functions π or μ is estimated sufficiently well (not necessarily both). The rate of convergence is a sum of two rates: one from standard non-parametric regression problems (depending on the bandwidth h), and another coming from estimation of the nuisance functions π and μ.

Theorem 2. Let $\bar{\pi}$ and $\bar{\mu}$ denote fixed functions to which $\hat{\pi}$ and $\hat{\mu}$ converge in the sense that $\|\hat{\pi}-\bar{\pi}\|_{\mathcal{Z}}=o_p(1)$ and $\|\hat{\mu}-\bar{\mu}\|_{\mathcal{Z}}=o_p(1)$, and let $a\in\mathcal{A}$ denote a point in the interior of the compact support $\mathcal{A}$ of A. Along with assumption 2 (positivity), make the following assumptions.

(a) Either $\bar{\pi}=\pi$ or $\bar{\mu}=\mu$.
(b) The bandwidth $h=h_n$ satisfies $h\rightarrow 0$ and $nh^3\rightarrow\infty$ as $n\rightarrow\infty$.
(c) K is a continuous symmetric probability density with support $[-1,1]$.
(d) $\theta(a)$ is twice continuously differentiable, and both $\varpi(a)$ and the conditional density of $\xi(\mathbf{Z};\bar{\pi},\bar{\mu})$ given $A=a$ are continuous as functions of a.
(e) The estimators $(\hat{\pi},\hat{\mu})$ and their limits $(\bar{\pi},\bar{\mu})$ are contained in uniformly bounded function classes with finite uniform entropy integrals (as defined in section 5 of the on-line appendix), with $1/\hat{\pi}$ and $1/\bar{\pi}$ also uniformly bounded.

Then

$$|\hat{\theta}_h(a) - \theta(a)| = O_p\left\{ \frac{1}{\surd(nh)} + h^2 + r_n(a)s_n(a) \right\}$$

where

$$\sup_{t:|t-a|\leqslant h} \|\hat{\pi}(t|\mathbf{L}) - \pi(t|\mathbf{L})\| = O_p\{r(n)\},$$

$$\sup_{t:|t-a|\leqslant h} \|\hat{\mu}(\mathbf{L},t) - \mu(\mathbf{L},t)\| = O_p\{s(n)\}$$

are the 'local' rates of convergence of $\hat{\pi}$ and $\hat{\mu}$ near $A = a$.

A proof of theorem 2 is given in the on-line appendix (section 6). The conditions required are all quite weak. Condition (a) is arguably the most important of the conditions and says that at least one of the estimators $\hat{\pi}$ or $\hat{\mu}$ must be consistent for the true π or μ in terms of the uniform norm. Since only one of the nuisance estimators is required to be consistent (not both), theorem 2 shows the double robustness of the proposed estimator $\hat{\theta}_h(a)$. Conditions (b), (c) and (d) are all common in standard non-parametric regression problems, whereas condition (e) involves the complexity of the estimators $\hat{\pi}$ and $\hat{\mu}$ (and their limits), and is a usual minimal regularity condition for problems involving nuisance functions.

Condition (b) says that the bandwidth parameter h decreases with sample size but not too quickly (so that $nh^3 \to \infty$). This is a standard requirement in local linear kernel smoothing (Fan and Gijbels, 1996; Wasserman, 2006; Li and Racine, 2007). Since $nh = nh^3/h^2$, it is implied that $nh \to \infty$; thus we can view nh as a kind of effective or local sample size. Roughly speaking, the bandwidth h needs to go to 0 to control bias, whereas the local sample size nh (and nh^3) needs to go to ∞ to control variance. We postpone more detailed discussion of the bandwidth parameter until a later subsection, where we detail how it can be chosen in practice by using cross-validation. Condition (c) puts some minimal restrictions on the kernel function. It is clearly satisfied for most common kernels, including the uniform kernel $K(u) = I(|u| \leqslant 1)/2$, the Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leqslant 1)$ and a truncated version of the Gaussian kernel $K(u) = I(|u| \leqslant 1)\phi(u)/\{2\Phi(1) - 1\}$ with ϕ and Φ the density and distribution functions for a standard normal random variable. Condition (d) restricts the smoothness of the effect curve $\theta(a)$, the density of $\varpi(a)$ and the conditional density given $A = a$ of the limiting pseudo-outcome $\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu})$. These are standard smoothness conditions imposed in non-parametric regression problems. By assuming more smoothness of $\theta(a)$, bias reducing (higher order) kernels could achieve faster rates of convergence and even approach the parametric root n rate (see for example Fan and Gijbels (1996), Wasserman (2006) and others).

Condition (e) puts a mild restriction on how flexible the nuisance estimators (and their corresponding limits) can be, although such uniform entropy conditions still allow a wide array of data-adaptive estimators and, importantly, do not require the use of parametric models. Andrews (1994) (section 4), van der Vaart and Wellner (1996) (sections 2.6–2.7) and van der Vaart (2000) (examples 19.6–19.12) have discussed a wide variety of function classes with finite uniform entropy integrals. Examples include standard parametric classes of functions indexed by Euclidean parameters (e.g. parametric functions satisfying a Lipschitz condition), smooth functions with uniformly bounded partial derivatives and Sobolev classes of functions, as well as convex combinations or Lipschitz transformations of any such sets of functions. The uniform entropy restriction in condition (e) is therefore not a very strong restriction in practice; however, it could be further weakened via sample splitting techniques (see chapter 27 of van der Laan and Rose (2011)).

The convergence rate that is given in the result of theorem 2 is a sum of two components. The first, $1/\sqrt{(nh)}+h^2$, is the rate that is achieved in standard non-parametric regression problems without nuisance functions. If $h \to 0$ slowly, then $1/\sqrt{(nh)}$ will tend to 0 quickly but h^2 will tend to 0 more slowly; similarly, if $h \to 0$ quickly, then h^2 will as well, but $1/\sqrt{(nh)}$ will tend to 0 more slowly. Balancing these two terms requires $h \sim n^{-1/5}$ so $1/\sqrt{(nh)} \sim h^2 \sim n^{-2/5}$. This is the optimal pointwise rate of convergence for standard non-parametric regression on a single covariate, for a twice continuously differentiable regression function.

The second component, $r_n(a)s_n(a)$, is the product of the local rates of convergence (around $A=a$) of the nuisance estimators $\hat{\pi}$ and $\hat{\mu}$ towards their targets π and μ. Thus, if the nuisance function estimates converge slowly (because of the curse of dimensionality), then the convergence rate of $\hat{\theta}_h(a)$ will also be slow. However, since the term is a product, we have two chances at obtaining fast convergence rates, showing the bias reducing benefit of doubly robust estimators. The usual explanation of double robustness is that, even if $\hat{\mu}$ is misspecified so that $s_n(a)=O(1)$, then as long as $\hat{\pi}$ is consistent, i.e. $r_n(a)=o(1)$, we shall still have consistency since $r_n(a)s_n(a)=o(1)$. But this idea also extends to settings when both $\hat{\pi}$ and $\hat{\mu}$ are consistent. For example suppose that $h \sim n^{-1/5}$ so that $1/\sqrt{(nh)}+h^2 \sim n^{-2/5}$, and suppose that $\hat{\pi}$ and $\hat{\mu}$ are locally consistent with rates $r_n(a)=n^{-2/5}$ and $s_n(a)=n^{-1/10}$. Then the product is $r_n(a)s_n(a)=O(n^{-1/2})=o(n^{-2/5})$ and the contribution from the nuisance functions is asymptotically negligible, in the sense that the estimator proposed has the same rate of convergence as an infeasible estimator with known nuisance functions. Contrast this with non-doubly-robust plug-in estimators whose rate of convergence generally matches that of the nuisance function estimator, rather than being faster (van der Vaart, 2014).

In section 8 of the on-line appendix we give some discussion of uniform consistency, which, along with weak convergence, will be pursued in more detail in future work.

3.4. Asymptotic normality of kernel estimator

In the next theorem we show that, if one or both of the nuisance functions are estimated at sufficiently fast rates, then the estimator proposed is asymptotically normal after appropriate scaling.

Theorem 3. Consider the same setting as theorem 2. Along with assumption 2 (positivity) and conditions (a)–(e) from theorem 2, also assume that

(f) the local convergence rates satisfy $r_n(a)s_n(a)=o_p\{1/\sqrt{(nh)}\}$.

Then

$$\sqrt{(nh)}\{\hat{\theta}_h(a)-\theta(a)+b_h(a)\} \xrightarrow{\mathrm{d}} N\left\{0,\ \frac{\sigma^2(a)\int K(u)^2\mathrm{d}u}{\varpi(a)}\right\}$$

where $b_h(a)=\theta''(a)(h^2/2)\int u^2K(u)\mathrm{d}u+o(h^2)$, and

$$\sigma^2(a)=\mathbb{E}\left[\frac{\tau^2(\mathbf{L},a)+\{\mu(\mathbf{L},a)-\bar{\mu}(\mathbf{L},a)\}^2}{\{\bar{\pi}(a|\mathbf{L})/\bar{\varpi}(a)\}^2/\{\pi(a|\mathbf{L})/\varpi(a)\}}\right]-\{\theta(a)-\bar{m}(a)\}^2$$

for $\tau^2(\mathbf{l},a)=\mathrm{var}(Y|\mathbf{L}=\mathbf{l},A=a)$, $\bar{\varpi}(a)=\mathbb{E}\{\bar{\pi}(a|\mathbf{L})\}$ and $\bar{m}(a)=\mathbb{E}\{\bar{\mu}(\mathbf{L},a)\}$.

The proof of theorem 3 is given in the on-line appendix (section 7). Conditions (a)–(e) are the same as in theorem 2 and were discussed earlier. Condition (f) puts a restriction on the local convergence rates of the nuisance estimators. This will in general require at least some semiparametric modelling of the nuisance functions. Truly non-parametric estimators of π and μ will typically converge at slow rates because of the curse of dimensionality and will generally not satisfy the rate requirement in the presence of multiple continuous covariates. Condition (f) basically ensures that estimation of the nuisance functions is irrelevant asymptotically; depending on the specific nuisance estimators that are used, it could be possible to give weaker but more complicated conditions that allow for a non-negligible asymptotic contribution while still yielding asymptotic normality.

Importantly, the rate of convergence that is required by the additional condition of theorem 3 is slower than the root n rate that is typically required in standard semiparametric settings where the parameter of interest is finite dimensional and Euclidean. For example, in a standard setting where the support $\mathcal{A}$ is finite, a sufficient condition for yielding the requisite asymptotic negligibility for attaining efficiency is $r_n(a)=s_n(a)=o(n^{-1/4})$; however, in our setting the weaker condition $r_n(a)=s_n(a)=o(n^{-1/5})$ would be sufficient if $h\sim n^{-1/5}$. Similarly, if one nuisance estimator $\hat{\pi}$ or $\hat{\mu}$ is computed with a correctly specified generalized additive model, then the other nuisance estimator would need only to be consistent (without a rate condition). This is because, under regularity conditions and with optimal smoothing, a generalized additive model estimator converges at rate $O_p(n^{-2/5})$ (Horowitz, 2009), so if the other nuisance estimator is merely consistent we have $r_n(a)s_n(a)=O(n^{-2/5})\,o(1)=o(n^{-2/5})$, which satisfies condition (f) as long as $h\sim n^{-1/5}$. In standard settings such flexible nuisance estimation would make a non-negligible contribution to the limiting behaviour of the estimator, preventing asymptotic normality and root n consistency.

Under the assumptions of theorem 3, the estimator proposed is asymptotically normal after appropriate scaling and centring. However, the scaling is by the square root of the local sample size $\surd(nh)$ rather than by the usual parametric rate $\surd n$. This slower rate of convergence is a cost of making fewer assumptions (equivalently, the cost of better efficiency would be less robustness); thus we have a typical bias–variance trade-off. As in standard non-parametric regression the estimator is consistent but not quite centred at $\theta(a)$; there is a bias term of order $O(h^2)$, denoted $b_h(a)$. In fact the estimator is centred at a smoothed version of the effect curve $\theta_h^*(a)=\mathbf{g}_{ha}(a)^{\mathrm{T}}\boldsymbol{\beta}_h(a)=\theta(a)+b_h(a)$. This is ubiquitous in non-parametric regression and complicates the process of computing confidence bands. It is sometimes assumed that the bias term is $o\{1/\surd(nh)\}$ and thus asymptotically negligible (e.g. by assuming $h=o(n^{-1/5})$ so that $nh^5\to 0$); this is called undersmoothing and technically allows the construction of valid confidence bands around $\theta(a)$. However, there is little guidance about how actually to undersmooth in practice, so it is mostly a technical device. We follow Wasserman (2006) and others by expressing uncertainty about the estimator $\hat{\theta}_h(a)$ by using confidence intervals centred at the smoothed data-dependent parameter $\theta_h^*(a)$. For example, under the conditions of theorem 3, pointwise Wald 95% confidence intervals can be constructed with $\hat{\theta}_h(a)\pm 1.96\hat{\sigma}/\surd n$, where $\hat{\sigma}^2$ is the $(1,1)$ element of the sandwich variance estimate $\mathbb{P}_n\{\hat{\varphi}_{ha}(\mathbf{Z})^{\otimes 2}\}$ based on the estimated efficient influence function for $\boldsymbol{\beta}_h(a)$ given by

$$\hat{\varphi}_{ha}(\mathbf{Z})=\hat{\mathbf{D}}_{ha}^{-1}\Bigg[\mathbf{g}_{ha}(A)K_{ha}(A)\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})-\mathbf{g}_{ha}(A)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_h(a)\}\\ +\int_{\mathcal{A}}\mathbf{g}_{ha}(t)K_{ha}(t)\,\{\hat{\mu}(\mathbf{L},t)-\hat{m}(t)\}\,\hat{\varpi}(t)\,\mathrm{d}t\Bigg]$$

for $\hat{\mathbf{D}}_{ha}=\mathbb{P}_n\{\mathbf{g}_{ha}(A)K_{ha}(A)\mathbf{g}_{ha}^{\mathrm{T}}\}$, $\hat{m}(t)=\mathbb{P}_n\{\hat{\mu}(\mathbf{L},t)\}$ and $\hat{\varpi}(t)=\mathbb{P}_n\{\hat{\pi}(t|\mathbf{L})\}$.

3.5. Data-driven bandwidth selection

The choice of smoothing parameter is critical for any non-parametric method; too much smoothing yields large biases and too little yields excessive variance. In this subsection we discuss how to use cross-validation to choose the relevant bandwidth parameter h. In general the method that we propose parallels those used in standard non-parametric regression settings and can give similar optimality properties.

We can exploit the fact that our method can be cast as a non-standard non-parametric regression problem and borrow from the wealth of literature on bandwidth selection there. Specifically, the logic behind theorem 3 (i.e. that nuisance function estimation can be asymptotically irrelevant) can be adapted to the bandwidth selection setting, by treating the pseudo-outcome $\xi(\mathbf{Z};\hat{\pi},\hat{\mu})$ as known and using for example the bandwidth selection framework from Härdle *et al.* (1988), who proposed a unified selection approach that includes generalized cross-validation, Akaike's information criterion and leave-one-out cross-validation as special cases, and showed the asymptotic equivalence and optimality of such approaches. In our setting, leave-one-out cross-validation is attractive because of its computational ease. The simplest analogue of leave-one-out cross-validation for our problem would be to select the optimal bandwidth from some set $\mathcal{H}$ with

$$\hat{h}_{\text{opt}} = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{n} \left\{ \frac{\hat{\xi}(\mathbf{Z}_i;\hat{\pi},\hat{\mu}) - \hat{\theta}_h(A_i)}{1-\hat{W}_h(A_i)} \right\}^2,$$

where $\hat{W}_h(a_i) = (1,0)\mathbb{P}_n\{\mathbf{g}_{ha_i}(A)K_{ha_i}(A)\mathbf{g}_{ha_i}(A)^{\mathrm{T}}\}^{-1}(1,0)^{\mathrm{T}}h^{-1}K(0)$ is the ith diagonal of the so-called smoothing or hat matrix. The properties of this approach can be derived by using logic which is similar to that in theorem 3, e.g. by adapting results from Li and Racine (2004). Alternatively we could split the sample, estimate the nuisance functions in one half and then do leave-one-out cross-validation in the other half, treating the pseudo-outcomes estimated in the other half as known. We expect that these approaches will be asymptotically equivalent to an oracle selector.

An alternative option would be to use the k-fold cross-validation approach of van der Laan and Dudoit (2003) or Díaz and van der Laan (2013). This would entail randomly splitting the data into k parts, estimating the nuisance functions and the effect curve on $k-1$ training folds, using these estimates to compute measures of risk on the kth test fold and then repeating across all k folds and averaging the risk estimates. One would then repeat this process for each bandwidth value h in some set $\mathcal{H}$ and pick that which minimized the estimated cross-validated risk. van der Laan and Dudoit (2003) gave finite sample and asymptotic results showing that the resulting estimator behaves similarly to an oracle estimator that minimizes the true unknown cross-validated risk. Unfortunately this cross-validation process can be more computationally intensive than the above leave-one-out method, especially if the nuisance functions are estimated with flexible computationally heavy methods. However, this approach will be crucial when incorporating general machine learning and moving beyond linear kernel smoothers.

4. Simulation study

We used simulation to examine the finite sample properties of our proposed methods. Specifically we simulated from a model with normally distributed covariates

$$\mathbf{L} = (L_1,\ldots,L_4)^{\mathrm{T}} \sim N(0,\mathbf{I}_4),$$

beta-distributed exposure

$$(A/20)|\mathbf{L} \sim \text{beta}\{\lambda(\mathbf{L}), 1-\lambda(\mathbf{L})\},$$
$$\text{logit}\{\lambda(\mathbf{L})\} = -0.8 + 0.1L_1 + 0.1L_2 - 0.1L_3 + 0.2L_4$$

and a binary outcome

$$Y|\mathbf{L}, A \sim \text{Bernoulli}\{\mu(\mathbf{L}, A)\},$$
$$\text{logit}\{\mu(\mathbf{L}, A)\} = 1 + (0.2, 0.2, 0.3, -0.1)\mathbf{L} + A(0.1 - 0.1L_1 + 0.1L_3 - 0.13^2A^2).$$

This set-up roughly matches the data example from the next section. Fig. 2 shows a plot of the effect curve $\theta(a) = \mathbb{E}\{\mu(\mathbf{L}, a)\}$ that is induced by the simulation set-up, along with treatment *versus* outcome data for one simulated data set (with $n = 1000$).

To analyse the simulated data we used three different estimators: a marginalized regression (plug-in) estimator $\hat{m}(a) = \mathbb{P}_n\{\hat{\mu}(\mathbf{L}, a)\}$ and two different versions of the local linear kernel estimator proposed. Specifically we used an inverse-probability-weighted approach which was first developed by Rubin and van der Laan (2006b), which relies solely on a treatment model estimator $\hat{\pi}$ (equivalent to setting $\hat{\mu} = 0$), and the standard doubly robust version that used both estimators $\hat{\pi}$ and $\hat{\mu}$. To model the conditional treatment density π we used logistic regression to estimate the parameters of the mean function $\lambda(\mathbf{l})$; we separately considered correctly specifying this mean function, and then also misspecifying the mean function by transforming the covariates with the same covariate transformations as in Kang and Schafer (2007). To estimate the outcome model μ we again used logistic regression, considering a correctly specified model and then a misspecified model that used the same transformed covariates as with π and also left out the cubic term in a (but kept all other interactions). To select the bandwidth we used the leave-one-out approach that was proposed in Section 3.5, which treats the pseudo-outcomes as

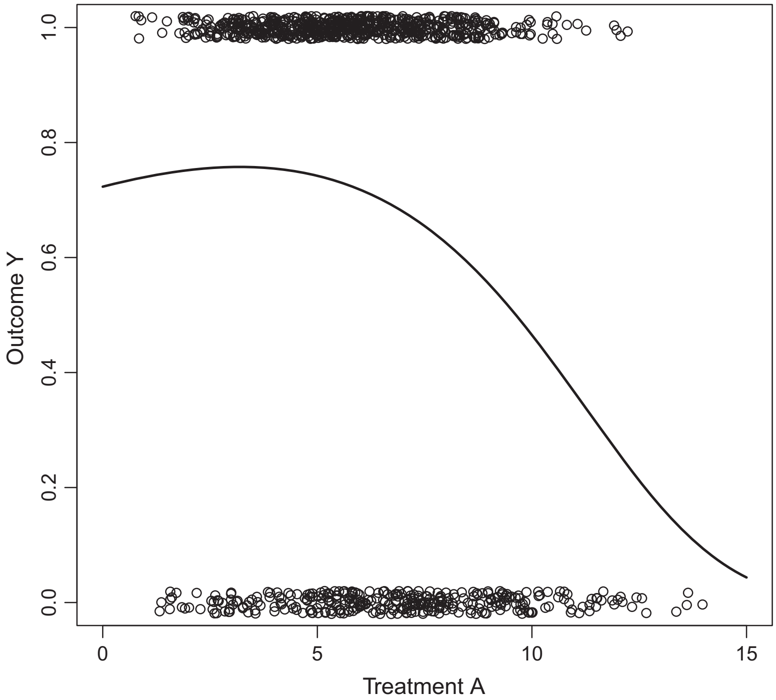


Fig. 2. Plot of effect curve $\theta(a)$ induced by the simulation set-up (———) with treatment and outcome data (○) from one simulated data set with $n = 1000$

known. For comparison we also considered an oracle approach that picked the bandwidth by minimizing the oracle risk $\mathbb{P}_n[\{\theta(A)-\hat{\theta}_h(A)\}^2]$. In both cases we found the minimum bandwidth value in the range $\mathcal{H}=[0.01, 50]$ by using numerical optimization.

We generated 500 simulated data sets for each of three sample sizes: $n=100$, 1000, 10000. To assess the quality of the estimates across simulations we calculated the empirical bias and root-mean-squared error at each point and integrated across the function with respect to the density of A. Specifically, letting $\hat{\theta}_s(a)$ denote the estimated curve at point a in simulation s ($s=1,\ldots,S$ with $S=500$), we estimated the integrated absolute mean bias and root-mean-squared error with

$$\widehat{\text{bias}} = \int_{\mathcal{A}^*} \left| \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}_s(a) - \theta(a) \right| \varpi(a)\,\mathrm{d}a,$$

$$\widehat{\text{RMSE}} = \int_{\mathcal{A}^*} \left[\frac{1}{S} \sum_{s=1}^{S} \{\hat{\theta}_s(a) - \theta(a)\}^2 \right]^{1/2} \varpi(a)\,\mathrm{d}a.$$

In these equations $\mathcal{A}^*$ denotes a trimmed version of the support of A, excluding 10% of mass at the boundaries. The integrands (except for the density) correspond to the usual definitions of absolute mean bias and root-mean-squared error for estimation of a single scalar parameter (e.g. the curve at a single point).

The simulation results are given in Table 1 (both the integrated bias and the root-mean-squared error are multiplied by 100 for easier interpretation). The estimators indicated are those with bandwidths selected by using the oracle risk. When both $\hat{\pi}$ and $\hat{\mu}$ were misspecified, all estimators gave substantial integrated bias and mean-squared errors (although the doubly robust estimator consistently performed better than the other estimators in this setting). Similarly, all estimators had relatively large mean-squared errors in the small sample size setting ($n=100$) because of a lack of precision, although differences in bias were similar to those at moderate and

Table 1. Integrated mean bias and root-mean-squared error (in parentheses) after 500 simulations

n	*Method*	*Results when correct model is as follows:*			
		Neither	*Treatment*	*Outcome*	*Both*
100	Regression	2.67 (5.54)	2.67 (5.54)	0.62 (5.25)	0.62 (5.25)
	Inverse probability weighted	2.26 (8.49)	1.64 (8.57)	2.26 (8.49)	1.64 (8.57)
	Inverse probability weighted†	2.26 (7.36)	1.58 (7.37)	2.26 (7.36)	1.58 (7.37)
	Doubly robust	2.23 (6.27)	1.01 (6.28)	1.12 (5.92)	1.10 (6.50)
	Doubly robust†	2.12 (5.48)	1.00 (5.36)	1.03 (5.08)	1.02 (5.65)
1000	Regression	2.62 (3.07)	2.62 (3.07)	0.06 (1.53)	0.06 (1.53)
	Inverse probability weighted	2.38 (3.97)	0.86 (2.94)	2.38 (3.97)	0.86 (2.94)
	Inverse probability weighted†	2.11 (3.44)	0.70 (2.34)	2.11 (3.44)	0.70 (2.34)
	Doubly robust	2.03 (3.11)	0.75 (2.39)	0.74 (2.53)	0.68 (2.25)
	Doubly robust†	1.84 (2.67)	0.64 (1.88)	0.61 (1.78)	0.58 (1.78)
10000	Regression	2.65 (2.70)	2.65 (2.70)	0.02 (0.47)	0.02 (0.47)
	Inverse probability weighted	2.36 (3.42)	0.33 (1.09)	2.36 (3.42)	0.33 (1.09)
	Inverse probability weighted†	2.24 (3.28)	0.35 (0.85)	2.24 (3.28)	0.35 (0.85)
	Doubly robust	1.81 (2.35)	0.26 (0.86)	0.20 (1.21)	0.25 (0.78)
	Doubly robust†	1.76 (2.27)	0.31 (0.68)	0.24 (1.10)	0.29 (0.64)

†Uses the oracle bandwidth.

small sample sizes ($n = 1000, 10000$). Specifically the regression estimator gave a small bias when $\hat{\mu}$ was correct and large bias when $\hat{\mu}$ was misspecified, whereas the inverse-probability-weighted estimator gave a small bias when $\hat{\pi}$ was correct and large bias when $\hat{\pi}$ was misspecified. However, the doubly robust estimator gave a small bias as long as either $\hat{\pi}$ or $\hat{\mu}$ was correctly specified, even if one was misspecified.

The inverse-probability-weighted estimator was least precise, although it had smaller mean-squared error than the misspecified regression estimator for moderate and large sample sizes. The doubly robust estimator was roughly similar to the inverse-probability-weighted estimator when the treatment model was correct, but it gave less bias and was more precise, and was similar to the regression estimator when the outcome model was correct (but slightly more biased and less precise). In general the estimators that are based on the oracle-selected bandwidth were similar to those using the simple leave-one-out approach but gave marginally less bias and mean-squared error for small and moderate sample sizes. The benefits of the oracle bandwidth were relatively diminished with larger sample sizes.

5. Application

In this section we apply the proposed methodology to estimate the effect of nurse staffing on hospital readmissions penalties, as discussed in Section 1. Originally McHugh *et al.* (2013) used a matching approach to control for hospital differences and found that hospitals with more nurse staffing were less likely to be penalized; this suggests increasing nurse staffing to help to curb excess readmissions. However, their analysis considered the effect of higher nurse staffing *versus* lower nurse staffing only and did not explore the full effect curve; it also relied solely on matching for covariate adjustment, i.e. it was not doubly robust.

In this analysis we use the proposed kernel smoothing approach to estimate the full effect curve flexibly, while also allowing for doubly robust covariate adjustment. We use the same data on 2976 acute care hospitals as in McHugh *et al.* (2013); full details are given there. The covariates $\mathbf{L}$ include hospital size, teaching intensity, not-for-profit status, urban *versus* rural location, patient race proportions, proportion of patients on Medicaid, average socio-economic status, operating margins, a measure of market competition and whether open-heart or organ transplant surgery is performed. The treatment A is nurse staffing hours, measured as the ratio of registered nurse hours to adjusted patient days, and the outcome Y indicates whether the hospital was penalized because of excess readmissions. Excess readmissions are calculated by the Centers for Medicare and Medicaid Services and aim to adjust for the fact that different hospitals see different patient populations. Without unmeasured confounding, the quantity $\theta(a)$ represents the proportion of hospitals that would have been penalized if all hospitals had changed their nurse staffing hours to level a. Otherwise $\theta(a)$ can be viewed as an adjusted measure of the relationship between nurse staffing and readmissions penalties.

For the conditional density $\pi(a|\mathbf{l})$ we assumed a model $A = \lambda(\mathbf{L}) + \gamma(\mathbf{L})\varepsilon$, where ε has mean 0 and unit variance given the covariates but otherwise has an unspecified density. We flexibly estimated the conditional mean function $\lambda(\mathbf{l}) = \mathbb{E}(A|\mathbf{L}=\mathbf{l})$ and variance function $\gamma(\mathbf{l}) = \text{var}(A|\mathbf{L}=\mathbf{l})$ by combining linear regression, multivariate adaptive regression splines, generalized additive models, the lasso and boosting, using the cross-validation-based Super Learner algorithm (van der Laan *et al.*, 2007), to reduce chances of model misspecification. A standard kernel approach was used to estimate the density of ε.

For the outcome regression $\mu(\mathbf{l}, a)$ we again used the Super Learner approach, combining logistic regression, multivariate adaptive regression splines, generalized additive models, the lasso and boosting. To select the bandwidth parameter h we used the leave-one-out approach

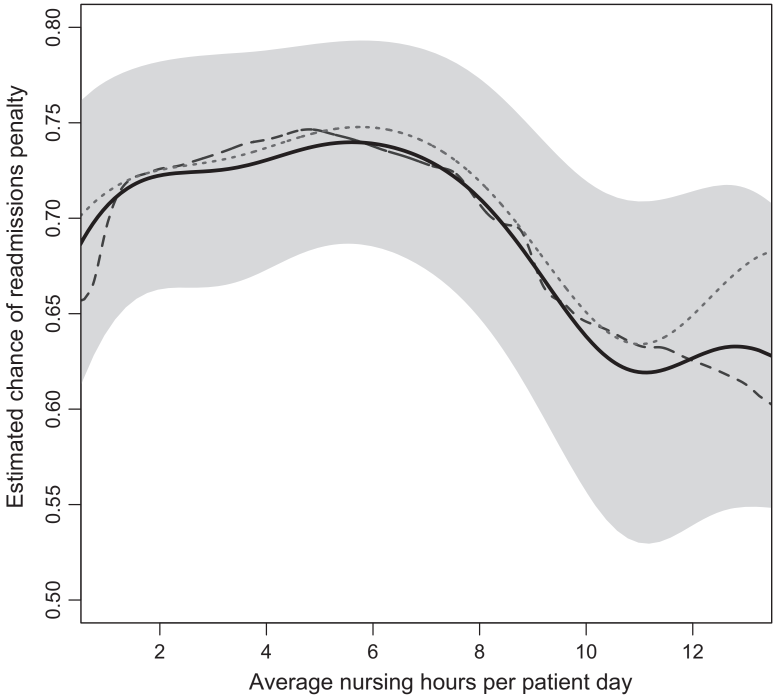


Fig. 3. Estimated effects of nurse staffing on readmissions penalties: — —, regression estimate; ------, inverse-probability-weighted estimate; ———, doubly robust estimate; ■, doubly robust estimate uncertainty interval

that was discussed in Section 3.5. We considered regression, inverse-probability-weighted and doubly robust estimators, as in the simulation study in Section 4. The two hospitals (less than 0.1%) with smallest inverse probability weights were removed as outliers. For the doubly robust estimator we also computed pointwise uncertainty intervals (i.e. confidence intervals around the smoothed parameter $\theta_h^*(a)$; see Section 3.4) by using a Wald approach based on the empirical variance of the estimating function values.

A plot showing the results from the three estimators (with uncertainty intervals for the doubly robust estimator proposed) is given in Fig. 3. In general the three estimators were very similar. For fewer than five average nurse staffing hours the adjusted chance of penalization was estimated to be roughly constant, around 73%, but at 5 h chances of penalization began to decrease, reaching approximately 60% when nurse staffing reached 11 h. This suggests that adding nurse staffing hours may be particularly beneficial in the 5–10-h range, in terms of reducing the risk of readmissions penalization; most hospitals (65%) lie in this range in our data.

6. Discussion

In this paper we developed a novel approach for estimating the average effect of a continuous treatment; importantly the approach allows for flexible doubly robust covariate adjustment without requiring any parametric assumptions about the form of the effect curve, and it can incorporate general machine learning and non-parametric regression methods. We presented a novel efficient influence function for a stochastic intervention parameter defined within a non-parametric model; this influence function motivated the approach proposed but may also be useful to estimate on its own. In addition we provided asymptotic results (including rates of convergence and asymptotic normality) for a particular kernel estimation version of our method, which only require the effect curve to be twice continuously differentiable, and allows for flexible data-adaptive estimation of nuisance functions. These results show the double robustness of the approach, since either a conditional treatment density or outcome regression model can be misspecified and the estimator will still be consistent as long as one such nuisance function is correctly specified. We also showed how double robustness can result in smaller second-order bias even when both nuisance functions are consistently estimated. Finally, we proposed a simple and fast data-driven cross-validation approach for bandwidth selection, found favourable finite sample properties of our proposed approach in a simulation study and applied the kernel estimator to examine the effects of hospital nurse staffing on excess readmissions penalty.

This paper integrates semiparametric (doubly robust) causal inference with non-parametric function estimation and machine learning, helping to bridge the 'huge gap between classical semiparametric models and the model in which nothing is assumed' (van der Vaart, 2014). In particular our work extends standard non-parametric regression by allowing for complex covariate adjustment and doubly robust estimation, and extends standard doubly robust causal inference methods by allowing for non-parametric smoothing. Many interesting problems arise in this gap between standard non-parametric and semiparametric inference, leading to many opportunities for important future work, especially for complex non-regular target parameters that are not pathwise differentiable. In the context of this paper, in future work it will be useful to study uniform distributional properties of our proposed estimator (e.g. weak convergence), as well as its role in testing and inference (e.g. for constructing tests that have power to detect a wide array of deviations from the null hypothesis of no effect of a continuous treatment).

Acknowledgements

Edward Kennedy was supported by National Institutes of Health grant R01-DK090385, Zongming Ma by National Science Foundation Career award DMS-1352060 and Dylan Small by National Science Foundation grant SES-1260782. The authors thank Marshall Joffe and Alexander Luedtke for helpful discussions, and two referees for very insightful comments and suggestions.

References

Andrews, D. W. (1994) Empirical process methods in econometrics. In *Handbook of Econometrics* (eds R. F. Engle and D. L. McFadden), vol. 4, pp. 2247–2294. Amsterdam: Elsevier.

Bickel, P. J., Klaassen, C. A., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Díaz, I. and van der Laan, M. J. (2012) Population intervention causal effects based on stochastic interventions. *Biometrics*, **68**, 541–549.

Díaz, I. and van der Laan, M. J. (2013) Targeted data adaptive estimation of the causal dose-response curve. *J. Causl Inf.*, **1**, 171–192.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. Boca Raton: CRC Press.

Galvao, A. F. and Wang, L. (2015) Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *J. Am. Statist. Ass.*, **110**, 1528–1542.

Gill, R. D. and Robins, J. M. (2001) Causal inference for complex longitudinal data: the continuous case. *Ann. Statist.*, **29**, 1785–1811.

Györfi, L., Kohler, M., Krzykaz, A. and Walk, H. (2002) *A Distribution-free Theory of Nonparametric Regression*. New York: Springer.

Härdle, W., Hall, P. and Marron, J. S. (1988) How far are automatically chosen regression smoothing parameters from their optimum? *J. Am. Statist. Ass.*, **83**, 86–95.

Hill, J. L. (2011) Bayesian nonparametric modeling for causal inference. *J. Computnl Graph. Statist.*, **20**, 217–240.

Hirano, K. and Imbens, G. W. (2004) The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives* (eds A. Gelman and X.-L. Meng), pp. 73–84. New York: Wiley.

Horowitz, J. L. (2009) *Semiparametric and Nonparametric Methods in Econometrics*. New York: Springer.

Imai, K. and van Dyk, D. A. (2004) Causal inference with general treatment regimes. *J. Am. Statist. Ass.*, **99**, 854–866.

Imbens, G. W. (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.*, **86**, 4–29.

Kang, J. D. and Schafer, J. L. (2007) Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.*, **22**, 523–539.

van der Laan, M. J. and Dudoit, S. (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. *Working Paper 130*. Division of Biostatistics, University of California at Berkeley, Berkeley.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) Super Learner. *Statist. Appl. Genet. Molec. Biol.*, **6**, article 25.

van der Laan, M. J. and Robins, J. M. (1998) Locally efficient estimation with current status data and time-dependent covariates. *J. Am. Statist. Ass.*, **93**, 693–701.

van der Laan, M. J. and Robins, J. M. (2003) *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

van der Laan, M. J. and Rose, S. (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.

van der Laan, M. J. and Yu, Z. (2001) Comments on inference for semiparametric models: some questions and an answer. *Statist. Sin.*, **11**, 910–917.

Li, Q. and Racine, J. S. (2004) Cross-validated local linear nonparametric regression. *Statist. Sin.*, **14**, 485–512.

Li, Q. and Racine, J. S. (2007) *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.

McHugh, M. D., Berez, J. and Small, D. S. (2013) Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing. *Hlth Aff.*, **32**, 1740–1747.

Neugebauer, R. and van der Laan, M. J. (2007) Nonparametric causal effects based on marginal structural models. *J. Statist. Plannng Inf.*, **137**, 419–434.

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* (eds M. E. Halloran and D. Berry), pp. 95–133. New York: Springer.

Robins, J. M. and Rotnitzky, A. (2001) Comments on inference for semiparametric models: some questions and an answer. *Statist. Sin.*, **11**, 920–936.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Rubin, D. B. and van der Laan, M. J. (2005) A general imputation methodology for nonparametric regression with censored data. *Working Paper 194*. Division of Biostatistics, University of California at Berkeley, Berkeley.

Rubin, D. B. and van der Laan, M. J. (2006a) Doubly robust censoring unbiased transformations. *Working Paper 208*. Division of Biostatistics, University of California at Berkeley, Berkeley.

Rubin, D. B. and van der Laan, M. J. (2006b) Extending marginal structural models through local, penalized, and additive learning. *Working Paper 212*. Division of Biostatistics, University of California at Berkeley, Berkeley.

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.

van der Vaart, A. W. (2000) *Asymptotic Statistics*. New York: Cambridge University Press.

van der Vaart, A. W. (2014) Higher order tangent spaces and influence functions. *Statist. Sci.*, **29**, 679–686.

van der Vaart, A. W. and van der Laan, M. J. (2006) Estimating a survival distribution with current status data and high-dimensional covariates. *Int. J. Biostatist.*, **2**, 1–40.

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

Wang, L., Rotnitzky, A. and Lin, X. (2010) Nonparametric regression with missing outcomes using weighted kernel estimating equations. *J. Am. Statist. Ass.*, **105**, 1135–1146.

Wasserman, L. (2006) *All of Nonparametric Statistics*. New York: Springer.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Nonparametric methods for doubly robust estimation of continuous treatments: Web Appendix’.