36-732 Final Project: Estimation of the Continuous Treatment Effect

Wanshan Li*

Department of Statistics& Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
wanshanl@andrew.cmu.edu

1 Introduction

Continuous causal effect, arising often in practice, is usually described by curves. This makes the estimation more difficult than the binary or multilevel treatments. Researchers are seeking for estimators for continuous treatment effect that are flexible to use and adaptive for high-dimensional confounders. Many different estimators and methods have been proposed towards this goal. According to the strength of assumptions, we can roughly classify them to three groups.

The first group need the strongest assumptions because they require correct specification of the outcome model to build the regression model of outcomes on covariates and treatment. The estimator based on the marginal structural model we discussed in class is one example in this group.

The second group make weaker assumptions because they do not directly specify the outcome model, but still need at least one correct model related to the outcome. There are two subgroups, one based on propensity score and the other based on the doubly robust estimator.

The former one, including [4], [5] and [2], assume that at least one conditional density related to the outcome is known. [4] propose a generalized propensity score for continuous treatment and use it to study the treatment effect. Since they does not discuss the non-parametric estimation of the conditional treatment density or the propensity score, they impose a parametric model for the conditional treatment density in their numerical experiments and real data analysis, though their method does not depend on the specific form of the propensity score. [2] discuss more general estimators based on semi-parametric models. Though their methodology allow more flexible estimators, they does not discuss and implement them and just use parametric models in their numerical parts, because they find the conditions for non-parametric estimators would be too strong.

The later one either assume a parametric dose-response curve ([9],[13]) or a working model that approximates the truth well ([8]). One example is the doubly robust estimator based on the marginal structural model we discussed in class. Estimators in this group have the nice property of double robustness, but still need some parametric specification of the outcome model or some approximation.

The third group, including [10], [11], [12], [1], [7], consider estimation and inference under nonparametric settings. These methods does not require parametric models, thus make more sense in general applications. Except for [12], they consider doubly robust estimation under the non-parametric setting, which allows for even more general model assumptions. However, the doubly robust estimator in [10], [11] are for weighted average dose-response functions, and the risk of some estimator in [1]. [7], the most recent work, considers the doubly robust estimator for the average dose-response function under nonparametric settings.

In this report, I will focus on [4], [5], [2], and [7].

*

xxst Conference on Neural Information Processing Systems (NIPS 20xx), xxxx, xx, xxx.

2 Discussion on Different Estimators/Mehtods

2.1 Methods Based on Specific Outcome Model

These methods include the estimators based on the marginal structural model $m(a;\beta)$ we discussed in class. For example in the IPW-type MSM estimator, we assume a marginal structural model $\mathbb{E}Y(a) = m(a;\beta)$ where β is a vector in a finite dimensional space, and want to find the solution $\hat{\beta}$ to

$$\mathbb{P}_n\left[h(A)\left\{\frac{Y - m(A;\beta)}{\pi(A|\mathbf{X})}\right\}\right] = 0.$$

We also discussed the doubly robust estimator based on MSM. However, the assumption of a specific parametric model for $m(a;\beta)$ can be too strong to use in reality.

2.2 Methods Based on Propensity Score

The fundamental paper is [4] and [5]. They generalize the unconfoundedness assumption for binary treatments and propose estimators based on that. [5] propose a general framework for estimation while [4] used some specific models to show how their method works. In both papers, they assume the following unconfoundedness of the treatment, and show some key properties of the generalized propensity score, given by theorem 2.1, 2.2 and 2.3.

- **Weak Unconfoundedness**. $Y(a) \perp\!\!\!\perp A|\mathbf{X}$ for all $a \in \mathcal{A}$.

The generalized propensity score is defined as

Definition 2.1 (Generalized Propensity Score). *Let $r(a,\mathbf{x})$ be the conditional density of the treatment given the covariates: $r(a,\mathbf{x}) = f_{A|\mathbf{X}}(a|\mathbf{x})$. Then the generalized propensity score is $R = r(A,\mathbf{X})$.*

The generalized propensity score has nice properties similar to the propensity score for binary treatment. For instance, we have

Theorem 2.1. *Suppose that assignment to the treatment is weakly unconfounded given pre-treatment variables $\mathbf{X}$. Then, for every a,*

$$f_A(a|\mathbf{X}) = f_A(a|r(a,\mathbf{X}),\mathbf{X}) = f_A(a|r(a,\mathbf{X})).$$

Loosely speaking, this result means that $\mathbf{X} \perp\!\!\!\perp \mathbb{1}\{A = a\}|r(a,\mathbf{X})$. Using it we can derive the following result.

Theorem 2.2. *Suppose that assignment to the treatment is weakly unconfounded given pre-treatment variables $\mathbf{X}$. Then, for every a,*

$$f_A(a|r(a,\mathbf{X}),Y(a)) = f_A(a|r(a,\mathbf{X})).$$

Such property provide us a convenient approach to estimate the treatment effect or the dose-response function (DRF) $\mu(a) = \mathbb{E}\{Y(a)\}$ at a particular level a. We first define

$$\beta(a,r) = \mathbb{E}(Y|A = a, R = r),$$

and then average $\beta(a,r)$ over the GPS

$$\mu(a) = \mathbf{E}\{\beta(a,r(a,\mathbf{X}))\}.$$

The support for such procedure is the following theorem:

Theorem 2.3. *Suppose the weak unconfoundedness assumption holds. Then*

$$(i)\ \beta(a,r) = \mathbb{E}\{Y(a)|r(a,\mathbf{X}) = r\} = \mathbb{E}\{Y|A = a, R = r\}.$$
$$(ii)\ \mu(a) \equiv \mathbb{E}\{Y(a)\} = \mathbf{E}\{\beta(a,r(a,\mathbf{X}))\}.$$

Their methods loose the restriction of a specific model on the outcome to a specific model on the conditional density or the propensity score. But such a specific model is still too strong and, in reality, one need to specify a model for the outcome first to get the model for the conditional density and the propensity score, like what they do in [4].

2.2.1 Semi-parametric Estimator

[2] proposed a semi-parametric estimator. They propose a general framework for generic moment restriction estimators (Z-estimators). In this framework, one needs to firstly define a generalized residual function $m(Y(a); \beta(a))$. Then the dose-response function at a is defined as the value $\beta(a) \in B \subset \mathbb{R}$ that solves the following moment condition

$$\mathbb{E}\{m(Y(a); \beta(a))\} = 0. \tag{1}$$

To ensure that $\beta(a)$ is uniquely-defined, we need to assume that $\beta(a)$, $a \in \mathcal{A}$, uniquely solve the identifying conditions above. Such a definition is quite general and can cover two commonly used cases:

Example 2.1. *Average and Quantile Dose-Response Functions*

- *Let $m(Y(a); \beta(a)) = Y(a) - \mu(a)$. Now the solution to the moment condition is $\mu_0(a) = \mathbb{E}\{Y(a)\}$, the average dose-response function.*
- *Let $m(Y(a); q_\tau(a)) = \tau - \mathbb{1}\{Y(a) < q_\tau(a)\}$, then the solution is $q_{\tau 0}(a) \in \inf\{q : \mathbb{P}(Y(a) \leq q) \geq \tau\}$, the unconditional q-th quantile dose-response function.*

To make their quantity of interest identifiable, they need following conditions

I.I For each $a \in \mathcal{A}$, $\boldsymbol{\beta}_0 a(a)$ uniquely solves $\mathbb{E}\{m(Y(a); \beta(a))\} = 0$, where $, : \mathbb{R} \times B \to \mathbb{R}$ is measurable.

I.II For all $a \in MA$, we have:

1. $Y(a) \perp\!\!\!\perp A|\mathbf{X}$;
2. $f_{A|\mathbf{X},Y}(a|\mathbf{x}, y) > 0$ for $a \in \mathcal{A}$, $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$.

I.III Assume that

1. There exists a function $e(y)$ with $\int e(y)dy < \infty$ such that $|m(y; \beta(a_0)) f_{A,Y|\mathbf{X}}(t_0 + \Delta t, t|\mathbf{x})| \leq |e(y)|$.
2. $\mathbb{E}[m(Y; \beta(t_0))|\mathbf{X}, T = t_0] = \lim_{\Delta t \downarrow 0} \mathbb{E}[m(Y; \beta(t_0))|\mathbf{X}, T \in [t_0, t_0 + \Delta t]]$. $\mathcal{A}$ is right open.

Notice that the ignorability condition $Y(a) \perp\!\!\!\perp A|\mathbf{X}$ here is stronger than the one in [7]. This is because here they need to consider general form of the function $m(Y(a); \beta(a))$. By looking at their proof one can see that if we only consider the average DRF, we can instead use the weaker condition $\mathbb{E}(Y(a)|\mathbf{X}, A) = \mathbb{E}(Y(a)|\mathbf{X})$. Denote $\mathbf{u} = (\mathbf{x}', y)'$ and $\mathbf{U} = (\mathbf{X}', Y)'$. There main theorem for the identification of the estimator is

Theorem 2.4. *Under conditions I-III, for each $a \in \mathcal{A}$, we have*

$$\mathbb{E}\{m(Y(a); \beta(a))\} = \mathbb{E}\{m(Y(a); \beta(a))\omega_0(\mathbf{U}; a)\},$$

where $\omega(\mathbf{u}; a) \equiv \frac{f_{A|\mathbf{X},Y}(a|\mathbf{x},y)}{f_{A|\mathbf{X}}(a|\mathbf{x})}$. Consequently, $\mathbb{E}\{m(Y(a); \beta(a))\omega_0(\mathbf{u}; a)\} = 0$ if and only if $\beta(a) = \beta_0(a)$.

By this theorem, to estimate $\beta(a)$ we just need two steps:

1. Estimate $\omega_0(\mathbf{Z}; a) = \frac{f_{A|\mathbf{X},Y}(a|\mathbf{x},y)}{f_{A|\mathbf{X}}(a|\mathbf{x})}$ and obtain an estimator $\hat{\pi}$.
2. For each $a \in \mathcal{A}$, find $\hat{\beta}(a)$ as a zero of the following condition

$$\mathbb{P}_n\{m(Y(a); \beta(a))\omega_0(\mathbf{U}; a)\} = \frac{1}{n}\sum_{i=1}^{n} m(Y_i(a); \beta(a))\hat{\omega}_0(\mathbf{U}_i; a) = 0. \tag{2}$$

As an example, the resulting estimators for the average and quantile dose-response functions are

Example 2.2. *Estimators*

- *Average DRF. Let $m(Y(a); \beta(a)) = Y(a) - \mu(a)$. Then*

$$\hat{\mu}(a) = \frac{\mathbb{P}_n\{\hat{\omega}(\mathbf{X}, Y, a)Y\}}{\mathbb{P}_n(\hat{\omega}(\mathbf{X}, Y, a))}$$

where $\hat{\omega}$ is the estimator of $\omega(\mathbf{U}, a)$.

- *Quantile DRF. Let* $m(Y(a); q_\tau(a)) = \tau - \mathbb{1}\{Y(a) < q_\tau(a)\}$. *Then*

$$\hat{q_\tau}(a) = \underset{q}{\operatorname{argmin}}\, \mathbb{P}_n \hat{\omega}_0(\mathbf{U}; a)\rho_\tau(Y - q),$$

where $\rho_\tau(u) = u(\tau - \mathbb{1}\{u \leq 0\})$ *is the check function often used in relevant literature.*

Remark One can see that the estimator $\hat{\mu}$ is quite similar to the IPW-type estimator based on the MSM, and the only difference is that in MSM the parameter β is a finite dimensional vector, but here β (i.e., μ) can be a parameter function living in infinite dimensional space. This is one of the reason why their result is more general and require more involved theory.

2.2.2 Asymptotic Properties

To ensure the consistency of $\hat{\beta}$, they list five conditions C.I-C.V.

Theorem 2.5 (Consistency)**.** *Suppose* $\mathbb{E}\{m(Y(a); \beta_0(a))\omega_0(\mathbf{u}; a)\} = 0$ *and conditions C.I-C.V hold. Then as* $n \to \infty$

$$\sup_{a \in \mathcal{A}} |\hat{\beta}(a) - \beta_0(a)| = o_{p^*}(1). \tag{3}$$

For the average DRF and quantile DRF, the conditions C.I-C.V can be simplified to AC.I, AC.II plus C.IV and QC.I-QC.III plus C.IV, respectively.

They list six conditions G.I-G.VI for the weak convergence of $\hat{\beta}$.

Theorem 2.6 (Weak Convergence)**.** *Suppose that* $|\mathbb{E}\{m(Y(a); \beta_0(a))\omega_0(\mathbf{u}; a)\}|_\infty = 0$*, that* $|\hat{\beta} - \beta_0|_\infty = o_{p^*}(1)$*, and that conditions G.I-G.VI are satisfied. Then in* $\ell^\infty(\mathcal{A})$*,*

$$\sqrt{nr_n}(\hat{\beta}(a) - \beta_0(a)) \rightsquigarrow Z_1^{-1}(\beta_0, \omega_0(\mathbf{U}; a))\mathbb{G}(a). \tag{4}$$

2.3 Nonparametric Methods

[10], [11], [12], [1] mainly focus on the non-parametric methods. Under the non-parametric settings, as we mentioned in class, even when there is no covariates, the convergence rate of the estimator $\hat{\theta}(a)$ to $\theta(a)$ can be much slower $O(n^{-\frac{2s}{2s+d}})$ than the parametric model ($O(n^{-1/2})$), especially for high-dimensional covariates. I will omit the details of these four papers for the ease of writing.

2.4 Doubly Robust Estimator under Nonparametric Settings

[7] proposed a doubly robust estimator that does require parametric assumptions for both the outcome and treatment processes. They just make some standard assumptions in Causal and nonparametric inference. The causality assumptions are for the identification of the average dose-response function $\theta(a) \equiv \mathbb{E}Y(a)$, including

- Assumption 1. Consistency: $A = a$ implies $Y = Y^a$.
- Assumption 2. Positivity: $\pi(a|\mathbf{x}) \geq \pi_{\min} > 0$ for all $\mathbf{x} \in \mathcal{X}$.
- Assumption 3. Ignorability: $\mathbb{E}(Y(a)|\mathbf{X}, A) = \mathbb{E}(Y(a)|\mathbf{X})$.

Notice that the ignorability assumption here is weaker than that in [2], because in terms of the notation in [2], here they just consider the special case $m(Y(a); \beta(a)) = Y(a) - \mu(a)$, i.e., the average dose-response function. Also, the positivity condition can be a strong condition in the continuous treatment regime.

Under assumption 1-3, the effect curve $\theta(a)$ can be identified as

$$\theta(a) = \mathbb{E}\{\mu(\mathbf{X}, a)\} = \int_{\mathcal{X}} \mu(\mathbf{x}, a)\mathrm{d}P(\mathbf{x}). \tag{5}$$

can be identified under the three assumptions as

$$\theta(a) = \mathbb{E}\{\mu(\mathbf{X}, a)\} = \int_{\mathcal{X}} \mu(\mathbf{x}, a)\mathrm{d}P(\mathbf{x}). \tag{6}$$

If $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})\} = \theta(a)$ then $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})\} = \psi$ where

$$\psi = \int_{\mathcal{A}} \int_{\mathcal{X}} \mu(\mathbf{x}, a)\lambda(a)\mathrm{d}P(\mathbf{x})\mathrm{d}a. \tag{7}$$

Theorem 2.7. *Under a non-parametric model, the efficient influence function for ψ defined in equation* (7) *is $\xi(\mathbf{Z};\pi,\mu) - \psi + \int_{\mathcal{A}}\{\mu(\mathbf{X},a) - \int_{\mathcal{X}} \mu(\mathbf{x},a)\mathrm{d}P(\mathbf{x})\}]\lambda(a)\mathrm{d}a$, where*

$$\xi(\mathbf{Z};\pi,\mu) = \frac{Y - \mu(\mathbf{X},A)}{\pi(A|\mathbf{X})} \int_{\mathcal{X}} \pi(A|\mathbf{x})\mathrm{d}P(\mathbf{x}) + \int_{\mathcal{X}} \mu(\mathbf{x},A)\mathrm{d}P(\mathbf{x}). \tag{8}$$

It can be proved that the function $\xi(\mathbf{Z};\pi,\mu)$ satisfy the double-robustness property, or $\mathbb{E}\{\xi(\mathbf{Z};\bar{\pi},\bar{\mu})|A = a\} = \theta(a)$ if either $\bar{\pi} = \pi$ or $\bar{\mu} = \mu$. Such property motivates a regression based approach to estimating the effect curve $\theta(a)$. Specifically we may estimate the nuisance functions (π,μ) first and then regress the estimated $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$ on A. The plug-in estimator $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$ is given by

$$\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) = \frac{Y - \mu(\mathbf{X},A)}{\hat{\pi}(A|\mathbf{X})} \int_{\mathcal{X}} \hat{\pi}(A|\mathbf{x})\mathrm{d}\mathbb{P}_n(\mathbf{x}) + \int_{\mathcal{X}} \hat{\mu}(\mathbf{x},A)\mathrm{d}\mathbb{P}_n(\mathbf{x}). \tag{9}$$

We can get an estimator for $\theta(a)$ by

1. Get the estimated nuisance functions $\hat{\pi}$ and $\hat{\mu}$ and the plug-in pseudo-outcome $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$.
2. Regress $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$ on treatment variable A

The regression can be conducted by off-the-shelf methods in non-parametric regression or machine learning literature. In summary, the double robustness of $\hat{\theta}(a)$ relies on $\xi(\mathbf{Z};\pi,\mu)$ and the "non-parametric method" refers to the non-parametric regression in step 2.

To shed some light on the intuition of this approach, let's compare it with the doubly robust estimator for binary treatment:

$$\hat{\mathbb{E}}(Y^1) = \mathbb{P}_n\left[\frac{Y - \hat{\mu}_A(\mathbf{X})}{\hat{\pi}(\mathbf{X})}A + \hat{\mu}_1(\mathbf{X})\right] = \frac{1}{n}\sum_{A_i=1} \frac{Y_i - \hat{\mu}_1(\mathbf{X}_i)}{\hat{\pi}(\mathbf{X}_i)} - \mathbb{P}_n\hat{\mu}_1(\mathbf{X}).$$

Let $A = 1$ in $\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu})$, we would get:

$$\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) = \frac{Y - \hat{\mu}_1(\mathbf{X})}{\hat{\pi}(\mathbf{X})}\mathbb{P}_n\hat{\pi}(\mathbf{X}) + \mathbb{P}_n\hat{\mu}_1(\mathbf{X}).$$

Suppose $\#\{i : A_i = 1\} = m$, then by iterated expectation we have $\mathbb{P}_n\hat{\pi}(\mathbf{X}) = \mathbb{P}_n A = m/n$. Now regressing $\hat{\xi}$ on A, we get

$$\hat{\theta}(1) = \frac{1}{m}\cdot\frac{m}{n}\sum_{A_i=1} \frac{Y_i - \hat{\mu}_1(\mathbf{X}_i)}{\hat{\pi}(\mathbf{X}_i)} - \mathbb{P}_n\hat{\mu}_1(\mathbf{X}),$$

which is exactly the doubly robust estimator in the binary case.

[7] discuss the local linear kernel regression for step 2, the estimator is $\hat{\theta}_h(a) = \mathbf{g}_{ha}(a)'\hat{\boldsymbol{\beta}}_h(a)$, where $\mathbf{g}_{ha}(a) = (1, (t-a)/h)'$ and

$$\hat{\boldsymbol{\beta}}_h(a) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\operatorname{argmin}}\, \mathbb{P}_n[K_{ha}(A)\{\hat{\xi}(\mathbf{Z};\hat{\pi},\hat{\mu}) - \mathbf{g}_{ha}(a)'\hat{\boldsymbol{\beta}}_h(a)\}], \tag{10}$$

for $K_{ha} = h^{-1}K\{(t-a)/h\}$ with K a standard kernel function and h a scalar bandwidth parameter.

2.4.1 Asymptotic Properties

Theorem 2.8 (Consistency of kernel estimator)**.** *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. We make the following three (sets of) assumptions:*

(a) Positivity.

(b) Either $\bar{\pi} = \pi$ or $\bar{\mu} = \mu$.

(c) Standard regularity conditions for kernel estimation.

Then

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

where

$$\sup_{t:|t-a|\leq h} \|\hat{\pi}(t|\mathbf{X}) - \pi(t|\mathbf{X})\| = O_p\{r(n)\},$$
$$\sup_{t:|t-a|\leq h} \|\hat{\mu}(t,\mathbf{X}) - \mu(t,\mathbf{X})\| = O_p\{s(n)\}$$

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

The "standard regularity conditions" used in this theorem are

1. The bandwidth $h = h_n$ satisfies $h \to 0$ and $nh^3 \to \infty$ as $n \to \infty$.
2. K is a continuous symmetric probability density with support $[-1, 1]$.
3. $\theta(a)$ is twice continuously differentiable, and both $\lambda(a)$ and the conditional density of $\xi(\mathbf{Z}; \bar{\pi}, \bar{\mu})$ given $A = a$ are continuous as functions of a.
4. 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, with $1/\hat{\pi}$ and $1/\bar{\pi}$ also uniformly bounded.

In their appendix, they also briefly show that under standard smoothness and bandwidth conditions, their estimator $\hat{\theta}(a)$ is uniformly consistent, in the sense that

$$\sup_{a\in\mathcal{A}} |\hat{\theta}(a) - \theta(a)| = O_p(\sqrt{\frac{\log n}{nh}} + h^2 + r_n^* s_n^*), \tag{11}$$

where r_n^* and s_n^* are the error rate of $\hat{\pi}$ and $\hat{\mu}$: $\sup_{a\in\mathcal{A}} \|\hat{\pi}(a|\mathbf{X}) - \pi(a|\mathbf{X})\| = O_p(r_n^*)$ and similar for $\hat{\mu}$.

Theorem 2.9 (Asymptotic normality of kernel estimator)**.** *Consider the same setting as theorem 2.8. Along with the assumptions from theorem 2.8, also assume that*

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

Then

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

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

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

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

3 Conclusion

Until now there are not many papers on the estimation of continuous treatment effect. In some sense that reason is that people face a dilemma: if we assume specific models for $Y(a)$, the assumption of the methodology could be too strong and can hardly fit the real case well; otherwise if we allow much flexibility of the dose-response function, we must introduce a semi-parametric or a non-parametric framework, which can makes the asymptotic behaviour of the estimator worse or make the theoretical analysis more difficult, just as we see in [2] and [7].

The worse asymptotic behaviour of the non-parametric estimator is an unavoidable problem, and is exactly the reason why in [2] they finally use a parametric model in numerical examples. [7] exploits the power of the doubly robust estimator and reduce the impact of the non-parametric rate of $\hat{\pi}$ and $\hat{\mu}$ to some extent.

For the results, [2] provide very complete results, including the uniform consistency and the weak convergence, but also require more conditions than [7]. [7] have estimators with better properties since they are doubly robust, and also provide some basic results on uniform consistency.

References

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

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

[3] J. L. Hill. Bayesian nonparametric modeling for causal inference. *J. Computnl Graph. Statist.*, 20:217–240, 2011.

[4] K. Hirano and G. W. Imbens. The propensity score with continuous treatments. In A. Gelman and X.-L. Meng, editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, pages 73–84. Wiley, New York, 2004.

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

[6] G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Statist.*, 84:4–29, 2004.

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

[8] R. Neugebauer and M. J. van der Laan. Nonparametric causal effects based on marginal structural models. *J. Statist. Planng Inf.*, 137:419–434, 2007.

[9] J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In M. E. Halloran and D. Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. Springer, New York, 2000.

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

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

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

[13] M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality.*, pages 95–133. Springer, 2003.