Learning On Semiparametric Models With Monotone Functions*

Mengshan Xu†
London School of Economics

November 12, 2020

Abstract

This paper studies a semiparametric estimator where the associated moment condition contains a nuisance monotone function estimated by isotonic regression. We show that the properties of the isotonic estimator satisfy the framework of Newey (1994), and that the associated sample moment function with a plug-in isotonic estimator is within a distance of $o_p(n^{-1/2})$ from its Neyman-orthogonalized sample moment function. As a result, the estimator is $\sqrt{n}$-consistent, asymptotically normally distributed, and tuning-parameter-free. Furthermore, in a number of relevant cases, the estimator is efficient.

The estimator we consider generalizes the estimation methods of existing semiparametric models with monotone nuisance functions, such as the monotone partially linear model and monotone single index model. We also apply the estimator to the case of inverse probability weighting, where the propensity scores are assumed to be monotone increasing. Simulations show that while the estimator we develop is more robust against misspecification than parametric plug-in estimators commonly adopted in applied work, it has similar performance to the latter under correct specifications. Compared to methods with other nonparametric plug-in estimators, the newly proposed method requires minimum smoothness conditions on nuisance functions. Furthermore, we establish the asymptotic validity of bootstrap, which ensures that the estimator is tuning-parameter-free in both estimation and inference.

*I am deeply indebted to my advisor, Taisuke Otsu, for his continuous support. I am very grateful to Juan Carlos Escanciano, Javier Hidalgo, Tatiana Komarova, Martin Pesendorfer, Steve Pischke, Chen Qiu, Marcia Schafgans, Canh Thien Dang, John Van Reenen, Stefan Wager, Yike Wang, and Weining Wang for very insightful discussions. I would like to thank Svetlana Chekmasova, Heidi Christina Thysen, Nicola Fontana, Maximilian Guennewig, William Matcham, Tillman Hoenig, and Lukasz Rachel for helpful conversations and comments. I would like to thank participants at CFE-CMStatistics 2019 conference and LSE Econometrics work-in-progress seminars.

†Address: Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK. Email: M.Xu8@lse.ac.uk.

1 Introduction

This paper is concerned with the following semiparametric estimation problem. Suppose we have a moment condition

$$E[m(Z, \beta_0, p_0(\cdot))] = 0, \tag{1}$$

where Z is a random vector defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P}_0)$, and $\beta_0 \in \mathfrak{B} \subset \mathbb{R}^k$ is a real-valued parameter of interest. $p_0(\cdot)$ is a monotone increasing nuisance function, which is the conditional mean of some function of data and β_0. (1) can be an unconditional moment restriction or the first-order condition of a maximization problem. Let $\{Z_i\}_{i=1}^n$ be independent realizations of Z. An estimator $\hat{\beta}$ can be solved from the sample moment condition of (1), with a plugged-in $\hat{p}(\cdot)$:

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta, \hat{p}(\cdot)) = 0, \tag{2}$$

where $\hat{p}(\cdot)$ is an isotonic estimator of $p_0(\cdot)$.

1.1 Motivation and challenges

Without the monotonicity assumption about $p_0(\cdot)$, the model (1) and its plug-in estimator based on (2) have been extensively studied, where $p_0(\cdot)$ is usually estimated by smoothing nonparametric methods such as sieve estimator or kernel estimator. See, e.g., van der Vaart (1991), Newey (1994), Andrews (1994), Ai and Chen (2003), and Chernozhukov et al. (2018), among others. Our interest in the case, where $p_0(\cdot)$ is monotone increasing and estimated by isotonic estimation, is motivated by the following reasons.

First, monotonicity is a natural shape restriction which can be justified in many applications in social science, economic studies, and medical research. Well-known examples in economics are that the demand function is usually monotone decreasing, and the supply function and utility functions are often monotone increasing. Furthermore, many functions derived from CDF functions inherit the monotonicity from the latter. For example, in a binary choice model

$$Y = \begin{cases} 1 & \text{if } X'\beta_0 > \varepsilon \\ 0 & \text{if } X'\beta_0 \leq \varepsilon \end{cases}, \tag{3}$$

we can express the conditional expectation $E(Y|X) = P(Y = 1|X) = F_\varepsilon(X'\beta_0)$, where $F_\varepsilon(\cdot)$ is the CDF of ε. If we assume $\varepsilon \sim N(0, 1)$, (3) becomes a probit model; if we assume $\varepsilon \sim Logistic(0, 1)$, (3) becomes a logit model. If we don't impose any distributional assumptions on ε, we can express (3) with a semiparametric model $Y = F_\varepsilon(X'\beta_0) + \nu$, with a nonparametric link function $F_\varepsilon(\cdot)$. It is monotone increasing by the nature of CDF.

Second, the well-known benefits of isotonic estimation make it a special type of nonparametric method: (i) the isotonic estimator is a tuning-parameter-free nonparametric estimator, (ii)

isotonic estimation imposes minimal assumptions on the smoothness of the true function. We will discuss further features of the isotonic estimator, which makes it particularly suitable for being a plug-in estimator in a semiparametric model.

A challenge of making inference of $\hat{\beta}$ based on (2) is the discreteness of the isotonic estimator $\hat{p}(\cdot)$, which could make the traditional inference procedure (see, e.g., Newey and McFadden, 1994) inapplicable. Particularly in the case where the estimator $\hat{p}(\cdot)$ depends on β, (2) no longer has a continuous total derivative w.r.t β even if $m(Z, \beta_0, p_0(\cdot))$ is differentiable w.r.t. β. Since $\hat{\beta}$ and $\hat{p}(\cdot)$ usually have to be estimated simultaneously in this case, the framework of Chen et al. (2003) cannot be applied here either. The recent developments in the monotone single index model provide us with tools for dealing with this problem. Groeneboom and Hendrickx (2018), Balabdaoui, Groeneboom, and Hendrickx (2019) (BGH hereafter), and Balabdaoui and Groeneboom (2020) developed a novel score-type approach for the monotone single index model. In this paper, we generalize their methods to the framework of the model (1). We show that under mild conditions, the semiparametric estimator $\hat{\beta}$ with a plug-in isotonic estimator satisfies the framework of Newey (1994), and the associated sample moment function is within a distance of $o_p(n^{-1/2})$ from its Neyman-orthogonalized sample moment function. As a result, the proposed estimator is $\sqrt{n}$-consistent, asymptotically normally distributed, and has many other desirable properties.

1.2 Examples and Literature

We give three examples of semiparametric models, which can be estimated with the procedure described in (1) and (2). If no monotonicity assumption is imposed on nuisance functions, these models have been extensively studied in the literature. See, e.g., Engle et al. (1986), Robinson (1988), and Stock (1991) for the partially linear model; Stoker (1986), Hall (1989), and Härdle, Hall, and Ichimura (1993) for the single index model; Robins and Rotnitzky (1995), Hahn (1998), Hirano et al. (2003), Bang and Robins (2005), and Imbens and Rubin (2015) for the inverse probability weighted (IPW) model and the augmented IPW estimators (AIPW) models, to name a few.

With monotonicity assumptions on nuisance functions, some results have been obtained for individual cases of semiparametric models in the past decades, including Example 1 and Example 2 below.

Example 1: Monotone partially linear model.

$$Y = D\beta_0 + p_0(X) + \varepsilon \quad \text{with } E[\varepsilon|X, D] = 0 \tag{4}$$

For monotone increasing $p_0(X)$, Huang (2002) estimates β_0 with the monotone least square method. If we set $p_0(X) = c + \sum_{j=1}^{k} m^j(X^j)$, where X^j is the j-th element of the k-dimensional vector X, we have the monotone additive model, studied in Cheng (2009) and Yu (2014).

Alternatively, β_0 can be estimated by solving the problem (1), with the moment condition

$$E\left[m(Z,\beta,p(\cdot))\right] = E\left[D(Y - D\beta - p(X))\right] = 0. \tag{5}$$

As illustrated in Chernozhukov et al. (2018), the simple plug-in method based on (5) could fail since this moment function is not Neyman-orthogonalized. In Section 2.1, we will show that if $p_0(\cdot)$ is monotone increasing and estimated with isotonic regression, the estimator $\hat{\beta}$ based on (5) is $\sqrt{n}$-consistent and has the same asymptotic variance as in Robinson (1988). We do not need to orthogonalize (5).

Example 2: Monotone single index model

$$Y = p_0(X'\beta_0) + \varepsilon \textbf{ with } E[\varepsilon|X] = 0 \tag{6}$$

If Y is a binary random variable taking values in $\{0,1\}$, this model can be derived from (3), and $p_0(X)$ is by nature monotone increasing. This model was studied by Cosslett (1983), Klein and Spady (1993), and Cosslett (2007), among others. For continuously distributed Y, if the parameter β_0 is the main interest, Han (1987) and Sherman (1993) showed its consistency and $\sqrt{n}-$normality respectively. If monotone increasing $p_0(X)$ is estimated with isotonic regression, Balabdaoui, Durot, and Jankowski (2019) studied (6) with the monotone least square method. Groeneboom and Hendrickx (2018), BGH, and Balabdaoui and Groeneboom (2020) estimated β_0 and $p_0(\cdot)$ by solving a score-type sample moment function[1]:

$$E\left[X\left\{Y - p(X'\beta)\right\}\right] = 0. \tag{7}$$

They show that solving (7) can simultaneously estimate β_0 and $p_0(\cdot)$, at $n^{-1/2}$-rate and $n^{-1/3}$-rate respectively. Note that (7) can be regarded as an individual case of the model (1) with $m(Z,\beta,p(\cdot)) = X\left\{Y - p(X'\beta)\right\}$.

Example 3: IPW and AIPW with monotone increasing propensity scores

We have a triple $Z = (Y,T,X)$, where T is a binary random variable indicating the treatment status. The propensity score is defined as $p_0(X) \overset{\text{def.}}{=} E(T|X) = P(T=1|X)$. Examples of IPW are:

(a) Missing At Random Model (MAR): Among the triple (Y,T,X), only $Z = (T,X,T\cdot Y)$ is observed. Under unconfoundedness and overlapping assumptions, we are interested in $E(Y) = E(\frac{Y\cdot T}{p_0(X)}) := \beta_0$. We can estimate β_0 by solving the problem (1), with the moment condition.

$$E[m(Z,\beta,p(\cdot))] = E(\frac{Y\cdot T}{p(X)} - \beta) = 0.$$

[1]Groeneboom and Hendrickx (2018) estimated the current status model by solving a profile maximum likelihood estimator. The score function of their log-likelihood function takes a similar form of (7).

(b) Average Treatment Effect Model (ATE): the triple $Z = (Y, T, X)$ is observed, where Y takes its values from a random vector $(Y(1), Y(0))$: we have $Y = Y(1)$ if only if $T = 1$, and $Y = Y(0)$ if only if $T = 0$. Under unconfoundedness and overlapping assumptions, we have the average treatment effect $\beta_0 = E(\frac{Y \cdot T}{p_0(X)} - \frac{Y \cdot (1-T)}{1-p_0(X)})$. We can estimate β_0 by solving the problem (1), with the moment condition

$$E[m(Z, \beta, p(\cdot)] = E(\frac{Y \cdot T}{p(X)} - \frac{Y \cdot (1-T)}{1-p(X)} - \beta) = 0.$$

Example of AIPW:

(c) Doubly robust MAR: in addition to the setting in (a), we also know $E(Y|X) = \psi_0(X)$. Under unconfoundedness and overlapping assumptions, we have the conditional expectation $E(Y|X) = E(\frac{Y \cdot T}{p_0(X)} - \frac{T-p_0(X)}{p_0(X)}\psi_0(X)) := \beta_0$. We can estimate β_0 by solving the problem (1), with the moment condition.

$$E[m(Z, \beta, p(\cdot))] = E(\frac{Y \cdot T}{p(X)} - \frac{T - p(X)}{p(X)}\psi(X)) - \beta) = 0. \tag{8}$$

Here we need to plug in the estimators of both $p(\cdot)$ and $\psi(\cdot)$.

IPW and AIPW with monotone increasing propensity scores have rarely been studied. The only exception we found is Qin et al. (2019). They applied the monotone single index model to estimate the propensity score $p(X) := \theta(X'\alpha)$ of an AIPW model, then plugged $\hat{p}(\cdot)$ and another parametric estimator of $\psi_0(\cdot)$ into (2). Their asymptotic results depend on the estimation of both $p_0(\cdot)$ and $\psi_0(\cdot)$. Another different but related paper is Westling et al. (2019). They studied a continuous version of AIPW. The monotonicity is imposed on the relation between the continuous dose of treatments and the outcomes, instead of on the propensity score. To the best of our knowledge, there is no paper estimating the IPW model with a plug-in isotonic estimator of the propensity score. In the following Section 2.2, we show that our method can give us a tuning-parameter free, $\sqrt{n}$-consistent, and asymptotically normal IPW estimator.

1.3 Contribution and structure of this paper

The main contributions of our paper are:

1. We develop a tuning-parameter-free semiparametric estimator of (1). It generalizes existing semiparametric models with monotone nuisance functions. Furthermore, we show its potential applicability by applying it to the case of IPW with monotone increasing propensity score.

2. We show that the sample moment function of the proposed estimator with a plug-in isotonic estimator is within a distance of $o_p(n^{-1/2})$ from its Neyman-orthogonalized sample moment function. Therefore, $\sqrt{n}$-consistency is guaranteed in many cases, without the need for estimating and adding the correction term. As a result, the tuning-parameter-free benefit

is twofold: we save the effort to choose tuning parameters to estimate both the monotone nuisance function and the correction term.

3. We show this estimator is efficient in the case $p_0(x)$ is a function of a scalar x. The semiparametric efficiency here is w.r.t. the unconditional moment condition (1). With x being a multi-dimensional vector, the estimator is $\sqrt{n}$-consistent under different structures combining monotonicity and multi-dimensional covariates.

4. Simulation results show that the proposed method is attractive: (i) while it is more robust against misspecification than parametric plug-in estimators commonly adopted in applied work, it has similar performance to the latter under correct specifications; (ii) compared to methods with other nonparametric plug-in estimators, the proposed estimator requires minimum smoothness conditions on nuisance functions.

5. We develop a bootstrap method to ensure that our semiparametric estimator is tuning-parameter-free in both estimation and inference.

This paper is organized as follows. In Section 2, we present the basic setup and study the theoretical properties of the proposed estimator. In Section 3, we discuss different possibilities of allowing multi-dimensional covariates in a monotone nuisance function, as well as the theoretical properties of the relevant estimators. In Section 4, we discuss the bootstrap inference. In Section 5, we perform simulation studies to illustrate the proposed method. All the proofs are presented in the Appendix.

2 Z-estimation with a plug-in isotonic estimator

We try to develop a general theory for Z-estimation with its plug-in nuisance parameter estimated by isotonic estimation. Let (Y, X) be a sub-vector of random vector Z. To show the idea clearly, we first let X be a random scalar in this section. In Section 3, we will allow X to be multi-dimensional covariates. Now we have (1) and

$$E(Y|X) = p_0(X), \tag{9}$$

where $p_0(\cdot)$ is a monotone increasing function in X. Condition (9) is needed to implement isotonic estimation since it is a method for the conditional mean. We are interested in estimating the parameter β_0. To illustrate the idea clearly, we focus on the just-identified case, where $\dim(\beta) = \dim(m)$. All the results can be extended to over-identified moment conditions with standard GMM procedures.

We can extend (2) further around $p_0(\cdot)$

$$\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\beta_0,\hat{p}(\cdot))}{\partial\beta}(\hat{\beta}-\beta_0) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,\hat{p}(\cdot)) + o_p(\hat{\beta}-\beta_0) \\
&= \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,p_0(X_i)) + \frac{1}{n}\sum_{i=1}^{n} D(Z_i,\beta_0)(\hat{p}(X_i)-p_0(X_i)) \\
&+ \frac{1}{n}\sum_{i=1}^{n} O_p(\hat{p}(X_i)-p_0(X_i))^2 + o_p(\hat{\beta}-\beta_0) \\
&:= \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,p_0(X_i)) + I + II + o_p(\hat{\beta}-\beta_0)
\end{aligned}$$

$D(z,\beta)$ is the functional derivative of $m(z,\beta,p(x))$ w.r.t. $p(\cdot)$.[2] $\sqrt{n}$-consistency of $\hat{\beta}$ requires both I and II to converge at least at $n^{-1/2}$-rate. If $||\hat{p}-p_0|| = o_p(n^{-1/4})$, we have $II = o_p(n^{-1/2})$. Many nonparametric estimators can achieve this rate with properly chosen tuning parameters. For isotonic estimator $\hat{p}(\cdot)$, we usually have

$$||\hat{p}-p_0||^2 = O_p((\log n)^2 n^{-2/3}) = o_p(n^{-1/2}). \tag{10}$$

(See, e.g., Lemma 5.15 in van de Geer, S., 2000). The condition is satisfied without involving any tuning parameter.

We can decompose I into

$$\begin{aligned}
I =& \frac{1}{n}\sum_{i=1}^{n} D(Z_i,\beta_0)(\hat{p}(X_i)-p_0(X_i)) \\
=& \frac{1}{n}\sum_{i=1}^{n}\left\{D(Z_i,\beta_0)(\hat{p}(X_i)-p_0(X_i)) - \int D(Z_i,\beta_0)(\hat{p}(X_i)-p_0(X_i))d\mathbb{P}_0\right\} \\
&+ \int D(Z_i,\beta_0)(\hat{p}(X_i)-p_0(X_i))d\mathbb{P}_0 \\
:=& III + IV
\end{aligned}$$

The condition $III = o_p(n^{-1/2})$ is often referred to as stochastic continuity. The condition $IV = 0$ (or $= o_p(n^{-1/2})$), is referred to as Neyman (Near-) orthogonality. If we have both stochastic continuity and Neyman (Near-) orthogonality, solving the moment condition (2) with plug-in $\hat{p}(\cdot)$ will not depend on the estimation of the nuisance function $p_0(\cdot)$. In the following sub-section, we adapt the definition of Neyman orthogonality (see, e.g., Chernozhukov et al., 2018) to our setting.

[2]To illustrate the idea, we assume a simple case, where $m(z,\beta_0,p_0(x))$ depends on $p_0(\cdot)$ only through its value on x. This assumption is not necessary. The formula can be written into the standard pathwise derivative form, as in Newey (1994).

2.1 Properties of the plug-in isotonic estimator

Definition 1. [Neyman orthogonality] Let T be a convex set, and $T_n \subset T$ be a nuisance realization set for $\hat{p}(\cdot)$. We say the moment function m satisfy Neyman orthogonality condition if we have $E[m(Z, \beta_0, p_0(X))] = 0$ and

$$E[D(Z, \beta_0)(p(X) - p_0(X))] = 0, \qquad \text{for all } p \in T_n$$

If m does not satisfy Neyman orthogonality condition, $\hat{\beta}$ obtained by solving its corresponding sample moment function (2) might suffer from some issues. In some cases, it is even no longer $\sqrt{n}$-consistent. The following is an example in Chernozhukov et al. (2018).

Example 1 continued: The partially linear model

$$Y = D\beta + p(X) + U \qquad E[U|X, D] = 0$$

implies the moment condition $E\left[D(Y - D\beta - p(X))\right] = 0$. But its moment function $m(Z, \beta, p(\cdot)) = D(Y - D\beta - p(X))$ is not Neyman orthogonal, since

$$E[\frac{\partial m(Z, \beta_0, p_0(\cdot))}{\partial p}(p(X) - p_0(X))] = E[D(p(X) - p_0(X))] \neq 0 \text{ in general}$$

Now we do not assume the monotonicity of $p_0(\cdot)$, and let $\hat{p}(\cdot)$ be an arbitrary nonparametric estimator. In this case, the plug-in estimator obtained by choosing $\hat{\beta}$, such that

$$\frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - D_i\hat{\beta} - \hat{p}(X_i)) = 0, \tag{11}$$

can fail to be $\sqrt{n}$-consistent. Let us rearrange (11)

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta_0) &= (\frac{1}{n}\sum_{i=1}^{n} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(Y_i - D_i\beta_0 - \hat{p}(X_i)) \\
&= (\frac{1}{n}\sum_{i=1}^{n} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(Y_i - D_i\beta_0 - p_0(X_i) + p_0(X_i) - \hat{p}(X_i)) \\
&= (\frac{1}{n}\sum_{i=1}^{n} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(U_i + p_0(X_i) - \hat{p}(X_i)) \\
&= (\frac{1}{n}\sum_{i=1}^{n} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_iU_i + (\frac{1}{n}\sum_{i=1}^{n} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(p_0(X_i) - \hat{p}(X_i)).
\end{aligned}$$

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(p_0(X_i) - \hat{p}(X_i))$ might explode since $\hat{p}(X_i)$ is a nonparametric estimator and usually converges slower than $n^{-1/2}$.

To fix this problem, people usually want to orthogonalize m, i.e., transform m into m^*, such

that

1. $E[m^*(Z, \beta_0, p_0(X))] = 0$ still holds, and
2. $E[D^*(Z, \beta_0)(p(X) - p_0(X))] = 0$ for all $p \in T_n$.

In general, people obtain orthogonalized moment function by subtracting from $m(Z, \beta_0, p_0)$ its projection on the linear space of its derivatives w.r.t $p_0(\cdot)$. For example, if m is a just-identified moment condition, then

$$m^*(Z, \beta, p) = (I_{d_m} - G_p(G_p'G_p)^{-1}G_p')m(Z, \beta, p),$$

where G_p is the functional derivative of $m(Z, \beta, p)$ w.r.t p. In our setting (9), where $p_0(X)$ is a conditional mean of Y, the orthogonalization can be achieved by applying Proposition 4 in Newey (1994):

$$m^{**}(Z, \beta, p) = m(Z, \beta, p) + E[D(Z, \beta)|X](Y - p(X)).$$

We can check the two conditions for the Neyman orthogonalization. For m^{**}:

1. $E[m^{**}(Z, \beta_0, p_0(X))] = 0 + E[E[D(Z, \beta_0)|X](Y - p_0(X))] = 0,$
2. and

$$\begin{aligned} E[D^{**}(Z, \beta_0)(p(X) - p_0(X))] &= E[\frac{\partial m^{**}(\mathrm{Z},\beta,p_0(\mathrm{X}))}{\partial p}(p(X) - p_0(X))] \\ &= E[D(Z, \beta_0)(p(X) - p_0(X))] - E[D(Z, \beta_0)|X][(p(X) - p_0(X))] \\ &= E[D(Z, \beta_0)|X][(p(X) - p_0(X))] - E[D(Z, \beta_0)|X][(p(X) - p_0(X))] \\ &= 0. \end{aligned}$$

The equality in Condition 1 and the third equality in Condition 2 follow from the law of iterated expectation.

In practice, we need to add an estimated correction term of $E[D(Z, \beta_0)|X](Y - p_0(X))$ into our sample moment function. In Example 1, this term is $\widehat{E[D_i|X_i]}(Y_i - D_i\hat{\beta} - \hat{p}(X_i))$. Then we have the same estimator as in Robinson (1988).

An interesting feature is that with the following Lemma 1, sample moment function with a plug-in isotonic estimator is within a distance of $o_p(n^{-1/2})$ from its Neyman-orthogonalized sample moment function.

Let us have the following assumptions:

A1 X is a random scalar taking value in the space $\mathcal{X}$. The space $\mathcal{X}$ is convex with non-empty interiors, and satisfies $\mathcal{X} \subset \mathcal{B}(0, R)$ for some $R > 0$.

A2 The true mean function $E(Y|X = x) = p_0(x)$ is monotone increasing in x. There exists $K_0 > 0$ such that $|p_0(x)| < K_0$ for all $x \in \mathcal{X}$.

A3 There exist $c_0 > 0$ and $M_0 > 0$ such that $E[|Y|^m|X = x] \leq m!M_0^{m-2}c_0$ for all integers $m \geq 2$ and almost every x.

A1 and A2 impose boundedness on the monotone function p_0 and the support of X. These conditions are used to control the entropy of the function classes that characterize (2). A3 is to restrict the size of the tail of $Y|X$. With A3, we can show that $\sup_{x \in X} \hat{p}(x) = O_p(\log n)$, which is used to obtain an entropy result associated with the $\sqrt{n}$-convergence rate.

Lemma 1. *$\hat{p}(\cdot)$ is an isotonic estimator of the conditional mean $E(Y|X)$. $\delta(X)$ is a bounded function of X with a finite total variation. Under A1, A2, and A3, we have $\frac{1}{n}\sum_{i=1}^n \delta(X_i)(Y_i - \hat{p}(X_i)) = o_p(n^{-1/2})$.*

The proof in Appendix is based on techniques applied in Groeneboom and Jongbloed (2014), Groeneboom and Hendrickx (2018), and BGH, combining the properties of the isotonic estimator and entropy results for monotone functions.

Let us assume

A4 For all $\beta \in \mathfrak{B}$, $E[D(Z, \beta)|X]$ is a bounded function of X with a finite total variation, and there exist $c_1 > 0$ and $M_1 > 0$ such that $E[|D(Z, \beta)|^m|X = x] \leq m!M_1^{m-2}c_1$ for all integers $m \geq 2$ and almost every x.

we have immediately:

$$\frac{1}{n}\sum_{i=1}^n E[D(Z, \beta_0)|X_i](Y_i - \hat{p}(X_i)) = o_p(n^{-1/2}).$$

Then we add the following assumption,

A5 The first-order expansion of $m(z, \beta, p(\cdot))$ w.r.t $p(\cdot)$ at $p^*(\cdot)$, $D(z, \beta, p(\cdot) - p^*(\cdot))$, is linear in $p(\cdot) - p^*(\cdot)$. Especially, $D(z, \beta, p(x) - p^*(x)) = D(z, \beta)\,(p(x) - p^*(x))$.

A5 enables us to analyze the impact of the estimation of the nuisance function $p(\cdot)$, it is similar to (4.1) and (4.2) in Newey (1994). Now we have

Proposition 1. ***(Sample moment function)*** *Assuming A1-A5, and $p_0(\cdot)$ is estimated with isotonic estimation and plugged into (2), then the semiparametric estimator $\hat{\beta}$ estimated based on this sample moment function is similar to that estimated based on its Neyman-orthogonalized sample moment function, in the sense that $\sqrt{n}(\hat{\beta} - \beta_0)$ has the same asymptotic distribution.*

Remark 1. This proposition shows that with isotonic plug-in estimator $\hat{p}(\cdot)$, the difference between the sample moment function $\frac{1}{n}\sum_{i=1}^n m(Z, \beta, \hat{p}(\cdot))$ and its orthogonalized version is $o_p(n^{-1/2})$. Therefore, there is no need to orthogonalize it in the estimation of $\hat{\beta}$. In this sense, the sample moment function can be regarded as "automatic" Neyman-orthogonalized. The term "automatic" should be understood only in the context of the estimation of $\hat{\beta}$. It does not claim that $m(z, \beta, p(\cdot))$ is Neyman-orthogonalized.

Example 1 Continued: Let $\hat{p}(X)$ is an isotonic estimator of $E[Y - D\beta|X]$ and assume $E[D|X]$ is a bounded function of X with a finite total variation. We have by Lemma 2 (A modified version Lemma 1 in the following Section 2.3, which can be applied to the case that $\hat{p}(\cdot)$ depends on β.)

$$\frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - D_i\hat{\beta} - \hat{p}(X_i)) = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n}(D_i - E[D_i|X_i])(Y_i - D_i\hat{\beta} - \hat{p}(X_i)) = o_p(n^{-1/2}).$$

Then we have

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta_0) &= (\frac{1}{n}\sum_{i=1}^{n}(D_i - E[D_i|X_i])D_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - E[D_i|X_i])(Y_i - D_i\beta_0 - \hat{p}(X_i)) + o_p(1)\\ &= (\frac{1}{n}\sum_{i=1}^{n}(D_i - E[D_i|X_i])D_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - E[D_i|X_i])U_i + o_p(1)\\ &+ (\frac{1}{n}\sum_{i=1}^{n}(D_i - E[D_i|X_i])D_i)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - E[D_i|X_i])(p_0(X_i) - \hat{p}(X_i))\end{aligned}$$

Now under mild conditions, we have $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(D_i - E[D_i|X_i])(p_0(X_i) - \hat{p}(X_i)) = o_p(1)$ and $\frac{1}{n}\sum_{i=1}^{n}(D_i - E[D_i|X_i])D_i \xrightarrow{p} E[(D_i - E[D_i|X_i])^2]$. Then we have $\sqrt{n}$-consistent $\hat{\beta}$. Also, $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \sigma_u^2 E(D - E[D|X])^{-2})$.

Remark 2. Huang (2012) showed the same asymptotic variance for the partially linear model with monotone nuisance function, with monotone least square methods. Here we revisit it from a different angle: we highlight the relation between isotonic plug-in estimator and Neyman orthogonalization. We start from an unorthogonalized moment function (11) and achieve the same result as in Robinson (1988), without adding the estimated correction term $\widehat{E[D_i|X_i]}(Y_i - D_i\hat{\beta} - \hat{p}(X_i))$. Therefore, the benefit of the isotonic plug-in estimator in terms of tuning-parameter-free is doubled: an isotonic plug-in estimator will save us not only one tuning parameter for the nuisance function $p(\cdot)$ but also other tuning parameters for estimating the nonparametric part in the correction term ($\widehat{E[D_i|X_i]}$ in this case).

2.2 Efficiency and the plug-in isotonic estimator

The correction term $E[D(Z, \beta_0)|X](Y - p_0(X)))$ is also associated with efficiency. As illustrated in Proposition 4 of Newey (1994), for unconditional moment condition $E[m(Z, \beta, p(X))] = 0$, where $p_0(X) = E(Y|X)$ for some sub-vector Y, the efficient influence function ψ is:

$$\psi(Z) = -\big[\frac{\partial E[m(Z, \beta_0, p(X))]}{\partial \beta}\big]^{-1}(m(Z, \beta_0, p_0(X)) + E[D(Z, \beta_0)|X](Y - p_0(X)))$$

If we could show for an isotonic plug-in estimator $\hat{p}(\cdot)$

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta_0, \hat{p}(X_i)) = \frac{1}{n}\sum_{i=1}^{n} \left[m(Z_i, \beta_0, p_0(x_i)) + E[D(Z, \beta_0)|X_i](y_i - p_0(x_i))\right] + o_p(n^{-1/2}),$$

we could show the efficiency. Let's assume the following assumptions:

A6 There are $b(z) > 0$ and $D(z, g)$ that (i) $||m(z, \beta, p) - m(z, \beta, p_0) - D(z, \beta, p - p_0)|| \leq b(z)||p - p_0||^2$; (ii) $E[b(z)] = o_p(n^{1/6}(\log n)^{-2})$, for all $\beta \in \mathfrak{B}$.

A7 There are ε, $b(z)$, $\tilde{b}(z) > 0$ and $p(\cdot)$ with $||p|| > 0$. Such that (i) for all $\beta \in \mathfrak{B}$, $m(z, \beta, p_0)$ is continuous at β and $m(z, \beta, p_0) \leq b(z)$; (ii) $||m(z, \beta, p) - m(z, \beta, p_0)|| \leq \tilde{b}(z)(||p - p_0||)^{\varepsilon}$.

A8 $E\{m(z, \beta, p_0)\} = 0$ has a unique solution on $\mathfrak{B}$ at β_0, and $\mathfrak{B}$ is compact.

A9 For $\beta \in$ interior($\mathfrak{B}$), (i) there are $p(\cdot)$ with $||p|| > 0$, ε and a neighborhood $\mathcal{N}$ of β_0 such that for all $||p - p_0|| \leq \varepsilon$, $m(z, \beta, p)$ is differentiable in β on $\mathcal{N}$; (ii) $M_\beta = -E\left\{\frac{\partial m(Z, \beta_0, p_0(X))}{\partial \beta}\right\}$ is nonsingular; (iii) $E[||m(z, \beta, p)||^2] < \infty$; (iv) Assumption A7 is satisfied with $m(z, \beta, p)$ there equal to each row of $\frac{\partial m(Z, \beta, p)}{\partial \beta}$.

A6 is an adaption of Newey's Assumption 5.1. This assumption requires that the high order term from a linear approximation is small. Combining (ii) in A6 and (10), we have the reminder term converging faster than $n^{-1/2}$. A7, A8, and A9 are adapted from Assumption 5.4, 5.5, and 5.6 in Newey (1994). They are general conditions for the consistency and asymptotical normality for the method of moment.

Let us define

$$M_\beta = -E\left\{\frac{\partial m(Z_i, \beta_0, p_0(X_i))}{\partial \beta}\right\}$$
$$M(Z) = E[D(Z, \beta_0)|X](Y - p_0(X)).$$

We have

Theorem 1. ***(Efficiency)*** *Assuming A1-A9, for unconditional moment condition $E[m(Z, \beta_0, p_0(X))] = 0$, $\hat{p}(\cdot)$ is an isotonic estimator of the conditional mean $E(Y|X) := p_0(X)$.*

Then $\hat{\beta}$ obtained by solving the sample moment condition (2) is $\sqrt{n}$-consistent and efficient, with

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

where

$$V = M_\beta^{-1} E[\{m(Z, \beta_0, p_0) + M(Z)\}\{m(z, \beta_0, p_0) + M(Z)\}']M_\beta^{-1},$$

The proof is in Appendix. It is based on a combination of techniques in Newey (1994), Hirano, Imbens, and Ridder (2000, 2003), Groeneboom and Jongbloed (2014), Groeneboom and Hendrickx (2018), and BGH.

We can apply Theorem 1 to the IPW model by using the isotonic regression to estimate the propensity score.

Example 3 (b) continued: For the ATE model, we have $m(Z, \beta, p(\cdot)) = \frac{Y \cdot T}{p_0(X)} - \frac{Y \cdot (1-T)}{1-p(X)} - \beta$. The $p_0(x)$ is the propensity score

$$p_0(x) = E[T|X = x] = Pr(T = 1|X = x).$$

Let $\hat{p}(\cdot)$ is the isotonic estimator of the propensity score. We are interested in the plug-in estimator $\hat{\beta}$:

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{Y_i \cdot T_i}{\hat{p}(X_i)} - \frac{Y_i \cdot (1 - T_i)}{1 - \hat{p}(X_i)} \right\} \tag{12}$$

Here we assume

C1 $T \perp (Y(1), Y(0))|X$, unconfoundedness.

C2 (i) The support $\mathcal{X}$ of X is convex and compact; (ii) the density of X is bounded from 0 on $\mathcal{X}$.

C3 (i) $E(Y(0)^2) < \infty$ and $E(Y(1)^2) < \infty$; (ii) $\mu_0(x) = E(Y(0)|X = x)$ and $\mu_1(x) = E(Y(1)|X = x)$ are continuously differentiable for all $x \in \mathcal{X}$.

C4 The true propensity score $p_0(x)$ satisfies: (i) $p_0(\cdot)$ is continuous and monotone increasing; (ii) there exist positive number $\underline{p}$ and $\bar{p}$, such that $1 > \bar{p} \geq p_0(x) \geq \underline{p} > 0$ for all $x \in \mathcal{X}$.

And we have

Corollary 1. *Suppose Assumptions C1-C4 hold. The average treatment effect estimator $\hat{\beta}$ is obtained by* (12)*. Then $\hat{\beta} \xrightarrow{p} \beta_0$, and*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Omega),$$

where $\Omega = Var(E[Y(1) - Y(0)]|X) + E[Var(Y(1)|X)/p_0(X)] + E[Var(Y(0)|X)/(1 - p_0(X))]$. $\hat{\beta}$ reaches the semiparametric efficiency bound.

2.3 The case that $\hat{p}(\cdot)$ depends on β

The isotonic estimator $\hat{p}(\cdot)$ can depend on β in some cases, as we have seen in the partially linear model. We use the notation $\hat{p}_\beta(\cdot)$ to represent such an estimator. In this case, we might have a problem of finding a root for (2). Since the isotonic estimator $\hat{p}_\beta(\cdot)$ is a step function, changes in β might also cause discontinuous changes of $\hat{p}_\beta(\cdot)$. Groeneboom and Hendrickx (2018) and BGH tried to solve this problem with a so-called zero-crossing root, a technique dealing with discrete

score-type functions. Then they found that it is non-trivial to show the existence of zero-crossing root in finite samples. Balabdaoui and Groeneboom (2020) proposed another method. They replaced the zero-crossing root of a score function with the minimizer of L^2 norm of it. They showed that this minimizer has the same properties as the zero-crossing root for the single index model. We extend their methods to the general case of the method of moments.

Let $\hat{p}_\beta(X)$ be an isotonic estimator of the conditional mean $E[T(Z,\beta)|X]$, where T is a known function of data Z and the parameter β. An example of this case can be the partially linear model, where $T(Z,\beta) = Y - X\beta$. A feasible version of the plug-in estimator of $\hat{\beta}$ w.r.t (2) can be

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left|\left| \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta, \hat{p}_\beta(X_i)) \right|\right|^2, \tag{13}$$

where $||\cdot||$ is the Euclidean norm. To implement our method, we need to assume the monotonicity holding in a neighbor of the true value β_0. Let A1' be the same as A1, and we modify Assumptions A2 and A3 :

A2' There exists $\delta_0 > 0$ such that $E(T(Z,\beta)|X) := p_\beta(X)$ is monotone increasing for each $\beta \in \mathcal{B}(\beta_0, \delta_0)$. There exists $K_0 > 0$ such that $|p_0(x)| < K_0$ for all $x \in \mathcal{X}$.

A3' There exist $c_0 > 0$ and $M_0 > 0$ such that $E[|T(Z,\beta)|^m|X = x] \leq m! M_0^{m-2} c_0$ for all integers $m \geq 2$ and almost every x and $\beta \in \mathcal{B}(\beta_0, \delta_0)$.

We have

Lemma 2. *For fixed β, $\hat{p}_\beta(X)$ is an isotonic estimator of the conditional mean $E(T(Z,\beta)|X)$. $\delta(X)$ is a bounded function of X with a finite total variation. Under A1' - A3', we have $\frac{1}{n}\sum_{i=1}^{n} \delta(X_i)(T(Z,\beta) - \hat{p}_\beta(X)) = o_p(n^{-1/2})$.*

To show the results of Lemma 2, we do not need to solve the root of a discrete moment function. Therefore, the proof is similar to that of Lemma 1.

Similarly, let A4' and A5' be the same as A4 and A5, we have

Proposition 2. ***(Sample moment function)*** *Assuming A1' - A5', and $p_0(\cdot)$ is estimated with isotonic estimation and plugged into the moment condition $m(Z,\beta,p(\cdot))$. Then the semiparametric estimator $\hat{\beta}$ estimated based on (13) is similar to that estimated based on the minimizer of the L^2 norm of its Neyman-orthogonalized sample moment function, in the sense that $\sqrt{n}(\hat{\beta} - \beta_0)$ has the same asymptotic distribution.*

Now let (i) A6' be the same as A6; (ii) A7' to A9' are modified versions of A7 to A9, where all the conditions in A7 to A9 satisfied with $m(z,\beta,p)$ there equal to $\{m(z,\beta,p) + E(D(Z,\beta_0)|x)T(z,\beta)\}$.

Theorem 2. ***(Efficiency)*** *Assuming A1' - A9', for unconditional moment condition $E[m(Z,\beta_0,p_0(X))] = 0$, $\hat{p}(\cdot)$ is an isotonic estimator of the conditional mean $E(T(Z,\beta)|X) := p_\beta(X)$.*

Then $\hat{\beta}$ obtained by (13) is $\sqrt{n}$-consistent and efficient.

3 Multi-dimensional X

The isotonic function is always a mapping from $\mathbb{R}^1$ to. In order to have wide applicability, the model should be able to deal with multivariate covariates. In this section, we consider two different ways to combine the plug-in isotonic estimator with multivariate covariates X: the monotone single index model and the monotone additive model.

3.1 Plug-in monotone Single Index Model

For a k-dimensional data sample X, A1 can be modified to

A1" X is a random vector taking value in the space $\mathcal{X}$. The space $\mathcal{X}$ is convex with non-empty interiors, and satisfies $\mathcal{X} \subset \mathcal{B}(0, R)$ for some $R > 0$.

We model the conditional mean function with $E(Y|X) = p_0(X) \equiv F_0(X'\alpha_0)$. α_0 is a k-dimensional vector with $||\alpha_0|| = 1$.[3]

In this case, we need to estimate both p_0 and α_0 in the first step, then plug them in (2).

To estimate F_0 and α_0, we can apply the method of BGH. For a fixed α

$$\hat{F}_\alpha = \arg\min_{F \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - F(X_i'\alpha)\}^2, \tag{14}$$

where $\mathcal{M}$ is the set of monotone increasing functions defined on $\mathbb{R}$. Then, $\hat{F}_\alpha(u)$ can be solved with isotonic regression on the data points $\{u_i\}_{i=1}^n := \{X_i'\alpha\}_{i=1}^n$.

Then $\hat{\alpha}$ can be estimated by minimizing the square sum of a score function. For example, the simple score estimator in Balabdaoui and Groeneboom (2020) and BGH is given by solving

$$\hat{\alpha} = \underset{\alpha}{\text{argmin}} || \frac{1}{n} \sum_{i=1}^{n} X_i' \{Y_i - \hat{F}_\alpha(X_i'\alpha)\} ||^2 \tag{15}$$

Balabdaoui and Groeneboom (2020) and BGH showed that under certain assumptions, $\hat{\alpha}$ is a $\sqrt{n}$-consistent estimator for α_0, and $E\left[\hat{F}_{\hat{\alpha}}(X_i'\hat{\alpha}) - F_0(X'\alpha_0)\right]^2 = O_P((\log n)^2 n^{-2/3})$. We also include those assumptions in our framework.

We can also allow $\hat{F}$ depend on β, as we did in Section 2.3. In this case, we should replace Y_i in (14) by $T(Z_i, \beta)$, and in the second step, we replace (15) with

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\text{argmin}} || \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \beta, \hat{F}_{\alpha,\beta}(X_i'\alpha)) ||^2.$$

To implement isotonic estimation to the link function F_0, we need that the monotonicity holds in the neighbors of the true values α_0 and β_0. For fixed α and β, we define $F_{\alpha,\beta}(u) = E(T(Z, \beta)|\alpha' X = u)$. The assumption A2 are modified to adapt to the current setting:

[3] In estimation, the constraint $||\alpha_0|| = 1$ can be dealt with reparameterization or the augmented Lagrange method by Balabdaoui and Groeneboom (2020). In this section, we discuss our model without discussing those technical details in estimation. See BGH and Balabdaoui and Groeneboom (2020) for more details.

A2" There exists $\delta_0 > 0$ that the true mean function $u \mapsto E[T(Z,\beta)|X'\alpha = u]$ is monotone increasing for each $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$ and $\beta \in \mathcal{B}(\beta_0, \delta_0)$.

Now let A3" be the same as A3'. We have

Lemma 3. *For fixed α and β, $\hat{F}_{\alpha,\beta}(\cdot)$ are solved by solving (14). $\delta(u)$ is a bounded function of u with a finite total variation. Under A1"-A3", we have $\frac{1}{n}\sum_{i=1}^{n}\delta(X_i'\alpha)(T(Z_i,\beta) - F_{\alpha,\beta}(X_i'\alpha)) = o_p(n^{-1/2})$.*

Moreover, we add Assumptions A10" and A11".

A10" For all $\alpha \neq \alpha_0$ with $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$ and $\beta \neq \beta_0$ with $\beta \in \mathcal{B}(\beta_0, \delta_0)$, we have

$$Cov\left\{[T(Z,\beta_0) - T(Z,\beta)] + (\alpha_0 - \alpha)'X, [T(Z,\beta_0) - T(Z,\beta)] + F_0(X'\alpha_0)|X'\alpha\right\} \neq 0 \text{ almost surely.}$$

A11" $E\left\{[D(Z_i,\beta_0) - E(D(Z_i,\beta_0)|X_i'\alpha_0)]\{X_i - E[X_i|X_i'\alpha_0]\}F_0^{(1)}(X_i'\alpha_0)\right\}$ is non-singular.

A10" and A11" are adapted from BGHs' A7 and A9. These two assumptions ensure the consistency and existence of limiting variances of our estimators.

Let (i) A3" and A6" be the same as A3' to A6'; (ii) A7" to A9" are modified versions of A7 to A9, where all the conditions in A7 to A9 satisfied with $m(z,\beta,p)$ there equal to $\{m(z,\beta,p) + E[D(Z,\beta_0)|x'\alpha_0]T(z,\beta)\}$. Furthermore, we define

$$\begin{aligned}
M_\alpha &= -E\left\{[D(Z_i,\beta_0) - E(D(Z_i,\beta_0)|X_i'\alpha_0)]\{X_i - E[X_i|X_i'\alpha_0]\}'F_0^{(1)}(X_i'\alpha_0)\right\} \\
M_\beta &= -E\left\{\frac{\partial m(Z_i,\beta_0,F_0(X_i'\alpha_0))}{\partial\beta} + E[D(Z_i,\beta_0)|X_i'\alpha_0]\frac{\partial T(Z_i,\beta_0)}{\partial\beta}\right\} \\
M(Z) &= E(D(Z,\beta_0)|X'\alpha_0)(T(Z,\beta_0) - F_0(X'\alpha_0)).
\end{aligned} \tag{16}$$

Then we have

Theorem 3. *Suppose Assumptions A1"-A10" hold, then*

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} N(0, V_\alpha) \text{ and } \sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V_\beta).$$

where

$$\begin{aligned}
V_\beta &= M_\beta^{-1}E[\{m(Z,\beta_0,p_0) + A(Z) + M(Z)\}\{m(z,\beta_0,p_0) + A(Z) + M(Z)\}']M_\beta^{-1}, \\
V_\alpha &= M_{\alpha,1}^{-1}E[\{m_1(Z,\beta_0,p_0) + B_1(Z) + M_1(Z)\}\{m_1(Z,\beta_0,p_0) + B_1(Z) + M_1(Z)\}']M_{\alpha,1}^{-1},
\end{aligned}$$

with $m_1(Z,\beta,F(X'\alpha)) \overset{def.}{=} X\{T(Z,\beta) - F(X'\alpha)\}$; $M_{\alpha,1}$, B_1, and M_1 are M_α, B, and M

corresponding to the moment function m_1; $A(Z)$ and $B(Z)$ are defined by

$$-M_\alpha(\hat{\alpha}-\alpha_0) := \frac{1}{n}\sum_{i=1}^{n} A(Z_i) + o_p(n^{-1/2}), \quad E\left[A(Z)\right] = 0,$$
$$-M_\beta(\hat{\beta}-\beta_0) := \frac{1}{n}\sum_{i=1}^{n} B(Z_i) + o_p(n^{-1/2}), \quad E\left[B(Z)\right] = 0.$$

Example 2 continued: The simple score estimator (SSE) for the monotone single index model of BGH can be regarded as an individual case of the estimator in Theorem 3, where $m(Z,\beta_0,F_0(X'\alpha_0)) = m_1(Z,\beta_0,F_0(X'\alpha_0)) = X\{Y - F_0(X'\alpha_0)\}$. Here β_0 is absent from the model, thus $B_1(Z) = 0$. We have

$$\begin{aligned}
T(Z,\beta_0) &= Y,\\
D(Z,\beta_0) &= -X,\\
E(D(Z,\beta_0)|X'\alpha_0) &= -E(X|X'\alpha_0),\\
M(Z) = M_1(Z) &= -E(X|X'\alpha_0)\left\{Y - F_0(X'\alpha_0)\right\}, \text{ and}\\
M_\alpha = M_{\alpha,1} &= -E\left\{[X - E[X|X'\alpha_0]]\{x - E[X|X'\alpha_0]]\}'F_0^{(1)}(X'\alpha_0)\right\}
\end{aligned}$$

Plugging these values into the formula of V_α, we can see it is the same as the asymptotical variance of SSE in BGH.

3.2 Plug-in monotone additive model

We can also model the conditional mean function with an additive structure. First we introduce some notations here. k is the dimension of the vector x_i. For $j = 1,2,...,k$, $m_0^j(\cdot)$ is a strict monotone increasing function of a scalar x_i^j. We use x_i^j to represent the j-th element of the observation i, with $j = 1,...,k$, and $i = 1,...,n$; we use boldfaced $\mathbf{x}_i$ to represent the k-dimensional vector of the observation i, $\mathbf{x}_i = \{x_i^1, x_i^2, ..., x_i^k\}$; we use the boldfaced $\mathbf{x}^j$ to represent the vector of all the j-row of our $n \times k$ matrix of covariates, $\mathbf{x}^j = \{x_1^j, x_2^j, ..., x_n^j\}'$, and the boldfaced $\mathbf{y} = \{y_1, y_2, ..., y_n\}'$. We use the capitals $\mathbf{Y}$,$\mathbf{X}_i^j$,$\mathbf{X}_i$, and $\mathbf{X}^j$ to represent the corresponding random variable or vectors. A slightly confusing notation is: we use X^j (non-bold typeface) to represent the j-th element of the k-dimensional random vector X, without specifying the index of the observation it belongs to.

The plug-in nuisance function is a conditional mean function of some random scalar, Y_i, say. It takes the form of

$$E(Y_i|\mathbf{X}_i) = c + m_0^1(X_i^1) + ...m_0^k(X_i^k) \tag{17}$$

Without loss of generality, we assume each m_0^j is supported on $[0,1]$. To identify each m_0^j,

we add the normalizing condition

$$\int_0^1 m_j(x_j) = 0 \text{ for } j = 1, 2, ..., k \tag{18}$$

The least square estimator of 17 can be defined as the minimizer of

$$\underset{c\in\mathbb{R}^1,\{m^j\}_{j=1}^k\in M_0}{\arg\min} \sum_{i=1}^{n} \left[Y_i - c - \sum_{j=1}^{k} m^j(X_i^j) \right] \tag{19}$$

where M_0 denotes the class of monotone increasing function satisfying (18). We use $\left\{\hat{m}^j(\cdot)\right\}_{j=1}^k$ to denote the estimator from 19. Its asymptotic properties were discussed by Mammen and Yu (2007). Cheng (2009) and Yu (2014) extended their results to the partially linear monotone additive model. The estimator $\left\{\hat{m}^j(\cdot)\right\}_{j=1}^k$ can be obtained with backfitting, an iterative procedure that updates each time a single sub-function with isotonic estimation while treating other sub-functions as fixed. See Mammen and Yu (2007) for a literature review of backfitting. The procedure is described here:

For a fixed sample $\{y_i, x_i\}_{i=1}^n$. To solve the problem 19, we can first solve the following problem

$$\min_{\mathrm{G}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} g_i^j)^2, \tag{20}$$

where G is a $k \times n$ matrix of real numbers g_i^j, and each of its column, $\mathbf{g}^j$, being an isotonic vector w.r.t to the ordered $\mathbf{x}^j$. For example, if $k = 3$ and $n = 3$, we have

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \ \mathbf{x} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 \\ x_2^1 & x_2^2 & x_2^3 \\ x_3^1 & x_3^2 & x_3^3 \end{pmatrix}, \text{ and the estimator } G = \begin{pmatrix} g_1^1 & g_1^2 & g_1^3 \\ g_2^1 & g_2^2 & g_2^3 \\ g_3^1 & g_3^2 & g_3^3 \end{pmatrix}.$$

If $x_2^1 > x_1^1 > x_3^1$, then the least square isotonic estimator should satisfy $g_2^1 > g_1^1 > g_3^1$. Given G solving the problem (20), the value of the estimated monotone function $\hat{m}$ on the point x_i^j can be assigned with $\hat{m}(x_i^j) = g_i^j - \bar{g}^j$, where $\bar{g}^j = \frac{1}{n}\sum_{i=1}^n g_i^j$ that is needed for the normalization, and the estimated constant is $\hat{c} = \sum_{j=1}^k \bar{g}^j$. Since there is a one-to-one relationship between g_i^j and x_i^j, we can rewrite $g_i^j = g^j(x_i^j)$, i.e, $g^j(\cdot)$ is a monotone function defined on $\mathbf{x}^j$.

Let $g_{i,[r]}^j(\cdot)$ denote the backfitting estimator of $g^j(\cdot)$ updated at the r-th round of the iteration. In the j-th step of the round r. We see that $g_{i,[r]}^j(\cdot)$ is obtained by regressing

$$\left\{ Y_i - g_{[r]}^1(X_i^1) - .... - g_{[r]}^{j-1}(X_i^{j-1}) - g_{[r-1]}^{j+1}(X_i^{j+1}) - ...g_{[r-1]}^k(X_i^k) \right\}_{i=1}^n$$

on $\{X_i^j\}_{i=1}^n$ with the isotonic regression. In each round and each step, we repeat this type of isotonic regression recursively for $r = 1, 2, ...$ and $j = 1, ..., k$. After some stopping condition is

satisfied, we can normalize these backfitting estimators and obtain $\hat{c}$ and $\hat{m}$.

Now we incorporate this method into the estimation of the nuisance function of the model (1). As in Section 2.3, we should also allow the estimation of (17) to depend on β by replacing Y_i by $T(Z_i, \beta)$.

W.l.o.g., A1" can be modified to

A1$^{(3)}$ X is a random vector taking value in the space $[0,1]^k$.

and A2 is modified to

A2$^{(3)}$ There exists $\delta_0 > 0$ and $K_0 > 0$ that the mean function $E[T(Z_i,\beta)|X_i = x_i] := p_\beta(x_i)$ is a sum of k monotone increasing functions $m_\beta(\cdot)$, i.e., $p_\beta(x_i) \equiv c_\beta + \sum_{j=1}^k m_\beta^j(x_i^j)$ each $\beta \in \mathcal{B}(\alpha_0, \delta_0)$.

Let A3$^{(3)}$ be the same as A3. Similarly, we have

Lemma 4. *For fixed β, $\hat{p}_\beta(X_i) \equiv \hat{c}_\beta + \sum_{j=1}^k \hat{m}_\beta(X_i^j)$ is a least square isotonic estimator of the conditional mean $E(T(Z_i,\beta)|X_i)$. $\delta(X)$ is a bounded function of X with a finite total variation. Under A1$^{(3)}$ - A3$^{(3)}$, we have $\frac{1}{n}\sum_{i=1}^n \delta(X_i)(T(Z_i,\beta) - \hat{p}_\beta(X_i)) = o_p(n^{-1/2})$.*

The proof is in Appendix. It is based on Theorem 2 of Mammen and Yu (2007), which states that for a given sample of size n, the backfitting estimator of the problem (20) will converge to the least square estimator of this problem, with r growing to ∞.

Now let (i) **A6$^{(3)}$** to **A9$^{(3)}$** are the same as A6' to A9'. We use p_0 to denote p_{β_0}, $p_0(x_i) = c_0 + \sum_{j=1}^k m_0^j(x_i^j)$. And we define

$$M_\beta = -E\left\{\frac{\partial m(Z_i,\beta_0,p_0(X_i))}{\partial\beta} + E[D(Z_i,\beta_0)|X_i]\frac{\partial T(Z_i,\beta_0)}{\partial\beta}\right\}, \text{ and}$$
$$M(Z_i) = E(D(Z,\beta_0)|X_i)(T(Z_i,\beta_0) - p_0(X_i)).$$

Theorem 4. *Assuming A1$^{(3)}$ - A9$^{(3)}$, for unconditional moment condition $E[m(Z,\beta_0,p_0(X))] = 0$, $\hat{p}_\beta(\cdot)$ is an isotonic estimator of the additive conditional mean $E(T(Z_i,\beta)|X_i) := p_\beta(X_i) \equiv c_\beta + \sum_{j=1}^k m_\beta^j(X_i^j)$.*

Then $\hat{\beta}$ obtained by (13) is $\sqrt{n}$-consistent and

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

where $V = M_\beta^{-1}E[\{m(Z,\beta_0,p_0) + M(Z)\}\{m(z,\beta_0,p_0) + M(Z)\}']M_\beta^{-1}$.

Example 2 continued: If we apply Theorem 4 to the partially linear monotone additive model

$$
\begin{aligned}
Y_i &= D_i\beta_0 + p_0(X_i) + \varepsilon \\
&= D_i\beta_0 + \sum_{j=1}^{k} m_0^j(X_i^j) + \varepsilon \quad \text{with } E[\varepsilon|X, D] = 0.
\end{aligned}
$$

we can choose $m(Z, \beta_0, F_0(X'\alpha_0)) = D_i \left\{ Y_i - \beta_0 D_i - \sum_{j=1}^{k} m_0^j(X_i^j) \right\}$. For simplicity we set $D_i \in \mathbb{R}^1$ then we have

$$
\begin{aligned}
T(Z_i, \beta_0) &= Y - \beta_0 D_i, \\
D(Z_i, \beta_0) &= -D_i, \\
E(D(Z_i, \beta_0)|X_i) &= -E(D_i|X_i), \\
\frac{\partial m(Z_i, \beta_0, p_0(X_i))}{\partial \beta} &= -D_i^2 \\
\frac{\partial T(Z_i, \beta_0)}{\partial \beta} &= -D_i \\
M_\beta &= E\left[D_i(D_i - E(D_i|X_i))\right] = E\left[(D_i - E(D_i|X_i))^2\right] \\
M(Z_i) &= -E(D_i|X_i) \left\{ Y - \beta_0 D_i - \sum_{j=1}^{k} m_0^j(X_i^j) \right\}
\end{aligned}
$$

then $V = \sigma^2 E[(D_i - E[D_i|X_i])^2]^{-1}$. This variance is larger than the one achieved in Cheng (2009), which is $\sigma^2 E[(D_i - \sum_{j=1}^{k} E[D_i|X_i^j])^2]^{-1}$, because he assumed the pairwise independence of X_i. We do not have this assumption.

4 Bootstrap inference

One advantage of the proposed estimator $\hat{\beta}$ is tuning-parameter-free. However, since $\hat{\beta}$ is a semiparametric estimator, its asymptotic variance involves conditional means. The estimation of variances might still require some smoothing methods. To obtain an estimator that is free from tuning parameters in both estimation and inference, we propose a bootstrap method to approximate the asymptotic distribution $\hat{\beta}$.

Groeneboom and Hendrickx (2017) showed the bootstrap validity of the single index parameter in the current status model. We generalize their result to the model (1).

The bootstrap procedure is:

1. $\{Z_i^*\}_{i=1}^n$ is a resample with replacement from $\{Z_i\}_{i=1}^n$.
2. $\hat{p}^*(\cdot)$ is an isotonic estimator w.r.t. $\{Z_i^*\}_{i=1}^n$.
3. $\hat{\beta}^*$ solves $\frac{1}{n}\sum_{i=1}^n m(Z_i^*, \beta, \hat{p}^*(\cdot)) = 0$ (or $\operatorname{argmin}_\beta \left|\left| \frac{1}{n}\sum_{i=1}^n m(Z_i^*, \beta, \hat{p}_\beta^*(\cdot)) \right|\right|^2$).

Theorem 5. *Let $\hat{\beta}^*$ be the bootstrap counterpart of $\hat{\beta}$ in Theorem 1, 2 or 3, which are estimated based on resamples from the empirical distribution of $\{Z_i\}_{i=1}^n$. Suppose the corresponding assumptions for these theorems hold. Then*

$$\sup_{t\in\mathbb{R}^k} |P^*\{\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \leq t\} - P_0\{\sqrt{n}(\hat{\beta} - \beta_0) \leq t\}| \xrightarrow{p} 0,$$

where P^* is the bootstrap distribution conditional on the data.

5 Simulation

In this section, we conduct four simulations for the proposed estimators.

5.1 Efficiency for IPW model with single covariates

We use two numerical results to show evidence that MAR model and ATE model with univariate propensity score can achieve the semi-parametric efficiency bound. This is in accordance with Corollary 1. We also show the bootstrap validity under each setting.

5.1.1 Missing at random model

Example 3 (a) continued: The associated moment condition for the MAR model is

$$E[m(Z, \beta_0, p_0(\cdot))] = E(\frac{Y \cdot T}{p_0(X)} - \beta_0) = 0.$$

Assuming that $p_0(\cdot)$ is a monotone increasing function, we are interested in the asymptotic properties of the plug-in estimator $\hat{\beta}$:

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i \cdot t_i}{\hat{p}(x_i)},$$

where $\hat{p}(\cdot)$ is the isotonic estimator of the propensity score

$$p_0(x) = E[T|X = x] = Pr(T = 1|X = x).$$

The semi-parametric bound for the estimate $\hat{\beta}$ is $\Omega = \mathrm{Var}(E[Y|X]) + E[\mathrm{Var}(Y|X)/p_0(X)]$. (See, e.g., Section 4.1 of Hirano, Imbens, and Ridder, 2000)

We set $X = 0.15 + 0.7Z$, Z and ν are independently uniformly distributed on $[0, 1]$, and

$$\begin{aligned} Y &= 2X + \varepsilon \\ \varepsilon &\sim N(0, 1) \\ T &= \begin{cases} 0 & \text{if } X < \nu \\ 1 & \text{if } X \geq \nu \end{cases} \end{aligned}$$

In this setting, we have

$$\beta_0 \equiv \int E(Y|X)dP(X) = E(2X) = 2 \times 0.5 = 1.$$

The efficient variance is

$$\begin{aligned}\Omega &= \mathrm{Var}(E[Y|X]) + E[\mathrm{Var}(Y|X)/p_0(X)] = \mathrm{Var}(2X) + E[1/p_0(X)] \\ &= 4 \cdot \frac{0.7^2}{12} + \int_{0.15}^{0.85} \frac{1}{x}\frac{1}{0.7}dx \approx 2.63\end{aligned}$$

The simulation results are in Table 1:

Table 1: MAR model

n	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\beta$	n	$\hat{\mu}^*_\beta$	$\hat{\sigma}^{2*}_\beta$
100	0.9966	2.9991	100	1.2044	1.3656
1000	0.9959	2.8373	1000	0.9879	2.8921
2000	0.9972	2.7514	2000	1.0721	2.4442
5000	0.9981	2.6845	5000	1.0259	2.4274
10000	0.9987	2.6625	10000	1.0233	2.6815
∞	1	2.63	∞	1	2.63

The left panel of Table 1 shows the simulation results based on 5000 Monte Carlo replications. The sample sizes are $n = 100$, 1000, 2000, 5000 and 10000. We present the Monte Carlo averages $\hat{\mu}_\beta$, and variances $\hat{\sigma}^2_\beta$ (multiplied by n) of the estimates of β_0. We can see with the sample size growing, both $\hat{\mu}_\beta$ and $\hat{\sigma}^2_\beta$ are converging to their theoretical limit.

In the right panel, we present the corresponding simulation results based on 5000 bootstrap samples. The sample sizes are the same. $\hat{\mu}^*_\beta$ and variances $\hat{\sigma}^{2*}_\beta$ are defined similarly. Since all the bootstrap samples are originated from one Monte Carlo sample, the pattern of biases and variances could be less stable than those in the left panel, as expected. Nevertheless, $\hat{\mu}^*_\beta$ and $\hat{\sigma}^{2*}_\beta$ are still converging to their theoretical limit.

5.1.2 Average Treatment Effect Model

Example 3 (b) continued: The efficient asymptotical variance for ATE model is $\Omega = \mathrm{Var}(E[Y(1)-Y(0)]|X)+E[\mathrm{Var}(Y(1)|X)/p_0(X)+E[\mathrm{Var}(Y(0)|X)/(1-p_0(X))]$. (See, e.g., Section 4.2 of Hirano, Imbens, and Ridder, 2000)

We set $X = 0.15 + 0.7Z$, Z and ν are independently uniformly distributed on $[0, 1]$, and

$$T = \begin{cases} 0 & \text{if } X < \nu \\ 1 & \text{if } X \geq \nu \end{cases}$$
$$Y = 0.5T + 2X + \varepsilon$$
$$\varepsilon \sim N(0, 1)$$

The average treatment effect

$$\beta_0 = 0.5$$

The efficient variance

$$\begin{aligned}
\Omega_2 &= \text{Var}(E[Y(1) - Y(0)]|X) + E[\text{Var}(Y(1)|X)/p_0(X)] + E[\text{Var}(Y(0)|X)/(1 - p_0(X))] \\
&= \text{Var}(0.5) + E[1/p_0(X)] + E[1/(1 - p_0(X))] \\
&= 0 + \int_{0.15}^{0.85} \frac{1}{x} \frac{1}{0.7} dx + \int_{0.15}^{0.85} \frac{1}{1-x} \frac{1}{0.7} dx \\
&\approx 2 \times 2.47 = 4.94
\end{aligned}$$

The simulation results are in Table 2:

Table 2: ATE model

n	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\beta$	n	$\hat{\mu}^*_\beta$	$\hat{\sigma}^{2*}_\beta$
100	0.4242	6.0707	100	0.6692	2.9584
1000	0.4846	5.3859	1000	0.4794	5.8949
2000	0.4900	5.2478	2000	0.5702	5.2076
5000	0.4943	4.9404	5000	0.5013	4.8445
10000	0.4964	4.9492	10000	0.4920	5.3305
∞	0.5	4.94	∞	0.5	4.94

All the simulation settings are similar to those of Table 1, so do the outcomes. In general, Monte Carlo averages and variances for both original and bootstrap samples converge to their theoretical limits. Overall, the simulation outcomes for both MAR and ATE are in accordance with our theoretical results in the previous section.

5.2 Comparison with parametric plug-in estimators

5.2.1 With correctly specified parametric model

Here we compare the performances of two average treatment effect estimators, whose propensity scores are estimated with probit estimation and isotonic estimation. We consider the following setting:

$$Y = X'\gamma_0 + T \cdot \beta_0 + \varepsilon$$
$$T = \begin{cases} 0 & \text{if } X'\alpha_0 < \nu \\ 1 & \text{if } X'\alpha_0 \geq \nu \end{cases}, \tag{21}$$

where $X \overset{\text{i.i.d.}}{\sim} U[-1,1]^3$. ε and v are independently distributed standard normal random variables. Under this setting, we have $Pr(T = 1|X = x) := p_0(x) = \Phi(x'\alpha_0)$, where Φ is the CDF of the standard normal distribution. $\alpha_0' = (1,1,1)/\sqrt{3}$, $\beta_0 = 0.5$ and $\gamma_0' = (0.1, 0.2, 0.3)$. The propensity score is correctly specified in a probit estimation. We are interested in the average treatment effect β_0.

Table 3: ATE of the model (21) with plug-in probit and isotonic estimators

	probit			normalized probit			isotonic		
n	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\beta$	MSE	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\beta$	MSE	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\beta$	MSE
100	0.5018	5.9972	5.9975	0.5045	5.7167	5.7187	0.4823	5.8732	5.9047
1000	0.5025	5.2794	5.2855	0.5025	4.9949	5.0010	0.4956	5.0885	5.1081
2000	0.4996	5.4129	5.4133	0.4997	5.0820	5.0822	0.4951	5.1846	5.2330
5000	0.5004	5.4781	5.4788	0.5006	5.2139	5.2154	0.4982	5.2466	5.2634
10000	0.5002	5.3383	5.3388	0.5004	5.0288	5.0303	0.4987	5.0643	5.0807

Table 3 shows the simulation results based on 5000 Monte Carlo replications. The sample sizes are $n = 100$, 1000, 2000, 5000, and 10000. The variances and MSE is scaled with n. In the left panel and the right panel, the ATE estimators $\hat{\beta}$ are calculated with (12), where the inversed propensity weights are not normalized. In the middle panel we normalize the weights to unity. The estimator in the middle panel is calculated by

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i \cdot T_i}{\hat{p}(X_i)} \Big/ \left(\sum_{i=1}^{n}\frac{T_i}{\hat{p}(X_i)}\right) - \frac{Y_i \cdot (1-T_i)}{1-\hat{p}(X_i)} \Big/ \left(\sum_{i=1}^{n}\frac{1-T_i}{1-\hat{p}(X_i)}\right)\right\}$$

From Table 3, we can see that the ATE with isotonic plug-in estimators (the right panel) outperforms the ATE with correctly specified parametric plug-in estimators without normalization (the left panel), in every sample size. If we normalize the parametrically estimated propensity scores, the probit models perform better, as pointed out by Imbens (2004). With the sample size growing, the performance of the ATE with isotonic plug-in estimators are converging to those with correctly specified parametric plug-in estimators with normalization (the middle panel). With $n = 10000$, they are very close to each other. We can conclude that our semiparametric method performs similarly to the parametric method under the correct model specification.

5.2.2 Robustness

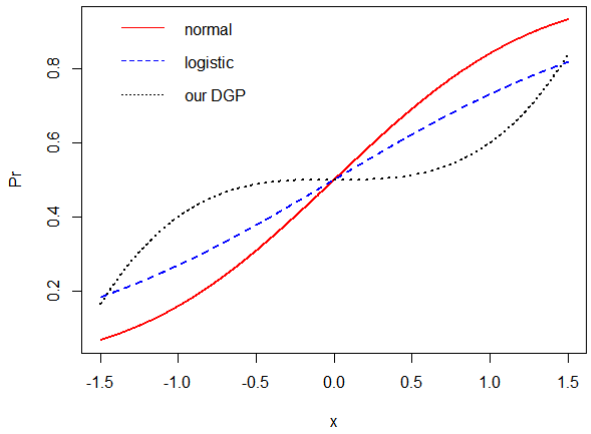
Compared to the popular choice of parametric models for propensity scores, such as the binary probit model or logit model, our semiparametric estimator is robust to the model specification. Considering the following setting:

$$Y = X^3 \cdot \gamma_0 + T \cdot \beta_0 + \varepsilon \tag{22}$$

$$\text{with } Pr(T = 1|X = x) = x^3/10 + 0.5, \tag{23}$$

where $\varepsilon \sim N(0, 1)$ and independent from X and T, $\gamma_0 = 1$, and $\beta_0 = 0.5$. $X \sim U[-1.5, 1.5]$. Figure 1 describes the function (23), the CDF of the standard normal distribution and the logistic function.

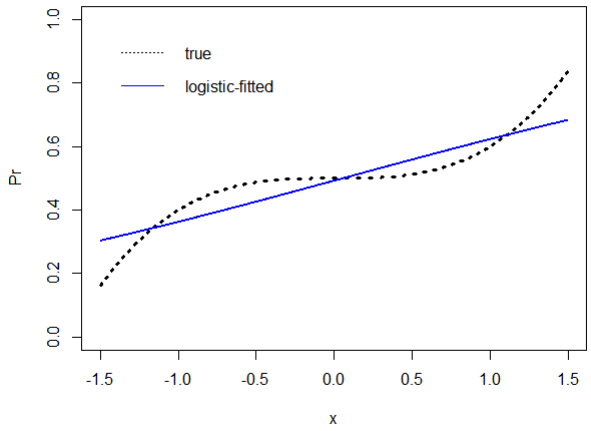
Figure 1: Normal CDF, logistic function, and the DGP (23)



The dotted black line is the DGP (23). The solid red line is the CDF of standard normal, $y = \Phi(x)$. The dashed blue line is the logistic function, $\Pr(T = 1|X = x) = \frac{\exp(a+bx)}{\exp(a+bx)+1}$. Three lines intersect at $[0, 1/2]$.

The idea of (23) is to find a monotone increasing function, which cannot be well approximated by the common choices of parametric models, such as the probit model or the logit model. The function (23) is convex for $x > 0$ and concave for $x < 0$. If we use $\Pr(T = 1|X = x) = \frac{\exp(a+bx)}{\exp(a+bx)+1}$ to approximate this function, we have an almost linear fitted line. See Figure 2

Figure 2: . The function (23) fitted with logistic function.



The dotted black line is the DGP (23). The solid red line is the CDF of standard normal, $y = \Phi(x)$. The dashed blue line is the logistic function, $y = \frac{\exp(a+bx)}{\exp(a+bx)+1}$. Three lines intersect at the point $[0, 1/2]$.

While this line roughly fits the quasi-linear part of the function (23) (the piece around zero), the departure becomes large for $|x| > 1.2$. If the outcome y has large values far from zero, as the case in (22), we might have large estimation bias. Table 4 confirms this conjecture.

Table 4: ATE estimated with logistic and isotonic plug-in estimator

	logistic			isotonic		
n	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\beta$	MSE	$\hat{\mu}_\beta$	$\hat{\sigma}^2_\beta$	MSE
1000	0.5930	5.6958	14.3380	0.4735	5.0426	5.7442
2000	0.6044	5.6533	27.4569	0.4824	4.8256	5.4446
5000	0.6153	5.5104	71.9331	0.4886	4.6304	5.2748

Table 3 shows the simulation results based on 5000 Monte Carlo replications. The sample sizes are $n = 1000$, 2000, and 5000. The variances and MSE's are scaled with n. In the left panel, the propensity score is estimated with the logistic function $y = \frac{\exp(a+bx)}{\exp(a+bx)+1}$; in the right panel, the propensity score is estimated with the isotonic estimation. We can see that the misspecified logit model cannot lead to satisfying estimators, and it presents stable biases and growing MSE's. Isotonic plug-in estimators do not suffer from this issue and have stable performances across different sample sizes.

5.3 Comparison with other non-parametric plug-in estimators: smoothness conditions

$\sqrt{n}-$consistency and efficiency can also be achieved with series or kernel plug-in estimators. However, tuning parameters should be carefully chosen, such that the high-order residual term and bias term could disappear at fast rates. Moreover, the smoothness conditions for the nui-

sance function can sometimes be demanding. For ATE estimators, Hirano, Imbens, and Ridder (2003) require that

$$p_0(\mathrm{x}) \text{ is continuously differentiable of order } s \geq 7.$$

Compared to our assumption:

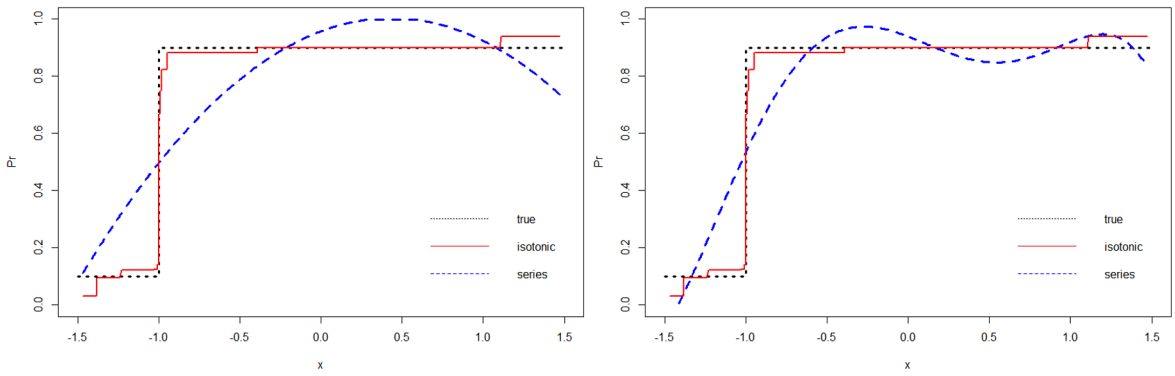
$$p_0(\mathrm{x}) \text{ is monotone increasing.}$$

We even do not need continuity. Consider

$$\begin{aligned} Y &= X \cdot \gamma_0 + T \cdot \beta_0 + \varepsilon \\ p_0(x) &= Pr(T = 1|X = x) = 0.1 + 0.8 \times 1(x > -1) \end{aligned} \tag{24}$$

where $\varepsilon \sim N(0, 1)$ and independent from X and T, $\gamma_0 = 1$, and $\beta_0 = 0.5$. $X \sim U[-1.5, 1.5]$. We see from (24) that $p_0(x)$ is a step probability function with a jump point at -1. Figure 3 describe $p_0(x)$ and curves fitted with series estimator and isotonic estimator.

Figure 3: The function (24) fitted with series estimators and isotonic estimators



The sample size $n = 1000$. The black dotted lines are the function (24). The blue dashed lines are fitted with series estimators. The red lines are fitted with isotonic estimators. In the left panel the series length $k = 3$. In the right panel the series length $k = 6$.

We see that series estimators cannot fit the discrete function (24) very well, while isotonic estimators do good jobs. The results are collected in Table 5. It compares ATE estimates with series and isotonic plug-in estimators based on 5000 Monte Carlo replications. The sample sizes are $n = 100$, 1000, 2000, 5000, and 10000. The MSE's are scaled with n. Series estimations are conducted with different series lengths ranging from 3 to 6.

Table 5: ATE estimated with series and isotonic plug-in estimator

	series								isotonic	
length	3		4		5		6		–	
n	$\hat{\mu}_\beta$	MSE	$\hat{\mu}_\beta$	MSE	$\hat{\mu}_\beta$	MSE	$\hat{\mu}_\beta$	MSE	$\hat{\mu}_\beta$	MSE
100	0.01	488.48	0.57	100.40	0.56	89.05	0.46	258.53	0.29	22.09
1000	-0.35	1637.11	0.43	72.82	0.44	73.97	0.42	229.40	0.42	19.28
2000	-0.49	3341.69	0.43	67.86	0.44	69.87	0.41	198.41	0.44	19.68
5000	-0.64	8470.42	0.43	82.86	0.45	68.90	0.37	241.87	0.46	20.59
10000	-0.73	17814.28	0.43	112.95	0.45	76.43	0.35	384.80	0.47	21.07

We can see that estimates with the series length 4 and 5 perform comparatively good, but their MSE's are still considerably larger than those with isotonic plug-in estimators, and the biases of them seem not to shrink with the sample size growing. In comparison, the estimates with isotonic plug-in estimators in the last two columns perform the best: MSE's are much lower, and with the sample size growing, biases are shrinking towards zero. Overall, Table 5 highlights two merits of our method: (i) it saves us the bother of selecting the tuning parameter that delivers the best result; (ii) its performances remain stable and well in the case of non-smooth nuisance functions.

6 Application

Since the work of LaLonde (1986), National Supported Work (NSW) data and its different variations were analyzed by many authors, including Dehejia and Wahba (1999, 2002), Smith and Todd (2004), Dehejia (2005). We follow the setting in Dehejia and Wahba (1999) (hereafter, DW). The data is downloaded from the website of Rajeev Dehejia (http://users.nber.org/˜rdehejia/).

6.1 Data description

The dataset is a combination of observations from NSW and two other datasets, Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS). In the NSW dataset, the treatment was randomly assigned, and thus the ATE estimator calculated from the NSW dataset can be regarded as unbiased and serve as a benchmark. Since no observation in PSID and CPS was treated, the dataset, which combines the treated observations from NSW and the observations from PSID and CPS, can be regarded as a non-experimental dataset. The comparison of estimators from the NSW dataset and this combined dataset can be used to evaluate the non-experimental methods.

DW presents estimators from combinations of the NSW treated group and different subsets of PSID and CPS. In our application, we use the PSID-2 as the control group, which is the second row of Table 3 in DW.

6.2 Estimation results

We choose the same set of covariates for the subset PSID-2 as DW. The details are in the description under DW's Table 3. Given these covariates, we estimate ATE and ATT with plug-in logistic estimators and isotonic estimators. In Table 6, we compare these four estimators with those obtained by DW for the same dataset.

Table 6: NSW-PSID2 estimation

Method	Propensity score	$\hat{\beta}$	$se(\hat{\beta})$
NWS random (benchmark)	—	1,794	633
DW's stratifying estimator	logistic	2,220	1,768
DW's matching estimator	logistic	1,455	2,303
IPW ATE estimator	logistic	1,888	2,175
IPW ATE estimator	isotonic	1,841	1,723
IPW ATT estimator	logistic	1,870	1,149
IPW ATT estimator	isotonic	1,802	1,496

The first three rows are from DW's Table 3. The last four rows are from our calculations. The standard errors are calculated with bootstrap.

All the estimators from non-experimental data have comparatively large standard deviations. This is in line with the results of other authors analyzing this dataset. Compared to other estimators, the ATE and ATT estimators with isotonic plug-in estimators seem to be closer to the benchmark estimator in the first row, than other non-experimental estimators. While the standard deviation of the ATT estimator with the isotonic plug-in estimator is larger than its counterpart with the logistic plug-in estimator, the standard deviation of the ATE estimator with the isotonic plug-in estimator is smaller than its counterpart. Overall, the application results support our estimation strategy.

7 Conclusion

We study a general framework of semiparametric estimation with plug-in isotonic estimators. We show that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal. In the univariate cases, the estimator is efficient. It generalizes the estimation methods of existing semiparametric models with monotone nuisance functions in the literature. Furthermore, we apply the estimator to the case of inverse probability weighting for ATE models, where the propensity scores are assumed to be monotone increasing. In this setting, the monotonicity assumption is a natural implication of the binary selection model and is satisfied by many parametric models widely adopted in applied work.

We show that while the proposed estimator has a similar performance to methods with parametric plug-in estimators under correct specifications, it is more robust against misspecification than the latter. Compared to methods with other nonparametric plug-in estimators,

the newly proposed method requires minimum smoothness conditions on nuisance functions. Finally, we establish the asymptotic validity of the bootstrap, which ensures that the estimator is tuning-parameter-free in both estimation and inference.

A Mathematical appendix

A.1 Proof of Lemma 1

The proof here is based on the supplementary material of BGH (hereafter BGH-supp). Similar techniques can also be found in Groeneboom & Jongbloed (2014) and Groeneboom & Hendrickx (2018).

Let $\{x_{n_j}\}_{j=1}^k$ be the subsequence of $\{x_i\}_{i=1}^n$ representing all the jump points of $\hat{p}(\cdot)$. By the construction of $\hat{p}(\cdot)$ (see, e.g., Lemmas 2.1 and 2.3 in Groeneboom and Jongbloed, 2014), we have $\sum_{i=n_j}^{n_{j+1}-1}\{y_i - \hat{p}(x_i)\} = 0$ for each $j = 1, \ldots, k$, which implies

$$\sum_{j=1}^{k} m_j \sum_{i=n_j}^{n_{j+1}-1} \{y_i - \hat{p}(x_i)\} = 0, \tag{25}$$

for any weights $\{m_j\}_{j=1}^k$. (See also Barlow and Brunk, 1972). We define the step function $\bar{\delta}_n(x)$:

$$\bar{\delta}_n(x) = \begin{cases} \delta(x_{n_j}) & \text{if } p_0(x) > \hat{p}(x_{n_j}) \text{ for all } x \in (x_{n_j}, x_{n_{j+1}}) \\ \delta(s) & \text{if } p_0(s) = \hat{p}(s) \text{ for some } s \in (x_{n_j}, x_{n_{j+1}}) \\ \delta(x_{n_{j+1}}) & \text{if } p_0(x) < \hat{p}(x_{n_j}) \text{ for all } x \in (x_{n_j}, x_{n_{j+1}}) \end{cases},$$

for $x \in [x_{n_j}, x_{n_{j+1}})$ with $j = 1, \ldots, k$ (if $j = k$, set $x_{n_{j+1}} = \max_i x_{n_i}$). By (25), it holds

$$\int \bar{\delta}_n(x)\{y - \hat{p}(x)\} d\mathbb{P}_n(z) = 0,$$

Thus, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} \delta(X_i)(Y_i - \hat{p}(X_i)) \\ = & \int \delta(x)\{y - \hat{p}(x)\} d\mathbb{P}_n(z) \\ = & \int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x)) d\mathbb{P}_n(z) \end{aligned} \tag{26}$$

By assumption, $\delta(x)$ is a bounded function with a finite total variation, so is $\bar{\delta}_n(x)$. Therefore, by a similar argument as in pp. 18-20 of BGH-supp, we have $\int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x)) d\mathbb{P}_n(z) = o_p(n^{-1/2})$. We see that (26) can be decomposed as:

$$
\begin{aligned}
&\int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x)) d\mathbb{P}_n(z) \\
=&\int [\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x)) d\left(\mathbb{P}_n(z) - \mathbb{P}_0(z)\right) \\
+&\int [\delta(x) - \bar{\delta}_n(x)](y - p_0(x)) d\mathbb{P}_0(z) \\
+&\int [\delta(x) - \bar{\delta}_n(x)](p_0(x) - \hat{p}(x)) d\mathbb{P}_0(z) \\
:=&I + II + III
\end{aligned}
$$

By Lemma 21 in BGH-supp, both $\delta(x) - \bar{\delta}_n(x)$ are bounded functions with finite total variations. With similar arguments in Groeneboom and Jongbloed (2014) we have some $C_0 > 0$, with all $x \in \mathcal{X}$

$$
|\delta(x) - \bar{\delta}_n(x)| \leq C_0 |p_0(x) - \hat{p}(x)| \tag{27}
$$

For I, let us define the following function classes

$$
\begin{aligned}
\mathcal{M}_{RK} &= \{\text{monotone increasing functions on } [-R, R] \text{ and bounded by } K\}, \\
\mathcal{G}_{RK} &= \{g : g(x) = p(x), x \in \mathcal{X}, p \in \mathcal{M}_{RK}\}, [\text{seems can be removed}] \\
\mathcal{D}_{RKv} &= \{d : d(x) = g_1(x) - g_2(x),\ (g_1, g_2) \in \mathcal{G}_{RK}^2,\ ||d(x)||_{P_0} \leq v\}, \\
\mathcal{H}_{RKv} &= \{h : h(y, x) = y d_1(x) - d_2(x),\ (d_1, d_2) \in \mathcal{D}_{RKv}^2,\ z \in \mathcal{Z}\}.
\end{aligned} \tag{28}
$$

And we have the integrand of I

$$
\begin{aligned}
&[\delta(x) - \bar{\delta}_n(x)](y - \hat{p}(x)) \\
=&[\delta(x) - \bar{\delta}_n(x)]y - [\delta(x) - \bar{\delta}_n(x)]\hat{p}(x)
\end{aligned} \tag{29}
$$

Let

$$
\mathcal{F}_a = \left\{ f : f(z) = [\delta(x) - \bar{\delta}_n(x)]y - [\delta(x) - \bar{\delta}_n(x)]\hat{p}(x),\ z \in \mathcal{Z} \right\}.
$$

We note:

(i) By Lemma 21 in BGH-supp, $[\delta(x) - \bar{\delta}_n(x)]$ is a bounded function of x with finite total variation.

(ii) By Assumption A3, we can show $\sup_{x \in \mathcal{X}} |\hat{p}(x)| = O_p(\log n)$ (See, e.g., Lemma 7.1 in Balabdaoui, Durot, and Jankowski, 2019). Therefore, there exists $K_1 > 0$, such that $\hat{p}(x) \in \mathcal{G}_{R(K_1 \log n)}$ with probability approaching one.

(iii) By (10) and (27), we have $||\delta(x) - \bar{\delta}_n(x)||_2 \leq C_1 (\log n) n^{-1/3}$, for some $C_1 > 0$. Thus, there exists a positive constant C_2 that is larger than twice the bound of $\delta(x)$, and $v_1 =$

$C_1(\log n)n^{-1/3}$, such that $[\delta(x) - \bar{\delta}_n(x)] \in \mathcal{D}_{RC_2v_1}$.

(iv) By (ii), a similar argument of (iii), (10), and Jensen's inequality, we have $[\delta(x) - \bar{\delta}_n(x)]\hat{p}(x) \in \mathcal{D}_{R(K_2 \log n)v_2}$ for a large enough constant $K_2 > 0$ and $v_2 = C_3(\log n)^2 n^{-1/3}$ for some $C_3 > 0$, with probability approaching one.

We choose $K = \max\{C_2, K_2 \log n\}$ and $v = \max\{v_1, v_2\}$. Now we have (29)$\in \mathcal{H}_{RKv}$.

Now we define some notations. Let $||\mathbb{G}_n||_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\sqrt{n}(\mathbb{P}_n - P_0)f|$,

$$H_B(\varepsilon, \mathcal{F}, ||\cdot||) = \log N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||)$$

be the entropy of the ε-bracketing number of the function class $\mathcal{F}$ under the norm $||\cdot||$, and

$$J_n(\delta, \mathcal{F}, ||\cdot||) \stackrel{\text{def.}}{=} \int_0^\delta \sqrt{1 + H_B(\varepsilon, \mathcal{F}, ||\cdot||)}d\varepsilon.$$

Let $||\cdot||_{B,P_0}$ be the Bernstein norm under a measure P_0. In this section, we use $J_n(\delta)$ to denote $J_n(\delta, \mathcal{F}, ||\cdot||_{B,P_0})$.

By similar arguments in Lemma 13 of BGH-supp (In our case we can ignore the single-index coefficients), we have, with probability approaching one:

$$H_B(\varepsilon, \tilde{\mathcal{F}}_a, ||\cdot||_{B,P_0}) \leq \frac{C_3}{\varepsilon}, \tag{30}$$

for some $C_3 > 0$, where $\tilde{\mathcal{F}}_a = (C_4 \log n)^{-1}\mathcal{F}_a$ with some $C_4 > 0$. Also, there exists a constant $C_5 > 0$ such that

$$||\tilde{f}||_{B,P_0} \leq C_5(\log n)n^{-1/3}, \tag{31}$$

for all $\tilde{f}_a \in \tilde{\mathcal{F}}_a$, with probability approaching one. We use $\mathscr{E}$ to denote the event that both (30) and (31) happen, and we have $\lim_{n\to\infty} P(\mathscr{E}) = 1$.

Let $\delta_n = C_5(\log n)n^{-1/3}$ and I_j be the j-th component of I. For any positive constants A and ν, there exist positive constants B_1, and B_2, for all n large enough, such that

$$
\begin{aligned}
P\{|I_j| > An^{-1/2}\} &\leq P\left\{|I_j| > An^{-1/2}, \mathscr{E}\right\} + P(\mathscr{E}^c) \\
&\leq P\left\{||\mathbb{G}_n||_{\mathcal{F}_a} > A, \mathscr{E}\right\} + \frac{\nu}{2} \\
&\leq \frac{E[||\mathbb{G}_n||_{\mathcal{F}_a} \quad |\mathscr{E}]}{A} + \frac{\nu}{2} \\
&= \frac{C_4 \log n}{A} E[||\mathbb{G}_n||_{\tilde{\mathcal{F}}_a} \quad |\mathscr{E}] + \frac{\nu}{2} \\
&\lesssim \frac{C_4 \log n}{A} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A}(\delta_n + 2B_1^{1/2}\delta_n^{1/2})\left(1 + \frac{\delta_n + 2B_1^{1/2}\delta_n^{1/2}}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_2}{(\log n)^{3/2}}\right) + \frac{\nu}{2} \\
&\lesssim \nu, \qquad (32)
\end{aligned}
$$

The second inequality follows from the definition of $\mathcal{F}_a$; The third inequality follows from the Markov inequality, the first equality follows from the definition of $\tilde{\mathcal{F}}_a$, the first wave inequality ($\lesssim$) comes from Lemma 3.4.3 of van der Vaart and Wellner (1996) and the definition of δ_n, the second wave inequality comes from (30) and Equation (.2) in BGH-supp, the third wave inequality follows from $\delta_n \lesssim \delta_n^{1/2}$ and the definition of δ_n. Therefore,

$$
I = o_p(n^{-1/2}). \qquad (33)
$$

For II, we have by the law of iterated expectation.

$$
II = \int [\delta(x) - \bar{\delta}_n(x)](y - p_0(x)) d\mathbb{P}_0(z) = 0
$$

For III, we have

$$
\begin{aligned}
III &= \int [\delta(x) - \bar{\delta}_n(x)](p_0(x) - \hat{p}(x)) d\mathbb{P}_0(z) \\
&\lesssim \int (p_0(x) - \hat{p}(x))^2 d\mathbb{P}_0(z) \\
&= O_p((\log)^2 n^{-2/3}) = o_p(n^{-1/2}),
\end{aligned}
$$

Where the first wave inequality follows from (30), the second equality follows from (10). Finally, we can conclude that

A.2 Proof of Proposition 1

Under A1-A4, we have $\frac{1}{n}\sum_{i=1}^{n} E[D(Z,\beta_0)|X_i](Y_i - \hat{p}(X_i)) = o_p(n^{-1/2})$. Then we have

$$\frac{1}{n}\sum_{i=1}^{n} m(z_i,\beta,\hat{p}(\cdot)) = 0 \tag{34}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \{m(z_i,\beta,\hat{p}(\cdot)) + E[D(Z,\beta_0)|X_i](Y_i - \hat{p}(X_i))\} = o_p(n^{-1/2}) \tag{35}$$

Let $\hat{\beta}$ be the solution of (34), and $\tilde{\beta}$ be the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \{m(z_i,\beta,\hat{p}(\cdot)) + E[D(Z,\beta_0)|X_i](Y_i - \hat{p}(X_i))\} = 0.$$

Then by (35), $\sqrt{n}(\hat{\beta} - \beta_0)$ and $\sqrt{n}(\tilde{\beta} - \beta_0)$ should have the same limit distribution.

A.3 Proof of Theorem 1

The proof is a combination of the techniques for isotonic regression applied in Groeneboom and Hendrickx (2018) and BGH, and the framework of Newey (1994).

Let $u = y - p_0(x)$ and $M(z) = \delta(x)u$. We verify the assumptions 5.1-5.6 in Newey (1994).

Step 1: Verify Assumption 5.1 in Newey (1994).

Assumption 5.1 (Newey, 1994): (i) There is a function $D(z,p)$ that is linear in p such that for all p with $||p - p_0||$ small enough,

$$||m(z,p) - m(z,p_0) - D(z,p-p_0)|| \leq b(z)||p - p_0||^2;$$

(ii) $E(b(z))\sqrt{n}||\hat{p} - p_0||^2 \xrightarrow{p} 0$

(i) is a restatement of A6 (i). (ii) can be derived by A6(ii) and the fact

$$||\hat{p} - p_0||^2 = O_p((\log n)^2 n^{-2/3})$$

(See, e.g., Lemma 5.15 in van de Geer, S., 2000).

Step 2: Verify Assumption 5.2 in Newey (1994).

Assumption 5.2 (Newey, 1994): $\frac{1}{n}\sum_{i=1}^{n} D(z,\hat{p}(x) - p_0(x)) - \int D(z,\hat{p}(x) - p_0(x))d\mathbb{P}_0(z) = o_p(n^{-1/2})$.

By A5, we have

$$\frac{1}{n}\sum_{i=1}^{n} D(Z,\beta_0,\hat{p}(x)-p_0(x))-\int D(z,\beta_0,\hat{p}(x)-p_0(x))d\mathbb{P}_0(z) = \int D(z,\beta_0)(p_0(x)-\hat{p}(x))d(\mathbb{P}_n-\mathbb{P}_0)(z) \tag{36}$$

let

$$\mathcal{F}_b = \big\{f : f(z) = D(z,\beta_0)(p_0(x)-\hat{p}(x)),\ x\in\mathcal{X}\big\}.$$

To avoid heavy notations, we re-define some constant terms in this section, such as A_i, C_i, K_i,δ_n v, etc.. They are not related to the same constants in other sections.

By similar arguments as in Section A.1, for some $C_1, C_2 > 0$, we have

$$p_0(x)-\hat{p}(x) \in \mathcal{D}_{R(C_1\log n)\left(C_2 n^{-1/3}\log n\right),} \tag{37}$$

with probability approaching one.

By Theorem 2.7.5 in van der Vaart and Wellner (1996) and Lemma 11 in BGH-supp, with $R, C, v > 0$, we have

$$H_B(\varepsilon, \mathcal{D}_{RCv}, ||\cdot||_{P_0}) \le \frac{AC}{\varepsilon},$$

for some $A > 0$. Now we define

$$\mathcal{H}^{(2)}_{RKv} = \{h : h(z) = D(z,\beta_0)d(x),\ d(\cdot)\in\mathcal{D}_{RCv},\ z\in\mathcal{Z}\}.$$

Let $\left(d^L, d^U\right)$ to be any ϵ-bracket of the function class $\mathcal{D}_{RKv}$.

Let us define

$$h^L = \begin{cases} D(z,\beta_0)d^L(x) & \text{if } D(z,\beta_0)\ge 0 \\ D(z,\beta_0)d^U(x) & \text{if } D(z,\beta_0)<0 \end{cases},$$

and

$$h^U = \begin{cases} D(z,\beta_0)d^U(x) & \text{if } D(z,\beta_0)\ge 0 \\ D(z,\beta_0)d^L(x) & \text{if } D(z,\beta_0)<0 \end{cases}.$$

We see that $\left(h^L, h^U\right)$ is a bracket of h, its size is

$$\begin{aligned}\int_{\mathcal{Z}} \left[h^U(z)-h^L(z)\right]^2 d\mathbb{P}_0(z) &= \int_{\mathcal{Z}} D(z,\beta_0)^2\left(d^U(x)-d^L(x)\right)^2 d\mathbb{P}_0(z) \\ &= \int_{\mathcal{X}} E\left[D(z,\beta_0)^2|x\right]\left(d^U(x)-d^L(x)\right)^2 d\mathbb{P}_0(x) \\ &= A_1\epsilon^2,\end{aligned}$$

for some $A_1 > 0$. The last equality follows from Assumption A4 and the definition of ϵ-bracket. Now for some $\tilde{A} > 0$, we have

$$H_B(\varepsilon, \mathcal{H}^{(2)}_{RCv}, ||\cdot||_{P_0}) \leq \frac{\tilde{A}C}{\varepsilon}. \tag{38}$$

Now we switch to Bernstein norm since we do not want to put a bound on $D(z, \beta_0)$. By the definition of Bernstein norm

$$\begin{aligned} ||h||^2_{B,P_0} &= 2\mathbb{P}_0\left[\exp(|h|) - |f| - 1\right] \\ &= 2\int \sum_{k=2}^{\infty} \frac{1}{k!}|h|^k d\mathbb{P}_0(z), \end{aligned}$$

by the extension of the natural exponential function. Now we try to bound the Bernstein norm of $\frac{h}{H}$, where H is some positive number we choose in the following steps to achieve a finite upper bound.

$$\begin{aligned} ||H^{-1}h||^2_{B,P_0} &= 2\int \sum_{k=2}^{\infty} \frac{1}{H^k}\frac{1}{k!}|D(z,\beta_0)d(x)|^k d\mathbb{P}_0(z) \\ &\leq 2\int \sum_{k=2}^{\infty} \frac{1}{H^k}\frac{1}{k!}|D(z,\beta_0)|^k|d(x)|^k d\mathbb{P}_0(z) \\ &\leq 2\sum_{k=2}^{\infty} \frac{1}{H^k}\frac{(2C)^{k-2}}{k!}k!M_1^{k-2}c_1\int |d(x)|^2 d\mathbb{P}_0(z) \\ &= \frac{2}{H^2}\sum_{k=2}^{\infty}\frac{(2M_1C)^{k-2}}{H^{k-2}}c_1\int |d(x)|^2 d\mathbb{P}_0(z) \\ &= \frac{2}{H^2}\sum_{k=2}^{\infty}\left(\frac{2M_1C}{H}\right)^{k-2}c_1 v^2 \\ &= \left(\frac{2}{H}\right)^2 c_1 v^2 \end{aligned}$$

The second inequality follows from Assumption A4 and the fact $d(\cdot) \in \mathcal{D}_{RCv}$, where c_1 and M_1 are the same constants in Assumption A4. (different from the capital C_1 defined before (37)) The third equality follows from the definition of v in$\mathcal{D}_{RCv}$. The last equality follows by choosing $H = 4M_1C$. Now we have

$$||\frac{h}{H}||_{B,P_0} \lesssim \frac{v}{H} \tag{39}$$

Now we set $C = C_1 \log n$, $v = C_2 n^{-1/3}\log n$

$$\mathcal{F}_b \subset \mathcal{H}^{(2)}_{R(C_1\log n)\left(C_2 n^{-1/3}\log n\right)}$$

and let $\tilde{H} = 4M_1C_1\log n$, then we have for some $C_3 > 0$,

$$\tilde{\mathcal{F}}_b = \tilde{H}^{-1}\mathcal{F}_b$$

Combined with (38) and (39), we have with probability approaching one

$$H_B(\varepsilon, \tilde{\mathcal{F}}_b, ||\cdot||_{B,P_0}) \leq \frac{C_3}{\varepsilon}, \tag{40}$$

for some $C_3 > 0$, and

$$\text{and } ||\tilde{f}||_{B,P_0} \leq C_4 n^{-1/3}, \tag{41}$$

for all $\tilde{f}_b \in \tilde{\mathcal{F}}_b$, for some $C_4 > 0$.

We use $\mathscr{E}_1$ to denote the event described in (40) and (41), S to denote the value of (36). and $\delta_n = C_4 n^{-1/3}$. Now For any $A_2 > 0$.

$$\begin{aligned}
P\{|S| > A_2 n^{-1/2}\} &\leq P\left\{|S| > A_2 n^{-1/2}, \mathscr{E}_1\right\} + P(\mathscr{E}_1^c) \\
&\leq P\left\{||\mathbb{G}_n||_{\mathcal{F}_b} > A_2, \mathscr{E}_1\right\} + \frac{\nu}{2} \\
&\leq \frac{E[||\mathbb{G}_n||_{\mathcal{F}_b} \quad |\mathscr{E}_1]}{A_2} + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2} E[||\mathbb{G}_n||_{\tilde{\mathcal{F}}_b} \quad |\mathscr{E}_1] + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2}(\delta_n + 2B_1^{1/2}\delta_n^{1/2})\left(1 + \frac{\delta_n + 2B_1^{1/2}\delta_n^{1/2}}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_2}{(\log n)^{3/2}}\right) + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2} n^{-1/6} B_2 + \frac{\nu}{2} \\
&\lesssim \nu,
\end{aligned} \tag{42}$$

Each steps are similar to those of (32). Thus, we have $\int D(z,\beta_0)(p_0(x) - \hat{p}(x))d(\mathbb{P}_n - \mathbb{P}_0)(z) = o_p(n^{-1/2})$. Newey's Assumption 5.2 satisfied.

Assumption 5.3: $\int D(z, \hat{p}(x) - p_0(x))d\mathbb{P}_0(z) = \frac{1}{n}\sum_{i=1}^n M(z_i) + o_p(n^{-1/2})$. [4]

We have

$$\begin{aligned}
\int D(z, \beta_0, \hat{p}(x) - p_0(x))d\mathbb{P}_0(z) &= E[D(Z, \beta_0, \hat{p}(X) - p_0(X)] \\
&= E\big\{D(Z, \beta_0)(\hat{p}(X) - p_0(X))\big\} \\
&= E[E(D(Z, \beta_0)|X)(\hat{p}(X) - p_0(X))] \\
&= \int \delta(x)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x).
\end{aligned}$$

[4] This is a simplified version of Assumption 5.3, which is mentioned in p.1366 in Newey (1994).

The second equality follows from A5. In the last equality we set $E(D(Z,\beta_0)|X=x) := \delta(x)$. Therefore, by plugging in $M(z) = \delta(x)u$

$$\begin{aligned}
&\int D(z,\hat{p}(x)-p_0(x))d\mathbb{P}_0(z) - \frac{1}{n}\sum_{i=1}^{n} M(Z_i) \\
&= \int \delta(x)(\hat{p}(x)-p_0(x))d\mathbb{P}_0(x) - \frac{1}{n}\sum_{i=1}^{n}\delta(X_i)(Y_i - p_0(X_i)) \\
&= \int \delta(x)(\hat{p}(x)-p_0(x))d\mathbb{P}_0(x) - \int \delta(x)(y-\hat{p}(x)+\hat{p}(x)-p_0(x))d\mathbb{P}_n(z) \\
&= \int -\delta(x)(y-\hat{p}(x))d\mathbb{P}_n(z) + \int -\delta(x)(\hat{p}(x)-p_0(x))d(\mathbb{P}_n-\mathbb{P}_0)(x) \\
&:= I + II. \qquad (43)
\end{aligned}$$

By Lemma 1, we have $I = o_p(n^{-1/2})$.

For II, by A4 and a similar argument as in p. 23 of BGH-supp, we have $II = o_p(n^{-1/2})$. Assumption 5.3 is satisfied.

Assumptions 5.4 to 5.6 are adapted as A7 to A9 in this paper. Then the consistency is proved by Lemma 5.2 of Newey (1994). Finally, we have by Lemma 5.3 of Newey (1994)

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,V),$$

where

$$V = M_\beta^{-1} E[\{m(Z,\beta_0,p_0)+M(Z)\}\{m(z,\beta_0,p_0)+M(Z)\}']M_\beta^{-1},$$

The efficiency is proved according to Proposition 4 of Newey (1994) (See also his Theorem 2.1).

A.4 Proof of Corollary 1

Let us check A1 to A9 of Theorem 1 for $m(Z,\beta_0,p(\cdot)) = \frac{Y\cdot T}{p_0(X)} - \frac{Y\cdot(1-T)}{1-p_0(X)} - \beta_0$.

C2 directly implies A1; C4 implies A2; A3 is satisfied by the fact that $T \in \{0,1\}$.

For A4, we have for the ATE model $E[D(Z,\beta)|X] = -(\frac{\mu_1(x)}{p_0(x)} + \frac{\mu_0(x)}{1-p_0(x)})$. It a bounded function of X with finite total variation by C2 and C3.

A5 is satisfied since we have $D(z,\beta,p(x)-p_0(x)) = \left(\frac{y\cdot t}{p_0(x)^2} + \frac{y\cdot(1-t)}{(1-p_0(x))^2}\right)(p(x)-p_0(x))$.

A6-A9 is satisfied by the same arguments in pp.26-33 of Hirano, Imbens, and Ridder (2000).

Therefore, we have all the assumptions for Theorem 1 satisfied. The asymptotical variance matrix Ω can be obtained in the same way as pp.34-35 of Hirano, Imbens, and Ridder (2000).

A.5 Proof of Lemma 2.

The additional complication caused by the possible dependence of $p(\cdot)$ on β does not affect this lemma. The proof is similar to that for Lemma 1 in Appendix A.1, with Y replaced by $T(Z,\beta)$.

A.6 Proof of Proposition 2.

The proof is similar to that of Proposition 1 in Appendix A.2.

A.7 Proof of Theorem 2

Here we might not be able to solve the sample moment condition (2)

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta,\hat{p}_\beta(\cdot)) = 0,$$

as we did in Theorem 1, since changing β will change the left-hand side discretely.

Now for $\beta \in \mathcal{B}(\beta_0,\delta_0)$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta,\hat{p}_\beta(X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i,\beta,p_\beta(X_i)) + D(Z,\beta)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i,\beta,p_\beta(X_i)) + D(Z,\beta)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z,\beta)|X_i)(T(Z_i,\beta) - \hat{p}_\beta(X_i)) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i,\beta,p_\beta(X_i)) + D(Z,\beta)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} \{E(D(Z,\beta)|X_i)(T(Z_i,\beta) - p_\beta(X_i)) + E(D(Z,\beta)|X_i)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i,\beta,p_\beta(X_i)) + E(D(Z,\beta)|X_i)(T(Z_i,\beta) - p_\beta(X_i))\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} [D(Z,\beta) - E(D(Z,\beta)|X_i)][(\hat{p}_\beta(X_i) - p_\beta(X_i)] + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i,\beta,p_\beta(X_i)) + E(D(Z,\beta)|X_i)(T(Z_i,\beta) - p_\beta(X_i))\} + o_p(n^{-1/2}) \qquad (44)
\end{aligned}$$

The first equality follows from A5' and A6'. The second equality follows from Lemma 2. The third equality and the fourth equality are some rearrangements. The last equality is by $\frac{1}{n}\sum_{i=1}^{n}[D(Z,\beta) - E(D(Z,\beta)|X_i)][(\hat{p}_\beta(X_i) - p_\beta(X_i)] = o_p(n^{-1/2})$, which can be proved by A4' and similar arguments in p.23 BGH-supp.

By (44) and the definition of $\hat{\beta}$ in (13), we have

$$\begin{aligned}
&||\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{p}_{\hat{\beta}}(X_i))|| \\
=&\inf_{\beta} ||\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta, \hat{p}_{\beta}(X_i))|| \\
\leq&\inf_{\beta} ||\frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_{\beta}(X_i)) + E(D(Z,\beta)|X_i)(T(Z_i,\beta) - p_{\beta}(X_i))\} + o_p(n^{-1/2})||.
\end{aligned}$$

The leading term in the last expression,

$$\frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_{\beta}(X_i)) + E(D(Z,\beta)|X_i)(T(Z_i,\beta) - p_{\beta}(X_i))\},$$

does not depend on the discrete estimator $\hat{p}(\cdot)$. It is a smooth moment function of β. Thus, under standard conditions on m, T, and p, we have

$$\inf_{\beta} ||\frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_{\beta}(X_i)) + E(D(Z,\beta)|X_i)(T(Z_i,\beta) - p_{\beta}(X_i))\}|| = 0,$$

and by (44) we have

$$||\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{p}_{\hat{\beta}}(X_i))|| = o_p(n^{-1/2}). \tag{45}$$

Let

$$\begin{aligned}
M_{n,\beta} &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial m(Z_i, \beta_0, p_0(X_i))}{\partial \beta} + E[D(Z_i,\beta_0)|X_i]\frac{\partial T(Z_i,\beta_0)}{\partial \beta}\right\}, \\
M_{\beta} &= -E\left\{\frac{\partial m(Z_i, \beta_0, p_0(X_i))}{\partial \beta} + E[D(Z_i,\beta_0)|X_i]\frac{\partial T(Z_i,\beta_0)}{\partial \beta}\right\}, \text{ and} \\
M(Z_i) &= E(D(Z,\beta_0)|X_i)(T(Z_i,\beta_0) - p_0(X_i)).
\end{aligned}$$

We have

$$
\begin{aligned}
o_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{p}_{\hat{\beta}}(X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{p}_{\hat{\beta}}(X_i)) + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z, \beta_0)|X_i)(T(Z_i, \hat{\beta}) - \hat{p}_{\hat{\beta}}(X_i)) \\
&= -M_{n,\beta}(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta_0, \hat{p}_{\hat{\beta}}(X_i)) + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z, \beta_0)|X_i)(T(Z_i, \beta_0) - \hat{p}_{\hat{\beta}}(X_i)) + o_p(\hat{\beta} - \beta_0) \\
&= -M_{\beta}(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta_0, p_0(X_i)) + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z_i, \beta_0)|X_i)(T(Z_i, \beta_0) - p_0(X_i)) + o_p(\hat{\beta} - \beta_0) \\
&= -M_{\beta}(\hat{\beta} - \beta_0) + \left\{ \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta_0, p_0(X_i)) + M(Z_i) \right\} \\
&+ o_p(n^{-1/2} + (\hat{\beta} - \beta_0)). \qquad (46)
\end{aligned}
$$

The first equality follows from (45). The second equality follows from Lemma 2. The third equality follows from the expansion around β_0 and the definition of $M_{n,\beta}$. The fourth equality follows from $M_{n,\beta} - M_\beta = o_p(1)$ and similar arguments in Step 1 and 2 of Appendix A.3. The last equality follows from the definition of $M(Z)$.

Based on (46), consistency of $\hat{\beta}$ can be similarly proved as in Lemma 5.2 in Newey (1994).

Finally, we have

$$
\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta_0) &= M_\beta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{m(Z_i, \beta_0, p_0(X_i)) + M(Z_i)\} + o_p(1) \\
&\xrightarrow{d} N(0, \Pi), \qquad (47)
\end{aligned}
$$

while $M_\beta^{-1}\{m(Z_i, \beta_0, p_0(X_i)) + M(Z_i)\}$ is the efficient influence function. (See pp.1357-1361 of Newey, 1994).

A.8 Proof of Lemma 3

The proof is very similar to pp. 18-20 of BGH-supp. We replace $E(X|S(\beta)'X)$ and Y_i in BGH-supp with $\delta(X'\alpha)$ and $T(Z_i, \beta)$ in our setting.

A.9 Proof of Theorem 3

Now the nuisance function $\hat{F}_{\hat{\alpha},\hat{\beta}}(x'\hat{\alpha})$ depends on $\hat{\alpha}$ and $\hat{\beta}$. By a similar argument to (45), we have

$$||\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}))|| = o_p(n^{-1/2}).$$

Based on A8" to A11", the consistency of $\hat{\alpha}$ and $\hat{\beta}$ can be shown by similar arguments as in Newey (1994) and Otsu and Xu (2019). (See also Theorem 5 of BGH).

Let us define

$$\begin{aligned}
E[\cdot|u] &= E[\cdot|X'\hat{\alpha} = u], \\
M_{n,\beta} &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial m(Z_i, \beta_0, p_0(X_i))}{\partial \beta} + E[D(Z_i,\beta_0)|X_i'\hat{\alpha}]\frac{\partial T(Z_i,\beta_0)}{\partial \beta}\right\}, \text{ and} \\
M_{\beta} &= -E\left\{\frac{\partial m(Z_i, \beta_0, F_0(X'\alpha_0))}{\partial \beta} + E[D(Z_i,\beta_0)|X_i'\alpha_0]\frac{\partial T(Z_i,\beta_0)}{\partial \beta}\right\}
\end{aligned}$$

We have

$$\begin{aligned}
o_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{m(z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) + E(D(Z_i,\beta_0)|X_i'\hat{\alpha})(T(Z_i,\hat{\beta}) - \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}))\right\} + o_p(n^{-1/2}) \\
&= -M_{n,\beta}(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, F_0(X_i'\alpha_0)) + o_p(n^{-1/2} + (\hat{\beta} - \beta_0)) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\{D(Z_i,\beta_0)(\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0)) + E(D(Z_i,\beta_0)|X_i'\hat{\alpha})(T(Z_i,\beta_0) - \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}))\right\} \\
&= -M_{\beta}(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, F_0) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0) - E(D(Z_i,\beta_0)|X_i'\hat{\alpha})](\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0))\right\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z_i,\beta_0)|X_i'\alpha_0)(T(Z_i,\beta_0) - F_0(X_i'\alpha_0)) + o_p(n^{-1/2} + (\hat{\beta} - \beta_0)) \qquad (48)
\end{aligned}$$

The second equality follows from Lemma 3. The third equality follows from extending $m(Z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}))+E(D(Z_i,\beta_0)|X_i'\hat{\alpha})T(Z_i,\hat{\beta})$ around β_0 and F_0, and by some rearrangements. The last equality follows from $M_{n,\beta} - M_{\beta} = o_p(1)$ and

$$\frac{1}{n}\sum_{i=1}^{n}\left[E(D(Z_i,\beta_0)|X_i'\alpha_0) - E(D(Z_i,\beta_0)|X_i'\hat{\alpha})\right](T(Z_i,\beta_0) - F_0) = o_p(n^{-1/2}),$$

which can be shown by a similar argument about (C.20) in pp.21-22 of BGH-supp.

The second term in the last equality of (48) can be rewritten into:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\hat{\alpha})](\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})-F_0(X_i'\alpha_0))\right\}\\
=&\frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\hat{\alpha})](\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})-F_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}))\right\}\\
+&\frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\hat{\alpha})](F_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})-F_0(X_i'\alpha_0))\right\}\\
:=&I_m+II_m
\end{aligned}$$

$I_m = o_p(n^{-1/2})$ by a similar argument about (C.22) in p.23 of BGH-supp.
For II_m, we have by Lemma 17 of BGH-supp.

$$\begin{aligned}
\frac{\partial}{\partial\alpha_j}F_\alpha(X'\alpha)\Big|_{\alpha=\alpha_0} &= \{x_j - E[X_j|X'\alpha_0 = x'\alpha_0]\}F^{(1)}_{0,\hat{\beta}}(x'\alpha_0),\\
&= \{x_j - E[X_j|X'\alpha_0 = x'\alpha_0]\}F^{(1)}_{0}(x'\alpha_0) + O_p(\hat{\beta}-\beta_0),
\end{aligned}$$

where α_j and x_j are j-th elements of α and x. Then we can extend II_m around α_0:

$$\begin{aligned}
II_m &= \frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\hat{\alpha})]\{X_i - E[X_i|X_i'\alpha_0]\}'F_0^{(1)}(X_i'\alpha_0) + O_p(\hat{\beta}-\beta_0)\right\}(\hat{\alpha}-\alpha_0)\\
&\quad + o_p(\hat{\alpha}-\alpha_0)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\hat{\alpha})]\{X_i - E[X_i|X_i'\alpha_0]]\}'F_0^{(1)}(X_i'\alpha_0])\right\}(\hat{\alpha}-\alpha_0) + o_p(\hat{\alpha}-\alpha_0)\\
&= E\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\alpha_0)]\{X_i - E[X_i|X_i'\alpha_0]]\}'F_0^{(1)}(X_i'\alpha_0)\right\}(\hat{\alpha}-\alpha_0) + o_p(\hat{\alpha}-\alpha_0)
\end{aligned} \tag{49}$$

The second equality follows from $\hat{\beta}-\beta_0 = o_p(1)$ The last equality follows from $\hat{\alpha}-\alpha_0 = o_p(1)$ and $E(D(Z_i,\beta_0)|X_i'\hat{\alpha}) - E(D(Z_i,\beta_0)|X_i'\alpha_0) = o_p(1)$. Now let us define

$$\begin{aligned}
M(Z) &= E(D(Z,\beta_0)|X'\alpha_0)(T(Z_i,\beta_0)-F_0(X'\alpha_0))\\
M_\alpha &= -E\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\alpha_0)]\{x_j - E[X_j|X_i'\alpha_0]\}'F_0^{(1)}(X_i'\alpha_0)\right\}.
\end{aligned} \tag{50}$$

Combining (49) and (50) with (48), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) \\
&= -M_\beta(\hat{\beta}-\beta_0) - M_\alpha(\hat{\alpha}-\alpha_0) + \frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, F_0) \\
&+ \frac{1}{n}\sum_{i=1}^{n} M(Z_i) + o_p(n^{-1/2} + (\hat{\beta}-\beta_0) + (\hat{\alpha}-\alpha_0)). \qquad (51)
\end{aligned}$$

Combining the fact $E\left[m(Z, \beta_0, F_0)\right] = 0$ and $E\left[M(Z)\right] = 0$ with the assumptions A3", A4", A9" and A11", we have $\frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, F_0) + \frac{1}{n}\sum_{i=1}^{n} M(Z_i) = O_p(n^{-1/2})$. Then (51) implies both $\hat{\alpha} - \alpha_0 = O_p(n^{-1/2})$ and $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$. Besides, from (51) we can see that $\hat{\alpha} - \alpha_0$ and $\hat{\beta} - \beta_0$ are asymptotically linear. Thus, we can rewrite the first term in the last row into:

$$-M_\alpha(\hat{\alpha}-\alpha_0) := \frac{1}{n}\sum_{i=1}^{n} A(Z_i) + o_p(n^{-1/2}),$$

with $E\left[A(Z_i)\right] = 0$. Similarly, we can rewrite

$$-M_\beta(\hat{\beta}-\beta_0) := \frac{1}{n}\sum_{i=1}^{n} B(Z_i) + o_p(n^{-1/2}),$$

with $E\left[B(Z_i)\right] = 0$.

Now we can rewrite (51) to obtain asymptotical expressions of $\hat{\alpha}$ and $\hat{\beta}$

Note that given β, $\hat{\alpha}$ is solved with the $\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} ||\frac{1}{n}\sum_{i=1}^{n} X_i'\{T(Z_i, \beta) - \hat{F}_\alpha(X_i'\alpha)\}||^2$. It corresponds to the moment condition

$$m_1(Z, \beta, F(X'\alpha)) \overset{\text{def.}}{=} X\left\{T(Z, \beta) - F(X'\alpha)\right\}$$

We can express $\sqrt{n}(\hat{\alpha} - \alpha_0)$ by replacing m in (51) by m_1. Then we have

$$\begin{aligned}
\sqrt{n}(\hat{\alpha}-\alpha_0) &= M_{\alpha,1}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m_1(Z, \beta_0, p_0) + B_1(Z) + M_1(Z)\} \\
&= M_{\alpha,1}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[X - E(X|X'\alpha_0)\right]\left\{T(Z_i, \beta_0) + \frac{\partial T(Z_i, \beta_0)}{\partial \beta}(\hat{\beta}-\beta_0) - F_0(X'\alpha_0)\right\}
\end{aligned}$$

where $M_{\alpha,1}$, B_1, and M_1 are M_α, B, and M corresponding to the moment function m_1 .

Then we have

$$\sqrt{n}(\hat{\alpha}-\alpha_0) \xrightarrow{d} N(0, V_\alpha) \text{ and } \sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, V_\beta),$$

where

$$V_\alpha = M_{\alpha,1}^{-1} E[\{m_1(Z,\beta_0,p_0) + B_1(Z) + M_1(Z)\}\{m_1(Z,\beta_0,p_0) + B_1(Z) + M_1(Z)\}']M_{\alpha,1}^{-1}$$
$$V_\beta = M_\beta^{-1} E[\{m(Z,\beta_0,p_0) + A(Z) + M(Z)\}\{m(z,\beta_0,p_0) + A(Z) + M(Z)\}']M_\beta^{-1}$$

A.10 Proof of Lemma 4

Let implement the iteration procedure described in p. 184 of Mammen and Yu (2007) and stop at r-th round and j-th elements. In the last step, we actually apply isotonic regression to regress $T(Z_i,\beta) - g_{[r]}^1(X_i^1) - .... - g_{[r]}^{j-1}(X_i^{j-1}) - g_{[r-1]}^{j+1}(X_i^{j+1}) - ...g_{[r-1]}^k(X_i^k) := \tilde{Y}_i$ on X_i^j, and the last sub-function updated in the iteration is $g_{[r]}^j(X_i^j)$. We can replace the Y_i in Lemma 1 with $\tilde{Y}_i$, and replace X_i in Lemma 1 with X_i^j. $\delta(X)$ is assumed to be a bounded function with a finite variation of X. Since X_i^j is an element of X_i, δ is also a bounded function of X_i^j as well. Therefore, all the arguments in the proof of Lemma 1 still hold. We have

$$\frac{1}{n}\sum_{i=1}^n \delta(X_i)(T(Z_i,\beta) - g_{[r]}^1(X_i^1) - .... - g_{[r]}^{j-1}(X_i^{j-1}) - g_{[r]}^j(X_i^j) - g_{[r-1]}^{j+1}(X_i^{j+1}) - ...g_{[r-1]}^k(X_i^k))$$
$$= o_p(n^{-1/2}). \tag{52}$$

By Theorem 2 of Mammen and Yu (2007), with $r \to \infty$, the backfitting estimator $\{g_{[r]}^j(\cdot)\}_{j=1}^k$ is converging to the least square isotonic estimator of the problem (20), $\{g^j(\cdot)\}_{j=1}^k$,i.e.,

$$\lim_{r\to\infty} g_{[r]}^j(\cdot) = g^j(\cdot) \text{ for all } j = 1,...,k \tag{53}$$

in a fixed sample. As mentioned in Section 3.2, the least square estimator of the problem (19) is obtained by normalizing $\{g^j(\cdot)\}_{j=1}^k$. Therefore, we have

$$\hat{c} + \sum_{j=1}^k \hat{m}^j(X_i^j) = \sum_{j=1}^k g^j(X_i^j) \tag{54}$$

Combining (52), (53), and (54), we have

$$\frac{1}{n}\sum_{i=1}^n \delta(X_i)(T(Z_i,\beta) - \hat{c} - \sum_{j=1}^k \hat{m}^j(X_i^j)) = o_p(n^{-1/2}).$$

A.11 Proof of Theorem 4

The following proof is mostly similar to that in Appendix A.7. The only difference is that we need to bind the L^2 norm of the additive monotone nuisance function, as discussed in Mammen and Yu (2007).

Now the nuisance function $\hat{p}_{\hat{\beta}}(X)$ depends on $\hat{\beta}$. By a similar argument to (45) we have

$$||\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{p}_{\hat{\beta}}(X_i)|| = o_p(n^{-1/2}).$$

$$\begin{aligned}
o_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{p}_{\hat{\beta}}(X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big\{ m(z_i, \hat{\beta}, \hat{p}_{\hat{\beta}}(X_i)) + E(D(Z_i, \beta_0)|X_i)(T(Z_i, \hat{\beta}) - \hat{p}_{\hat{\beta}}(X_i)) \Big\} + o_p(n^{-1/2}) \\
&= -M_{n,\beta}(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, p_0(X_i)) + o_p(n^{-1/2} + (\hat{\beta} - \beta_0)) \\
&+ \frac{1}{n}\sum_{i=1}^{n} \Big\{ D(Z_i, \beta_0)(\hat{p}_{\hat{\beta}}(X_i) - p_0(X_i)) + E(D(Z_i, \beta_0)|X_i)(T(Z_i, \beta_0) - \hat{p}_{\hat{\beta}}(X_i)) \Big\} \\
&= -M_{\beta}(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, p_0) \\
&+ \frac{1}{n}\sum_{i=1}^{n} \Big\{ [D(Z_i, \beta_0) - E(D(Z_i, \beta_0)|X_i)](\hat{p}_{\hat{\beta}}(X_i) - p_0(X_i)) \Big\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z_i, \beta_0)|X_i)(T(Z_i, \beta_0) - p_0(X_i)) + o_p(n^{-1/2} + (\hat{\beta} - \beta_0)) \qquad (55) \\
&= -M_{\beta}(\hat{\beta} - \beta_0) + \left\{ \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta_0, p_0(X_i)) + M(Z_i) \right\} + o_p(n^{-1/2} + (\hat{\beta} - \beta_0)). \qquad (56)
\end{aligned}$$

The second equality follows from Lemma 4. The third equality follows from the expansion around β_0 and the definition of $M_{n,\beta}$. The fourth equality follows from $M_{n,\beta} - M_{\beta} = o_p(1)$. The last equality follows from the similar arguments in p.187 of Mammen and Yu (2007) (see also Theorem 9.2 in van de Geer, 2000) and Step 1 and 2 of Appendix A.3.

With A7$^{(3)}$ and A8$^{(3)}$, consistency of $\hat{\beta}$ can be similarly proved as in Lemma 5.2 in Newey (1994).

Finally, we have

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta_0) &= M_{\beta}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ m(Z_i, \beta_0, p_0(X_i)) + M(Z_i) \} + o_p(1) \\
&\xrightarrow{d} N(0, V).
\end{aligned}$$

A.12 Proof of Theorem 5

The proof is based on Groeneboom and Hendrickx (2017) (hereafter GH). Here we prove the counterpart for Theorem 2. It can be easily modified to fit the settings of Theorem 1 and Theorem 3 by changing the relevant notations.

Let Z^* is the bootstrap sample of the data. $\hat{\beta}^*$ and $\hat{p}^*(\cdot)$ are the corresponding estimators for the parameter and the nuisance monotone function. By similar arguments to (44) and (45), we have

$$||\frac{1}{n}\sum_{i=1}^{n} m(z_i^*, \hat{\beta}^*, \hat{p}^*_{\hat{\beta}^*}(x_i^*))|| = o_{P_M}(n^{-1/2}), \tag{57}$$

where P_M is defined in p. 3450 of GH. Let

$$M^*_{n,\beta} = -\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial m(Z_i^*, \beta_0, p_0(X_i^*))}{\partial\beta} + \frac{\partial\left\{E[D(Z_i^*,\beta_0)|X_i^*]T(Z_i^*,\beta_0)\right\}}{\partial\beta}\right\}, \text{ and}$$
$$M_\beta = -E\left\{\frac{\partial m(Z_i, \beta_0, p_0(X_i))}{\partial\beta} + \frac{\partial\left\{E[D(Z_i,\beta_0)|X_i]T(Z_i,\beta_0)\right\}}{\partial\beta}\right\}.$$

Step 1: Show

$$M_\beta(\hat{\beta}^* - \beta_0) = \frac{1}{n}\sum_{i=1}^{n}\{m(Z_i^*, \beta_0, p_0(X_i^*)) + M(Z_i^*)\} + o_{P_M}(n^{-1/2} + (\hat{\beta}^* - \beta_0)) \tag{58}$$

By extending (57) we have

$$\begin{aligned}
o_{P_M}(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i^*, \hat{\beta}^*, \hat{p}^*_{\hat{\beta}^*}(X_i^*)) \\
&= \frac{1}{n}\sum_{i=1}^{n} m(Z_i^*, \hat{\beta}^*, \hat{p}^*_{\hat{\beta}^*}(X_i^*)) + o_{P_M}(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z,\beta_0)|X_i^*)(T(Z_i^*, \hat{\beta}^*) - \hat{p}^*_{\hat{\beta}^*}(X_i^*)) \\
&= -M^*_{n,\beta}(\hat{\beta}^* - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i^*, \beta_0, \hat{p}^*_{\hat{\beta}^*}(X_i^*)) + o_{P_M}(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z,\beta_0)|X_i^*)(T(Z_i^*, \beta_0) - \hat{p}^*_{\hat{\beta}^*}(X_i^*)) + o_{P_M}(\hat{\beta}^* - \beta_0) \\
&= -M_\beta(\hat{\beta}^* - \beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i^*, \beta_0, p_0(X_i^*)) + o_{P_M}(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z_i^*,\beta_0)|X_i^*)(T(Z_i^*, \beta_0) - p_0(X_i^*)) + o_{P_M}(\hat{\beta}^* - \beta_0) \\
&= -M_\beta(\hat{\beta} - \beta_0) + \frac{1}{n}\sum_{i=1}^{n}\{m(Z_i^*, \beta_0, p_0(X_i^*)) + M(Z_i^*)\} + o_{P_M}(n^{-1/2} + (\hat{\beta}^* - \beta_0))
\end{aligned}$$

All steps are similar to what we have in (46). In the fourth equality, we use $M^*_{n,\beta} - M_\beta = o_p(1)$, and the conditional bootstrapped L_2-result:

$$\frac{1}{n}\sum_{i=1}^{n}\{\hat{p}^*_{\hat{\beta}^*}(X_i^*) - p_0(X_i^*)\}^2 = O_{P_M}((\log n)^2 n^{-2/3}) = o_{P_M}(n^{-1/2}).$$

See (6.21) in GH and Proposition 4 in BGH. Now we have shown (58).

Step 2: Rearrangement

(58) can be rearranged to

$$M_\beta(\hat{\beta}^* - \beta_0) = \left\{ \frac{1}{n}\sum_{i=1}^n m(Z_i^*, \beta_0, p_0(X_i^*)) - \frac{1}{n}\sum_{i=1}^n m(Z_i, \beta_0, p_0(X_i)) \right\} + \left\{ \frac{1}{n}\sum_{i=1}^n M(Z_i^*) - \frac{1}{n}\sum_{i=1}^n M(Z_i) \right\}.$$
$$+ \frac{1}{n}\sum_{i=1}^n \{m(Z_i, \beta_0, p_0(X_i) + M(Z_i)\} + o_{P_M}(n^{-1/2} + (\hat{\beta}^* - \beta_0)) \quad (59)$$

Then we could subtract (46) from (59) and get

$$M_\beta(\hat{\beta}^* - \hat{\beta}) = \left\{ \frac{1}{n}\sum_{i=1}^n m(Z_i^*, \beta_0, p_0(X_i^*)) - \frac{1}{n}\sum_{i=1}^n m(Z_i, \beta_0, p_0(X_i)) \right\} + \left\{ \frac{1}{n}\sum_{i=1}^n M(Z_i^*) - \frac{1}{n}\sum_{i=1}^n M(Z_i) \right\}$$
$$+ o_{P_M}((\hat{\beta}^* - \beta_0) + n^{-1/2}),$$

Note the bootstrap mean $E^*[m(Z_i^*, \beta_0, p_0(X_i^*))] = \frac{1}{n}\sum_{i=1}^n m(Z_i, \beta_0, p_0(X_i))$ and $E^*[M(Z_i^*)] = \frac{1}{n}\sum_{i=1}^n M(Z_i)$. Then we have by CLT

$$\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \xrightarrow{d} N(0, \Pi),$$

where Π is defined in (47).

References

[1] Balabdaoui, F., Durot, C. and Jankowski, H., (2019). Least squares estimation in the monotone single index model. *Bernoulli*, 25(4B), pp.3276-3310.

[2] Balabdaoui, F., Groeneboom, P. and K. Hendrickx (2019) Score estimation in the monotone single index model, *Scandinavian Journal of Statistics.*, 46, 517-544.

[3] Balabdaoui, F., & Groeneboom, P. (2020). Profile least squares estimators in the monotone single index model. *arXiv preprint arXiv:2001.05454*.

[4] Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61 (4)*, 962{973}.

[5] Barlow, R., & Brunk, H. (1972). The Isotonic Regression Problem and Its Dual. *Journal of the American Statistical Association*, 67(337), 140-147.

[6] Bartholomew, D. J., A Test of Homogeneity for Ordered Alternatives I and II, *Biometrika*, 46, Nos. 1 and 2 (1959), 36-48, 329-81.

[7] Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71, 1591-1608

[8] Cheng, G. (2009). Semiparametric additive isotonic regression, *Journal of Statistical Planning and Inference*, 139, 1980-1991.

[9] Cheng, G., Zhao, Y. and Li, B., (2012). Empirical likelihood inferences for the semiparametric additive isotonic regression. *Journal of Multivariate Analysis*, 112, pp.172-182.

[10] Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 51 765–782. MR712369 (85a:62174)

[11] Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* 78 837–842.

[12] Engle, R. F., Granger, C. W., Rice, J., & Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394), 310-320.

[13] Gaffke, N. and Mathar, R. (1989). A cyclic projection algorithm via duality. *Metrika* 36 29–54.

[14] Goldstein, L. and K. Messer (1992), Optimal Plug-in Estimators for Nonparametric Functional Estimation, *Annals of Statistics*, 20, 1306–1328.

[15] Groeneboom, P. and K. Hendrickx (2018). Current status linear regression, *Annals of Statistics*, 46, 1415-1444.

[16] Hahn, J. (1998), On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects, *Econometrica*,66, 315-331.

[17] Hall, P. (1989). On projection pursuit regression. *Ann. Statist.* 17 573–588. MR0994251

[18] Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics* 35 303–316.

[19] Härdle, W., Hall, P. And Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* 21 157–178. MR1212171

[20] Hirano, K., Imbens, G. W., and Ridder, G. (2000). Efficient estimation of average treatment effects using the estimated propensity score., *NBER Technical Working Paper No. 251.*

[21] Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score., *Econometrica*, 71(4):1161-1189.

[22] Horowitz, J. L. (2009). Semiparametric and nonparametric methods in econometrics, *New York: Springer,* (Vol. 12).

[23] Horvitz, D.G., and Thompson, D.J.,. A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association,* 47 (1952), 663-685.

[24] Huang, J. (2002) A note on estimating a partly linear model under monotonicity constraints, *Journal of Statistical Planning and Inference*, 107, 343-351

[25] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics,* 86(1), 4-29.

[26] Imbens, G., Newey, W., & Ridder, G. (2006). Mean-squared-error Calculations for Average Treatment Effects, *Institute of Economic Policy Research (IEPR),* (No. 06.57).

[27] Imbens, G. W. and D. B. Rubin (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. *Cambridge: Cambridge University Press*

[28] Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* 61 387–421. MR1209737 (93k:62214)

[29] Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349-1382.

[30] Otsu, T. and Xu, M. (2019) Score estimation of monotone partially linear index model. Working Paper, London School of Economics, London, 2019

[31] Qin, J., Yu, T., Li, P., Liu, H., & Chen, B. (2019). Using a monotone single-index model to stabilize the propensity score in missing data problems and causal inference. *Statistics in medicine,* 38(8), 1442-1458.

[32] Robins, J., and A. Rotnitzky (1995), Semiparametric Efficiency in Multivariate Regression Models with Missing Data, *Journal of the American Statistical Association*, 90, 122-129.

[33] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931-954.

[34] Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* 54 1461–1481. MR0868152

[35] Stock, J. H. (1991). Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, 77-98

[36] Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61 123–137. MR1201705

[37] van de Geer, S. (2000). Empirical Processes in M-Estimation. *Cambridge University Press.*

[38] Westling, T., Gilbert, P., & Carone, M. (2018). Causal isotonic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):719—747.

[39] Yu, K. (2014). On partial linear additive isotonic regression. *Journal of the Korean Statistical Society*, 43(1), 1