The London School of Economics and Political Science

Essays in Semiparametric Estimation and Inference with Monotonicity Constraints

Mengshan Xu

A thesis submitted to the Department of Economics
for the degree of Doctor of Philosophy

London, July 2021

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

A version of Chapter 1 was published in the Journal of Nonparametric Statistics, Volume 32, issue 4, October 2020, p. 838-863.

I declare that my thesis consists of 22248 words.

Statement of conjoint work

I confirm that Chapter 1 was jointly co-authored with Professor Taisuke Otsu, and Chapter 2 was jointly co-authored with Professor Taisuke Otsu and Dr Keisuke Takahata. I contributed at least 50% of the work in each case.

Mengshan Xu

Abstract

Chapter 1 studies semiparametric estimation of partially linear single index models with a monotone link function. Our estimator is an extension of the score-type estimator developed by Balabdaoui, Groeneboom, and Hendrickx (2019) for monotone single index models, which profiles out the unknown link function by isotonic regression. We show that our estimator for the finite-dimensional components is tuning-parameter-free, $\sqrt{n}$-consistent, and asymptotically normal. Furthermore, by introducing a single smoothing parameter, we propose an asymptotically efficient estimator for the finite-dimensional components.

Chapter 2 proposes an empirical likelihood inference method for monotone index models. We construct the empirical likelihood function based on the modified score function of a monotone index model, where the monotone link function is estimated by isotonic regression. It is shown that the empirical likelihood ratio statistic converges to a weighted chi-squared distribution. We suggest inference procedures based on an adjusted empirical likelihood statistic that is asymptotically pivotal, and a bootstrap calibration with recentering. A Monte-Carlo simulation study illustrates the usefulness of the proposed inference methods.

The models in Chapter 1 and 2 can be regarded as special cases of the framework analyzed in Chapter 3, which studies a general 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). As a result, the proposed 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 generalizes the estimation methods of existing semiparametric models with monotone nuisance functions. 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 the proposed estimator has desired properties. Furthermore, we establish the asymptotic validity of the bootstrap, which ensures that the estimator is tuning-parameter-free in both estimation and inference.

Acknowledgements

I am deeply indebted to my supervisor, Taisuke Otsu, for his continuous support. I am very grateful to Wei Cui, Canh Thien Dang, Juan Carlos Escanciano, Kirill Evdokimov, Joachim Freyberger, Javier Hidalgo, Alois Kneip, Tatiana Komarova, Geert Mesters, Martin Pesendorfer, Steve Pischke, Chen Qiu, Stephen Ross, Christoph Rothe, Marcia Schafgans, João Santos Silva, John Van Reenen, Stefan Wager, Yike Wang, Weining Wang, and Chen Zhou for very insightful discussions. I would like to thank participants at CFE-CMStatistics 2019 conference, LSE Econometrics work-in-progress seminars, and job market seminars at Pompeu Fabra University, the University of Surrey, the University of Mannheim, Erasmus University Rotterdam, the University of Bonn, and the University of Connecticut.

I would like to thank my fellow PhD students at LSE, Svetlana Chekmasova, Martina Fazio, Nicola Fontana, Maximilian Guennewig, Tillman Hoenig, William Matcham, Tsogsag Nyamdavaa, Lukasz Rachel, Heidi Thysen, and Céline Zipfel, for their very helpful conversations and comments.

I would like to thank professional services staff at the LSE, especially Mark Wilbor, Sharon Peate, and Anna Watmuff, for their patient and outstanding administrative support.

Finally, my parents and my wife Xiwen Chen have provided me with constant support throughout this long journey. I owe them everything for their unconditional trust and endless patience.

Contents

List of Figures

List of Tables

Chapter 1

Score estimation of monotone partially linear index model

1.1 Introduction

This paper is concerned with the monotone partially linear single index (PLSI) model

$$Y = X'\beta_0 + \psi_0(Z'\alpha_0) + \epsilon, \qquad E[\epsilon|X, Z] = 0, \tag{1.1}$$

where $Y \in \mathbb{R}$ is a response variable, $X \in \mathcal{X} \subseteq \mathbb{R}^k$ and $Z \in \mathcal{Z} \subseteq \mathbb{R}^d$ are covariates, $\epsilon \in \mathbb{R}$ is an error term, α_0 and β_0 are finite dimensional parameters, and $\psi_0 : \mathbb{R} \to \mathbb{R}$ is an unknown monotone increasing function. For identification, we assume that Z does not contain a constant and α_0 belongs to the d-dimensional unit sphere $\mathcal{S}_{d-1} = \{\alpha \in \mathbb{R}^d : ||\alpha|| = 1\}$.

Since a seminal work by Carroll *et al.* (1997), the model (1.1) (without the monotonicity assumption about ψ_0) has been studied by many authors, including Xia, Tong and Li (1999), Yu and Ruppert (2002), Xia and Härdle (2006), Wang *et al.* (2010), and Ma and Zhu (2013), among others. The model (1.1) is very flexible. If α_0 is known, it becomes a partially linear model. If $\beta_0 = 0$, it becomes a single index model. See, e.g., Wang *et al.* (2010) for a review on these models. Estimation of the model (1.1) typically requires some nonparametric smoothing method to evaluate the unknown function ψ_0, which involves tuning parameters,

such as bandwidth and series length parameters.

In this paper, we consider the situation where ψ_0 is known to be monotone. Instead of assuming certain degree of smoothness as in the above cited papers, we impose a shape restriction on ψ_0, and propose a $\sqrt{n}$-consistent estimator for the parameters (α_0, β_0) that is free from tuning parameters. Furthermore, we establish the asymptotic validity of a bootstrap inference method based the proposed estimator, which is also free from tuning parameters.

A natural approach to incorporate monotonicity into nonparametric estimation is to employ the isotonic regression technique (see, e.g., Groeneboom and Jongbloed, 2014, for a review). For example, one may consider the least square estimation for the model (1.1), say $\min_{\alpha,\beta}[\min_{\psi\in\mathcal{M}} \sum_{i=1}^n \{Y_i - X_i'\beta + \psi(Z_i'\alpha)\}^2]$, where $\mathcal{M}$ the set of monotone increasing functions. In this case, we can apply the isotonic regression technique for each (α, β), and then minimize the concentrated criterion function with respect to (α, β). However, because of lack of smoothness of the isotonic regression estimator for ψ_0, it is not clear whether such a profile least square estimator for (α_0, β_0) will be $\sqrt{n}$-consistent or asymptotically normal. This point was clarified by Balabdaoui, Groeneboom and Hendrickx (2019) (BGH hereafter) and Groeneboom and Hendrickx (2018) for single index (and current status) models.

For this problem, BGH and Groeneboom and Hendrickx (2018) developed a novel score estimation approach for single index models, say $Y = \psi_0(Z'\alpha_0) + \epsilon$. Their basic idea is to construct a feasible score equation $\sum_{i=1}^n Z_i\{Y_i - \psi_\alpha(Z_i'\alpha)\} = 0$ where ψ_α is estimated by isotonic regression for given α. Then the estimator for α_0 is obtained by the solution of the feasible score equation. BGH showed that their score estimator for α_0 is $\sqrt{n}$-consistent and asymptotically normal. Furthermore, BGH proposed an asymptotically efficient estimator for α_0 by evaluating an optimal score equation. Groeneboom and Hendrickx (2018) and Groeneboom and Hendrickx (2017) studied the score-type estimator for current status models and its bootstrap validity, respectively.

In this paper, we extend the score estimation approach developed by BGH and Groeneboom and Hendrickx (2018) to the monotone PLSI model in (1.1). We

show that the proposed score-type estimator for (α_0, β_0) is $\sqrt{n}$-consistent and asymptotically normal. Also, by estimating nonparametrically the efficient score function, we derive an asymptotically efficient estimator for (α_0, β_0) whose asymptotic variance coincides with the efficient variance matrix in Carroll *et al.* (1997). Finally, we establish the validity of a bootstrap inference method based on the score-type estimator. Similar to the existing papers on (not necessarily monotone) PLSI models cited above, the extension from single index or current status models to the PLSI model is not a trivial task. In particular, the presence of linear indices both inside and outside the nonparametric monotone function complicates the theoretical development.

This paper complements the literature on score-type estimation for semiparametric models with isotonic nuisance parameter estimates. Groeneboom and Hendrickx (2018) and BGH argued that score-type estimation and monotone least square estimation are not equivalent methods; they showed theoretically and numerically that the score-type estimator behaves at least as well as (or even better than) the monotone least square in single index models. The present paper shows analogous advantages continue to hold in PLSI models. Huang (2002), Cheng (2009), and Yu (2014) studied asymptotic properties of the monotone least square estimator, but it was unclear whether the score-type estimator could also achieve the $\sqrt{n}$-convergence rate and semiparametric efficiency. Our paper fills this gap.

Furthermore, the results in this paper can be considered as extensions of the ones for monotone partially linear models (Huang, 2002, and Cheng, 2009). However, since the partially linear model does not involve unknown parameters (i.e., α_0) in the argument of the unknown function ψ_0, the theoretical development is very different from ours.

This paper is organized as follows. In Section 2, we introduce our score-type estimator for the model (1.1) and present its asymptotic properties. We also propose an asymptotically efficient estimator for (α_0, β_0) and bootstrap inference method. Section 3 presents some Monte-Carlo simulation evidence to illustrate the finite sample performance of our estimators and bootstrap method.

1.2 Main results

1.2.1 Estimation method

Let us first introduce our estimator for the PLSI model in (1.1). In particular, we extend the score estimation approach by BGH to estimate the parameters (α_0, β_0) in (1.1). Consider a parameterization $\mathbb{S}$ from a subset of $\mathbb{R}^{d-1}$ to $\mathcal{S}_{d-1}$ such that for each α in a neighborhood of α_0 on $\mathcal{S}_{d-1}$, there exists a unique $\gamma \in \mathbb{R}^{d-1}$ satisfying $\alpha = \mathbb{S}(\gamma)$.[1] Then the reparameterized model (1.1) is written as

$$Y = X'\beta_0 + \psi_0(Z'\mathbb{S}(\gamma_0)) + \epsilon, \qquad E[\epsilon|X, Z] = 0.$$

To motivate our estimation approach, we tentatively assume that ψ_0 is known. In this case, the population score equation for $\theta_0 = (\beta_0', \gamma_0')'$ is

$$E\left[\begin{pmatrix} X \\ \mathbb{J}(\gamma_0)'Z\psi_0'(Z'\mathbb{S}(\gamma_0)) \end{pmatrix}\{Y - X'\beta_0 - \psi_0(Z'\mathbb{S}(\gamma_0))\}\right] = 0, \tag{1.2}$$

where ψ_0' is the derivative of ψ_0 and $\mathbb{J}(\gamma)$ is the Jacobian of $\mathbb{S}(\gamma)$. Thus, it is natural to construct an estimator of θ_0 by taking an empirical counterpart of (1.2) and inserting estimators for ψ_0' and ψ_0. However, when we estimate ψ_0 by the isotonic regression method, the resulting estimator of ψ_0 is typically discontinuous and it is not clear how to evaluate the derivative ψ_0' without introducing smoothing parameters. To address this issue, we follow the idea in BGH and Groeneboom and Hendrickx (2018) and focus on the following modified population score equation

$$E\left[\begin{pmatrix} X \\ \mathbb{J}(\gamma_0)'Z \end{pmatrix}\{Y - X'\beta_0 - \psi_0(Z'\mathbb{S}(\gamma_0))\}\right] = 0. \tag{1.3}$$

[1] Examples of such parametrization are the spherical coordinate system $\mathbb{S} : [0,\pi]^{d-2} \times [0, 2\pi] \to \mathcal{S}_{d-1}$ with

$$\begin{aligned}\mathbb{S}(\gamma) =& (\cos(\gamma_1), \sin(\gamma_1)\cos(\gamma_2), \sin(\gamma_1)\sin(\gamma_2)\cos(\gamma_3), \\ & \ldots, \sin(\gamma_1)\cdots\sin(\gamma_{d-2})\cos(\gamma_{d-1}), \sin(\gamma_1)\cdots\sin(\gamma_{d-2})\sin(\gamma_{1-2}))',\end{aligned}$$

and the half sphere $\mathbb{S} : \{\gamma \in [0,1]^{d-1} : ||\gamma|| \leq 1\} \to \mathcal{S}_{d-1}$ with

$$\mathbb{S}(\gamma) = (\gamma_1, \ldots, \gamma_{d-1}, \sqrt{1 - \gamma_1^2 - \cdots - \gamma_{d-1}^2})'.$$

Since the error term ϵ is orthogonal to any function of (X, Z) under $E[\epsilon|X, Z] = 0$, (1.3) is also a valid score equation, and we construct an estimator for θ_0 based on this equation.

In particular, for each $\theta = (\beta', \gamma')'$, we estimate the monotone function ψ_0 by the least squares

$$\hat{\psi}_{n\theta} = \arg\min_{\psi \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - X_i'\beta - \psi(Z_i'\mathbb{S}(\gamma))\}^2,$$

where $\mathcal{M}$ is the set of monotone increasing functions defined on $\mathbb{R}$. The function $\hat{\psi}_{n\theta}$ can be obtained by isotonic regression (see, e.g., Groeneboom and Jongbloed, 2014, for a review). Then our estimator $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')'$ of θ_0 is given by the zero-crossing root of the score function[2]

$$\phi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ \mathbb{J}(\gamma)' Z_i \end{pmatrix} \{Y_i - X_i'\beta - \hat{\psi}_{n\theta}(Z_i'\mathbb{S}(\gamma))\}, \tag{1.4}$$

and α_0 is estimated by $\hat{\alpha} = \mathbb{S}(\hat{\gamma})$. The reason for the definition based on the zero-crossing is due to the fact that $\hat{\psi}_{n\theta}$ is a discrete function taking finite different values. Thus, we might be unable to solve $\phi_n(\theta) = 0$ exactly.[3] As $n \to \infty$, the zero-crossing solution should become an exact solution. In practice, we can minimize the square sum of the right hand side of (1.4) to obtain a good approximation of the zero-crossing.

Remark 1.1. [Technical intuition for the difference between the score estimation and least square approaches] Our discussion is based on Groeneboom and Hendrickx (2018, pp. 1419-1420). Let $\Gamma_n(\theta)$ be some objective function for θ and $\Gamma(\theta)$ is its population counterpart. The M-estimator is defined as a maximizer of $\Gamma_n(\theta)$. The $\sqrt{n}$-consistency of the estimator is typically derived from a quadratic expansion $\Gamma(\theta) - \Gamma(\theta_0) \leq -c||\theta - \theta_0||^2$ for some $c > 0$ in a neighborhood of θ_0

[2]We say that θ^* is a zero-crossing of a real-valued function $\zeta : \Theta \to \mathbb{R}$ if each open neighborhood of θ^* contains points $\theta_1, \theta_2 \in \Theta$ such that $\zeta(\theta_1)\zeta(\theta_2) \leq 0$. This definition can be extended to a vector of functions, where a zero-crossing vector has each of its component to be a zero-crossing in the corresponding dimension.

[3]Similar to other estimators by BGH or Groeneboom and Hendrickx (2018), our zero-crossing estimator $\hat{\theta}$ may not be unique. Indeed there are many flat parts in $\phi_n(\theta)$, and the intersection of $\phi_n(\theta)$ and zero could be an interval. In this case, any point on this interval will satisfy the results in Theorems 1.1 and 1.3 below.

combined with the approximation to the objective function

$$\Gamma_n(\theta)-\Gamma_n(\theta_0) = \Gamma(\theta)-\Gamma(\theta_0)+O_p(n^{-1/2}||\theta-\theta_0||)+o_p(||\theta-\theta_0||^2)+O_p(n^{-1}), \quad (1.5)$$

uniformly over a shrinking neighborhood of θ_0. However, when we apply this argument to the (profile) least square objective function $\frac{1}{n}\sum_{i=1}^n\{Y_i - X_i'\beta - \hat{\psi}_{n\theta}(Z_i'\mathbb{S}(\gamma))\}^2$, it seems to have an extra term of order $O_p(n^{-2/3})$ in (1.5) due to discontinuity of $\hat{\psi}_{n\theta}$ in θ (although there is no rigorous proof). If there is such an extra term, we expect that the least square estimator for θ will not achieve $\sqrt{n}$-consistency.[4] On the other hand, it turns out that our score (or Z-) estimating equation $\phi_n(\theta)$ can be approximated by $\phi_n(\theta) = \phi'(\theta_0)(\theta - \theta_0) + O_p(n^{-1/2})$ uniformly over a shrinking neighborhood of θ_0, where $\phi'(\theta_0)$ is the derivative of the population counterpart of $\phi_n(\theta)$ displayed in (1.3). In short, the difference between the score estimation and least square approaches is due to different orders of the remainders in the Z- and M-estimation approaches in this setup.

Remark 1.2. [Comparison with smoothing approach] Let us take Xia and Härdle (2006) as an example for the conventional smoothing approach to estimate the PLSI model (without monotonicity on ψ_0) and compare with our estimation approach. A common feature is that both methods estimate the nonparametric function ψ_0 with fixed θ, and then optimize or solve for $\hat{\theta}$ in a two step or recursive strategy. The main difference is that we use the isotonic regression to estimate the monotone function ψ_0, but Xia and Härdle (2006) employ a weighted local linear regression to estimate ψ_0 for each fixed θ. Our score-type estimation method does not require any tuning parameter to estimate ψ_0, while a smoothing parameter is innate in Xia and Härdle (2006). The technical arguments are very different as well. Our consistency and asymptotic normality proofs below heavily rely on properties of the monotone function class and associated empirical processes. On the other hand, the argument in Xia and Härdle (2006) is to show how the linear regression for θ_0 averages out the estimation errors from the local linear regression for ψ_0 based on the U-statistic theory to achieve the $\sqrt{n}$-consistency of their estimator for θ_0.

[4]We note that even for single index models, the convergence rate and asymptotic distribution of the least square estimator, $\arg\min_\gamma \{\min_{\psi\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n\{Y_i - \psi(Z_i'\mathbb{S}(\gamma))\}^2\}$, is an open problem.

1.2.2 Asymptotic properties of estimator

We now investigate asymptotic properties of the estimator $\hat{\theta}$. Let $\mathbb{I}_k$ be the $k \times k$ identity matrix, $\|\cdot\|$ be the Euclidean norm, $\mathcal{B}(a_0, A) = \{a : \|a - a_0\| \leq A\}$ be a ball around a_0 of radius A, and

$$T_0 = \begin{bmatrix} \mathbb{I}_k & 0 \\ 0 & \mathbb{J}(\gamma_0)' \end{bmatrix}, \qquad V_{x,z} = \begin{pmatrix} x - E[X|z'\mathbb{S}(\gamma_0)] \\ z - E[Z|z'\mathbb{S}(\gamma_0)] \end{pmatrix},$$
$$V_{x,z,\psi'} = \begin{pmatrix} x - E[X|z'\mathbb{S}(\gamma_0)] \\ \{z - E[Z|z'\mathbb{S}(\gamma_0)]\}\psi_0'(z'\mathbb{S}(\gamma_0)) \end{pmatrix}.$$

We impose the following assumptions.

Assumption.

A1 *The spaces $\mathcal{X}$ and $\mathcal{Z}$ are convex with non-empty interiors, and satisfy $\mathcal{X} \subset \mathcal{B}(0, R)$ and $\mathcal{Z} \subset \mathcal{B}(0, R)$ for some $R > 0$.*

A2 *There exists $K_0 > 0$ such that $|\psi_0(u)| < K_0$ for all $u \in \{z'\alpha : z \in \mathcal{Z}, \alpha \in \mathcal{S}_{d-1}\}$.*

A3 *There exists $\delta_0 > 0$ such that the function $\psi_\theta(u) = \psi_{\alpha,\beta}(u) = E[Y - X'\beta | Z'\alpha = u]$ is monotone increasing on $I_\alpha = \{z'\alpha, z \in \mathcal{Z}\}$ for each $\theta \in \mathcal{B}(\theta_0, \delta_0)$.*

A4 *For $W = X$ or Z, the mapping $u \mapsto E[W|Z'\alpha = u]$ defined on I_α is bounded and has a finite total variation.*

A5 *There exist $c_0 > 0$ and $M_0 > 0$ such that $E[|Y - X'\beta|^m | Z = z] \leq m! M_0^{m-2} c_0$ for all integers $m \geq 2$, each β satisfying $(\beta', \gamma')' \in \mathcal{B}(\theta_0, \delta_0)$ and almost every $z \in \mathcal{Z}$ (according to the true distribution).*

A6 $\mathrm{Cov}[(\beta_0 - \beta)'X + Z'(\mathbb{S}(\gamma_0) - \mathbb{S}(\gamma)),\ (\beta_0 - \beta)'X + \psi_0(Z'\mathbb{S}(\gamma_0)) | Z'\mathbb{S}(\gamma)] \neq 0$ *almost surely for each $\theta \neq \theta_0$.*

A7 *$B = T_0 \int V_{x,z} V_{x,z,\psi'}' dP_0(x,z) T_0'$ and $B_E = T_0 \int V_{x,z,\psi'} V_{x,z,\psi'}' dP_0(x,z) T_0'$ are non-singular.*

A1 and A2, which are similar to the assumptions A1 and A2 in BGH, impose boundedness on the support of covariates and the monotone function ψ_0. These conditions are used to control the entropy of the function classes that characterize (1.4). We note that Xia and Härdle (2006) and Wang *et al.* (2010) imposed similar conditions. A3, which is an adaptation of BGH's A3, requires monotonicity of ψ_θ in a neighborhood of θ_0. This assumption is used to establish the consistency of the estimator $\hat{\psi}_{n\theta}(z'\mathbb{S}(\gamma))$ for each $\theta \in \mathcal{B}(\theta_0, \delta_0)$. For example, A3 is satisfied with $\psi_0(u) = u^3$, $\alpha_0 = \mathbb{S}(\gamma_0) = (2^{-1/2}, 2^{-1/2})'$, and $Z_1, Z_2 \sim U[1,2]$, which are independent of X.[5] A4 is imposed to control the entropy of function classes to achieve the $\sqrt{n}$-convergence rate. This assumption can be derived from BGH's A4 and A5. A5 is a modified version of BGH's A6. This assumption is introduced to show that $\max_{\theta \in \mathcal{B}(\theta_0, \delta_0)} \sup_{z \in \mathcal{Z}} \hat{\psi}_{n\theta}(z'\mathbb{S}(\gamma)) = O_p(\log n)$, which is used to obtain an entropy result associated with the $\sqrt{n}$-convergence rate.[6] A6 and A7 are to ensure the consistency and existence of limiting variances of the simple score and efficient score estimators, respectively. A6 is related to BGH's A7 after taking expansion of $\mathbb{S}(\gamma_0) - \mathbb{S}(\gamma)$ around $\gamma = \gamma_0$.

Under these assumptions, the asymptotic properties of the simple estimator $\hat{\theta}$ are presented as follows.

Theorem 1.1. *Suppose Assumptions A1-A7 hold true. Then $\hat{\theta}$ exists with prob-*

[5]More precisely, take δ_0 small enough so that elements of $\alpha = (\alpha_1, \alpha_2) = \mathbb{S}(\gamma)$ satisfying $(\beta', \gamma')' \in \mathcal{B}(\theta_0, \delta_0)$ are always positive. Then we have $I_\alpha = [\alpha_1 + \alpha_2, 2\alpha_1 + 2\alpha_2]$. If $\alpha_1 \leq \alpha_2$ (the case of $\alpha_1 > \alpha_2$ is analyzed in the same manner), the computation of $\psi_{\alpha,\beta}(u)$ is split into four cases (i) $\alpha_1 + \alpha_2 < u \leq 2\alpha_1 + \alpha_2$, (ii) $2\alpha_1 + \alpha_2 < u \leq \alpha_1 + 2\alpha_2$, (iii) $\alpha_1 + 2\alpha_2 < u < 2\alpha_1 + 2\alpha_2$, and (iv) $u = \alpha_1 + \alpha_2$ or $u = 2\alpha_1 + 2\alpha_2$. For (i), a direct calculation yields

$$\psi_{\alpha,\beta}(u) = \frac{\alpha_1}{u - \alpha_1 - \alpha_2} \int_1^{\frac{u-\alpha_2}{\alpha_1}} \{2^{-1/2} z_1 + 2^{-1/2} \alpha_2^{-1}(u - z_1\alpha_1)\}^3 dz_1 - E[X]'(\beta - \beta_0).$$

By taking derivative, we obtain $\frac{d\psi_{\alpha,\beta}(u)}{du} = 3u^2 + O(||\alpha - \alpha_0||)$. For (ii), (iii), and (iv), similar arguments also imply $\frac{d\psi_{\alpha,\beta}(u)}{du} = 3u^2 + O(||\alpha - \alpha_0||)$. Therefore, by taking δ_0 small enough, we obtain $\frac{d\psi_{\alpha,\beta}(u)}{du} > 0$ for all $u \in I_\alpha$, so A3 is satisfied.

[6]For example, for given β satisfying $(\beta', \gamma')' \in \mathcal{B}(\theta_0, \delta_0)$ and $z \in \mathcal{Z}$, we can show that $W_\beta = Y - X'\beta$ satisfies $E[|W_\beta|^m | Z = z] \leq m! M_0^{m-2} c_0$ for all integers $m \geq 2$ when the conditional density function of $W_\beta | Z$ takes the form of $f_{W_\beta|Z}(w|z) = h(w, \vartheta_{2,\beta,z}) \exp\{\vartheta_{2,\beta,z}^{-1} w\ell(\vartheta_{1,\beta,z}) - \vartheta_{2,\beta,z}^{-1} B(\ell(\vartheta_{1,\beta,z}))\}$). Here, $\vartheta_{1,\beta,z}$ is the conditional mean (may vary with β and z), $\vartheta_{2,\beta,z}$ is a conditional dispersion parameter (may vary with β and z), ℓ is a real valued function with a strictly positive first derivative on an open interval, B is a real valued function, and h is a normalizing function. This can be shown by adapting Balabdaoui, Durot and Jankowski (2019, Proposition 9.2) for the conditional case.

ability approaching one, $\hat{\theta} \xrightarrow{p} \theta_0$*, and*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Pi),$$

where $\Pi = B^{-1}T_0\Sigma T_0'(B^{-1})'$*,* $\Sigma = \mathrm{Var}(V_{X,Z}\epsilon)$*, and* $V_{X,Z}$ *is* $V_{x,z}$ *evaluated at* $(x, z) = (X, Z)$*.*

This theorem says that our score-type estimator $\hat{\theta}$ for the monotone PLSI model is $\sqrt{n}$-consistent and asymptotically normal without any tuning parameter.[7] The asymptotic variance Π can be estimated by (i) replacing P_0 with the empirical measure $\mathbb{P}_n$, (ii) replacing γ_0 with its estimator $\hat{\gamma}$, (iii) replacing ψ_0' with $\hat{\psi}'_{nh,\theta}$ in (1.7) below, (iv) replacing ϵ with the residuals based on our estimator, and (v) replacing the conditional expectations with kernel estimators.[8] Our result can be considered as an extension of BGH for the monotone PLSI model. Technically a major difference from BGH is the treatment on the mapping $\psi_\theta(\cdot)$, which involves an additional term from the linear component $X'\beta$ (i.e., the second term of (A.1) in Appendix). Most entropy results in our proof are modified to accommodate this additional term.

We note that the estimator $\hat{\theta}$ is derived from the modified population score equation in (1.3) instead of the original one in (1.2). Consequently, the asymptotic variance Π of $\hat{\theta}$ is not the efficient variance for the PLSI model. If we allow one tuning parameter, we can evaluate the efficient score function in (1.2) as

$$\xi_{nh}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} X_i \\ \mathbb{J}(\gamma)'Z_i\hat{\psi}'_{nh,\theta}(Z_i'\mathbb{S}(\gamma)) \end{pmatrix} \{Y_i - X_i'\beta - \hat{\psi}_{n\theta}(Z_i'\mathbb{S}(\gamma))\}, \tag{1.6}$$

where

$$\hat{\psi}'_{nh,\theta}(u) = \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\hat{\psi}_{n\theta}(x), \tag{1.7}$$

[7] Due to discontinuity in $\hat{\psi}_{n\theta}$, we can only guarantee the existence of $\hat{\theta}$ with probability approaching one. Similar to other zero-crossing estimators using isotonic regression, its existence for a given sample size is an open question.

[8] For example, the conditional expectation $\mu(z) = E[X|z'\mathbb{S}(\gamma_0)]$ in $V_{x,z}$ and $V_{x,z,\psi'}$ can be estimated by

$$\hat{\mu}(z) = \frac{\sum_{i=1}^n K\left(\frac{Z_i'\mathbb{S}(\hat{\gamma}) - z'\mathbb{S}(\hat{\gamma})}{b}\right) X_i}{\sum_{i=1}^n K\left(\frac{Z_i'\mathbb{S}(\hat{\gamma}) - z'\mathbb{S}(\hat{\gamma})}{b}\right)},$$

where K is a kernel function (e.g., Gaussian and Epanechnikov) and b is a bandwidth.

is an estimator for the derivative of ψ_θ (defined in A3) with a kernel function K and bandwidth h. Let $\tilde{\theta} = (\tilde{\beta}', \tilde{\gamma}')'$ be the zero-crossing of (1.6).[9] For this estimator, we add the following assumptions.

Assumption.

A8 *$\psi_\theta(z'\alpha)$ is twice continuously differentiable on $I_\alpha = \{z'\alpha, z \in \mathcal{Z}\}$ for each $\theta \in \mathcal{B}(\theta_0, \delta_0)$.*

A9 *$K(\cdot)$ is a symmetric twice differentiable kernel function with compact support $[-1, 1]$. Furthermore, $h \asymp n^{-1/7}$.*

A8 is an additional condition to control the entropy for classes of functions to achieve the $\sqrt{n}$-consistency of $\tilde{\theta}$. A9 contains assumptions for the kernel function K and bandwidth h to evaluate $\hat{\psi}'_{nh,\theta}$ in (1.7). The condition $h \asymp n^{-1/7}$ is also imposed in BGH.

The asymptotic properties of the estimator $\tilde{\theta}$ are presented as follows.

Theorem 1.2. *Suppose Assumptions A1-A9 hold true. Then $\tilde{\theta}$ exists with probability approaching one, $\tilde{\theta} \xrightarrow{p} \theta_0$, and*

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N(0, \Pi_E),$$

where $\Pi_E = B_E^{-1} T_0 \Sigma T_0' (B_E^{-1})'$, $\Sigma = \mathrm{Var}(V_{X,Z,\psi'}\epsilon)$, and $V_{X,Z,\psi'}$ is $V_{x,z,\psi'}$ evaluated at $(x, z) = (X, Z)$.

If we additionally assume $\mathrm{Var}(\epsilon|X, Z) = \mathrm{Var}(\epsilon) = \sigma^2$ (i.e., the error term ϵ is homoskedastic), then Σ can be written as $\Sigma = \sigma^2 \int V_{x,z,\psi'} V'_{x,z,\psi'} dP_0(x, z)$. Therefore, the asymptotic variance becomes $\Pi_E = B_E^{-1}$, which coincides with the efficient variance matrix derived in Carroll *et al.* (1997) and Xia and Härdle (2006). The asymptotic variance Π_E can be estimated in the same manner as Π.

[9] Similar to $\hat{\theta}$, the zero-crossing estimator $\tilde{\theta}$ may not be unique. If the intersection of $\xi_{nh}(\theta)$ and zero is an interval, any point on this interval satisfies the result in Theorem 1.2.

1.2.3 Bootstrap inference

One advantage of the proposed estimator $\hat{\theta}$ is that it is free from tuning parameters, such as bandwidths and series lengths. On the other hand, since its asymptotic variance Π involves conditional means, inference using estimation of Π requires some smoothing method. To obtain an inference procedure which is free from tuning parameters, we propose a bootstrap method to approximate the distribution of the score-type estimator $\hat{\theta}$. Groeneboom and Hendrickx (2017) established the bootstrap validity of their score estimator for the parametric part in a current status model. We extend their result to the monotone PLSI model.

Let $\hat{\theta}^*$ be the bootstrap counterpart of $\hat{\theta}$ defined in Section 1.2.1 based on resamples from the empirical distribution of $\{Y_i, X_i, Z_i\}_{i=1}^n$. The validity of the bootstrap approximation is obtained as follows.

Theorem 1.3. *Suppose Assumptions A1-A7 hold true. Then*

$$\sup_{t \in \mathbb{R}^{k+d-1}} |P^*\{\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq t\} - P_0\{\sqrt{n}(\hat{\theta} - \theta_0) \leq t\}| \xrightarrow{p} 0,$$

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

The bootstrap confidence interval and standard error can be obtained by this result. Note that computation of $\hat{\theta}^*$ and the resulting bootstrap inference are free from tuning parameters.

1.3 Monte-Carlo Simulations

In this section, we conduct a Monte-Carlo simulation study to illustrate the finite sample performance of the proposed estimators.

1.3.1 Simple score and efficient score estimators

We consider the following partial linear model:

$$\begin{aligned} Y &= X\beta_0 + \psi_0(Z'\alpha_0) + \epsilon, \\ \psi_0(u) &= u^3, \qquad \beta_0 = 1, \qquad \alpha_0' = (1,1)/\sqrt{2} \approx (0.7071, 0.7071), \end{aligned}$$

where $X \sim N(0,1)$ and $\epsilon \sim N(0,1)$. For Z, we consider two data generating processes: (i) $Z \sim U[1,2]^2$ (in Table 1.1) and (ii) $Z \sim N(0, \mathbb{I}_2)$ with the 2×2 identity matrix (in Table 1.2). The sample sizes are $n = 100$, 500, and 1000. The number of Monte Carlo replications is 1000. Tables 1.1 and 1.2 present the Monte Carlo averages $(\hat{\mu}_\beta, \hat{\mu}_{\alpha_1}, \hat{\mu}_{\alpha_2})$ and variances $(\hat{\sigma}^2_\beta, \hat{\sigma}^2_{\alpha_1}, \hat{\sigma}^2_{\alpha_2})$ (multiplied by n) of the estimates $(\hat{\beta}, \hat{\alpha}_1, \hat{\alpha}_2)$ and $(\tilde{\beta}, \tilde{\alpha}_1, \tilde{\alpha}_2)$ for Cases (i) and (ii), respectively.

In the tables, SSE is the simple score estimator obtained by solving the zero-crossing of (1.4), and ESE is the efficient score estimator obtained by solving the zero-crossing of (1.6). SSE_L and ESE_L are the Lagrange versions of SSE and ESE suggested by BGH and Groeneboom (2018).[10] All these methods are implemented by the Hooke-Jeeves algorithm to search a minimizer of the sum of squared score components. In the reported simulation results, we follow BGH and use the true values as starting values. Preliminary simulation suggests that the results are not sensitive to local changes for the starting values. For comparison, we include monotone least square methods (LSE in the tables). We also include the smoothing method by Xia and Härdle (2006) into our comparison (S_LSE in the tables). Xia and Härdle (2006) showed that the optimal bandwidth for their methods is of order $n^{-1/5}$. BGH showed that the optimal bandwidth for their efficient estimator is of order $n^{-1/7}$, and suggested to use $h = \hat{r}n^{-1/7}$, where $\hat{r}$ is the range of $Z'\alpha$, as bandwidth. Here we follow BGH's practice. We choose $\hat{r}n^{-1/7}$ as bandwidth for ESE and $\hat{r}n^{-1/5}$ for S_LSE.

The theoretical asymptotic variances are calculated for SSE, ESE, and S_LSE.

[10] More precisely, the estimator SSE_L is obtained by a zero-crossing of

$$\phi_n^L(\theta) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_i\{Y_i - X_i'\beta - \hat{\psi}_{n\alpha}(Z_i'\alpha)\} \\ \frac{1}{n}(1-\alpha'\alpha)\sum_{i=1}^n X_i\{Y_i - X_i'\beta - \hat{\psi}_{n\alpha}(Z_i'\alpha)\} \end{bmatrix},$$

and ESE_L is defined analogously.

Both ESE and S_LSE achieve semiparametric efficiency and therefore they should have the same limit. The asymptotic variance of LSE is unknown in the literature (see, Balabdaoui, Durot and Jankowski, 2019, for a detail). It can be shown that for both settings, $Z \sim U[1,2]^2$ and $Z \sim N(0, \mathbb{I}_2)$, we have $E[X|z'\alpha] = 0$ and $E[Z|z'\alpha] = \frac{\sqrt{2}}{2} z'\alpha(1,1)'$. The asymptotic variances of $(\hat{\beta}, \hat{\alpha})$ and $(\tilde{\beta}, \tilde{\alpha})$ (the estimators without reparameterization) can be obtained with Lemma 7 in BGH and numerical integral. In particular, we have $\sqrt{n}\{(\hat{\beta}', \hat{\alpha}')' - (\beta_0', \alpha_0')'\} \xrightarrow{d} N(0, V)$ and $\sqrt{n}\{(\tilde{\beta}', \tilde{\alpha}')' - (\beta_0', \alpha_0')'\} \xrightarrow{d} N(0, V_E)$, where

$$\text{Case (i)} \quad : \quad V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.0324 & -0.0324 \\ 0 & -0.0324 & 0.0324 \end{pmatrix}, \qquad V_E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.0315 & -0.0315 \\ 0 & -0.0315 & 0.0315 \end{pmatrix}.$$

$$\text{Case (ii)} \quad : \quad V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.0555 & -0.0555 \\ 0 & -0.0555 & 0.0555 \end{pmatrix}, \qquad V_E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.0185 & -0.0185 \\ 0 & -0.0185 & 0.0185 \end{pmatrix}.$$

Tables 1.1 and 1.2 show that the estimation biases are reasonably small for the both estimators even for $n = 100$. For the single index part ($\hat{\alpha}_1$ and $\hat{\alpha}_2$), ESE performs better than SSE in terms of efficiency, which is in accordance with the implication of Theorems 1.1 and 1.2. As the sample size increases, SSE_L and ESE_L become almost identical to SSE and ESE, respectively. LSE performs differently in two cases. In Table 1.2, LSE performs better than SSE but worse than ESE. In Table 1.1, LSE performs worse than SSE.

In general, all the variances of SSE and ESE are approaching to their theoretical limits. It seems that the approaching rates are faster in Case (i) than those in Case (ii). S_LSE is approaching the limit in Case (i), but stays away from the limit in Case (ii). Note that Case (ii) violates the assumption that the support of Z is compact required in both Xia and Härdle (2006) and our estimators. Therefore, some irregular behaviors of those estimators might be expected in Case (ii). Nevertheless, SSE and ESE seem to be more stable even if the support of Z is not compact.

Table 1.1: Monte-Carlo simulation results for Case (i) $Z \sim U[1,2]^2$

Methods	n	$\hat{\mu}_\beta$	$\hat{\mu}_{\alpha_1}$	$\hat{\mu}_{\alpha_2}$	$\hat{\sigma}^2_\beta$	$\hat{\sigma}^2_{\alpha_1}$	$\hat{\sigma}^2_{\alpha_2}$
SSE	100	0.9982	0.7068	0.7068	1.3401	0.0415	0.0416
	500	0.9982	0.7068	0.7073	1.0277	0.0364	0.0364
	1000	1.0002	0.7069	0.7073	1.1306	0.0322	0.0322
	∞	1	0.7071	0.7071	1	0.0324	0.0324
ESE	100	0.9984	0.7067	0.7069	1.3743	0.0404	0.0404
	500	0.9983	0.7068	0.7073	1.0252	0.0360	0.0359
	1000	1.0001	0.7069	0.7073	1.1310	0.0319	0.0319
	∞	1	0.7071	0.7071	1	0.0315	0.0315
SSE_L	100	0.9982	0.7072	0.7064	1.3425	0.0420	0.0421
	500	0.9982	0.7068	0.7073	1.0296	0.0363	0.0363
	1000	1.0002	0.7069	0.7073	1.1288	0.0323	0.0323
	∞	1	0.7071	0.7071	1	0.0324	0.0324
ESE_L	100	0.9982	0.7070	0.7066	1.3502	0.0408	0.0410
	500	0.9982	0.7069	0.7072	1.0262	0.0361	0.0360
	1000	1.0001	0.7069	0.7073	1.1336	0.0318	0.0318
	∞	1	0.7071	0.7071	1	0.0315	0.0315
LSE	100	0.9972	0.7074	0.7058	1.3967	0.0703	0.0699
	500	0.9984	0.7067	0.7073	1.0330	0.0754	0.0752
	1000	1.0002	0.7069	0.7072	1.1253	0.0740	0.0739
	∞	1	0.7071	0.7071	n/a	n/a	n/a
S_LSE	100	1.0022	0.7071	0.7065	1.2891	0.0441	0.0443
	500	1.0005	0.7069	0.7072	1.2213	0.0362	0.0361
	1000	1.0023	0.7069	0.7072	1.2053	0.0348	0.0348
	∞	1	0.7071	0.7071	1	0.0315	0.0315

Table 1.2: Monte-Carlo simulation results for Case (ii) $Z \sim N(0, \mathbb{I}_2)$

Methods	n	$\hat{\mu}_\beta$	$\hat{\mu}_{\alpha_1}$	$\hat{\mu}_{\alpha_2}$	$\hat{\sigma}^2_\beta$	$\hat{\sigma}^2_{\alpha_1}$	$\hat{\sigma}^2_{\alpha_2}$
SSE	100	0.9981	0.7035	0.7075	1.3620	0.2310	0.2301
	500	1.0001	0.7065	0.7074	1.1481	0.1087	0.1086
	1000	0.9998	0.7079	0.7062	1.0532	0.0932	0.0937
	∞	1	0.7071	0.7071	1	0.0555	0.0555
ESE	100	0.9988	0.7049	0.7080	1.4422	0.0943	0.0940
	500	1.0000	0.7069	0.7072	1.1333	0.0356	0.0355
	1000	0.9999	0.7075	0.7067	1.0531	0.0309	0.0310
	∞	1	0.7071	0.7071	1	0.0185	0.0185
SSE_L	100	0.9981	0.7037	0.7072	1.3625	0.2352	0.2347
	500	1.0000	0.7065	0.7074	1.1467	0.1090	0.1091
	1000	0.9998	0.7079	0.7062	1.0548	0.0936	0.0941
	∞	1	0.7071	0.7071	1	0.0555	0.0555
ESE_L	100	0.9974	0.7054	0.7074	1.4086	0.0967	0.0973
	500	1.0000	0.7070	0.7071	1.1357	0.0355	0.0355
	1000	0.9999	0.7075	0.7066	1.0589	0.0310	0.0311
	∞	1	0.7071	0.7071	1	0.0185	0.0185
LSE	100	0.9978	0.7063	0.7061	1.3306	0.1269	0.1281
	500	1.0001	0.7071	0.7069	1.1441	0.0815	0.0815
	1000	0.9998	0.7077	0.7064	1.0595	0.0726	0.0729
	∞	1	0.7071	0.7071	n/a	n/a	n/a
S_LSE	100	1.0052	0.7058	0.7034	6.2528	0.3584	0.3599
	500	0.9972	0.7067	0.7065	7.0103	0.3560	0.3589
	1000	1.0022	0.7069	0.7068	6.9869	0.3878	0.3878
	∞	1	0.7071	0.7071	1	0.0185	0.0185

1.3.2 Bootstrap

As mentioned in Section 1.2.3, the purpose of our bootstrap method is to obtain an inference method that is free of tuning parameters. Therefore, we focus on SSE here, since ESE requires at least one tuning parameter. Since the results are

analogous, we only consider Case (ii) above. Most notations in Table 1.3 are as defined in the previous subsection. Results for SSE are replicated from Table 1.2. SSE_b is the bootstrap counterpart of the estimator by SSE, and the number of the bootstrap replications is 500.

Table 1.3 shows that as the sample size increases, the distribution of SSE_b approaches to that of SSE, which is in accordance with the implication of Theorem 1.3.

Table 1.3: Monte-Carlo simulation results for bootstrap counterparts

Methods	n	$\hat{\mu}_\beta$	$\hat{\mu}_{\alpha_1}$	$\hat{\mu}_{\alpha_2}$	$\hat{\sigma}^2_\beta$	$\hat{\sigma}^2_{\alpha_1}$	$\hat{\sigma}^2_{\alpha_2}$
SSE	100	0.9982	0.7068	0.7068	1.3401	0.0415	0.0416
	500	0.9982	0.7068	0.7073	1.0277	0.0364	0.0364
	1000	1.0002	0.7069	0.7073	1.1306	0.0322	0.0322
SSE_b	100	1.0599	0.7078	0.7059	1.2614	0.0488	0.0507
	500	0.9729	0.6970	0.7170	1.0354	0.0286	0.0270
	1000	0.9952	0.7092	0.7049	1.1236	0.0359	0.0364

Overall, the Monte-Carlo simulation results are encouraging to support our estimation strategy.

Chapter 2

Empirical likelihood inference for monotone index model

2.1 Introduction

Single index models are widely used in statistics since they compromise interpretability of index coefficients in the parametric part and flexibility of regression modeling in the nonparametric part (see, ch. 8 of Li and Racine, 2007, for a review). Many estimation methods have been proposed for single index models, such as the semiparametric least squares estimator (Härdle, Hall and Ichimura, 1993; Ichimura, 1993), M-estimator (Klein and Spady, 1993), binary threshold choice model (Matzkin, 1992), and average derivative estimator (Powell, Stock and Stoker, 1989). Although these estimation methods have desirable theoretical properties under certain regularity conditions, they typically require some nonparametric smoothing method to evaluate the unknown link function, which involves tuning parameters, such as bandwidth and series length parameters, and the optimal choices of them are substantial (theoretical and practical) problems.

The monotone single index model, in which monotonicity is imposed on the link function, has been studied in recent years. Balabdaoui, Durot and Jankowski (2019) showed that the least square estimator of a monotone single index model generally converges at the cube root rate, but its asymptotic distribution is still unknown. The main difficulty for deriving the asymptotic distribution of the least

square estimator arises from the non-differentiability of the objective function; in a monotone single index model, the link function, which is an infinite-dimensional nuisance parameter, is generally estimated by a nonparametric approach such as isotonic regression, while the index part is parametrically modeled as a linear combination of the covariates. Then the derivative of the objective function with respect to the index coefficients is intractable due to the non-smoothness of the estimated nuisance parameter.

To overcome this issue, Groeneboom and Hendrickx (2018) developed a score-type estimator for the current status model, which is a special case of monotone single index models. Their approach is based on the estimating equation which is the same as the first-order condition of the least square estimator except that it ignores the derivative of the estimated link function. They proved $\sqrt{n}$-consistency and asymptotic normality of their estimator without any tuning parameter. Their result was extended to general monotone single index models by Balabdaoui, Groeneboom and Hendrickx (2019), where they derived $\sqrt{n}$-consistency and asymptotic normality for the parametric component and an $n^{1/3}/\log n$ convergence rate for the nonparametric estimator of the link function.

Although the score estimation approach is remarkable, the main drawback is that it requires smoothing parameters to estimate the asymptotic variance to implement hypothesis testing and interval estimation. Because the estimating function in the score-type approach is dependent on the estimated link function, some conditional expectation is involved in the asymptotic variance. Besides, the partial derivative of the link function is also included in the asymptotic variance even though the estimated link function is not smooth. Therefore, smoothing methods, such as the kernel smoothing, are employed to estimate such quantities, which require us to select multiple smoothing parameters and make statistical inference cumbersome.

To address this problem, we propose an empirical likelihood inference method based on the score-type approach for monotone index models. We show that the empirical likelihood statistic based on the estimating equation of Balabdaoui, Groeneboom and Hendrickx (2019) converges in distribution to the weighted chi-squared distribution. Even in our empirical likelihood approach, the conditional

expectation as mentioned above appears in the asymptotic distribution. To circumvent selection of smoothing parameters, we adapt the bootstrap calibration method proposed by Hjort, McKeague and van Keilegom (2009) to our context. Because of the estimating equation with the estimated nuisance parameter plugged-in, a classical naive bootstrap method is not asymptotically valid. Hjort, McKeague and van Keilegom (2009) provided a modified bootstrap method by recentering and reweighting to deal with such a situation. Combining the empirical likelihood and modified bootstrap methods, our approach provides a simple and theoretically justified method for statistical inference in monotone single index models.

The remainder of this paper is organized as follows. Section 2 presents our basic setup, methodology, and theoretical results. In Section 3, we conduct a Monte-Carlo simulation study to illustrate the proposed method. All proofs are contained in the appendix.

2.2 Main result

We closely follow the setup and notation of Balabdaoui, Groeneboom and Hendrickx (2019) (hereafter BGH). Consider the monotone index model

$$Y = \psi_0(X'\alpha_0) + \epsilon, \qquad E[\epsilon|X] = 0, \tag{2.1}$$

where Y is a scalar response variable, X is a d-dimensional vector of covariates, ϵ is an error term, α_0 is a k-dimensional vector of parameters, and $\psi_0 : \mathbb{R} \to \mathbb{R}$ is an unknown monotone increasing function. For identification, we assume that α_0 belongs to the d-dimensional unit sphere $\mathcal{S}_{d-1} = \{\alpha \in \mathbb{R}^d : ||\alpha|| = 1\}$. We are interested in conducting statistical inference (i.e., interval estimation and hypothesis testing) on α_0 based on the empirical likelihood approach.

Let $\mathbb{S} : \mathbb{R}^{d-1} \to \mathcal{S}_{d-1}$ be a parameterization such that for each α in a neighborhood of α_0 on $\mathcal{S}_{d-1}$, there exists a unique $\beta \in \mathbb{R}^{d-1}$ which satisfies $\alpha = \mathbb{S}(\beta)$. To motivate the score-type approach of BGH, we tentatively assume that ψ_0 is

known. The population score equation for the least square estimation of β_0 is

$$E\left[\mathbb{J}(\beta_0)'X\psi_0^{(1)}(X'\mathbb{S}(\beta_0))\{Y-\psi_0(X'\mathbb{S}(\beta_0))\}\right]=0, \tag{2.2}$$

where $\psi_0^{(1)}$ is the derivative of ψ_0 and $\mathbb{J}(\beta)$ is the Jacobian of $\mathbb{S}(\beta)$. Thus, it is natural to construct an estimator of β_0 by taking an empirical counterpart of (2.2) and inserting estimators for $\psi_0^{(1)}$ and ψ_0. However, when we estimate ψ_0 by the isotonic regression method, the resulting estimator of ψ_0 is typically discontinuous and it is not clear how to evaluate the derivative $\psi_0^{(1)}$ without introducing smoothing parameters. To address this issue, BGH and Groeneboom and Hendrickx (2018) considered the modified population score equation

$$E\left[\mathbb{J}(\beta_0)'X\{Y-\psi_0(X'\mathbb{S}(\beta_0))\}\right]=0. \tag{2.3}$$

In particular, for point estimation of α_0, BGH proposed to solve the following score-type equation

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{J}(\hat{\beta})'X_i\{Y_i-\hat{\psi}_{\hat{\beta}}(X_i'\mathbb{S}(\hat{\beta}))\}=0, \tag{2.4}$$

with respect to $\hat{\beta}$, and estimate α_0 by $\hat{\alpha}=\mathbb{S}(\hat{\beta})$, where for given β, $\hat{\psi}_\beta$ is obtained by the isotonic regression

$$\hat{\psi}_\beta=\arg\min_{\psi\in\mathcal{M}}\sum_{i=1}^{n}\{Y_i-\psi(X_i'\mathbb{S}(\beta))\}^2, \tag{2.5}$$

and $\mathcal{M}$ is the set of monotone increasing functions defined on $\mathbb{R}$.

In this paper, we employ the score-type equation in (2.3) as a moment function and propose the following empirical likelihood statistic

$$\ell(\beta_0)=-2\max_{\{p_i\}_{i=1}^n}\sum_{i=1}^{n}\log(np_i)\qquad\text{s.t.}\ \sum_{i=1}^{n}p_i=1,\quad\sum_{i=1}^{n}p_i\hat{g}_i(\beta_0)=0, \tag{2.6}$$

where

$$\hat{g}_i(\beta)=\mathbb{J}(\beta)'X_i\{Y_i-\hat{\psi}_\beta(X_i'\mathbb{S}(\beta))\}.$$

By the Lagrange multiplier argument, its dual form is obtained as

$$\ell(\beta_0) = 2\sum_{i=1}^{n} \log(1 + \hat{\lambda}'\hat{g}_i(\beta_0)), \tag{2.7}$$

where the Lagrange multiplier $\hat{\lambda}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{g}_i(\beta_0)}{1 + \hat{\lambda}'\hat{g}_i(\beta_0)} = 0. \tag{2.8}$$

In practice, we use the dual representation in (2.7) to implement statistical inference. To study the asymptotic properties of the empirical likelihood statistic $\ell(\beta_0)$, we impose the following assumptions. Let $\|\cdot\|$ be the Euclidean norm and $\mathcal{B}(a_0, A) = \{a : \|a - a_0\| \leq A\}$ be a ball around a_0 of radius A.

Assumption.

A1 *$\{Y_i, X_i\}_{i=1}^n$ is an iid sample generated by (2.1). The support $\mathcal{X}$ of X is convex with a nonempty interior, and $\mathcal{X} \subset \mathcal{B}(0, R)$ for some $R > 0$. The Lebesgue density of X has a bounded derivative on $\mathcal{X}$. There exist positive constants c and C such that $E[|Y|^m|X = x] \leq cm!C^{m-2}$ for all integers $m \geq 2$ and almost every $x \in \mathcal{X}$.*

A2 *ψ_0 is monotone increasing and there exists $K_0 > 0$ such that $|\psi_0(u)| \leq K_0$ for all $u \in \{x'\alpha_0 : x \in \mathcal{X}\}$.*

These assumptions are adaptations of Assumptions A1-A6 in BGH. Compared to BGH, our assumptions are simpler because we do not need to control the behavior of the score function outside the true parameter $\alpha_0 = \mathbb{S}(\beta_0)$. Assumption A1 is on the distribution form of the data. The support condition in A1 may be relaxed by assuming X to follow a sub-Gaussian distribution. The moment condition in A1, which is analogous to BGH's A6, is required to guarantee $\max_{1\leq i\leq n} |Y_i| = O_p(\log n)$ to control the entropy of a class of score functions. Assumption A2 is on the true link function ψ_0. Compared to BGH which considers point estimation, we only need to impose boundedness, which is a mild requirement.

Under these assumptions, our main result is presented as follows.

Theorem 2.1. *Under Assumptions A1-A2, it holds*

$$\ell(\beta_0) \xrightarrow{d} Z'V^{-1}Z,$$

where $Z \sim N(0, \Sigma)$ *with* $\Sigma = \mathbb{J}(\beta_0)'E[\epsilon^2(X - E[X|X'\mathbb{S}(\beta_0)])(X - E[X|X'\mathbb{S}(\beta_0)])']\mathbb{J}(\beta_0)$ *and* $V = \mathbb{J}(\beta_0)'E[\epsilon^2 XX']\mathbb{J}(\beta_0)$.

Remark 2.1. This theorem says that the empirical likelihood statistic $\ell(\beta_0)$ is not asymptotically pivotal and converges to a weighted chi-squared distribution $w_1\chi^2_{1,1} + \cdots + w_{d-1}\chi^2_{1,d-1}$, where $w_1, \ldots, w_{d-1}$ are the eigenvalues of $\Sigma^{-1}V$ and $\chi^2_{1,1}, \ldots, \chi^2_{1,d-1}$ are independent χ^2_1 random variables. This lack of asymptotic pivotalness is caused by the mismatch in the asymptotic variance Σ of the score function $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{g}_i(\beta_0)$ and the limit V of the sample variance $\hat{V} = \frac{1}{n}\sum_{i=1}^n \hat{g}_i(\beta_0)\hat{g}_i(\beta_0)'$. In the literature of empirical likelihood, weighted chi-squared limiting distributions often emerge when the score (or moment) functions involve estimated nuisance parameters (e.g., Qin and Jing, 2001; Xue and Zhu, 2006; Hjort, McKeague, and van Keilegom, 2009).

Remark 2.2. One way to conduct statistical inference based on $\ell(\beta_0)$ is to estimate the critical values of $w_1\chi^2_{1,1} + \cdots + w_{d-1}\chi^2_{1,d-1}$ based on some estimators of Σ and V. Based on (B.3), V is consistently estimated by $\hat{V}$. On the other hand, Σ can be estimated by

$$\hat{\Sigma} = \mathbb{J}(\beta_0)'\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2\{X_i - \hat{m}(X_i'\mathbb{S}(\beta_0))\}\{X_i - \hat{m}(X_i'\mathbb{S}(\beta_0))\}\mathbb{J}(\beta_0),$$

where $\hat{\epsilon}_i = Y_i - \hat{\psi}_{\beta_0}(X_i'\mathbb{S}(\beta_0))$ and $\hat{m}(\cdot)$ is a nonparametric estimator of $m(\cdot) = E[X|X'\mathbb{S}(\beta_0) = \cdot]$. An alternative way for statistical inference is to adjust the empirical likelihood statistic $\ell(\beta_0)$ to recover the asymptotic pivotalness. Based on Rao and Scott (1981) (see also Xue and Zhu, 2006), the above theorem implies

$$\ell_A(\beta_0) = \frac{d-1}{\text{trace}(\hat{\Sigma}^{-1}\hat{V})}\ell(\beta_0) \xrightarrow{d} \chi^2_{d-1}. \tag{2.9}$$

Then the confidence region of $\alpha_0 = \mathbb{S}(\beta_0)$ can be obtained by $\{\mathbb{S}(\beta) : \ell_A(\beta) \leq q_a\}$, where q_a is the $(1-a)$-th quantile of the χ^2_{d-1} distribution.

Remark 2.3. A drawback of the asymptotic inference method presented in the previous remark is that it requires a selection of a tuning parameter to implement

the nonparametric estimator $\hat{m}(\cdot)$. In order to obtain an inference procedure which is free from tuning parameters, we adapt the bootstrap method of Hjort, McKeague, and van Keilegom (2009) as follows.

1. Based on the original sample $\{Y_i, X_i\}_{i=1}^n$, compute $\hat{\beta}$ as in (2.4), and then compute

$$M_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^n \hat{g}_i(\mathbb{S}(\hat{\beta})), \qquad \bar{V} = \frac{1}{n}\sum_{i=1}^n \hat{g}_i(\mathbb{S}(\hat{\beta}))\hat{g}_i(\mathbb{S}(\hat{\beta}))'.$$

2. Draw $\{Y_i^*, X_i^*\}_{i=1}^n$ from the original sample $\{Y_i, X_i\}_{i=1}^n$ with equal weights. Then compute

$$M_n^*(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^n \mathbb{J}(\hat{\beta})' X_i^* \{Y_i^* - \hat{\psi}_{\hat{\beta}}^*(X_i^{*\prime}\mathbb{S}(\hat{\beta}))\},$$

   where $\hat{\psi}_{\hat{\beta}}^* = \arg\min_{\psi\in\mathcal{M}} \sum_{i=1}^n \{Y_i^* - \psi(X_i^{*\prime}\mathbb{S}(\hat{\beta}))\}^2$.

3. The bootstrap counterpart of $\ell(\beta_0)$ is given by

$$\ell^* = n\{M_n^*(\hat{\beta}) - M_n(\hat{\beta})\}'\bar{V}^{-1}\{M_n^*(\hat{\beta}) - M_n(\hat{\beta})\}. \tag{2.10}$$

Under the additional assumptions A3-A5 in the appendix, the validity of this bootstrap approximation is obtained as follows.

Theorem 2.2. *Under Assumptions A1-A5, it holds*

$$\sup_{t\geq 0} |P^*\{\ell^* \leq t\} - P_0\{\ell(\beta_0) \leq t\}| \xrightarrow{p} 0,$$

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

2.3 Monte-Carlo Simulation

We conduct a Monte-Carlo simulation study to investigate the finite sample performance of the proposed inference methods. We consider the following data

generation process:

$$\begin{aligned} Y &= \psi_0(X'\alpha_0) + \epsilon, \quad \psi_0(u) = u^3, \quad \alpha_0 = (1,1,1)'/\sqrt{3} \\ \epsilon &\sim N(0,1), \quad X \sim N(0, I_3), \end{aligned}$$

where I_3 is the 3×3 identity matrix. We consider sample sizes $n = 100, 500, 1000$. The number of Monte Carlo replications is 1000. We consider two testing methods discussed in Remarks 2 and 3. For the adjusted statistic in (2.9), we estimate $m(\cdot) = E[X|X'\mathbb{S}(\beta_0) = \cdot]$ by the Nadaraya-Watson estimator, and choose the bandwidths based on the expected Kullback-Leibler cross-validation (Hurvich, Simonoff and Tsai, 1998). To test the null hypothesis $H_0 : \alpha_0 = (1,1,1)'/\sqrt{3}$, we calculate the test statistic (2.9) and compare it with the 95 percentile of the χ^2_{d-1} distribution. For the bootstrap-calibrated test statistic (2.10), we compute $\hat{\beta}$ as in BGH (the computer code is available at Groeneboom's website), and generate 499 bootstrap samples, and calculate the bootstrap counterpart ℓ^* in (2.10).

Table 2.1 presents the rejection frequencies of the above empirical likelihood tests for the null $H_0 : \alpha_0 = (1,1,1)'/\sqrt{3}$ when the true values of α_0 are (N) $\alpha_0 = (1,1,1)'/\sqrt{3}$, (A1) $\alpha_0 = (1.03,1,1)'/\sqrt{1.03^2+2}$, (A2) $\alpha_0 = (1.05,1,1)'/\sqrt{1.05^2+2}$, and (A3) $\alpha_0 = (1.10,1,1)'/\sqrt{1.10^2+2}$. (N) is for the size properties, and (A1)-(A3) are to evaluate power properties.

The column "$\hat{\alpha}_1$" reports the Monte Carlos averages and standard deviations of the first element of the BGH estimator $\hat{\alpha}$. It shows that the mean is close to the truth, $\alpha_{01} = 1/\sqrt{3} \simeq 0.577$, while the standard deviation becomes smaller with the sample size. From the columns (N), we can see that both the adjusted and bootstrap empirical likelihood tests have reasonable size properties. Both tests become powerful as the sample size increases and the true values of α_0 are more distinct from the null values (i.e., from A1 to A3). Also, we find that overall the bootstrap test rejects slightly more often than the adjusted test.

Table 2.1: Rejection frequencies (in percentage %)

n	Adjusted				Bootstrap				$\hat{\alpha}_1$	
	N	A1	A2	A3	N	A1	A2	A3	mean	s.d.
100	4.7	4.9	6.1	8.7	8.1	8.3	9.0	13.9	0.577	0.0528
500	4.2	7.5	15.9	51.1	6.6	10.0	18.1	53.3	0.576	0.0166
1000	7.4	14.8	31.5	86.1	5.6	18.2	34.9	87.8	0.577	0.0113

Overall, our Monte-Carlo simulation results are encouraging.

2.4 Conclusion of Chapter 1 and Chapter 2

In Chapter 1 and 2, we study the estimation and inference methods of the monotone partially linear index model and the monotone single index model. In the following Chapter 3, we will study a general Z-estimator with plug-in isotonic estimators, which can encompass the estimation methods of the models in the first two chapters as special cases.

Chapter 3

Semiparametric estimation with plug-in isotonic estimators

3.1 Introduction

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

$$E[m(Z, \beta_0, p_0(\cdot))] = 0, \tag{3.1}$$

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. (3.1) 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 (3.1), with a plugged-in $\hat{p}(\cdot)$:

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta, \hat{p}(\cdot)) = 0, \tag{3.2}$$

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

3.1.1 Isotonic estimator

Suppose that the conditional expectation $E(Y|X) = p_0(X)$ is monotone increasing. For an i.i.d random sample $\{Y_i, X_i\}_{i=1}^n$, the isotonic estimator is the minimizer of the sum of squared errors:

$$\min_{p \in \mathcal{M}} \sum_{i=1}^{n} (Y_i - p(X_i))^2,$$

where $\mathcal{M}$ is the class of monotone increasing function. The minimizer can be calculated with Pool Adjacent Violators Algorithm (Barlow et al., 1972), or equivalently by solving the greatest convex minorant of the cumulative sum diagram $\left\{(0,0), (i, \sum_{j=1}^{i} Y_j), i = 1, ..., n\right\}$. See Ayer et al. (1955), Barlow et al. (1972), and the textbook of Groeneboom and Jongbloed (2014) for details.

3.1.2 Motivation and challenges

Without the monotonicity assumption about $p_0(\cdot)$, the model (3.1) and its plug-in estimator based on (3.2) 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

$$Y = \begin{cases} 1 & \text{if } X'\beta_0 > \varepsilon \\ 0 & \text{if } X'\beta_0 \leq \varepsilon \end{cases}. \tag{3.3}$$

We can express the conditional expectation $P_0(X) \equiv E(Y|X) = P(Y = 1|X) = F_\varepsilon(X'\beta_0)$, where $F_\varepsilon(\cdot)$ is the CDF of ε. If we assume $\varepsilon \sim N(0, 1)$, (3.3) becomes a probit model; if we assume $\varepsilon \sim Logistic(0, 1)$, (3.3) becomes a logit model. If we don't impose any distributional assumptions on ε, we can express (3.3) 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. All these features will be inherited by the corresponding semiparametric estimator.

Third, as a nonparametric estimator, the isotonic estimator has some drawbacks: (i) the isotonic estimator has a comparatively slower convergence rate of $n^{-1/3}$, while other nonparametric estimators can achieve better rates under moderate smoothness conditions; (ii) the isotonic estimator is a discrete estimator, which imposes problems in many applications. Interestingly, these drawbacks can become merits in the semiparametric estimator with isotonic plug-in estimator: the discrete feature is associated with the tuning-parameter-free property; the low convergence rate is associated with a smaller bias, and this small bias combined with monotonicity leads to a nice performance in the second stage semiparametric estimator. In contrast, a plug-in kernel estimator with optimally chosen bandwidth might lead to inefficiency in the semiparametric estimator. (Bickel and Ritov, 2003).

A challenge of making inference of $\hat{\beta}$ based on (3.2) 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, (3.2) no longer has a continuous total derivative w.r.t β even if $m(Z, \beta, p_0(\cdot))$ 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 (3.1). 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.

3.1.3 Examples and literature

We give examples of semiparametric models, which can be estimated with the procedure described in (3.1) and (3.2). 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; Carroll *et al.* (1997), Xia and Härdle (2006), and Wang *et al.* (2010) for the partially linear 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 to Example 3 below.

Example 1: Monotone partially linear model.

$$Y = D\beta_0 + p_0(X) + \varepsilon \quad \text{with } E[\varepsilon|X, D] = 0. \tag{3.4}$$

For monotone increasing $p_0(X)$, Huang (2002) estimated 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 partially

linear model, studied in Cheng (2009) and Yu (2014).

Alternatively, β_0 can be estimated by solving the problem (3.1), with the moment condition

$$E\left[m(Z,\beta,p(\cdot))\right] = E\left[D(Y - D\beta - p(X))\right] = 0. \tag{3.5}$$

As illustrated in Chernozhukov et al. (2018), the simple plug-in method based on (3.5) could fail sometimes since this moment function is not Neyman-orthogonalized. In Section 3.2.1, we will show that if $p_0(\cdot)$ is monotone increasing and estimated with isotonic regression, the estimator $\hat{\beta}$ based on (3.5) is $\sqrt{n}$-consistent and has the same asymptotic variance as that in Robinson (1988). We do not need to orthogonalize (3.5).

Example 2: Monotone single index model

$$Y = p_0(X'\beta_0) + \varepsilon \quad \text{with } E[\varepsilon|X] = 0. \tag{3.6}$$

In this example and the next example, $p_0(\cdot)$ is a monotone increasing link function of its index. If Y is a binary random variable taking values in $\{0,1\}$, this model can be derived from (3.3), and $p_0(\cdot)$ is by nature monotone increasing. This model was studied by Cosslett (1983, 1987), Matzkin (1992), 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 (3.6) 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] of:

$$E\left[X\left\{Y - p(X'\beta)\right\}\right] = 0. \tag{3.7}$$

They showed that solving (3.7) can simultaneously estimate β_0 and $p_0(\cdot)$, at $n^{-1/2}$-

[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 (3.7).

rate and $n^{-1/3}$-rate respectively. Note that (3.7) can be regarded as an individual case of the model (3.1) with $m(z, \beta, p(\cdot)) = x\{y - p(x'\beta)\}$.

Example 3: Monotone partially linear index model

$$Y = D'\beta_0 + p_0(X'\alpha_0) + \epsilon, \qquad E[\epsilon|D, X] = 0.$$

Here we let $Z = (Y, D, X)$, and $\theta = (\alpha', \beta')' \in \Theta$. This model combines the features of the model (3.4) and the model (3.6). For monotone increasing $p_0(\cdot)$, Xu and Otsu (2020) extended BGH's approach and showed that a score-type estimator, based on the moment condition

$$E\left[\binom{D}{X}\{Y - D'\beta - p(X'\alpha)\}\right] = 0,$$

can achieve the $\sqrt{n}$-consistency and asymptotically normality for θ_0. Their method can also be regarded as an individual case of the model (3.1) with $m(z, \theta, p(\cdot)) = \binom{d}{x}\{y - d'\beta - p(x'\alpha)\}$.

Example 4: IPW and AIPW with monotone increasing propensity scores

Here we let $Z = (Y, T, X)$, where T is a binary random variable indicating the treatment status. The propensity score is defined as $p_0(X) := E(T|X) = P(T = 1|X)$. 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 (3.1), with the moment condition.

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

(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 (3.1), 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 (3.1), with the moment condition.

$$E[m(Z, \beta, p(\cdot))] = E(\frac{Y \cdot T}{p(X)} - \frac{T - p(X)}{p(X)}\psi(X)) - \beta) = 0. \tag{3.8}$$

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 exceptions we found are Qin et al. (2019) and Yuan et al. (2021). They apply the monotone single index model to estimate the propensity score $p(X) = \pi(X'\alpha)$ of an AIPW model, then plug $\hat{p}(\cdot)$ and another estimator of $\psi_0(\cdot)$ into the sample counterpart of (3.8). Their asymptotic results depend on the consistent estimations of both $p_0(\cdot)$ and $\psi_0(\cdot)$, which are different from our settings. Another different but related paper is Westling et al. (2019). They studied a continuous version of AIPW, where 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 3.2.2, we show that our method can give us a tuning-parameter free, $\sqrt{n}$-consistent, and asymptotically normal IPW estimator.

3.1.4 Contribution and structure of the paper

The main contributions of this chapter are:

1. We develop a tuning-parameter-free semiparametric estimator of (3.1). It

generalizes existing semiparametric models with monotone nuisance functions, including those discussed in Chapter 1 and Chapter 2. 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 (3.1). With x being a multi-dimensional vector, the estimator is $\sqrt{n}$-consistent under different structures combining monotonicity and multi-dimensional covariates.

4. Monte-Carlo 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 Monte-Carlo simulation studies to illustrate the proposed method. All the proofs are presented in the appendix.

3.2 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 3.3, we will allow X to be multi-dimensional covariates. Now we have (3.1) and

$$E(Y|X) = p_0(X), \tag{3.9}$$

where $p_0(\cdot)$ is a monotone increasing function in X. Condition (3.9) 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.

First, we extend (3.2) around β_0, then around $p_0(\cdot)$. In the following part, for any differentiable function $g(\theta, z)$, we denote $\frac{dg(\theta,z)}{d\theta}|_{\theta=\theta_0}$ and $\frac{\partial g(\theta,z)}{\partial\theta}|_{\theta=\theta_0}$ by $\frac{dg(\theta_0,z)}{d\theta}$ and $\frac{\partial g(\theta_0,z)}{\partial\theta}$.

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(Z_i,\beta_0,\hat{p}(\cdot))}{\partial\beta}(\hat{\beta}-\beta_0)\\
&=\frac{1}{n}\sum_{i=1}^{n}m(Z_i,\beta_0,\hat{p}(\cdot))+o_p(\hat{\beta}-\beta_0)\\
&=\frac{1}{n}\sum_{i=1}^{n}m(Z_i,\beta_0,p_0(X_i))+\frac{1}{n}\sum_{i=1}^{n}D(Z_i,\beta_0)(\hat{p}(X_i)-p_0(X_i))\\
&+\frac{1}{n}\sum_{i=1}^{n}O_p(\hat{p}(X_i)-p_0(X_i))^2+o_p(\hat{\beta}-\beta_0)\\
&=\frac{1}{n}\sum_{i=1}^{n}m(Z_i,\beta_0,p_0(X_i))+I+II+o_p(\hat{\beta}-\beta_0).
\end{aligned} \tag{3.10}$$

$D(z,\beta)$ 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 $||\hat{p}-p_0|| = o_p(n^{-1/4})$,

[2]Note that D here is a function of z and β. It should be differentiated from the random variable D in the examples discussed in the introduction.

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

$$||\hat{p} - p_0||^2 = O_p((\log n)^2 n^{-2/3}) = o_p(n^{-1/2}). \tag{3.11}$$

(See, e.g., Theorem 9.2 and Lemma 5.15 in van de Geer, S., 2000). The condition is satisfied without involving any tuning parameter.

We can decompose I into

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

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 (3.2) 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 discuss the link between Neyman orthogonality (see, e.g., Chernozhukov et al., 2018) and the plug-in isotonic estimator.

3.2.1 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 (3.2) 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 estimator. In this case, the plug-in estimator obtained by choosing $\hat{\beta}$, such that

$$\frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - D_i\hat{\beta} - \hat{p}(X_i)) = 0, \tag{3.12}$$

can fail to be $\sqrt{n}$-consistent. Let us rearrange (3.12)

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

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i(p_0(X_i) - \hat{p}(X_i))$ might explode since it is an average of n terms that do not have zero mean.

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

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

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

where G_p is the functional derivative of $m(Z, \beta, p)$ w.r.t p. In our setting (3.9), where $p_0(X)$ is a conditional mean of Y, the orthogonalization can be achieved by applying Proposition 4 of Newey (1994):

$$m_1^o(Z, \beta, p) = m(Z, \beta, p) + E[D(Z, \beta)|X](Y - p(X)).$$

We can check the two conditions for the Neyman orthogonalization. For m_1^o:

1. $E[m_1^o(Z, \beta_0, p_0(X))] = 0 + E[E[D(Z, \beta_0)|X](Y - p_0(X))] = 0,$
2. and

$$\begin{aligned}
&E[D_1^o(Z, \beta_0)(p(X) - p_0(X))] \\
&= E[\frac{\partial m_1^o(\mathrm{Z},\beta,p_0(\mathrm{X}))}{\partial p}(p(X) - p_0(X))] \\
&= E[D(Z, \beta_0)(p(X) - p_0(X))] - E[D(Z, \beta_0)|X][(p(X) - p_0(X))] \\
&= E[D(Z, \beta_0)|X][(p(X) - p_0(X))] - E[D(Z, \beta_0)|X][(p(X) - p_0(X))] \\
&= 0.
\end{aligned}$$

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 3.1, 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 (3.2). 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 in the second-stage semiparametric estimator.

Lemma 3.1. *$\hat{p}(\cdot)$ 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*

$$\frac{1}{n} \sum_{i=1}^{n} \delta(X_i)(Y_i - \hat{p}(X_i)) = o_p(n^{-1/2}). \tag{3.13}$$

Remark 3.1. 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. Heuristically, the intuition can be explained with the OLS estimator:

1. The first-order condition of OLS estimation of $Y = X\beta + \varepsilon$ is that $X'\hat{\varepsilon}$ is equal to zero. The regression residuals $\hat{\varepsilon}$ is, in other words, the projection residual from projecting Y onto the linear space spanned by the columns of X. We have the projection residuals on the right-hand-side, and vectors from the projected space, X, on the left-hand. Their inner product is zero.

2. A similar case is the Lemma 3.1 with isotonic estimators. At the right-hand side of $\delta(X_i)(Y_i - \hat{p}(X_i))$ in (3.13), we have the regression residual of the isotonic regression, which can be regarded as the projection residual

of projecting Y onto the space of monotone increasing functions of X. On the left-hand-side is some function of X, which is assumed to be bounded and with finite total variations. Any bounded function with finite total variations can be decomposed into a sum of two monotone functions.

3. Then we have again the residuals of projecting Y onto the space of monotone functions on the right-hand side and the monotone functions on the left-hand side. It is not exactly zero because, on the right-hand, we have residuals of projecting Y onto the space of monotone piecewise constant function (isotonic estimator). It is not perfectly matched to the monotone functions (but not necessarily piecewise constant) on the left-hand side. The proof can be reduced to show what is left (the approximation error of monotone piecewise constant functions to monotone functions, times the residuals of isotonic estimation) converges to zero faster than $n^{-1/2}$. And the monotonicity plays a role here.

Now 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 for each row of $D(Z,\beta)$ ($D_j(Z,\beta)$ with $j \in \{1:k\}$), $E[|D_j(Z,\beta)|^m|X=x] \leq m!M_1^{m-2}c_1$ 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) of Newey (1994). A5 will be implied by the condition that $m(z,\beta,p(x))$ is differentiable in $p(x)$, for almost every x and z.

Now we have

Proposition 3.1. ***(Sample moment function)*** *Assuming A1-A5, and $p_0(\cdot)$ is estimated with isotonic estimation and plugged into* (3.2)*, 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 3.2. 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 for the estimation of β_0. In this sense, the sample moment function can be regarded as "automatic" Neyman-orthogonalized.

Remark 3.3. The term "automatic" should be understood only in the context of the estimation of β_0. It does not claim that the original moment function $m(z, \beta, p(\cdot))$ is Neyman-orthogonalized. In general, it is not. However, if the monotone nuisance function is estimated with isotonic estimation, the impact of the first-stage isotonic estimation on the moment function (multiplied by $\sqrt{n}$) will be asymptotically equivalent to a correction term, which would properly orthogonalize the original moment function.

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 3.2 (A modified version Lemma 3.1 in the following Section 3.2.3, 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, $\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{d}N(0,\sigma_u^2E(D-E[D|X])^{-2})$.

Remark 3.4. Huang (2012) showed the same asymptotic variance for the partially linear model with monotone nuisance function, with the monotone least square method. 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 (3.12) 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 estimating 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).

3.2.2 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:

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

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 $\beta \in \mathfrak{B}$, where $\mathfrak{B}$ is compact.

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.

A9 For $\beta \in$ interior($\mathfrak{B}$), (i) there are $\varepsilon > 0$ and a neighborhood $\mathscr{N}$ of β_0 such that for all $||p - p_0|| \leq \varepsilon$, $m(z, \beta, p)$ is differentiable in β on $\mathscr{N}$; (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 $m(z, \beta, p)$ equaling 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 (3.11), we have II in (3.10) converging to zero 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

$$M(Z) = E[D(Z, \beta_0)|X](Y - p_0(X)).$$

Then we have

Theorem 3.1. ***(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 (3.2) *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 3.1 to the IPW model by using the isotonic regression to estimate the propensity score.

Example 4 (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 $\hat{\beta}$:

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{Y_i \cdot T_i}{\hat{p}(X_i)} - \frac{Y_i \cdot (1 - T_i)}{1 - \hat{p}(X_i)} \right\}. \tag{3.14}$$

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 numbers $\underline{p}$ 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 3.1. *Suppose Assumptions C1-C4 hold. The average treatment effect estimator $\hat{\beta}$ is obtained by* (3.14)*. 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.*

3.2.3 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 (3.2). 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 its L^2-norm. 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 $p_\beta(X)$ be an isotonic estimator of the conditional mean $E[T(Z,\beta)|X]$, where T is a known function of data Z and the given parameter β. Let $\hat{p}_\beta(\cdot)$ be the isotonic estimator of $p_\beta(\cdot)$. Note $p_{\beta_0}(\cdot) = p_0(\cdot)$. 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 (3.2) can be

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left|\left| \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \beta, \hat{p}_\beta(X_i)) \right|\right|^2, \tag{3.15}$$

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 for each $\beta \in \mathcal{B}(\beta_0, \delta_0)$, $E(T(Z,\beta)|X = x) = p_\beta(x)$ is monotone increasing in x and differentiable in β. 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 3.2. *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 3.2, we do not need to solve the root of a discrete moment function. Therefore, the proof is similar to that of Lemma 3.1.

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

Proposition 3.2. ***(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* (3.15) *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 $m(z,\beta,p)$ equaling to $m(z,\beta,p_\beta)$:

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

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

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

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

Theorem 3.2. ***(Efficiency)*** *Assuming A1' - A9', $\hat{\beta}$ obtained by* (3.15) *is $\sqrt{n}$-consistent and efficient.*

3.3 Multi-dimensional X

The isotonic function is always a mapping from $\mathbb{R}$ to $\mathbb{R}$. 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.

3.3.1 Plug-in monotone single index Model

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

A1" X is a random vector taking value in the space $\mathcal{X} \subset \mathbb{R}^{k_\alpha}$. 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 $E(Y|X) = p_0(X) \equiv F_0(X'\alpha_0)$. For identification, α_0 is a k_α-dimensional vector normalized with $||\alpha_0|| = 1$.[3] We have

[3]In the 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 study our model without discussing those technical details. See BGH and Balabdaoui and Groeneboom (2020) for more details.

$\alpha_0 \in \mathcal{S}_{k_\alpha - 1}$, the unit $(k_\alpha - 1)$-dimensional sphere.

In this case, we need to estimate both p_0 and α_0 in the first step, then plug them into (3.2).

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

$$\hat{F}_\alpha = \arg\min_{F \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - F(X_i'\alpha)\}^2, \tag{3.16}$$

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

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} || \frac{1}{n} \sum_{i=1}^{n} X_i' \{Y_i - \hat{F}_\alpha(X_i'\alpha)\} ||^2. \tag{3.17}$$

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 3.2.3. In this case, we should replace Y_i in (3.16) by $T(Z_i, \beta)$

$$\hat{F}_{\alpha,\beta} = \arg\min_{F \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \{T(Z_i, \beta) - F(X_i'\alpha)\}^2, \tag{3.18}$$

where $T(Z_i, \beta)$ is differentiable in β. In the second step, we replace (3.17) with

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

Let k_β be the dimension of β and the moment condition m. 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. We denote $\theta = (\alpha', \beta')' \in \Theta \equiv \mathcal{S}_{k_\alpha - 1} \times \mathbb{R}^{k_\beta}$.

For fixed θ, we define $F_\theta(u) = F_{\alpha,\beta}(u) = E(T(Z,\beta)|\alpha' X = u)$. Let $F_\theta(\cdot) = F_\theta(\cdot'\alpha)$, and $F_0(\cdot) = F_{\theta_0}(\cdot'\alpha_0)$. The assumption A2 is adapted to the current setting:

A2" There exists $\delta_0 > 0$ that for each $\theta \in \mathcal{B}(\theta_0, \delta_0)$, the true mean function $u \mapsto E[T(Z,\beta)|X'\alpha = u]$ is monotone increasing in u and differentiable in θ. There exists $K_0 > 0$ such that $|F_0(\cdot)| < K_0$ for all $x \in \mathcal{X}$.

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

Lemma 3.3. *For fixed $\theta \in \mathcal{B}(\theta_0, \delta_0)$, $\hat{F}_\theta(\cdot)$ is obtained by solving* (3.18)*. $\delta(u)$ is a bounded function of u with a finite total variation. Under A1"-A3", we have $\frac{1}{n}\sum_{i=1}^n \delta(X_i'\alpha)(T(Z_i,\beta) - \hat{F}_\theta(X_i'\alpha)) = o_p(n^{-1/2})$.*

A4 is modified to

A4" For all $\theta \in \Theta$, $u \mapsto E[D(Z,\beta)|X'\alpha = u]$ is a bounded function of u with a finite total variation. There exist $c_1 > 0$ and $M_1 > 0$ such that for each row of $D(Z,\beta)$ ($D_j(Z,\beta)$ with $j \in \{1 : k_\beta\}$), $E[|D_j(Z,\beta)|^m|X = x] \leq m! M_1^{m-2} c_1$ for all integers $m \geq 2$ and almost every x.

Let A5" be the same as A5'. A6" to A7" are modified versions of A6 to A7:

A6" There are $b(z) > 0$ and $D(z,g)$ that (i) $||m(z,\beta,F_\theta) - m(z,\beta,F_0) - D(z,\beta,F_\theta - F_0)|| \leq b(z)||F_\theta - F_0||^2$; (ii) $E[b(Z)] = o_p(n^{1/6}(\log n)^{-2})$, for all $\theta \in \Theta$, where Θ is compact.

A7" There are ε, $b(z)$, $\tilde{b}(z) > 0$ and $F(\cdot)$ with $||F|| > 0$. Such that (i) for all $\theta \in \Theta$, $m(z,\beta,F_\theta)$ is continuous at θ and $m(z,\beta,F_\theta) \leq b(z)$; (ii) $||m(z,\beta,F) - m(z,\beta,F_\theta)|| \leq \tilde{b}(z)(||F - F_\theta||)^\varepsilon$.

Let $m_1(z,\beta,F_\theta) = x\,(T(z,\beta) - F_\theta(x'\alpha))$ and $m^*(z,\beta,F_\theta) = \begin{pmatrix} m(z,\beta,F_\theta) \\ m_1(z,\beta,F_\theta) \end{pmatrix}$.

Furthermore, we define

$$M_\alpha = -E\left\{[D(Z,\beta_0) - E(D(Z,\beta_0)|X'\alpha_0)]\{X - E[X|X'\alpha_0]\}'F_0^{(1)}(X'\alpha_0)\right\},$$
$$M_\beta = -E\left\{\frac{\partial m(Z,\beta_0,F_0(X_i'\alpha_0))}{\partial\beta} + E[D(Z,\beta_0)|X'\alpha_0]\frac{\partial T(Z,\beta_0)}{\partial\beta}\right\},$$
$$M_\theta = -E\left\{\frac{dm^*(Z,\beta_0,F_{\theta_0})}{d\theta}\right\}, \tag{3.19}$$
$$M(Z) = E(D(Z,\beta_0)|X'\alpha_0)(T(Z,\beta_0) - F_0(X'\alpha_0)), \tag{3.20}$$

and denote $M_{\alpha,1}$ as M_α corresponding to the moment function m_1. Then we have the modified A8 and A9:

A8" $E\{m^*(z,\beta,F_\theta)\} = 0$ has a unique solution on Θ at θ_0.

A9" For $\theta \in$ interior(Θ), (i) there are $\varepsilon > 0$ and a neighborhood $\mathscr{N}$ of β_0 such that for all $||F - F_0|| \leq \varepsilon$, $m(z,\beta,F)$ is differentiable in β on $\mathscr{N}$; (ii) M_β is nonsingular; (iii) $M_{\alpha,1}$ has rank $k_\alpha - 1$, and M_θ has rank $k_\alpha + k_\beta - 1$ (iv) $E[||m^*(Z,\beta,F_\theta)||^2] < \infty$; (v) Assumption A7 is satisfied with $m(z,\beta,p)$ equaling to each row of $\frac{dm^*(z,\beta,F_\theta)}{d\theta}$.

Note β in A9"(i) is only about the second argument in m, since $T(z,\beta)$ is assumed to be differentiable in β. Then we have

Theorem 3.3. *Suppose Assumptions A1"-A9" 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

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

where $M_{\alpha,1}^-$ is the Moore-Penrose inverse of $M_{\alpha,1}$, and A, B_1, and M_1 are defined in Appendix C.9.

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 3.3, 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

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

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

3.3.2 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

$$E(Y_i|\mathbf{X}_i) = c + m_0^1(X_i^1) + ...m_0^k(X_i^k). \tag{3.21}$$

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

$$\int_0^1 m_0^j(x^j)dx^j = 0 \text{ for } j = 1, 2, ..., k. \tag{3.22}$$

The least square estimator of 3.21 can be defined as the minimizer of

$$\underset{c\in\mathbb{R}^1,\{m^j\}_{j=1}^k\in M_0}{\arg\min} \sum_{i=1}^n \left[Y_i - c - \sum_{j=1}^k m^j(X_i^j) \right], \tag{3.23}$$

where M_0 denotes the class of monotone increasing function satisfying (3.22). We use $\{\hat{m}^j(\cdot)\}_{j=1}^k$ to denote the estimator from (3.23). 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 $\{\hat{m}^j(\cdot)\}_{j=1}^k$ 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 $\{y_i, \mathbf{x}_i\}_{i=1}^n$. To solve the problem (3.23), we can first solve the following problem

$$\min_{\mathrm{G}} \sum_{i=1}^n (y_i - \sum_{j=1}^k g_i^j)^2, \tag{3.24}$$

where G is a $n \times k$ 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 isotonic estimator $\mathbf{g}^1$ should satisfy $g_2^1 > g_1^1 > g_3^1$. Given G solving the problem (3.24), the value of the estimated monotone function $\hat{m}$ at 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 $g_{[r]}^j(\cdot)$ 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 $g_{[r]}^j(\cdot)$ 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 (3.1). As in Section 3.2.3, we can also allow the estimation of the additive monotone nuisance function to depend on β, i.e., we can replace Y_i by $T(Z_i, \beta)$.

Without loss of generality, 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 3.4. *For fixed β, $\hat{p}_\beta(X_i) \equiv \hat{c}_\beta + \sum_{j=1}^{k} \hat{m}_\beta(X_i^j)$ is an additive 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 (3.24) will converge to the least square estimator of this problem, with r growing to ∞.

Now let (i) A4$^{(3)}$ to A9$^{(3)}$ are the same as A4' 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

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

Theorem 3.4. *Assuming A1$^{(3)}$ - A9$^{(3)}$, for unconditional moment condition $E[m(Z,\beta_0,p_0(X))] = 0$, $\hat{p}_\beta(\cdot)$ is an additive isotonic estimator of the 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 (3.15) *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 1 continued: If we apply Theorem 3.4 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 sim-

plicity we set $D_i \in \mathbb{R}^1$ then we have

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

Then $V = \sigma^2 E[(D_i - E[D_i|X_i])^2]^{-1}$. This variance is larger than that 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.

3.4 Bootstrap inference

An 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 of $\hat{\beta}$.

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 (3.1).

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 3.5. *Let $\hat{\beta}^*$ be the bootstrap counterpart of $\hat{\beta}$ in Theorem 3.1, 3.2 or 3.3, 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*

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

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

3.5 Monte-Carlo Simulations

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

3.5.1 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 3.1. We also show the bootstrap validity under each setting.

3.5.1.1 Missing at random model

Example 4 (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 $\Omega = \text{Var}(E[Y|X])+E[\text{Var}(Y|X)/p_0(X)]$. (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

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

The Monte-Carlo simulation results are in Table 3.1:

Table 3.1: MAR model

n	$\hat{\mu}_\beta$	$n \cdot \hat{\sigma}_\beta^2$	n	$\hat{\mu}_\beta^*$	$n \cdot \hat{\sigma}_\beta^{2*}$
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 3.1 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, across the same set of sample sizes. $\hat{\mu}^*_\beta$ 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 looks 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.

3.5.1.2 Average treatment effect model

Example 4 (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

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

The Monte-Carlo simulation results are in Table 3.2:

Table 3.2: ATE model

n	$\hat{\mu}_\beta$	$n \cdot \hat{\sigma}^2_\beta$	n	$\hat{\mu}^*_\beta$	$n \cdot \hat{\sigma}^{2*}_\beta$
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 3.1. In general, Monte-Carlo averages and variances for both original and bootstrap samples converge to their theoretical limits.

Table 3.3: Bootstrap coverage rates

n	90% CI	95% CI
100	0.852	0.913
1000	0.885	0.942
2000	0.878	0.940
5000	0.893	0.947
∞	0.90	0.95

Table 3.3 shows the bootstrap coverage rates. We draw 2000 Monte-Carlo simulations, and for each simulation we draw 500 bootstrap samples. The coverage

rates are calculated with these 2000 sets of confidence intervals for both 90% and 95% confidence levels. From Table 3.3, we see clear trends that the bootstrap coverage rates are converging to their theoretical limits. Overall, the simulation outcomes for MAR, ATE, and bootstrap are in accordance with our theoretical results in the previous section.

3.5.2 Comparison with parametric plug-in estimators

3.5.2.1 With correctly specified parametric models

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:

$$\begin{aligned} Y &= X'\gamma_0 + T \cdot \beta_0 + \varepsilon, \\ T &= \begin{cases} 0 & \text{if } X'\alpha_0 < \nu \\ 1 & \text{if } X'\alpha_0 \geq \nu \end{cases}, \end{aligned} \tag{3.25}$$

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 3.4: ATE of the model (3.25) with plug-in probit and isotonic estimators

	probit			normalized probit			isotonic		
n	$\hat{\mu}_\beta$	$n \cdot \hat{\sigma}_\beta^2$	n·MSE	$\hat{\mu}_\beta$	$n \cdot \hat{\sigma}_\beta^2$	n·MSE	$\hat{\mu}_\beta$	$n \cdot \hat{\sigma}_\beta^2$	n·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 3.4 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's are scaled with n. In the left panel and the right panel, the ATE estimators $\hat{\beta}$ are calculated with (3.14), where the inversed propensity weights are not normalized. In the middle panel, we normalize the weights to unity, i.e.,

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

From Table 3.4, 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.

3.5.2.2 Robustness

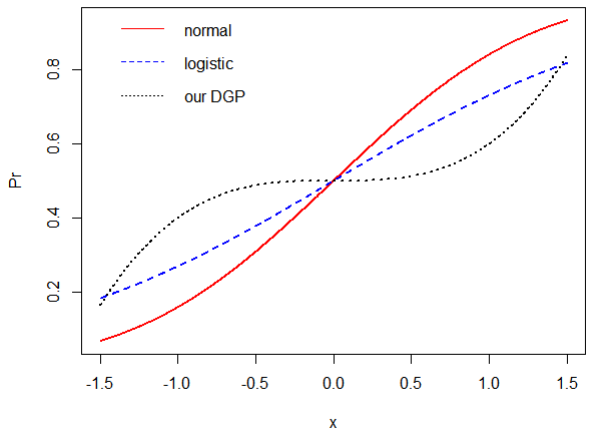
Compared to the popular choice of parametric models for propensity scores, such as the binary probit model or logit model, the proposed semiparametric estimator is robust to the model specification. Considering the following setting:

$$Y = X^3 \cdot \gamma_0 + T \cdot \beta_0 + \varepsilon \tag{3.26}$$

$$\text{with } Pr(T = 1|X = x) = x^3/10 + 0.5, \tag{3.27}$$

where $\varepsilon \sim N(0, 1)$ and is independent from X and T, $\gamma_0 = 1$, and $\beta_0 = 0.5$. $X \sim U[-1.5, 1.5]$. Figure 3.1 compares the function (3.27), the CDF of the standard normal distribution and the logistic function.

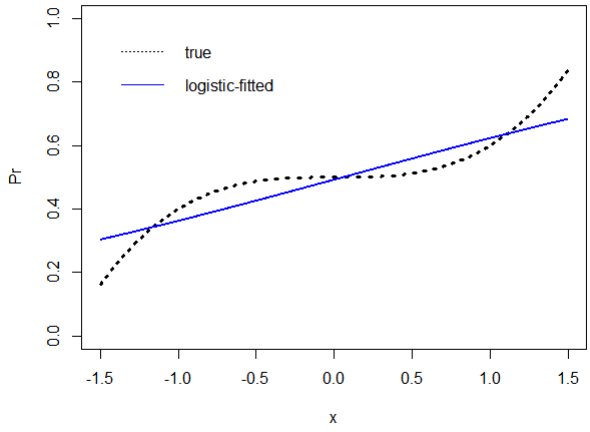
Figure 3.1: Normal CDF, logistic function, and the DGP (3.27)



The dotted black line is the DGP (3.27). The solid red line is the CDF of standard normal, $y = \Phi(a + x)$. The dashed blue line is the logistic function, $\Pr(T = 1|X = x) = \frac{\exp(a+bx)}{\exp(a+bx)+1}$. In this figure for both parametric models, $a = 0$ and $b = 1$. Three lines intersect at $[0, 1/2]$.

The idea of (3.27) 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 (3.27) 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 3.2

Figure 3.2: The function (3.27) fitted with logistic function



The dotted black line is the DGP (3.27). The dashed blue line is fitted with the logistic function, $y = \frac{\exp(a+bx)}{\exp(a+bx)+1}$.

While this line roughly fits the quasi-linear part of the function (3.27) (the part

around zero), the difference becomes large for $|x| > 1.2$. If the outcome y has large values far from zero, as the case in (3.26), we might have large estimation bias. Table 3.5 confirms this conjecture.

Table 3.5: ATE estimated with logistic and isotonic plug-in estimator

	logistic			isotonic		
n	$\hat{\mu}_\beta$	$n \cdot \hat{\sigma}^2_\beta$	n·MSE	$\hat{\mu}_\beta$	$n \cdot \hat{\sigma}^2_\beta$	n·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 3.5 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 $Pr(T = 1|X = x) = \frac{\exp(a+bx)}{\exp(a+bx)+1}$; 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. The right panel with isotonic plug-in estimators does not suffer from this issue and have stable performances across different sample sizes.

3.5.3 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 disappear at fast rates. Moreover, the smoothness conditions for the nuisance function can sometimes be demanding. For ATE estimators, Hirano, Imbens, and Ridder (2003) require that

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

Compared to our assumption:

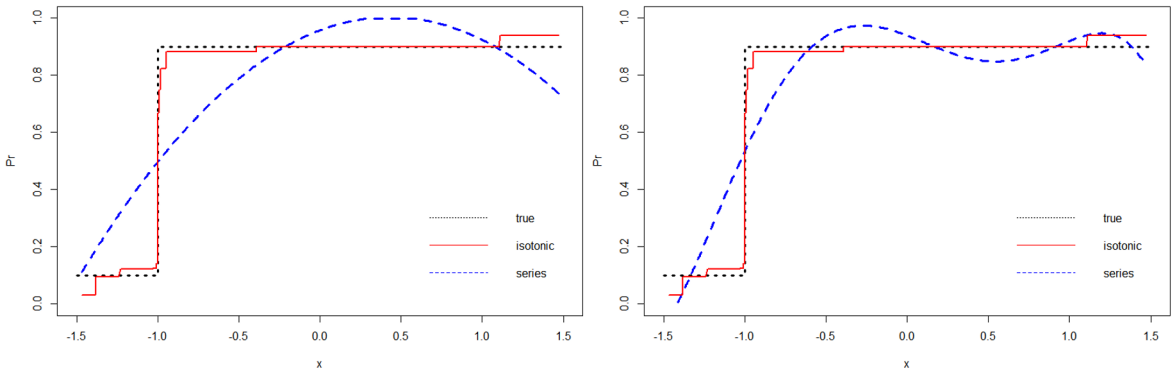
$$p_0(\text{x}) \text{ is monotone increasing.}$$

We even do not need continuity. Let's consider

$$\begin{aligned} Y &= X \cdot \gamma_0 + T \cdot \beta_0 + \varepsilon, \\ p_0(x) &= Pr(T = 1|X = x) = 0.1 + 0.8 \times 1(x > -1), \end{aligned} \tag{3.28}$$

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 (3.28) that $p_0(x)$ is a step probability function with a jump point at -1. Figure 3.3 describe $p_0(x)$ and curves fitted with series estimator and isotonic estimator.

Figure 3.3: The function (3.28) fitted with series estimators and isotonic estimators



The sample size $n = 1000$. The black dotted lines are the function (3.28). The blue dashed lines are series estimators. The red lines are 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 (3.28) very well, while isotonic estimators do good jobs.[4] The results are collected in Table 3.6. 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

[4]We acknowledge that parametric sigmoid-CDF-type link functions, such as Gaussian and Logistic functions, can also approximate the step function (3.28) well if the scale of the sigmoid shrinks to 0. However, we would like to point out that (i) in this subsection, we mainly focus on the comparison with non-parametric plug-in estimator; (ii) the parametric models might no longer work well if there are multiple jump points.

different series lengths ranging from 3 to 6.

Table 3.6: ATE estimated with series and isotonic plug-in estimator

	series								isotonic	
length	3		4		5		6		–	
n	$\hat{\mu}_\beta$	n·MSE	$\hat{\mu}_\beta$	n·MSE	$\hat{\mu}_\beta$	n·MSE	$\hat{\mu}_\beta$	n·MSE	$\hat{\mu}_\beta$	n·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 well, 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 right panel perform the best: MSE's are much lower, and with the sample size growing, biases are shrinking towards zero. Overall, Table 3.6 highlights two merits of the proposed 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 monotone nuisance functions.

3.6 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 (2005), and 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/).

3.6.1 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.

3.6.2 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 3.7, we compare these four estimators with those obtained by DW for the same dataset.

Table 3.7: 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 non-experimental estimators, the ATE and ATT estimators with isotonic plug-in estimators seem to be closer to the benchmark estimator in the first row. 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.

3.7 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 case, 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 characterize 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.

Appendix A

Proofs for Chapter 1

A.1 Proof of Theorem 1.1

Notation: We use the following notation. Let $||\mathbb{G}_n||_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\sqrt{n}(\mathbb{P}_n - P_0)f|$, $||\cdot||_{B,P_0}$ be the Bernstein norm under a measure P_0,

$$H_B(\varepsilon, \mathcal{F}, ||\cdot||_{B,P_0}) = \log N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||_{B,P_0}),$$

be the entropy of the ε-bracketing number of the function class $\mathcal{F}$ under $||\cdot||_{B,P_0}$, and

$$J_n(\delta) = J_n(\delta, \mathcal{F}, ||\cdot||_{B,P_0}) = \int_0^\delta \sqrt{1 + H_B(\varepsilon, \mathcal{F}, ||\cdot||_{B,P_0})}d\varepsilon.$$

A.1.1 Proof of existence and consistency

For fixed α and β (γ is also fixed by the uniqueness of reparameterization $\mathbb{S}(\cdot)$, so is θ). Let $\psi_\theta(u) = E[Y - X'\beta | Z'\alpha = u]$, which can be written as (by $E[\epsilon|Z] = 0$)

$$\psi_\theta(u) = E[\psi_0(Z'\alpha_0)|Z'\alpha = u] + (\beta_0 - \beta)'E[X|Z'\alpha = u]. \tag{A.1}$$

A similar argument to Theorem 5 of BGH implies that $\hat{\theta}$ exists with probability approaching one. We now show the consistency of $\hat{\theta}$. Since $\hat{\theta} = \hat{\theta}_n$ is estimated in a compact set, there exists a subsequence $\{\hat{\theta}_{n_k}\}_{k\in\mathbb{N}}$ of $\{\hat{\theta}_n\}_{n\in\mathbb{N}}$ almost surely

converging to some point $\theta^* = (\beta^{*\prime}, \gamma^{*\prime})'$. By Proposition 4 in BGH combined with $\hat{\theta}_{n_k} \xrightarrow{as} \theta^*$, we have

$$\int \left\{ \hat{\psi}_{n_k \hat{\theta}_{n_k}}(z'\mathbb{S}(\hat{\gamma}_{n_k})) - \psi_{\theta^*}(z'\mathbb{S}(\gamma_*)) \right\}^2 dP_0(z) \xrightarrow{p} 0.$$

Also by Proposition 9 in supplementary material of BGH (hereafter BGH-supp), the zero-crossing $\hat{\theta}$ becomes a root of the continuous limiting function, i.e.,

$$\phi_{n_k}(\hat{\theta}_{n_k}) \xrightarrow{p} \phi(\theta^*) = 0,$$

as $k \to \infty$, where $\phi(\theta) = \int \begin{pmatrix} x \\ J(\gamma)'z \end{pmatrix} \{y - x'\beta - \psi_\theta(z'\mathbb{S}(\gamma))\} dP_0(x, y, z)$, and the equality follows from the definition of zero-crossing and the continuity of $\psi_\theta(\cdot)$. Then we have

$$\begin{aligned}
0 &= (\theta_0 - \theta^*)'\phi(\theta^*) \\
&= (\theta_0 - \theta^*)' \int \begin{pmatrix} x \\ \mathbb{J}(\gamma^*)'z \end{pmatrix} \begin{Bmatrix} x'\beta_0 + \psi_0(z'\mathbb{S}(\gamma_0)) - x'\beta^* \\ -\{E[\psi_0(Z'\mathbb{S}(\gamma_0))|z'\mathbb{S}(\gamma^*)] + (\beta_0 - \beta^*)'E[X|z'\mathbb{S}(\gamma^*)]\} \end{Bmatrix} dP_0(x, z) \\
&= \begin{pmatrix} \beta_0 - \beta^* \\ \gamma_0 - \gamma^* \end{pmatrix}' \int \begin{pmatrix} x - E[X|z'\mathbb{S}(\gamma^*)] \\ \mathbb{J}(\gamma^*)'\{z - E[Z|z'\mathbb{S}(\gamma^*)]\} \end{pmatrix} \begin{Bmatrix} (\beta_0 - \beta^*)'\{x - E[X|z'\mathbb{S}(\gamma^*)]\} \\ +\psi_0(z'\mathbb{S}(\gamma_0)) - E[\psi_0(Z'\alpha_0)|z'\mathbb{S}(\gamma^*)] \end{Bmatrix} dP_0(x, z) \\
&= E\left[\mathrm{Cov}[(\beta_0 - \beta^*)'X + (\gamma_0 - \gamma^*)'\mathbb{J}(\gamma^*)'Z,\ (\beta_0 - \beta^*)'X + \psi_0(Z'\mathbb{S}(\gamma_0))|Z'\mathbb{S}(\gamma^*)]\right] \\
&= E\left[\mathrm{Cov}[(\beta_0 - \beta^*)'X + Z'(\mathbb{S}(\gamma_0) - \mathbb{S}(\gamma^*)) + o(\gamma_0 - \gamma^*),\ (\beta_0 - \beta^*)'X + \psi_0(Z'\mathbb{S}(\gamma_0))|Z'\mathbb{S}(\gamma^*)]\right] \\
&= E\left[\mathrm{Cov}[(\beta_0 - \beta^*)'X + Z'(\mathbb{S}(\gamma_0) - \mathbb{S}(\gamma^*)),\ (\beta_0 - \beta^*)'X + \psi_0(Z'\mathbb{S}(\gamma_0))|Z'\mathbb{S}(\gamma^*)]\right] + o(\gamma_0 - \gamma^*),
\end{aligned}$$

where the second equality follows from (A.1), the third equality follows from the law of iterated expectations, the fifth equality follows from an expansion of $\mathbb{S}(\gamma_0)$ around $\gamma_0 = \gamma_*$, and the last equality follows from A1. Therefore, by A6, $0 = (\theta_0 - \theta^*)'\phi(\theta^*)$ holds true only if $\theta^* = \theta_0$, and the consistency of $\hat{\theta}$ follows.

A.1.2 Proof of asymptotic normality

The proof is split into several steps.

Step 1: Derive a decomposition of $\phi_n(\hat{\theta})$

For each $\theta = (\beta', \gamma')'$, let $u_i = z_i'\mathbb{S}(\gamma)$ and $\{u_{n_j,\theta}\}_{j=1}^k$ be the subsequence of $\{u_i\}_{i=1}^n$ representing all the jump points of $\hat{\psi}_{n\theta}(\cdot)$. By the construction of $\hat{\psi}_{n\theta}(\cdot)$ (see, Lemmas 2.1 and 2.3 in Groeneboom and Jongbloed, 2014), we have $\sum_{i=n_j}^{n_{j+1}-1}\{y_i - x_i'\beta - \hat{\psi}_{n\theta}(u_i)\} = 0$ for each $j = 1, \ldots, k$, which means

$$\sum_{j=1}^{k} m_j \sum_{i=n_j}^{n_{j+1}-1} \{y_i - x_i'\beta - \hat{\psi}_{n\theta}(u_i)\} = 0, \tag{A.2}$$

for any weights $\{m_j\}_{j=1}^k$. As in BGH, we define for $W = X$ or Z,

$$\bar{E}_{n,\theta}[W|u] = \bar{E}_{n,\theta}[W|z'\mathbb{S}(\gamma)] = \begin{cases} E[W|Z'\mathbb{S}(\gamma) = u_{n_j}] & \text{if } \psi_\theta(u) > \hat{\psi}_{n\theta}(u_{n_j}) \text{ for all } u \in (u_{n_j}, u_{n_{j+1}}) \\ E[W|Z'\mathbb{S}(\gamma) = s] & \text{if } \psi_\theta(u) = \hat{\psi}_{n\theta}(s) \text{ for some } s \in (u_{n_j}, u_{n_{j+1}}) \\ E[W|Z'\mathbb{S}(\gamma) = u_{n_{j+1}}] & \text{if } \psi_\theta(u) < \hat{\psi}_{n\theta}(u_{n_j}) \text{ for all } u \in (u_{n_j}, u_{n_{j+1}}), \end{cases} \tag{A.3}$$

for $u \in [u_{n_j}, u_{n_{j+1}})$ with $j = 1, \ldots, k$ (if $j = k$, set $u_{n_{j+1}} = \max_i u_{n_i}$). By (A.2), it holds

$$\int \bar{E}_{n,\hat{\theta}}[W|z'\mathbb{S}(\gamma)]\{y - x'\hat{\beta} - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}d\mathbb{P}_n(x, y, z) = 0, \tag{A.4}$$

for $W = X$ and Z. Thus, $\phi_n(\hat{\theta})$ can be decomposed as

$$\begin{aligned} \phi_n(\hat{\theta}) &= T_n \int V_{I,n}^{x,z}\{y - x'\hat{\beta} - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\}d\mathbb{P}_n(x, y, z) \\ &+ T_n \int V_{II,n}^{x,z}\{y - x'\hat{\beta} - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\}d\mathbb{P}_n(x, y, z) \\ &= T_n(I + II), \end{aligned} \tag{A.5}$$

where $T_n = \begin{bmatrix} \mathbb{I}_k & 0 \\ 0 & \mathbb{J}(\hat{\gamma})' \end{bmatrix}$,

$$V_{I,n}^{x,z} = \begin{pmatrix} x - E[X|z'\mathbb{S}(\hat{\gamma})] \\ z - E[Z|z'\mathbb{S}(\hat{\gamma})] \end{pmatrix}, \quad V_{II,n}^{x,z} = \begin{pmatrix} E[X|z'\mathbb{S}(\hat{\gamma})] - \bar{E}_{n,\hat{\theta}}[X|z'\mathbb{S}(\hat{\gamma})] \\ E[Z|z'\mathbb{S}(\hat{\gamma})] - \bar{E}_{n,\hat{\theta}}[Z|z'\mathbb{S}(\hat{\gamma})] \end{pmatrix}.$$

Step 2: Show $II = o_p(n^{-1/2}) + o_p(\hat{\theta} - \theta_0)$

Note that the term II can be decomposed as

$$\begin{aligned} II &= \int V_{II,n}^{x,z}\{y - x'\hat{\beta} - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}d(\mathbb{P}_n - P_0)(x,y,z) \\ &\quad + \int V_{II,n}^{x,z}\{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}dP_0(x,y,z) + \int V_{II,n}^{x,z}\{\psi_{\hat{\theta}}(\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}dP_0(x,y,z) \\ &= II_a + II_b + II_c. \end{aligned}$$

First, we consider II_a. Note that Lemma 13 of BGH-supp and Lemma A.1 imply the following (A.6) and (A.7), with probability approaching one:

$$H_B(\varepsilon, \tilde{\mathcal{F}}_a, ||\cdot||_{B,P_0}) \leq \frac{C_1}{\varepsilon}, \tag{A.6}$$

for some $C_1 > 0$, where $\tilde{\mathcal{F}}_a = (C_2 \log n)^{-1}\mathcal{F}_a$ with some $C_2 > 0$ and $\mathcal{F}_a$ is defined in (A.29) below. Also, there exists a constant $C_3 > 0$ such that

$$||\tilde{f}||_{B,P_0} \leq C_3(\log n)n^{-1/3}, \tag{A.7}$$

for all $\tilde{f} \in \tilde{\mathcal{F}}_a$. Let $\delta_n = C_3(\log n)n^{-1/3}$ and $II_{a,j}$ be the j-th component of II_a. For any positive constants A and ν, there exist positive constants K_1, B_1, and B_2, such that $K = K_1 \log n$ and

$$\begin{aligned} & P\{|II_{a,j}| > An^{-1/2}\} = P\left\{|II_{a,j}| > An^{-1/2}, \sup_{\theta\in\mathcal{B}(\theta_0,\delta_0)}\sup_{z\in\mathcal{Z}}|\hat{\psi}_{n\theta}(z)| \leq K\right\} + \frac{\nu}{2} \\ \leq\ & P\left\{||\mathbb{G}_n||_{\mathcal{F}_a} > A, \sup_{\theta\in\mathcal{B}(\theta_0,\delta_0)}\sup_{z\in\mathcal{Z}}|\hat{\psi}_{n\theta}(z)| \leq K\right\} + \frac{\nu}{2} \\ \leq\ & \frac{E[||\mathbb{G}_n||_{\mathcal{F}_a} \quad |\sup_{\theta\in\mathcal{B}(\theta_0,\delta_0)}\sup_{z\in\mathcal{Z}}|\hat{\psi}_{n\theta}(z)| \leq K]}{A} + \frac{\nu}{2} \\ =\ & \frac{1}{AC_2\log n}E[||\mathbb{G}_n||_{\tilde{\mathcal{F}}_a} \quad |\sup_{\theta\in\mathcal{B}(\theta_0,\delta_0)}\sup_{z\in\mathcal{Z}}|\hat{\psi}_{n\theta}(z)| \leq K] + \frac{\nu}{2} \\ \lesssim\ & \frac{1}{AC_2\log n}J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\ \lesssim\ & \frac{\log n}{A}(\delta_n + 2B_1^{1/2}\delta_n^{1/2})\left(1 + \frac{\delta_n + 2B_1^{1/2}\delta_n^{1/2}}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\ \lesssim\ & \frac{1}{A}(\log n)^{3/2}n^{-1/6}\left(1 + \frac{B_2}{(\log n)^{3/2}}\right) + \frac{\nu}{2} \\ \lesssim\ & \nu, \end{aligned}$$

for all n large enough, where the first equality follows from Lemma 8 in BGH-supp, the first inequality follows from the definition of $\mathcal{F}_a$ (in (A.29)), the second inequality follows from the Markov inequality, the second equality follows from the definition of $\tilde{\mathcal{F}}_a$, the first wave inequality ($\lesssim$) follows from van der Vaart and Wellner (1996, Lemma 3.4.3) and the definition of δ_n, the second wave inequality follows from (A.6) 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,

$$II_a = o_p(n^{-1/2}). \tag{A.8}$$

Next, we consider II_b. Note that (see Lemma 17 in BGH-supp)

$$\frac{\partial}{\partial \alpha_j} E[\psi_0(Z'\alpha_0)|Z'\alpha = z'\alpha]\Big|_{\alpha=\alpha_0} = \{z_j - E[Z_j|Z'\alpha = z'\alpha_0]\}\psi_0'(z'\alpha_0), \tag{A.9}$$

for $j = 1, \ldots, d$. Using an expansion around $\hat{\gamma} = \gamma_0$ with (A.9) and $E[\psi_0(Z'\mathbb{S}(\gamma_0))|z'\mathbb{S}(\gamma_0)] = \psi_0(z'\mathbb{S}(\gamma_0))$, we have

$$E[\psi_0(Z'\mathbb{S}(\gamma_0))|z'\mathbb{S}(\hat{\gamma})] = \psi_0(z'\mathbb{S}(\gamma_0)) + (\hat{\gamma}-\gamma_0)'\mathbb{J}(\hat{\gamma})'\{z - E[Z|z'\mathbb{S}(\gamma_0)]\}\psi_0'(z'\mathbb{S}(\gamma_0)) + o_p(\hat{\gamma}-\gamma_0). \tag{A.10}$$

Then we have

$$\begin{aligned}
II_b &= \int V_{II,n}^{x,z} \left\{ \begin{array}{c} (\beta_0 - \hat{\beta})'\{x - E[X|z'\mathbb{S}(\hat{\gamma})]\} \\ +\psi_0(z'\mathbb{S}(\gamma_0)) - E[\psi_0(Z'\alpha_0)|z'\mathbb{S}(\hat{\gamma})] \end{array} \right\} dP_0(x,z) \\
&= \int V_{II,n}^{x,z} \left\{ \begin{array}{c} (\beta_0 - \hat{\beta})'\{x - E[X|z'\mathbb{S}(\hat{\gamma})]\} \\ -(\hat{\gamma} - \gamma_0)'\mathbb{J}(\gamma_0)'\{z - E[Z|z'\mathbb{S}(\gamma_0)]\}\psi_0'(z'\mathbb{S}(\gamma_0)) + o_p(\hat{\gamma} - \gamma_0) \end{array} \right\} dP_0(x,z) \\
&= -\int V_{II,n}^{x,z} \left(\begin{array}{c} x - E[X|z'\mathbb{S}(\gamma_0)] \\ \mathbb{J}(\gamma_0)'\{z - E[Z|z'\mathbb{S}(\gamma_0)]\}\psi_0'(z'\mathbb{S}(\gamma_0)) \end{array} \right)' dP_0(x,z) \left(\begin{array}{c} \hat{\beta} - \beta_0 \\ \hat{\gamma} - \gamma_0 \end{array} \right) + o_p(\hat{\gamma} - \gamma_0) \\
&= o_p(\hat{\theta} - \theta_0),
\end{aligned} \tag{A.11}$$

where the first equality follows from $E[\epsilon|X, Z] = 0$ and (A.1), the second equality follows from (A.10), and the last equality comes from $\int V_{II,n}^{x,z} dP_0(x,z) = o_p(1)$ and boundedness of the functions $x - E[X|z'\mathbb{S}(\gamma_0)]$ and $\mathbb{J}(\gamma_0)'\{\{z - E[Z|z'\mathbb{S}(\gamma_0)]\}\psi_0'(z'\mathbb{S}(\gamma_0))\}$.

Finally, we consider II_c. Since $E[W|z'\mathbb{S}(\gamma)]$ has totally bounded derivative for

$W = X$ and Z by A4, there exists $C_0 > 0$ such that

$$|E[W|Z'\mathbb{S}(\gamma) = u] - \bar{E}_{n,\theta}[W|Z'\mathbb{S}(\gamma) = u] \leq C_0|\psi_\theta(u) - \hat{\psi}_{n\theta}(u)|, \tag{A.12}$$

for each $\theta \in \mathcal{B}(\theta_0, \delta_0)$ and $u \in I_\alpha$. By this, we obtain

$$\begin{aligned} ||II_c|| &= || \int V_{II,n}^{x,z} \{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\} dP_0(x,z)|| \\ &\lesssim \int \{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}^2 dP_0(z) \\ &= O_p((\log)^2 n^{-2/3}) = o_p(n^{-1/2}), \end{aligned} \tag{A.13}$$

uniformly in $\theta \in \mathcal{B}(\theta_0, \delta_0)$, where the second equality follows from Proposition 4 in BGH. Combining (A.8), (A.11), and (A.13), we conclude that

$$II = o_p(n^{-1/2}) + o_p(\hat{\theta} - \theta_0).$$

Step 3: Decompose I

The term I can be decomposed as

$$\begin{aligned} I &= \int V_{I,n}^{x,z} \{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} dP_0(x,y,z) \\ &+ \int V_{I,n}^{x,z} \{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} d(\mathbb{P}_n - P_0)(x,y,z) \\ &\quad + \int V_{I,n}^{x,z} \{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\} d\mathbb{P}_n(x,y,z) \\ &= I_a + I_b + I_c. \end{aligned} \tag{A.14}$$

In the following steps, we show that

$$T_n I_a = -T_0 \int V_{x,z} V'_{x,z\psi'} dP_0(x,z) T_0'(\hat{\theta} - \theta_0) + o_p(\hat{\theta} - \theta_0), \tag{A.15}$$

$$\begin{aligned} T_n I_b &= T_0 \int V_{x,z} \{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) \\ &\quad + o_p(\hat{\theta} - \theta_0) + o_p(n^{-1/2}), \end{aligned} \tag{A.16}$$

$$I_c = o_p(n^{-1/2}). \tag{A.17}$$

Step 4: Show (A.15)

$$
\begin{aligned}
I_a &= \int V_{I,n}^{x,z} \left\{ \begin{array}{c} (\beta_0 - \hat{\beta})'\{x - E[X|z'\mathbb{S}(\hat{\gamma})]\} \\ +\psi_0(z'\mathbb{S}(\gamma_0)) - E[\psi_0(Z'\alpha_0)|z'\mathbb{S}(\hat{\gamma})] \end{array} \right\} dP_0(x,z) \\
&= \int V_{I,n}^{x,z} \left\{ \begin{array}{c} (\beta_0 - \hat{\beta})\{x - E[X|z'\mathbb{S}(\hat{\gamma})]\} \\ -(\hat{\gamma} - \gamma_0)'\mathbb{J}(\gamma_0)'\{z - E[Z|z'\mathbb{S}(\gamma_0)]\}\psi_0'(z'\mathbb{S}(\gamma_0)) + o_p(\hat{\gamma} - \gamma_0) \end{array} \right\} dP_0(x,z) \\
&= -\int V_{x,z} V_{x,z,\psi'}' dP_0(x,z) T_0' \left(\begin{array}{c} \hat{\beta} - \beta_0 \\ \hat{\gamma} - \gamma_0 \end{array} \right) + o_p(\hat{\gamma} - \gamma_0), \qquad \text{(A.18)}
\end{aligned}
$$

where the the first equality follows from $E[\epsilon|X,Z] = 0$ and (A.1), and some rearrangement, the second equality follows from (A.10), and the last equality follows from the definition of $V_{x,z,\psi'}$ and the fact that for $W = X$ or Z, we have $E[W|z'\mathbb{S}(\hat{\gamma})] - E[W|z'\mathbb{S}(\gamma_0)] = O_p(\hat{\gamma} - \gamma_0)$. Now, (A.15) follows by

$$T_n - T_0 = O_p(\hat{\gamma} - \gamma_0). \qquad \text{(A.19)}$$

Step 5: Show (A.16)

Decompose

$$
\begin{aligned}
T_n I_b &= T_n \int V_{I,n}^{x,z}\{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} d(\mathbb{P}_n - P_0)(x,y,z) \\
&= (T_n - T_0) \int V_{I,n}^{x,z}\{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} d(\mathbb{P}_n - P_0)(x,y,z) \\
&\quad + T_0 \int V_{I,n}^{x,z}\{x'\beta_0 - x'\hat{\beta} + \psi_0(z'\mathbb{S}(\gamma_0)) - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\} d(\mathbb{P}_n - P_0)(x,y,z) \\
&\quad + T_0 \int (V_{I,n}^{x,z} - V_{x,z})\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) \\
&\quad + T_0 \int V_{x,z}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) \\
&= (T_n - T_0) I_{b1} + T_0 I_{b2} + T_0 I_{b3} \\
&\quad + T_0 \int V_{x,z}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z).
\end{aligned}
$$

First, consider I_{b1}. Note that Lemma 13 BGH-supp and Lemma A.2 imply the following (A.20) and (A.21):

$$H_B(\varepsilon, \mathcal{F}_{b1}, ||\cdot||_{B,P_0}) \leq \frac{C_1}{\varepsilon}, \tag{A.20}$$

for some $C_1 > 0$, where $\mathcal{F}_{b1}$ is defined in (A.31). Also, there exists a constant $C_2 > 0$ such that

$$||f||_{B,P_0} \leq C_2, \tag{A.21}$$

for all $f \in \mathcal{F}_{b1}$. Let $I_{b1,j}$ be the j-th component of I_{b1}. For any $A > 0$, there exists a positive constant C such that

$$P\{|I_{b1,j}| > An^{-1/2}\} \leq \frac{1}{A} E[||\mathbb{G}_n||_{\mathcal{F}_{b1}}] \lesssim \frac{1}{A} J_n(C_2) \left(1 + \frac{J_n(C_2)}{\sqrt{n} C_2^2}\right) \lesssim \frac{C}{A},$$

for all n large enough, where the first inequality follows from the definition of $\mathcal{F}_{b1}$ and the Markov inequality, the first wave inequality follows from van der Vaart and Wellner (1996, Lemma 3.4.3), and the second wave inequality follows from (A.20), (A.21), and Equation (.2) in BGH. Thus, we have

$$I_{b1} = O_p(n^{-1/2}). \tag{A.22}$$

Next, consider I_{b2}. Let $I_{b2,j}$ be the j-th component of I_{b2}. For any positive constants A, ν, and η, there exist positive constants C', C_3, C_4, and C_5 such that

$$\begin{aligned} & P\{|I_{b2,j}| > An^{-1/2}\} \leq \frac{1}{A} E[||\mathbb{G}_n||_{\mathcal{F}_{b2}} | \mathfrak{B}_\eta] + \frac{\nu}{2} \lesssim \frac{1}{A} J_n(C'\eta) \left(1 + \frac{J_n(C'\eta)}{\sqrt{n}(C'\eta)^2} C_3\right) + \frac{\nu}{2} \\ \lesssim \; & \frac{1}{A} C_4 \eta^{1/2} \left(1 + \frac{C_5(1+\eta^{1/2})}{\sqrt{n}(C'\eta)^{3/2}} C_3\right) + \frac{\nu}{2}, \end{aligned} \tag{A.23}$$

for all n large enough, where the event $\mathfrak{B}_\eta$ is defined in Lemma A.3. The first inequality follows from Lemma A.3, the definition of $\mathcal{F}_{b2}$ in (A.33), and the Markov inequality, the first wave inequality follows from van der Vaart and Wellner (1996, Lemma 3.4.2) and Lemma A.3 (by choosing C' and η as therein), C_3 is a constant envelope of $\mathcal{F}_{b2}$, and the second wave inequality follows from Lemma A.3 and Equation (.2) in BGH-supp. Since we can choose η arbitrarily small, it holds

$$I_{b2} = o_p(n^{-1/2}). \tag{A.24}$$

Finally, consider I_{b3}. This is similar to the case of I_{b1} but with one difference, $V_{I,n}^{x,z} - V_{x,z} = o_p(1)$. Therefore we can use the same methods as for I_{b2} to find a upper bound of the L_2-norm (as we did in the proof of Lemma A.3 and (A.23).) Thus, we have

$$I_{b3} = o_p(n^{-1/2}). \tag{A.25}$$

Combining (A.22), (A.24), and (A.25) with (A.19), we obtain (A.16).

Step 6: Show (A.17)

Decompose

$$\begin{aligned} I_c &= \int V_{I,n}^{x,z}\{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}dP_0(x,y,z) \\ &\quad + \int V_{I,n}^{x,z}\{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}d(\mathbb{P}_n - P_0)(x,y,z) \\ &= I_{c1} + I_{c2}, \end{aligned}$$

For I_{c1}, the law of iterated expectation yields

$$I_{c1} = E\left[E\left[\left(\begin{array}{c} X - E[X|Z'\mathbb{S}(\hat{\gamma})] \\ Z - E[Z|Z'\mathbb{S}(\hat{\gamma})] \end{array}\right)\middle| Z'\mathbb{S}(\hat{\gamma})\right]\{\psi_{\hat{\theta}}(Z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(Z'\mathbb{S}(\hat{\gamma}))\}\right] = 0. \tag{A.26}$$

Now consider I_{c2}. For any positive constants A and ν, there exist positive constants C_1, C_2, and C' such that

$$\begin{aligned} P\{|I_{c2}| > An^{-1/2}\} &\leq \frac{C_1}{A}(\log n)^{1/2}\eta_n^{1/2}\left(1 + \frac{C_1(\log n)^{3/2}\eta_n^{1/2}}{\sqrt{n}\eta_n^2}\right) + \frac{\nu}{2} \\ &\leq \frac{C_2}{A}(\log n)n^{-1/6} + \frac{\nu}{2} \leq \nu, \end{aligned}$$

for all n large enough and $\eta_n = C'(\log n)n^{-1/3}$, where the first inequality follows by Lemma A.4 and a similar argument to (A.23), and the second inequality follows from the definition of η_n. Thus, we have $I_{c2} = o_p(n^{-1/2})$, and obtain (A.17).

Step 7: Conclusion

From Steps 1-6, we obtain

$$\begin{aligned} 0 &= \phi_n(\hat{\theta}) \\ &= -T_0 \int V_{x,z} V'_{x,z,\psi'} dP_0(x,z) T'_0(\hat{\theta} - \theta_0) \\ &\quad + T_0 \int V_{x,z}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) + o_p(n^{-1/2}) + o_p(\hat{\theta} - \theta_0). \end{aligned}$$

With B defined in A7, the central limit theorem implies

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta_0) &= \sqrt{n} B^{-1} T_0 \int V_{x,z}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) \\ &\quad + o_p(1 + \sqrt{n}(\hat{\theta} - \theta_0)) \qquad \text{(A.27)} \\ &\xrightarrow{d} N(0, \Pi). \end{aligned}$$

A.1.3 Lemmas

In this subsection, we use the following notations:

$$\begin{aligned} \mathcal{M}_{RK} &= \{\text{monotone non-decreasing functions on } [-R, R] \text{ and bounded by } K\}, \\ \mathcal{G}_{RK} &= \{g : g(z) = \psi_\theta(\alpha' z), z \in \mathcal{Z}, (\psi, \theta) \in \mathcal{M}_{RK} \times \mathcal{B}(\theta_0, \delta_0)\}, \\ \mathcal{D}_{RKv} &= \{d : d(z) = g_1(z) - g_2(z),\ (g_1, g_2) \in \mathcal{G}^2_{RK},\ ||d(z)||_{P_0} \le v\}, \\ \mathcal{H}_{RKv} &= \{h : h(\tilde{y}, z) = \tilde{y} d_1(z) - d_2(z),\ (d_1, d_2) \in \mathcal{D}^2_{RKv},\ (\tilde{y}, z) \in \mathbb{R} \times \mathcal{Z}\}. \qquad \text{(A.28)} \end{aligned}$$

A.1.3.1 Lemma for II_a

Let W_j be the j-th component of X or Z. Then decompose

$$\begin{aligned} & \{E[W_j|z'\mathbb{S}(\hat{\gamma})] - \bar{E}_{n,\hat{\theta}}[W_j|z'\mathbb{S}(\hat{\gamma})]\}\{y - x\hat{\beta} - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\} \\ = & \{E[W_j|z'\mathbb{S}(\hat{\gamma})] - \bar{E}_{n,\hat{\theta}}[W_j|z'\mathbb{S}(\hat{\gamma})]\}\{y - x\hat{\beta}\} \\ & - \{E[W_j|z'\mathbb{S}(\hat{\gamma})] - \bar{E}_{n,\hat{\theta}}[W_j|z'\mathbb{S}(\hat{\gamma})]\}\hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) \\ = & d_1(z)\{y - x\hat{\beta}\} - d_2(z). \end{aligned}$$

Let

$$\mathcal{F}_a = \left\{ f : f(x,y,z) = d_1(z)\{y - x\hat{\beta}\} - d_2(z),\ (x,y,z) \in \mathcal{X} \times \mathbb{R} \times \mathcal{Z} \right\}, \quad \text{(A.29)}$$

be a function class of the integrand of II_a. To control the term II_a, we use the following lemma.

Lemma A.1. *For some $K' \simeq \log n$ and positive constant v, it holds*

$$\mathcal{F}_a \subset \mathcal{H}_{RK'v},$$

with probability approaching one.

Proof. We use the following facts.

a) By A4, $E[W_j|z'\mathbb{S}(\hat{\gamma})]$ is a bounded function with a finite total variation.

b) $\bar{E}_{n,\hat{\theta}}[W_j|z'\mathbb{S}(\hat{\gamma})]$ is a discrete version of $E[W_j|z'\mathbb{S}(\hat{\gamma})]$ takes finite different values from it, so it is also bounded and has a finite total variation.

c) By Lemma 8 in BGH-supp, $\max_{\hat{\theta}\in\mathcal{B}(\theta_0,\delta_0)} \sup_{z\in\mathcal{Z}} |\hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))| = O_p(\log n)$. Thus, there exists $K = K_1 \log n$ such that $\hat{\psi}_{n\hat{\theta}} \in \mathcal{M}_{RK}$ with probability approaching to 1.

d) By Proposition 4 in BGH and (A.12), $||E[W_j|z'\mathbb{S}(\hat{\gamma})] - \bar{E}_{n,\hat{\theta}}[W_j|z'\mathbb{S}(\hat{\gamma})]||_2 \leq C_1(\log n)n^{-1/3}$ for some $C_1 > 0$.

e) The addition or multiplication of two functions with finite total variations is a function with a finite total variation.

Then by Jordan's decomposition and a), b), d), and e), there exist a positive constant C_0 larger than twice the bound of $E[W_j|z'\mathbb{S}(\hat{\gamma})]$ and $v_1 = C_1(\log n)n^{-1/3}$ such that

$$d_1(\cdot) \in \mathcal{D}_{RC_0v_1}, \quad \text{(A.30)}$$

with probability approaching 1. Additionally, c) and d) imply $d_2(\cdot) \in \mathcal{D}_{RK'v}$ with $K' = K_2 \log n$ for a large enough constant $K_2 > 0$ and $v = C_2(\log n)^2 n^{-1/3}$ for some $C_2 > 0$. Now, since $v_1 \lesssim v$ and $C_0 \lesssim K'$, setting $\tilde{y} = y - x\hat{\beta}$ in the definition of $\mathcal{H}_{RKv}$ in (A.28) yields the conclusion. □

A.1.3.2 Lemma for I_{b1}

Let W_j (and w_j) be the j-th component of X or Z (x or z), $\tilde{y} = y - x\hat{\beta}$ as in Lemma A.1, and

$$\mathcal{F}_{b1} = \left\{ f : f(w_j, y, z) = \{w_j - E[W_j|z'\mathbb{S}(\hat{\gamma})]\}\{\tilde{y} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\},\ (w_j, y, z) \in \mathcal{W}_j \times \mathbb{R} \times \mathcal{Z} \right\}, \tag{A.31}$$

be a function class of the j-th component of the integrand of I_{b1}. To control the term I_{b1}, we use the following lemma.

Lemma A.2. *For some positive constants C and v, it holds*

$$\mathcal{F}_{b1} \subset \mathcal{H}_{RCv},$$

with probability approaching 1.

Proof. We use the following facts.

a) w_j is bounded by $[-R, R]$.

b) By A4, $E[W_j|z'\mathbb{S}(\hat{\gamma})]$ is a function bounded by $[-R, R]$ and has a finite total variation.

c) By A1, A3, and (A.1), $\psi_{\hat{\theta}}$ is a bounded monotone function.

Let $d_1(z'\mathbb{S}(\hat{\gamma})) = E[W_j|z'\mathbb{S}(\hat{\gamma})]$ and $d_2(z'\mathbb{S}(\hat{\gamma})) = E[W_j|z'\mathbb{S}(\hat{\gamma})]\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))$. Any function in $\mathcal{F}_{b1}$ can be expressed as

$$\begin{aligned} & \{w_j - E[W_j|z'\mathbb{S}(\hat{\gamma})]\}\{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} \\ = \ & w_j\{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} + d_1(z'\mathbb{S}(\hat{\gamma}))(y - x'\hat{\beta}) - d_2(z'\mathbb{S}(\hat{\gamma})). \end{aligned} \tag{A.32}$$

By b) and c), we have

$$d_1(\cdot) \in \mathcal{D}_{RC_0v_1},$$

for C_0 defined in (A.30), which is larger than twice the bound of $E[W_j|z'\mathbb{S}(\hat{\gamma})]$, and some v_1, which is larger than the L_2-norm of a constant function R (the

upper bound in A1) on a compact support. Additionally, we have

$$d_2(\cdot) \in \mathcal{D}_{RC_1v_2},$$

for some positive constants C_1 and v_2. Therefore, by setting $\tilde{y} = y - x\hat{\beta}$ in the definition of $\mathcal{H}_{RKv}$ in (A.28), the second and third terms in (A.32) satisfy

$$d_1(z'\mathbb{S}(\hat{\gamma}))(y - x'\hat{\beta}) - d_2(z'\mathbb{S}(\hat{\gamma})) \in \mathcal{H}_{RC_1v_1}.$$

With similar steps we have:

$$w_j\{y - x'\hat{\beta} - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} \in \mathcal{H}_{RC_1'v_1'},$$

for some positive constants C_1' and v_1'. By choosing $C \geq \max(C_1, C_1')$ and $v \geq \max(v_1, v_1')$, the conclusion follows. □

A.1.3.3 Lemma for I_{b2}

Let

$$\mathcal{F}_{b2} = \Big\{ f : f(w_j, x, z) = \{w_j - E[W_j|z'\mathbb{S}(\hat{\gamma})]\}\{x'\beta_0 - x'\hat{\beta} + \psi_0(z'\mathbb{S}(\gamma_0) - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})\} \\ , \ (w_j, x, z) \in \mathcal{W}_j \times \mathcal{X} \times \mathcal{Z} \Big\}, \tag{A.33}$$

be a function class of the integrand of $I_{b2,j}$, the j-th component of I_{b2}. To control the term I_{b2}, we use the following lemma.

Lemma A.3.

For any positive constant η, we define the event $\mathfrak{B}_\eta$ as

$$\mathfrak{B}_\eta = \left\{ \sup_{x,z \in \mathcal{X}\times\mathcal{Z}, \hat{\theta} \in \mathcal{B}(\theta_0, \delta_0)} |x'\beta_0 - x'\hat{\beta} + \psi_0(z'\mathbb{S}(\gamma_0)) - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))| \leq \eta \right\}.$$

1. *For some $C > 0$, it holds $H_B(\varepsilon, \mathcal{F}_{b2}, ||\cdot||_{P_0}) \leq \frac{C}{\varepsilon}$.*
2. *For any positive constants ν and η, it holds $P(\mathfrak{B}_\eta) \geq 1 - \frac{\nu}{2}$ for all n large enough.*

3. *In case of the event* $\mathfrak{B}_\eta$*, there exists* $C' > 0$ *such that* $||f||_2 \leq C'\eta$ *for all* $f \in \mathcal{F}_{b2}$.

Proof. Both $E[W_j|z'\mathbb{S}(\hat{\gamma})]$ and $\psi_0(z'\mathbb{S}(\gamma_0)) - \psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))$ are bounded functions with finite total variations. Thus, they should have entropy of order $\frac{C_1}{\varepsilon}$ for some $C_1 > 0$. Also, both w_j and $(x'\beta_0 - x'\hat{\beta})$ are bounded. Thus, they should have entropy of order $\frac{C_2}{\varepsilon}$ for some $C_2 > 0$ (see, Example 19.7 in van der Vaart, 2000). Combining these results, the statement (1) follows. The consistency of $\hat{\theta}$ and Lemma 19 of BGH-supp imply the statement (2). The statement (3) follows from the definition of $\mathcal{F}_{b2}$. □

A.1.3.4 Lemma for I_{c2}

Let

$$\mathcal{F}_{c2} = \big\{ f : f(w_j, z) = \{w_j - E[W_j|z'\mathbb{S}(\hat{\gamma})]\}\{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\},\ (w_j, z) \in \mathcal{W}_j \times \mathcal{Z} \big\},$$

be a function class of the integrand of $I_{c2,j}$, the j-th component of I_{c2}. To control the term I_{c2}, we use the following lemma.

Lemma A.4.

1. *For some* $C > 0$*, it holds* $H_B(\varepsilon, \mathcal{F}_{c2}, ||\cdot||_{P_0}) \leq \frac{C \log n}{\varepsilon}$ *with probability approaching 1.*
2. *There exists a* $C' > 0$ *such that* $||f||_{P_0} \leq C'(\log n) n^{-1/3}$ *for all* $f \in \mathcal{F}_{c2}$.

Proof. We use the following facts.

a) w_j is bounded by $[-R, R]$.

b) By A4, $E[W_j|z'\mathbb{S}(\hat{\gamma})]$ is a function bounded by $[-R, R]$ and has a finite total variation.

c) By A1, A3, and (A.1), $\psi_{\hat{\theta}}$ is a bounded monotone function.

d) By Lemma 8 in BGH-supp, $\sup_{z\in\mathcal{Z}}|\hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))| = O_p(\log n)$. Therefore there exists $K = K_1 \log n$ such that $\hat{\psi}_{n\hat{\theta}} \in \mathcal{M}_{RK}$ with probability approaching to 1.

So, in the case that $\hat{\psi}_{n\hat{\theta}} \in \mathcal{M}_{RK}$:

1) $\{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}$ is bounded by $K + R$ with a finite variation.

2) $E[W_j|z'\mathbb{S}(\hat{\gamma})]\{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}$ is bounded by $R(K + R)$ with a finite variation, and the function class has an entropy of order $\frac{C_1 \log n}{\varepsilon}$ for some $C_1 > 0$.

3) From Lemma 10 of BGH-supp (by taking w_j as β in that lemma) and 1) above, the function class of $w_j\{\psi_{\hat{\theta}}(z'\mathbb{S}(\hat{\gamma})) - \hat{\psi}_{n\hat{\theta}}(z'\mathbb{S}(\hat{\gamma}))\}$ has an entropy of order $\frac{C_2 \log n}{\varepsilon}$ for some $C_2 > 0$.

From 2) and 3), the conclusion follows. □

A.2 Proof of Theorem 1.2

Existence and consistency of $\tilde{\theta}$ can be shown similarly as in Appendix A.1.1. The rest of the proof is split into several steps.

Step 1: Derive a decomposition of $\xi_{nh}(\tilde{\theta})$

In the same spirit of Step 1 of Appendix A.1.2, we introduce a piecewise constant function $\bar{\rho}_{n,\theta}$. Let $\{u_{n_j}\}_{j=1}^{k}$ be all the jump points of the monotone LSE $\hat{\psi}_{n\theta}(u)$.

We define for $u \in [u_{n_j}, u_{n_{j+1}})$ (if $j = k$, set $u_{n_{j+1}} = \max_i u_{n_i}$)

$$\bar{\rho}_{n,\theta}(W|u) = \bar{\rho}_{n,\theta}(W|Z'\mathbb{S}(\gamma))$$

$$= \begin{cases} \bar{\rho}_{n,\theta}(X|u) = \begin{cases} E[X|Z'\mathbb{S}(\gamma) = u_{n_j}] & \text{if } \psi_\theta(u) > \hat{\psi}_{n\theta}(u_{n_j}) \text{ for all } u \in (u_{n_j}, u_{n_{j+1}}), \\ E[X|Z'\mathbb{S}(\gamma) = s] & \text{if } \psi_\theta(u) = \hat{\psi}_{n\theta}(s) \text{ for some } s \in (u_{n_j}, u_{n_{j+1}}), \\ E[X|Z'\mathbb{S}(\gamma) = u_{n_{j+1}}] & \text{if } \psi_\theta(u) < \hat{\psi}_{n\theta}(u_{n_j}) \text{ for all } u \in (u_{n_j}, u_{n_{j+1}}), \end{cases} \\ \bar{\rho}_{n,\theta}(Z|u) = \begin{cases} E[Z|Z'\mathbb{S}(\gamma) = u_{n_j}]\psi'_\theta(u_{n_j}) & \text{if } \psi_\theta(u) > \hat{\psi}_{n\theta}(u_{n_j}) \text{ for all } u \in (u_{n_j}, u_{n_{j+1}}), \\ E[Z|Z'\mathbb{S}(\gamma) = s]\psi'_\theta(s) & \text{if } \psi_\theta(u) = \hat{\psi}_{n\theta}(s) \text{ for some } s \in (u_{n_j}, u_{n_{j+1}}), \\ E[Z|Z'\mathbb{S}(\gamma) = u_{n_{j+1}}]\psi'_\theta(u_{n_{j+1}}) & \text{if } \psi_\theta(u) < \hat{\psi}_{n\theta}(u_{n_j}) \text{ for all } u \in (u_{n_j}, u_{n_{j+1}}). \end{cases} \end{cases}$$

Similar to (A.12), we have for each $\theta \in \mathcal{B}(\theta_0, \delta_0)$

$$|E[Z|Z'\mathbb{S}(\gamma) = u]\psi'_\theta(u) - \bar{\rho}_{n,\theta}(Z|u)| \leq C_0|\psi_\theta(u) - \hat{\psi}_{n\theta}(u)|. \tag{A.34}$$

Similar to (A.4), we have

$$\int \bar{\rho}_{n,\tilde{\theta}}(W|z'\mathbb{S}(\gamma))\{y - x'\tilde{\beta} - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\}d\mathbb{P}_n(x,y,z) = 0,$$

for $W = X$ and Z. Thus, $\xi_{nh}(\tilde{\theta})$ can be decomposed as

$$\begin{aligned} \xi_{nh}(\tilde{\theta}) &= T_n \int V^{x,z}_{I,nh,\psi'}\{y - x'\tilde{\beta} - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\}d\mathbb{P}_n(x,y,z) \\ &\quad + T_n \int V^{x,z}_{II,n}\{y - x'\tilde{\beta} - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\}d\mathbb{P}_n(x,y,z) \\ &= T_n(I^E + II^E), \end{aligned}$$

where $T_n = \begin{bmatrix} \mathbb{I}_k & 0 \\ 0 & \mathbb{J}(\tilde{\gamma})' \end{bmatrix}$, and

$$V^{x,z}_{I,nh,\psi'} = \begin{pmatrix} x - E[X|z'\mathbb{S}(\tilde{\gamma})] \\ z\hat{\psi}'_{nh,\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - E[Z|z'\mathbb{S}(\tilde{\gamma})]\psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) \end{pmatrix}, \quad V^{x,z}_{I,n,\psi'} = \begin{pmatrix} x - E[X|z'\mathbb{S}(\tilde{\gamma})] \\ [z - E[Z|z'\mathbb{S}(\tilde{\gamma}))]\psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) \end{pmatrix},$$

$$V_{x,z,\psi'} = \begin{pmatrix} x - E[X|z'\mathbb{S}(\gamma_0)] \\ [z - E[Z|z'\mathbb{S}(\gamma_0))]\psi'_0(z'S(\gamma_0)) \end{pmatrix},$$

$$V^{x,z}_{II,n} = \begin{pmatrix} E[X|z'\mathbb{S}(\tilde{\gamma})] - \bar{\rho}_{n,\tilde{\theta}}(X|z'\mathbb{S}(\tilde{\gamma})) \\ E[Z|z'\mathbb{S}(\tilde{\gamma})]\psi'_\theta(z'\mathbb{S}(\tilde{\gamma}) - \bar{\rho}_{n,\tilde{\theta}}(Z|z'\mathbb{S}(\tilde{\gamma})) \end{pmatrix}.$$

Note: T_n and $V^{x,z}_{II,n}$ are redefined for $\tilde{\theta}$ in Appendix A.2.

Step 2: Show $II^E = o_p(n^{-1/2}) + o_p(\tilde{\theta} - \theta_0)$

Decompose

$$\begin{aligned} II^E &= \int V_{II,n}^{x,z}\{y - x'\tilde{\beta} - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\}d(\mathbb{P}_n - P_0)(x,y,z) \\ &\quad + \int V_{II,n}^{x,z}\{y - x'\tilde{\beta} - \psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\}dP_0(x,y,z) \\ &\quad + \int V_{II,n}^{x,z}\{\psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\}dP_0(x,y,z) \\ &= II_a^E + II_b^E + II_c^E. \end{aligned}$$

First, we consider II_a^E. By A8, $\psi'_\theta(z'\mathbb{S}(\tilde{\gamma}))$ is uniformly bounded with a bounded total variation. Therefore, $E[Z|z'\mathbb{S}(\tilde{\gamma})]\psi'_\theta(z'\mathbb{S}(\tilde{\gamma}))$ is also uniformly bounded with a bounded total variation, and all the arguments in Step 2 of Appendix A.1.2 can be applied to show $II_a^E = o_p(n^{-1/2})$.

Next, we consider II_b^E. For the redefined $V_{II,n}^{x,z}$, we still have $\int V_{II,n}^{x,z}dP_0(x,z) = o_p(1)$ and boundedness of the functions $x - E[X|z'\mathbb{S}(\gamma_0)]$ and $\mathbb{J}(\gamma_0)'\{\{z - E[Z|z'\mathbb{S}(\gamma_0)]\}\psi'_0(z'\mathbb{S}(\gamma_0))\}$. Thus the same argument as in in Step 2 of Appendix A.1.2 yields $II_b^E = o_p(\tilde{\theta} - \theta_0)$.

Finally, we consider II_c^E. By (A.12) and (A.34), the same argument in Step 2 of Appendix A.1.2 implies $II_c^E = o_p(n^{-1/2})$. Combining these results, we obtain $II^E = o_p(n^{-1/2}) + o_p(\tilde{\theta} - \theta_0)$.

Step 3: Decompose I^E

Note that

$$\begin{aligned} I^E &= T_n \int V_{I,nh,\psi'}^{x,z}\{y - x'\tilde{\beta} - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\}d\mathbb{P}_n(x,y,z) \\ &= \int V_{I,nh,\psi'}^{x,z}\{y - x'\tilde{\beta} - \psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\}dP_0(x,y,z) \\ &\quad + \int V_{I,nh,\psi'}^{x,z}\{y - x'\tilde{\beta} - \psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\}d(\mathbb{P}_n - P_0)(x,y,z) \\ &\quad + \int V_{I,nh,\psi'}^{x,z}\{\psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\}d\mathbb{P}_n(x,y,z) \\ &= I_a^E + I_b^E + I_c^E. \end{aligned}$$

In the following steps, we show that

$$
\begin{aligned}
T_n I_a^E &= -T_0 \int V_{x,z,\psi'} V'_{x,z\psi'} dP_0(x,z) T_0'(\tilde{\theta}-\theta_0) + o_p(\tilde{\theta}-\theta_0), && \text{(A.35)}\\
T_n I_b^E &= T_0 \int V_{x,z,\psi'} \{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) \\
&\quad + o_p(\tilde{\theta}-\theta_0) + o_p(n^{-1/2}), && \text{(A.36)}\\
I_c^E &= o_p(n^{-1/2}). && \text{(A.37)}
\end{aligned}
$$

Step 4: Show (A.35)

Decompose

$$
\begin{aligned}
I_a^E &= \int V_{I,n,\psi'}^{x,z} \{y - x'\tilde{\beta} - \psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\} dP_0(x,y,z) \\
&\quad + \int \begin{pmatrix} 0 \\ z[\hat{\psi'}_{nh,\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))] \end{pmatrix} \{y - x'\tilde{\beta} - \psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\} dP_0(x,y,z) \\
&= I_{a1}^E + I_{a2}^E.
\end{aligned}
$$

By a similar argument as in (A.18), we have

$$
I_{a1}^E = -\Big\{ \int V_{x,z,\psi'} V'_{x,z,\psi'} dP_0(x,z) \Big\} T_0'(\tilde{\theta}-\theta_0) + o_p(\tilde{\theta}-\theta_0).
$$

and

$$
I_{a2}^E = -\left\{ \int \begin{pmatrix} 0 \\ z\{\hat{\psi'}_{nh,\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\} \end{pmatrix} V'_{x,z,\psi'} dP_0(x,z) \right\} T_0'(\tilde{\theta}-\theta_0) + o_p(\tilde{\theta}-\theta_0).
$$

From $\hat{\psi'}_{nh,\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) = o_p(1)$, $V'_{x,z,\psi'} = O_p(1)$, and the compact supports of x and z, it holds $I_{a2}^E = o_p(\tilde{\theta}-\theta_0)$. Thus, we obtain (A.35).

Step 5: Show (A.36)

Decompose

$$\begin{aligned} T_n I_b^E &= T_n \int V_{I,n,\psi'}^{x,z} \{y - x'\tilde{\beta} - \psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\} d(\mathbb{P}_n - P_0)(x,y,z) \\ &\quad + T_n \int \begin{pmatrix} 0 \\ z\{\tilde{\psi}'_{nh,\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\} \end{pmatrix} \{y - x'\tilde{\beta} - \psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})\} d(\mathbb{P}_n - P_0)(x,y,z) \\ &= T_n I_{b1}^E + T_n I_{b2}^E. \end{aligned}$$

By similar steps as in Step 5 of Appendix A.1.2 combined with A8, we can derive

$$T_n I_{b1}^E = T_0 \int V_{x,z,\psi'} \{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) + o_p(\tilde{\theta} - \theta_0) + o_p(n^{-1/2}).$$

By Lemma 23 in BGH-supp, the analysis for $T_n I_{b2}^E$ is similar to the one for I_{b3} in Step 5 of Appendix A.1.2. Therefore, we have $T_n I_{b2}^E = o_p(n^{-1/2})$, and (A.36) is obtained.

Step 6: Show (A.37)

Decompose

$$\begin{aligned} I_c^E &= \int V_{I,nh,\psi'}^{x,z} \{\psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\} dP_0(x,y,z) \\ &\quad + \int V_{I,nh,\psi'}^{x,z} \{\psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\} d(\mathbb{P}_n - P_0)(x,y,z) \\ &= I_{c1}^E + I_{c2}^E. \end{aligned}$$

For I_{c1}^E, note that

$$\begin{aligned} I_{c1}^E &= \int V_{I,n,\psi'}^{x,z} \{\psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\} dP_0(x,y,z) \\ &\quad + \int \begin{pmatrix} 0 \\ z\{\hat{\psi}'_{nh,\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\} \end{pmatrix} \{\psi_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma})) - \hat{\psi}_{n\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))\} dP_0(x,y,z) \\ &= \int \begin{pmatrix} 0 \\ E[Z|u] \left\{ \frac{1}{h} \int K\left(\frac{u-x}{h}\right) d\hat{\psi}_{n\tilde{\theta}}(x) - \psi'_{\tilde{\theta}}(u) \right\} \end{pmatrix} \{\psi_{\tilde{\theta}}(u) - \hat{\psi}_{n\tilde{\theta}}(u)\} dP_0(u), \end{aligned} \tag{A.38}$$

where the last equality follows from a similar argument in (A.26), a change of variables $u = z'\mathbb{S}(\tilde{\gamma})$, and the definition of $\hat{\psi}_{nh,\tilde{\theta}}(u)$. We know $E[Z|u] = O(1)$ and $\int\{\psi_{\tilde{\theta}}(u) - \hat{\psi}_{n\tilde{\theta}}(u)\}^2 dP_0(u) = O_p((\log n)^2 n^{-2/3})$ by Proposition 4 in BGH. Also note that

$$\begin{aligned} & \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\hat{\psi}_{n\tilde{\theta}}(x) - \psi'_{\tilde{\theta}}(u) \\ = & \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d(\hat{\psi}_{n\tilde{\theta}}(x) - \psi_{\tilde{\theta}}(x)) + \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\psi_{\tilde{\theta}}(x) - \psi'_{\tilde{\theta}}(u) \\ = & -\frac{1}{h^2}\int K'\left(\frac{u-x}{h}\right)(\hat{\psi}_{n\tilde{\theta}}(x) - \psi_{\tilde{\theta}}(x))dx + \frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\psi_{\tilde{\theta}}(x) - \psi'_{\tilde{\theta}}(u), \end{aligned} \tag{A.39}$$

where the second equality follows from integration by parts and A9. With small h, $\frac{1}{h^2}\int K'\left(\frac{u-x}{h}\right)(\hat{\psi}_{n\tilde{\theta}}(x) - \psi_{\tilde{\theta}}(x))dx \sim \frac{1}{h}(\hat{\psi}_{n\tilde{\theta}}(u) - \psi_{\tilde{\theta}}(u))$. And $\frac{1}{h}\int K\left(\frac{u-x}{h}\right) d\psi_{\tilde{\theta}}(x) - \psi'_{\tilde{\theta}}(u)$ is a typical bias term of a kernel estimator, which is of order h^2 by A9. Plugging (A.39) into (A.38), the Cauchy-Schwarz inequality and A9 imply

$$I^E_{c1} = O_p((\log n)^2 n^{-2/3}) \cdot O_p(n^{1/7}) + O_p((\log n) n^{-1/3}) \cdot O_p(n^{-2/7}) = o_p(n^{-1/2}). \tag{A.40}$$

For I^E_{c2}, A8 and Lemma 23 in BGH-supp imply that both $z\hat{\psi}'_{nh,\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))$ and $E[Z|z'\mathbb{S}(\tilde{\gamma})]\psi'_{\tilde{\theta}}(z'\mathbb{S}(\tilde{\gamma}))$ are bounded with finite total variation. By a similar argument to Step 6 of Appendix A.1.2, we have $I^E_{c2} = o_p(n^{-1/2})$. Combined with (A.40), we obtain (A.37).

Step 7: Conclusion

From Steps 1-6 above, we obtain

$$\begin{aligned} 0 &= \xi_{nh}(\tilde{\theta}) \\ &= -T_0 \int V_{x,z,\psi'} V'_{x,z,\psi'} dP_0(x,z) T'_0(\tilde{\theta} - \theta_0) \\ &\quad + T_0 \int V_{x,z,\psi'}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\} d(\mathbb{P}_n - P_0)(x,y,z) + o_p(\tilde{\theta} - \theta_0) + o_p(n^{-1/2}). \end{aligned}$$

With B_E defined in A7, the central limit theorem implies

$$\begin{aligned}\sqrt{n}(\tilde{\theta}-\theta_0) &= \sqrt{n}B_E^{-1}T_0\int V_{x,z,\psi'}\{y-x'\beta_0-\psi_0(z'\mathbb{S}(\gamma_0)\}d(\mathbb{P}_n-P_0)(x,y,z)\\ &\quad +o_p(1+\sqrt{n}(\tilde{\theta}-\theta_0))\\ &\xrightarrow{d} N(0,\Pi_E).\end{aligned}$$

A.3 Proof of Theorem 1.3

Here we adapt the relevant proof in Groeneboom and Hendrickx (2017) (hereafter GH) to the monotone partially linear single index model. Let $\phi_n^*(\cdot)$ be the score function in the bootstrap sample. By definition (1.4),

$$\phi_n^*(\hat{\theta}^*)=\int\begin{pmatrix}x\\ \mathbb{J}(\hat{\gamma}^*)'z\end{pmatrix}\{y-x'\hat{\beta}^*-\hat{\psi}^*{}_{n\hat{\theta}^*}(z'\mathbb{S}(\hat{\gamma}^*))\}d\hat{\mathbb{P}}_n(x,y,z),$$

where $\hat{\mathbb{P}}_n$ is the empirical measure. Suppose

$$\begin{aligned}\phi_n^*(\hat{\theta}^*) &= -B(\hat{\theta}^*-\theta_0)+T_0\int V_{x,z}\{y-x'\beta_0-\psi_0(z'\mathbb{S}(\gamma_0)\}d(\hat{\mathbb{P}}_n-\mathbb{P}_n)(x,y,z) \qquad (A.41)\\ &\quad +T_0\int V_{x,z}\{y-x'\beta_0-\psi_0(z'\mathbb{S}(\gamma_0)\}d(\mathbb{P}_n-P_0)(x,y,z)+o_{P_M}(n^{-1/2}+(\hat{\theta}^*-\theta_0)).\end{aligned}$$

where P_M is defined in p. 3450 of GH. Then with $\phi_n^*(\hat{\theta}^*)=0$ and (A.27), we have

$$\begin{aligned}\sqrt{n}(\hat{\theta}^*-\hat{\theta}) &= \sqrt{n}B^{-1}T_0\int V_{x,z}\{y-x'\beta_0-\psi_0(z'\mathbb{S}(\gamma_0)\}d(\hat{\mathbb{P}}_n-\mathbb{P}_n)(x,y,z)\\ &\quad +o_{P_M}(1+\sqrt{n}(\hat{\theta}^*-\theta_0))\\ &\xrightarrow{d} N(0,\Pi),\end{aligned}$$

and the conclusion follows by Theorem 1.1.

It remains to prove (A.41). Similarly to Proposition 4 in BGH and (6.21) in GH, we can obtain the L^2-rate as

$$\sup_\theta\int\{\hat{\psi}^*{}_{n\theta}(z'\mathbb{S}(\gamma))-\psi_\theta(z'\mathbb{S}(\gamma)\}^2d\hat{\mathbb{P}}_n(x,y,z)=O_{P_M}(n^{-2/3}).$$

Define

$$T_n^* = \begin{bmatrix} \mathbb{I}_k & 0 \\ 0 & \mathbb{J}(\hat{\gamma})^{*\prime} \end{bmatrix}, \quad V_{I^*,n}^{x,z} = \begin{pmatrix} x - E[X|z'\mathbb{S}(\hat{\gamma}^*)] \\ z - E[Z|z'\mathbb{S}(\hat{\gamma}^*)] \end{pmatrix}, \quad V_{II^*,n}^{x,z} = \begin{pmatrix} E[X|z'\mathbb{S}(\hat{\gamma}^*)] - \bar{E}_{n,\hat{\theta}}[X|z'\mathbb{S}(\hat{\gamma}^*)] \\ E[Z|z'\mathbb{S}(\hat{\gamma}^*)] - \bar{E}_{n,\hat{\theta}}[Z|z'\mathbb{S}(\hat{\gamma}^*)] \end{pmatrix},$$

where $\bar{E}^*_{n,\theta}[W|u]$ is similarly defined as in (A.3). With similar arguments in Steps 1 and 2 in Section A.1.2, we can show that

$$\phi_n^*(\hat{\theta}^*) = T_n^* \int V_{I^*,n}^{x,z}\{y - x'\hat{\beta}^* - \hat{\psi^*}_{n\hat{\theta}^*}(z'\mathbb{S}(\hat{\gamma}^*))\}d\hat{\mathbb{P}}_n(x,y,z) + o_{P_M}(n^{-1/2} + (\hat{\theta}^* - \theta_0)). \tag{A.42}$$

For the first term of (A.42),

$$\begin{aligned} & T_n^* \int V_{I^*,n}^{x,z}\{y - x'\hat{\beta}^* - \hat{\psi^*}_{n\hat{\theta}^*}(z'\mathbb{S}(\hat{\gamma}^*))\}d\hat{\mathbb{P}}_n(x,y,z) \\ = \; & T_n^* \int V_{I^*,n}^{x,z}\{y - x'\hat{\beta}^* - \hat{\psi^*}_{n\hat{\theta}^*}(z'\mathbb{S}(\hat{\gamma}^*))\}d(\hat{\mathbb{P}}_n - \mathbb{P}_n)(x,y,z) \\ & + T_n^* \int V_{I^*,n}^{x,z}\{y - x'\hat{\beta}^* - \hat{\psi^*}_{n\hat{\theta}^*}(z'\mathbb{S}(\hat{\gamma}^*))\}d\mathbb{P}_n(x,y,z) \\ = \; & T_n^* I^* + T_n^* II^*. \end{aligned}$$

$T_n^* I^*$ is the bootstrap version of $T_n I_b$ in (A.16). Therefore, with a similar arguments in Step 5 of Section A.1.2, we have

$$T_n^* I^* = T_0 \int V_{x,z}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\}d(\hat{\mathbb{P}}_n - \mathbb{P}_n)(x,y,z) + o_{P_M}(n^{-1/2} + (\hat{\theta}^* - \theta_0)). \tag{A.43}$$

$T_n^* II^*$ is actually the first item of (A.5), $T_n I$, evaluated at $\hat{\theta}^*$. It can be decomposed as in (A.14). With similar argument from Step 3 to Step 6 in Section A.1.2, we have

$$\begin{aligned} T_n^* II^* \; = \; & -T_0 \int V_{x,z} V'_{x,z,\psi'} dP_0(x,z) T_0'(\hat{\theta}^* - \theta_0) \\ & + T_0 \int V_{x,z}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\}d(\mathbb{P}_n - P_0)(x,y,z) + o_P(n^{-1/2} + (\hat{\theta}^* - \theta_0)) \\ = \; & -B(\hat{\theta}^* - \theta_0) + T_0 \int V_{x,z}\{y - x'\beta_0 - \psi_0(z'\mathbb{S}(\gamma_0)\}d(\mathbb{P}_n - P_0)(x,y,z) \\ & + o_{P_M}(n^{-1/2} + (\hat{\theta}^* - \theta_0)), \end{aligned} \tag{A.44}$$

where the last equality follows from the definition of B and the fact that any item of order $o_P(n^{-1/2} + (\hat{\theta}^* - \theta_0))$ will be of order $o_{P_M}(n^{-1/2} + (\hat{\theta}^* - \theta_0))$.

Combining (A.42), (A.43), and (A.44), we have (A.41).

Appendix B

Proofs for Chapter 2

B.1 Proof of Theorem 2.1

Here we denote $\hat{g}_{0i} = \hat{g}_i(\beta_0)$, $\mathbb{S}_0 = \mathbb{S}(\beta_0)$, and $\mathbb{J}_0 = \mathbb{J}(\beta_0)$.

Note that (i) X has a bounded support (by Assumption A1), (ii) $\max|Y_i| = O_p(\log n)$ (by Assumption A2 and Lemma 7.1 of Balabdaoui, Durot and Jankowski, 2019), and
(iii) $\sup_{x\in\mathcal{X}} |\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)| = O_p(\log n)$ by Lemma 8 of the supplementary material of BGH (hereafter BGH-supp). Combining these results, it holds

$$\max_{1\le i\le n} |\hat{g}_{0i}| = O_p(\log n). \tag{B.1}$$

Thus, an expansion of (2.8) around $\hat{\lambda} = 0$ using the same argument in Owen (1991, proof of Theorem 2) based on (B.1) implies

$$\hat{\lambda} = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{g}_{0i}\hat{g}_{0i}{}'\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{g}_{0i} + o_p(n^{-1/2}). \tag{B.2}$$

A second-order expansion of (2.7) around $\hat{\lambda} = 0$ using (B.2) yields

$$\begin{aligned}\ell(\beta_0) &= 2\hat{\lambda}' \sum_{i=1}^{n} \hat{g}_{0i} - \hat{\lambda}' \left[\sum_{i=1}^{n} \hat{g}_{0i}\hat{g}_{0i}'\right] \hat{\lambda} + o_p(1) \\ &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_{0i}\right)' \left[\frac{1}{n} \sum_{i=1}^{n} \hat{g}_{0i}\hat{g}_{0i}'\right]^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_{0i}\right) + o_p(1).\end{aligned}$$

Then it is enough for the conclusion to show that

$$\frac{1}{n} \sum_{i=1}^{n} \hat{g}_{0i}\hat{g}_{0i}' \xrightarrow{p} V = \mathbb{J}_0' E[\epsilon^2 XX'] \mathbb{J}_0, \tag{B.3}$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{g}_{0i} \xrightarrow{d} N(0, \Sigma). \tag{B.4}$$

We first show (B.3). Decompose

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^{n} \hat{g}_{0i}\hat{g}_{0i}' &= \mathbb{J}_0' \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 X_i X_i'\right] \mathbb{J}_0 + \mathbb{J}_0' \left[\frac{1}{n} \sum_{i=1}^{n} \{\psi_0(X_i'\mathbb{S}_0) - \hat{\psi}_{\beta_0}(X_i'\mathbb{S}_0)\}^2 X_i X_i'\right] \mathbb{J}_0 \\ &\quad + \mathbb{J}_0' \left[\frac{2}{n} \sum_{i=1}^{n} \epsilon_i \{\psi_0(X_i'\mathbb{S}_0) - \hat{\psi}_{\beta_0}(X_i'\mathbb{S}_0)\} X_i X_i'\right] \mathbb{J}_0. \end{aligned} \tag{B.5}$$

By the law of large numbers, the first term of (B.5) converges to V; by Proposition 4 of BGH and Assumption A1, the second term converges to zero; by p.23 of BGH-supp and Assumption A1, the third term converges to zero. Combining these results, we obtain (B.3).

We now show (B.4). Let $\mathbb{P}_n$ be the empirical measure of $\{X_i, Y_i\}_{i=1}^n$, P_0 be the true measure of (X, Y), and

$$E[X|x'\mathbb{S}_0] = E[X|X'\mathbb{S}_0 = u] \text{ evaluated at } u = x'\mathbb{S}_0.$$

Decompose

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{g}_{0i} &= \mathbb{J}_0'\int x\{y-\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)\}d\mathbb{P}_n(y,x)\\ &= \mathbb{J}_0'\int\{x-E[X|x'\mathbb{S}_0]\}\{y-\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)\}d\mathbb{P}_n(y,x)\\ &\quad+\mathbb{J}_0'\int\{E[X|x'\mathbb{S}_0]-\bar{E}_n(x'\mathbb{S}_0)\}\{y-\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)\}d\mathbb{P}_n(y,x)\\ &\quad+\mathbb{J}_0'\int\bar{E}_n(x'\mathbb{S}_0)\{y-\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)\}d\mathbb{P}_n(y,x)\\ &= \mathbb{J}_0'(I+II+III),\end{aligned}$$

where

$$\bar{E}_n(u)=\begin{cases}E[X|x'\mathbb{S}_0=\tau_{i,\mathbb{S}_0}] & \text{if } \psi_0(u)>\hat{\psi}_{\beta_0}(u) \text{ for all } u\in(\tau_i,\tau_{i+1}),\\ E[X|x'\mathbb{S}_0=s] & \text{if } \psi_0(s)=\hat{\psi}_{\beta_0}(s) \text{ for some } s\in(\tau_i,\tau_{i+1}),\\ E[X|x'\mathbb{S}_0=\tau_{i+1,\mathbb{S}_0}] & \text{if } \psi_0(u)<\hat{\psi}_{\beta_0}(u) \text{ for all } u\in(\tau_i,\tau_{i+1}),\end{cases} \tag{B.6}$$

and $\tau_{i,\mathbb{S}_0}$ is the sequence of jump points of $\hat{\psi}_{\beta_0}$. By the definition of $\bar{E}_n(x'\mathbb{S}_0)$, it holds $III=0$ (see, (C.10) in BGH-supp).

For II, decompose

$$\begin{aligned}II &= \int\{E[X|x'\mathbb{S}_0]-\bar{E}_n(x'\mathbb{S}_0)\}\{y-\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)\}d(\mathbb{P}_n-P_0)(y,x)\\ &\quad+\int\{E[X|x'\mathbb{S}_0]-\bar{E}_n(x'\mathbb{S}_0)\}\{y-\psi_{\beta_0}(x'\mathbb{S}_0)\}dP_0(y,x)\\ &\quad+\int\{E[X|x'\mathbb{S}_0]-\bar{E}_n(x'\mathbb{S}_0)\}\{\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)-\psi_0(x'\mathbb{S}_0)\}dP_0(y,x)\\ &= II_a+II_b+II_c.\end{aligned} \tag{B.7}$$

The same argument as in pp. 19-20 of BGH-supp guarantees $II_a=o_p(n^{-1/2})$ and $II_b=o_p(n^{-1/2})$. For II_c, using (C.11) of BGH-supp and Proposition 4 of BGH, we have

$$\begin{aligned}\|II_c\| &\le C\int\{\hat{\psi}_{\beta_0}(x'\mathbb{S}_0)-\psi_0(x'\mathbb{S}_0)\}^2dP_0(y,x)\\ &= O_p((\log n)^2n^{-2/3})=o_p(n^{-1/2}),\end{aligned}$$

for some $C > 0$. Therefore, we obtain

$$II = o_p(n^{-1/2}). \tag{B.8}$$

For I, decompose

$$\begin{aligned} I &= \int \{x - E[X|x'\mathbb{S}_0]\}\{y - \psi_0(x'\mathbb{S}_0)\} d\mathbb{P}_n(y,x) \\ &\quad + \int \{x - E[X|x'\mathbb{S}_0]\}\{\psi_0(x'\mathbb{S}_0) - \hat{\psi}_{\beta_0}(x'\mathbb{S}_0)\} d\mathbb{P}_n(y,x) \\ &= I_a + I_b. \end{aligned}$$

From pp. 21-22 of BGH-supp, we can show that $I_b = o_p(n^{-1/2})$. Therefore,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \hat{g}_{0i} &= \mathbb{J}_0' \int \{x - E[X|x'\mathbb{S}_0]\}\{y - \psi_0(x'\mathbb{S}_0)\} d\mathbb{P}_n(y,x) + o_p(n^{-1/2}) \\ &= \mathbb{J}_0' \frac{1}{n}\sum_{i=1}^{n} \{X_i - E[X_i|X_i'\mathbb{S}_0]\}\epsilon_i + o_p(n^{-1/2}), \end{aligned} \tag{B.9}$$

and the central limit theorem implies (B.4). Therefore, the conclusion is obtained.

B.2 Proof of Theorem 2.2

Based on Hjort, McKeague and van Keilegom (2009), it is sufficient for the conclusion to show that

$$\bar{V} \xrightarrow{P_0} \mathbb{J}_0' E[\epsilon^2 XX'] \mathbb{J}_0, \tag{B.10}$$

$$\sqrt{n}\{M_n^*(\hat{\beta}) - M_n(\hat{\beta})\} \xrightarrow{d} N(0, \Sigma), \tag{B.11}$$

where $\hat{\beta}$ is obtained by solving (2.4). For the validity of bootstrap, we add the following assumptions.

A3 There exists $\delta_0 > 0$ such that the mapping $u \mapsto E[Y|X'\alpha = u]$ is monotone increasing on $I_\alpha = \{z'\alpha, z \in \mathcal{Z}\}$ for each $\alpha \in \mathcal{B}(\alpha_0, \delta_0)$.

A4 For all $\beta \neq \beta_0$ with $\mathbb{S}(\beta) \in \mathcal{B}(\alpha_0, \delta_0)$, $Cov\big[(\beta_0 - \beta)'\mathbb{J}(\beta)'X, \psi_0(\mathbb{S}(\beta_0)'X) | \mathbb{S}(\beta)'X\big] \neq 0$ almost surely.

A5 $\mathbb{J}_0' E[\psi_0^{(1)}(X'\alpha_0) Var(X|X'\alpha_0)]\mathbb{J}_0$ is non-singular.

By BGH, it can be shown that under A1-A5, $\hat{\beta}$ is consistent and $\sqrt{n}(\hat{\beta} - \beta_0)$ is asymptotically normal. Let $\psi_\beta(u) = E[Y|X'\mathbb{S}(\beta) = u]$. For (B.10), note that

$$\begin{aligned}
\bar{V} &= \mathbb{J}(\hat{\beta})' \left[\frac{1}{n} \sum_{i=1}^{n} X_i \{\epsilon_i + \psi_{\hat{\beta}}(X_i'\mathbb{S}(\hat{\beta})) - \hat{\psi}_{\hat{\beta}}(X_i'\mathbb{S}(\hat{\beta}))\}^2 X_i' \right] \mathbb{J}(\hat{\beta}) + o_p(1) \\
&= \{\mathbb{J}_0 + o_p(1)\}' \left\{ \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 X_i X_i' + o_p(1) \right\} \{\mathbb{J}_0 + o_p(1)\},
\end{aligned}$$

where the first equality follows from $\psi_0(x'\mathbb{S}(\beta_0)) - \psi_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta})) = O_p(\hat{\beta} - \beta_0)$ for almost every x (by p. 26 and Lemma 17 of BGH-supp) and the consistency of $\hat{\beta}$; the second equality follows from a combination of Proposition 4 of BGH, Assumption A1 and A3, p.23 of BGH-supp, and the consistency of $\hat{\beta}$. Thus, by the law of large numbers, we obtain (B.10).

We now prove (B.11). Note that $M_n^*(\hat{\beta}) - M_n(\hat{\beta}) = M_n^*(\hat{\beta})$ by (2.4). Let $\hat{\mathbb{P}}_n$ be the empirical measure of the bootstrap resample. Decompose

$$\begin{aligned}
M_n^*(\hat{\beta}) &= \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}(\hat{\beta}))\}\{y - \hat{\psi}_{\hat{\beta}}^*(x'\mathbb{S}(\hat{\beta}))\} d\hat{\mathbb{P}}_n \\
&\quad + \mathbb{J}(\hat{\beta})' \int \{E(X|x'\mathbb{S}(\hat{\beta})) - \bar{E}_n^*(x'\mathbb{S}(\hat{\beta}))\}\{y - \hat{\psi}_{\hat{\beta}}^*(x'\mathbb{S}(\hat{\beta}))\} d\hat{\mathbb{P}}_n \\
&\quad + \mathbb{J}(\hat{\beta})' \int \bar{E}_n^*(x'\mathbb{S}(\hat{\beta}))\{y - \hat{\psi}_{\hat{\beta}}^*(x'\mathbb{S}(\hat{\beta}))\} d\hat{\mathbb{P}}_n \\
&= I^* + II^* + III^*, \qquad \text{(B.12)}
\end{aligned}$$

where $\bar{E}_n^*(\cdot)$ is defined similarly to (B.6) with respect to $\hat{\psi}_{\hat{\beta}}^*$. Again, we have $III^* = 0$ by the definition of $\bar{E}_n^*(\cdot)$. For II^*, similar to (B.8) and p. 3481 of Groeneboom and Hendrickx (2017) (GH hereafter), we have $II^* = o_{P_M}(n^{-1/2})$, where P_M is defined in p. 3450 of GH.

For I^*, decompose

$$\begin{aligned} I^* &= \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}(\hat{\beta}))\}\{y - \hat{\psi}^*_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta}))\} d(\hat{\mathbb{P}}_n - \mathbb{P}_n) \\ &\quad + \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}(\hat{\beta}))\}\{y - \hat{\psi}_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta}))\} d\mathbb{P}_n \\ &\quad + \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}(\hat{\beta}))\}\{\hat{\psi}_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta})) - \hat{\psi}^*_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta}))\} d\mathbb{P}_n \\ &= I^*_a + I^*_b + I^*_c. \end{aligned}$$

For I^*_b, (2.4) and pp. 19-20 of BGH-supp combined with $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$ imply

$$\begin{aligned} I^*_b &= \mathbb{J}(\hat{\beta})' \int x\{y - \hat{\psi}_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta}))\} d\mathbb{P}_n \\ &\quad - \mathbb{J}(\hat{\beta})' \int \{E(X|x'\mathbb{S}(\hat{\beta})) - \bar{E}_n(x'\mathbb{S}(\hat{\beta}))\}\{y - \hat{\psi}_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta}))\} d\mathbb{P}_n \\ &= o_p(n^{-1/2}). \end{aligned}$$

For I^*_c, (6.21) in GH and pp. 21-22 of BGH-supp yield

$$\begin{aligned} I^*_c &= \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}(\hat{\beta}))\}\{\psi_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta})) - \hat{\psi}^*_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta}))\} d\mathbb{P}_n \\ &\quad + \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}(\hat{\beta}))\}\{\hat{\psi}_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta})) - \psi_{\hat{\beta}}(x'\mathbb{S}(\hat{\beta}))\} d\mathbb{P}_n \\ &= o_p(n^{-1/2}). \end{aligned}$$

Finally, for I^*_a, we have

$$\begin{aligned} I^*_a &= \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}_0)\}\{y - \psi_0(x'\mathbb{S}_0)\} d(\hat{\mathbb{P}}_n - \mathbb{P}_n) + o_{P_M}(n^{-1/2} + (\hat{\beta} - \beta_0)) \\ &= \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}_0)\}\epsilon d(\hat{\mathbb{P}}_n - \mathbb{P}_n) + o_{P_M}(n^{-1/2}), \end{aligned}$$

where the first equality follows from a similar argument to (6.25) in GH, and the second equality follows from a rearrangement and $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$. Combining these results, we have

$$M^*_n(\hat{\beta}) - M_n(\hat{\beta}) = \mathbb{J}(\hat{\beta})' \int \{x - E(X|x'\mathbb{S}_0)\}\epsilon d(\hat{\mathbb{P}}_n - \mathbb{P}_n) + o_{P_M}(n^{-1/2}).$$

Comparing this and (B.9), the central limit theorem yields (B.11). Therefore, the conclusion follows.

Appendix C

Proofs for Chapter 3

C.1 Proof of Lemma 3.1

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

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

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 (C.1), it holds

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

Thus, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\delta(X_i)(Y_i-\hat{p}(X_i)) \\
=&\int \delta(x)\{y-\hat{p}(x)\}d\mathbb{P}_n(z) \\
=&\int [\delta(x)-\bar{\delta}_n(x)](y-\hat{p}(x))d\mathbb{P}_n(z). \qquad \text{(C.2)}
\end{aligned}$$

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 (C.2) 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 $x\in\mathcal{X}$

$$|\delta(x)-\bar{\delta}_n(x)| \le C_0|p_0(x)-\hat{p}(x)|. \qquad \text{(C.3)}$$

For I, let us define the following function classes

$$\begin{aligned}
\mathcal{M}_{RK} &= \{\text{monotone increasing functions on } [-R,R] \text{ and bounded by } K\}, \\
\mathcal{G}_{RK} &= \{g: g(x)=p(x), x\in\mathcal{X}, p\in\mathcal{M}_{RK}\}, \\
\mathcal{D}_{RKv} &= \{d: d(x)=g_1(x)-g_2(x),\ (g_1,g_2)\in\mathcal{G}_{RK}^2,\ ||d(x)||_{P_0}\le v\}, \\
\mathcal{H}_{RKv} &= \{h: h(y,x)=yd_1(x)-d_2(x),\ (d_1,d_2)\in\mathcal{D}_{RKv}^2,\ z\in\mathcal{Z}\}. \qquad \text{(C.4)}
\end{aligned}$$

And we have the integrand of I

$$\begin{aligned}&[\delta(x)-\bar{\delta}_n(x)](y-\hat{p}(x))\\ =&[\delta(x)-\bar{\delta}_n(x)]y-[\delta(x)-\bar{\delta}_n(x)]\hat{p}(x).\end{aligned} \tag{C.5}$$

Let

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

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 (3.11) and (C.3), 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), (3.11), 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 $v_2=C_3(\log n)^2n^{-1/3}$ 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 (C.5)$\in$ $\mathcal{H}_{RKv}$.

Define $||\mathbb{G}_n||_{\mathcal{F}}=\sup_{f\in\mathcal{F}}|\sqrt{n}(\mathbb{P}_n-P_0)f|$. Let $N_{[]}(\varepsilon,\mathcal{F},||\cdot||)$ be the ε-bracketing number of the function class $\mathcal{F}$ under the norm $||\cdot||$, and

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

be the entropy of $N_{[]}(\varepsilon,\mathcal{F},||\cdot||)$, and

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

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:

$$H_B(\varepsilon, \tilde{\mathcal{F}}_a, ||\cdot||_{B,P_0}) \leq \frac{C_3}{\varepsilon}, \tag{C.6}$$

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

$$||\tilde{f}||_{B,P_0} \leq C_5(\log n)n^{-1/3}, \tag{C.7}$$

for all $\tilde{f}_a \in \tilde{\mathcal{F}}_a$, with probability approaching one. We use $\mathscr{E}$ to denote the event that both (C.6) and (C.7) 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

$$\begin{aligned}
P\{|I_j| > An^{-1/2}\} &\leq P\left\{|I_j| > An^{-1/2}, \mathscr{E}\right\} + P(\mathscr{E}^c) \\
&\leq P\left\{||\mathbb{G}_n||_{\mathcal{F}_a} > A, \mathscr{E}\right\} + \frac{\nu}{2} \\
&\leq \frac{E[||\mathbb{G}_n||_{\mathcal{F}_a} \quad |\mathscr{E}]}{A} + \frac{\nu}{2} \\
&= \frac{C_4 \log n}{A} E[||\mathbb{G}_n||_{\tilde{\mathcal{F}}_a} \quad |\mathscr{E}] + \frac{\nu}{2} \\
&\lesssim \frac{C_4 \log n}{A} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A}(\delta_n + 2B_1^{1/2}\delta_n^{1/2})\left(1 + \frac{\delta_n + 2B_1^{1/2}\delta_n^{1/2}}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_2}{(\log n)^{3/2}}\right) + \frac{\nu}{2} \\
&\lesssim \nu,
\end{aligned} \tag{C.8}$$

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 (C.6) 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,

$$I = o_p(n^{-1/2}). \tag{C.9}$$

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 (C.3), the second equality follows from (3.11).

Combining the rates for I, II, and III, the conclusion follows.

C.2 Proof of Proposition 3.1

Under A1-A4 and Lemma 3.1, 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

$$\frac{1}{n}\sum_{i=1}^n m(z_i, \beta, \hat{p}(\cdot)) = 0 \tag{C.10}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^n \{m(z_i, \beta, \hat{p}(\cdot)) + E[D(Z, \beta_0)|X_i](Y_i - \hat{p}(X_i))\} = o_p(n^{-1/2}). \tag{C.11}$$

Let $\hat{\beta}$ be the solution of (C.10), 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 (C.11), the difference of $\sqrt{n}(\hat{\beta}-\beta_0)$ and $\sqrt{n}(\tilde{\beta}-\beta_0)$ is $o_p(1)$.

C.3 Proof of Theorem 3.1

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 Assumptions 5.1-5.6 of Newey (1994).

Step 1: Verify Assumption 5.1 of 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., Theorem 9.2 and Lemma 5.15 in van de Geer, S., 2000).

Step 2: Verify Assumption 5.2 of 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

$$\frac{1}{n}\sum_{i=1}^{n} D(Z,\beta_0,\hat{p}(x)-p_0(x)) - \int D(z,\beta_0,\hat{p}(x)-p_0(x))d\mathbb{P}_0(z)$$
$$= \int D(z,\beta_0)(p_0(x)-\hat{p}(x))d(\mathbb{P}_n-\mathbb{P}_0)(z) \quad \text{(C.12)}$$

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 subsection, such as A_i, C_i, K_i, δ_n, and v, etc.. They are not related to those constants with the same names in other sections.

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

$$p_0(x) - \hat{p}(x) \in \mathcal{D}_{R(C_1 \log n)\left(C_2 n^{-1/3}\log n\right)}, \quad \text{(C.13)}$$

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

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

for some $A > 0$. Now we define

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

Now we let $D(z,\beta_0) \in \mathbb{R}^1$. This is just to simplify the notation of the following proof, i.e., the following steps hold for any $D_j(z,\beta_0)$ with $j \in \{1:k\}$, the j-th row of $D(z,\beta_0)$.

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

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

and

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

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

$$H_B(\varepsilon, \mathcal{H}^{(2)}_{RCv}, ||\cdot||_{P_0}) \leq \frac{\tilde{A}C}{\varepsilon}. \tag{C.14}$$

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

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

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

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

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 (C.13)) 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

$$
||\frac{h}{H}||_{B,P_0} \lesssim \frac{v}{H}. \tag{C.15}
$$

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 (C.14) and (C.15), we have with probability approaching one

$$
H_B(\varepsilon, \tilde{\mathcal{F}}_b, ||\cdot||_{B,P_0}) \leq \frac{C_3}{\varepsilon}, \tag{C.16}
$$

for some $C_3 > 0$, and

$$
\text{and } ||\tilde{f}||_{B,P_0} \leq C_4 n^{-1/3}, \tag{C.17}
$$

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 (C.16) and (C.17), and use S to denote the value of (C.12). Let $\delta_n = C_4 n^{-1/3}$. Now For any $A_2 > 0$.

$$\begin{aligned}
P\{|S| > A_2 n^{-1/2}\} &\leq P\left\{|S| > A_2 n^{-1/2}, \mathscr{E}_1\right\} + P(\mathscr{E}_1^c) \\
&\leq P\left\{||\mathbb{G}_n||_{\mathcal{F}_b} > A_2, \mathscr{E}_1\right\} + \frac{\nu}{2} \\
&\leq \frac{E[||\mathbb{G}_n||_{\mathcal{F}_b} \quad |\mathscr{E}_1]}{A_2} + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2} E[||\mathbb{G}_n||_{\tilde{\mathcal{F}}_b} \quad |\mathscr{E}_1] + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2} J_n(\delta_n)\left(1 + \frac{J_n(\delta_n)}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2}(\delta_n + 2B_1^{1/2}\delta_n^{1/2})\left(1 + \frac{\delta_n + 2B_1^{1/2}\delta_n^{1/2}}{\sqrt{n}\delta_n^2}\right) + \frac{\nu}{2} \\
&\lesssim \frac{1}{A}(\log n)^{3/2} n^{-1/6}\left(1 + \frac{B_2}{(\log n)^{3/2}}\right) + \frac{\nu}{2} \\
&\lesssim \frac{\log n}{A_2} n^{-1/6} B_2 + \frac{\nu}{2} \\
&\lesssim \nu, \qquad \text{(C.18)}
\end{aligned}$$

Each steps are similar to those of (C.8). Thus, we have $\int D(z, \beta_0)(p_0(x) - \hat{p}(x))d(\mathbb{P}_n - \mathbb{P}_0)(z) = o_p(n^{-1/2})$, and Newey's Assumption 5.2 is satisfied.

Assumption 5.3 (Newey, 1994):
$\int D(z, \hat{p}(x) - p_0(x))d\mathbb{P}_0(z) = \frac{1}{n}\sum_{i=1}^n M(z_i) + o_p(n^{-1/2})$. [1]

We have

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

The first equality follows from A5. In the last equality, we set $E(D(Z, \beta_0)|X = x) = \delta(x)$.

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

Therefore, by plugging in $M(z) = \delta(x)u$

$$\begin{aligned}
&\int D(z, \hat{p}(x) - p_0(x))d\mathbb{P}_0(z) - \frac{1}{n}\sum_{i=1}^{n} M(Z_i) \\
&= \int \delta(x)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x) - \frac{1}{n}\sum_{i=1}^{n} \delta(X_i)(Y_i - p_0(X_i)) \\
&= \int \delta(x)(\hat{p}(x) - p_0(x))d\mathbb{P}_0(x) - \int \delta(x)(y - \hat{p}(x) + \hat{p}(x) - p_0(x))d\mathbb{P}_n(z) \\
&= \int -\delta(x)(y - \hat{p}(x))d\mathbb{P}_n(z) + \int -\delta(x)(\hat{p}(x) - p_0(x))d(\mathbb{P}_n - \mathbb{P}_0)(x) \\
&= I + II. \qquad \text{(C.19)}
\end{aligned}$$

By Lemma 3.1, 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 $II = o_p(n^{-1/2})$. Thus, Newey's Assumption 5.3 is satisfied.

Newey's Assumptions 5.4 to 5.6 are adapted as A7 to A9 in this paper. Then the consistency can be proved by similar arguments as in 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).

C.4 Proof of Corollary 3.1

Let us check A1 to A9 of Theorem 3.1 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 $T \in \{0, 1\}$. (Y in A3 is T in Corollary 3.1). 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 3.1 satisfied. The asymptotic variance matrix Ω can be obtained in the same way as pp.34-35 of Hirano, Imbens, and Ridder (2000).

C.5 Proof of Lemma 3.2.

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 3.1 in Appendix C.1, with Y replaced by $T(Z,\beta)$.

C.6 Proof of Proposition 3.2.

The proof is similar to that of Proposition 3.1 in Appendix C.2.

C.7 Proof of Theorem 3.2

Here we might not be able to solve the sample moment condition (3.2)

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

as we did in Theorem 3.1, since changing β might change the left-hand side discretely.

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

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta, \hat{p}_\beta(X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_\beta(X_i)) + D(Z, \beta)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_\beta(X_i)) + D(Z, \beta)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z, \beta)|X_i)(T(Z_i, \beta) - \hat{p}_\beta(X_i)) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_\beta(X_i)) + D(Z, \beta)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} \{E(D(Z, \beta)|X_i)(T(Z_i, \beta) - p_\beta(X_i)) + E(D(Z, \beta)|X_i)[(\hat{p}_\beta(X_i) - p_\beta(X_i)]\} \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_\beta(X_i)) + E(D(Z, \beta)|X_i)(T(Z_i, \beta) - p_\beta(X_i))\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} [D(Z, \beta) - E(D(Z, \beta)|X_i)][(\hat{p}_\beta(X_i) - p_\beta(X_i)] + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta, p_\beta(X_i)) + E(D(Z, \beta)|X_i)(T(Z_i, \beta) - p_\beta(X_i))\} + o_p(n^{-1/2}).
\end{aligned}
\tag{C.20}
$$

The first equality follows from A5' and A6'. The second equality follows from Lemma 3.2. 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 (C.20) and the definition of $\hat{\beta}$ in (3.15), 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,

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

does not depend on the discrete estimator $\hat{p}(\cdot)$. It is a smooth moment function of β. Thus, under A9', with n large enough, we have

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

and by (C.20) we have

$$||\frac{1}{n}\sum_{i=1}^{n}m(Z_i,\hat{\beta},\hat{p}_{\hat{\beta}}(X_i))||=o_p(n^{-1/2}). \tag{C.21}$$

Let

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

Note we also have

$$M_\beta=-E\left(\frac{dm(Z,\beta,p_\beta)}{d\beta}|_{\beta=\beta_0}\right),$$

since

$$\begin{aligned}
&\frac{dm(Z,\beta,p_\beta)}{d\beta}|_{\beta=\beta_0}\\
&=\lim_{\beta\to\beta_0}\left(\frac{\partial m(Z,\beta_0,p_0)}{\partial\beta}\frac{\beta-\beta_0}{\beta-\beta_0}+\frac{\partial m(Z,\beta_0,p_0)}{\partial p}\frac{p_\beta(X)-p_0(X)}{\beta-\beta_0}+o_p(\beta-\beta_0)\right)\\
&=\frac{\partial m(Z,\beta_0,p_0)}{\partial\beta}+D(Z,\beta_0)p'_{\beta_0}(X),
\end{aligned} \tag{C.22}$$

where $p'_{\beta_0}(x)=\frac{d\left(p_\beta(x)\right)}{d\beta}$. Its existence is by A2'.

Then

$$
\begin{aligned}
E\left(\frac{dm(Z,\beta,p_\beta)}{d\beta}|_{\beta=\beta_0}\right) &= E\left(\frac{\partial m(Z,\beta_0,p_0)}{\partial\beta} + D(Z,\beta_0)p'_{\beta_0}(X)\right) \\
&= E\left(\frac{\partial m(Z,\beta_0,p_0)}{\partial\beta}\right) + E\left\{E\left[D(Z,\beta_0)|X\right]p'_{\beta_0}(X)\right\} \\
&= E\left(\frac{\partial m(Z,\beta_0,p_0)}{\partial\beta}\right) + E\left(E(D(Z,\beta_0)|X)\frac{\partial T(Z,\beta)}{\partial\beta}\right),
\end{aligned}
$$

where the third equality follows from the definition of $p_\beta(X)$ and Law of iterated expectation: from definition $p_\beta(X) = E\left[T(z,\beta)|X\right]$, we have $E\left[T(z,\beta) - p_\beta(X)|X\right] = 0$, then $E\left(\frac{\partial T(z,\beta_0)}{\partial\beta}|X\right) = E\left(p'_{\beta_0}(X)|X\right)$.

Now we have

$$
\begin{aligned}
o_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\hat{\beta},\hat{p}_{\hat{\beta}}(X_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\hat{\beta},\hat{p}_{\hat{\beta}}(X_i)) + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z,\beta_0)|X_i)(T(Z_i,\hat{\beta}) - \hat{p}_{\hat{\beta}}(X_i)) \\
&= -M_{n,\beta}(\hat{\beta}-\beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,\hat{p}_{\hat{\beta}}(X_i)) + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z,\beta_0)|X_i)(T(Z_i,\beta_0) - \hat{p}_{\hat{\beta}}(X_i)) + o_p(\hat{\beta}-\beta_0) \\
&= -M_\beta(\hat{\beta}-\beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,p_0(X_i)) + o_p(n^{-1/2}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z_i,\beta_0)|X_i)(T(Z_i,\beta_0) - p_0(X_i)) + o_p(\hat{\beta}-\beta_0) \\
&= -M_\beta(\hat{\beta}-\beta_0) + \left\{\frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,p_0(X_i)) + M(Z_i)\right\} \\
&+ o_p(n^{-1/2} + (\hat{\beta}-\beta_0)).
\end{aligned} \tag{C.23}
$$

The first equality follows from (C.21). The second equality follows from Lemma 3.2. 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 C.3. The last equality follows from the definition of $M(Z)$.

With Assumption A7' and A8', the consistency of $\hat{\beta}$ can be proved by similar arguments as in Lemma 5.2 of Newey (1994).

Finally, we have

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta_0) &= M_\beta^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(Z_i,\beta_0,p_0(X_i))+M(Z_i)\}+o_p(1)\\ &\xrightarrow{d} N(0,\Pi),\end{aligned} \tag{C.24}$$

where $\Pi := M_\beta^{-1}\text{Var}\{m(Z,\beta_0,p_0(X))+M(Z)\}M_\beta^{-1}$. Note $M_\beta^{-1}\{m(Z_i,\beta_0,p_0(X_i))+M(Z_i)\}$ is the efficient influence function. (See pp.1357-1361 of Newey, 1994). Thus, Π is the efficient variance matrix.

C.8 Proof of Lemma 3.3

The proof is similar to that on pp. 18-20 of BGH-supp and that for Lemma 3.1. 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.

C.9 Proof of Theorem 3.3

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 (C.21), we have

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

With Assumption A7" and A8", the consistency of $\hat{\theta}$ can be proved by similar arguments as in Lemma 5.2 of Newey (1994).

Let us define

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

We have

$$
\begin{aligned}
o_p(n^{-1/2}) &= \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\hat{\beta},\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{m(z_i,\hat{\beta},\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) + E(D(Z_i,\beta_0)|X_i'\hat{\alpha})(T(Z_i,\hat{\beta}) - \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}))\right\} + o_p(n^{-1/2}) \\
&= -M_{n,\beta}(\hat{\beta}-\beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(z_i,\beta_0,F_0(X_i'\alpha_0)) + o_p(n^{-1/2} + (\hat{\beta}-\beta_0)) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\{D(Z_i,\beta_0)(\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0)) + E(D(Z_i,\beta_0)|X_i'\hat{\alpha})(T(Z_i,\beta_0) - \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}))\right\} \\
&= -M_{\beta}(\hat{\beta}-\beta_0) + \frac{1}{n}\sum_{i=1}^{n} m(Z_i,\beta_0,F_0) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0) - E(D(Z_i,\beta_0)|X_i'\hat{\alpha})](\hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha}) - F_0(X_i'\alpha_0))\right\} \\
&+ \frac{1}{n}\sum_{i=1}^{n} E(D(Z_i,\beta_0)|X_i'\alpha_0)(T(Z_i,\beta_0) - F_0(X_i'\alpha_0)) + o_p(n^{-1/2} + (\hat{\beta}-\beta_0)).
\end{aligned}
\tag{C.25}
$$

The second equality follows from Lemma 3.3. 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 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 (C.25) 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:

$$\begin{aligned}
II_m &= \frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\hat{\alpha})]\{X_i-E[X_i|X_i'\alpha_0]\}'F_0^{(1)}(X_i'\alpha_0)+O_p(\hat{\beta}-\beta_0)\right\}(\hat{\alpha}-\alpha_0)\\
&\quad + o_p(\hat{\alpha}-\alpha_0)\\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{[D(Z_i,\beta_0)-E(D(Z_i,\beta_0)|X_i'\hat{\alpha})]\{X_i-E[X_i|X_i'\alpha_0]]\}'F_0^{(1)}(X_i'\alpha_0)\right\}(\hat{\alpha}-\alpha_0)+o_p(\hat{\alpha}-\alpha_0)\\
&= E\left\{[D(Z,\beta_0)-E(D(Z,\beta_0)|X'\alpha_0)]\{X-E[X|X'\alpha_0]]\}'F_0^{(1)}(X'\alpha_0)\right\}(\hat{\alpha}-\alpha_0)+o_p(\hat{\alpha}-\alpha_0).
\end{aligned} \tag{C.26}$$

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

$$\begin{aligned}
M(Z) &= E(D(Z,\beta_0)|X'\alpha_0)(T(Z,\beta_0)-F_0(X'\alpha_0))\\
M_\alpha &= -E\left\{[D(Z,\beta_0)-E(D(Z,\beta_0)|X'\alpha_0)]\{X-E[X|X'\alpha_0]\}'F_0^{(1)}(X'\alpha_0)\right\}.
\end{aligned} \tag{C.27}$$

Combining (C.26) and (C.27) with (C.25), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \hat{\beta}, \hat{F}_{\hat{\alpha},\hat{\beta}}(X_i'\hat{\alpha})) \\
&= -M_\beta(\hat{\beta} - \beta_0) - M_\alpha(\hat{\alpha} - \alpha_0) + \frac{1}{n}\sum_{i=1}^{n} m(z_i, \beta_0, F_0) \\
&+ \frac{1}{n}\sum_{i=1}^{n} M(Z_i) + o_p(n^{-1/2} + (\hat{\beta} - \beta_0) + (\hat{\alpha} - \alpha_0)). \qquad \text{(C.28)}
\end{aligned}$$

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", and A9", 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 (C.25) and (C.28) imply $\hat{\alpha} - \alpha_0 = O_p(n^{-1/2})$ and $\hat{\beta} - \beta_0 = O_p(n^{-1/2})$. Besides, from (C.28) 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 (C.28) to obtain asymptotical expressions of $\hat{\alpha}$ and $\hat{\beta}$.

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

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

We can express $\sqrt{n}(\hat{\alpha} - \alpha_0)$ by replacing m in (C.28) by m_1. Then we have

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

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

function m_1. $M^-_{\alpha,1}$ is the Moore-Penrose inverse of $M_{\alpha,1}$.

Combining with (C.28), 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

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

C.10 Proof of Lemma 3.4

Let's 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 $T(Z_i,\beta) - g^1_{[r]}(X^1_i) - .... - g^{j-1}_{[r]}(X^{j-1}_i) - g^{j+1}_{[r-1]}(X^{j+1}_i) - ...g^k_{[r-1]}(X^k_i) = \tilde{Y}_i$ on X^j_i, and the last sub-function updated in the iteration is $g^j_{[r]}(X^j_i)$. We can replace the Y_i in Lemma 3.1 with $\tilde{Y}_i$, and replace X_i in Lemma 3.1 with X^j_i. $\delta(X)$ is assumed to be a bounded function with a finite variation of X. Since X^j_i is an element of X_i, δ is also a bounded function of X^j_i as well. Therefore, all the arguments in the proof of Lemma 3.1 still hold. We have

$$\frac{1}{n}\sum_{i=1}^{n}\delta(X_i)(T(Z_i,\beta) - g^1_{[r]}(X^1_i) - .... - g^{j-1}_{[r]}(X^{j-1}_i) - g^j_{[r]}(X^j_i) - g^{j+1}_{[r-1]}(X^{j+1}_i) - ...g^k_{[r-1]}(X^k_i))$$
$$= o_p(n^{-1/2}). \tag{C.29}$$

By Theorem 2 of Mammen and Yu (2007), with $r \to \infty$, the backfitting estimator $\{g^j_{[r]}(\cdot)\}^k_{j=1}$ is converging to the least square isotonic estimator of the problem (3.24), $\{g^j(\cdot)\}^k_{j=1}$, i.e.,

$$\lim_{r\to\infty} g^j_{[r]}(\cdot) = g^j(\cdot) \text{ for all } j = 1, ..., k \tag{C.30}$$

in a fixed sample. As mentioned in Section 3.3.2, the least square estimator of the problem (3.23) is obtained by normalizing $\{g^j(\cdot)\}^k_{j=1}$. Therefore, we have

$$\hat{c} + \sum_{j=1}^{k} \hat{m}^j(X_i^j) = \sum_{j=1}^{k} g^j(X_i^j). \tag{C.31}$$

Combining (C.29), (C.30), and (C.31), we have

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

C.11 Proof of Theorem 3.4

The following proof is mostly similar to that in Appendix C.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 (C.21) we have

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

Then

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

The second equality follows from Lemma 3.4. 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 C.3.

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

Finally, we have

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

C.12 Proof of Theorem 3.5

The proof is based on Groeneboom and Hendrickx (2017) (hereafter GH). Here we prove the counterpart for Theorem 3.2. It can be easily adapted to the settings of Theorem 3.1 and Theorem 3.3 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 (C.20) and (C.21), we have

$$||\frac{1}{n} \sum_{i=1}^{n} m(Z_i^*, \hat{\beta}^*, \hat{p}^*_{\hat{\beta}^*}(X_i^*))|| = o_{P_M}(n^{-1/2}), \tag{C.32}$$

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

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

Step 1: Show

$$M_\beta(\hat{\beta}^* - \beta_0) = \frac{1}{n}\sum_{i=1}^{n}\{m(Z_i^*, \beta_0, p_0(X_i^*)) + M(Z_i^*)\} + o_{P_M}(n^{-1/2} + (\hat{\beta}^* - \beta_0)). \tag{C.33}$$

By extending (C.32) 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 (C.23). In the fourth equality, we use $M^*_{n,\beta} - M_\beta = o_p(1)$, and the conditional bootstrapped L_2-result:

$$\frac{1}{n}\sum_{i=1}^{n}\{\hat{p}^*_{\hat{\beta}^*}(X_i^*) - p_0(X_i^*)\}^2 = O_{P_M}((\log n)^2 n^{-2/3}) = o_{P_M}(n^{-1/2}). \tag{C.34}$$

See (6.21) in GH and Proposition 4 in BGH. Now we have shown (C.33). The consistency follows from Assumption A7', A8', and C.34.

Step 2: Rearrangement

(C.33) can be rearranged to

$$M_\beta(\hat{\beta}^* - \beta_0) = \left\{ \frac{1}{n}\sum_{i=1}^{n} m(Z_i^*, \beta_0, p_0(X_i^*)) - \frac{1}{n}\sum_{i=1}^{n} m(Z_i, \beta_0, p_0(X_i)) \right\} + \left\{ \frac{1}{n}\sum_{i=1}^{n} M(Z_i^*) - \frac{1}{n}\sum_{i=1}^{n} M(Z_i) \right\}$$
$$+ \frac{1}{n}\sum_{i=1}^{n} \{m(Z_i, \beta_0, p_0(X_i) + M(Z_i)\} + o_{P_M}(n^{-1/2} + (\hat{\beta}^* - \beta_0)). \tag{C.35}$$

Then we could subtract (C.23) from (C.35) 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 (C.24).

Bibliography

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 641-647.

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

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

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

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

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

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

Bickel, P. J., & Ritov, Y. A. (2003). Nonparametric estimators which can be "plugged-in". *Annals of Statistics*, 31(4), 1033-1053.

Carroll, R. J., Fan, J., Gijbels, I. and M. P. Wand (1997) Generalized partially linear single-index models, *Journal of the American Statistical Association*, 92, 477-489.

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

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

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.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 51 765–782.

Cosslett, S. R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica* 55 559-585.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448), 1053-1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1), 151-161.

Dehejia, R. (2005). Practical propensity score matching: a reply to Smith and Todd. *Journal of Econometrics*, 125(1-2), 355-364.

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

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.

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

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

Groeneboom, P. and K. Hendrickx (2017) The nonparametric bootstrap for the current status model, *Electronic Journal of Statistics*, 11, 3446-3484.

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

Groeneboom, P. and Jongbloed, G. (2014). Nonparametric estimation under shape constraints. *Cambridge University Press.*

Groeneboom, P. (2018). The Lagrange approach in the monotone single index model. *arXiv preprint arXiv:1812.01380.*

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

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

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

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

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.*

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.

Hjort, N. L., McKeague, I. W. and I. van Keilegom (2009) Extending the scope of empirical likelihood, *Annals of Statistics*, 37, 1079-1111.

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

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.

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

Hurvich, C. M., Simonoff, J. S. and C.-L. Tsai (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society*, B, 60, 271-293.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics*, 58, 71-120.

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

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

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

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

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604-620.

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

Ma, Y. and L. Zhu (2013) Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates, *Journal of the Royal Statistical Society*, B, 75, 305-322.

Matzkin RL. (1992). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica* 60:239–70

Mukerjee, H. (1988). Monotone nonparametric regression, *Annals of Statistics*, 16, 741-750.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2), 99-135.

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

Newey, W. K., Hsieh, F., & Robins, J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72(3), 947-962.

Owen, A. (1991) Empirical likelihood for linear models, *Annals of Statistics*, 19, 1725-1747.

Powell, J. L., Stock, J. H. and T. M. Stoker (1989) Semiparametric estimation of index coefficients, *Econometrica*, 57, 1403-1430.

Qin, G. S. and B. Y. Jing (2001) Censored partial linear models and empirical likelihood, *Journal of Multivariate Analysis*, 78, 37-61.

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.

Rao, J. N. K. and A. J. Scott (1981) The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence in two-way tables, *Journal of the American Statistical Association*, 76, 221-230.

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

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

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2), 305-353.

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

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.

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

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

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

Wang, J.-L., Xue, L., Zhu, L. and Y. S. Chong (2010) Estimation for a partial-linear single-index model, *Annals of Statistics*, 38, 246-274.

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

Xia, Y. and W. Härdle (2006) Semi-parametric estimation of partially linear single-index models, *Journal of Multivariate Analysis*, 97, 1162-1184.

Xia, Y., Tong, H. and W. Li (1999) On extended partially linear single-index models, *Biometrika*, 86, 831-842.

Xu, M., & Otsu, T. (2020). Score estimation of monotone partially linear index model. *Journal of Nonparametric Statistics*, 32(4), 838-863.

Xue, L.-G. and L. Zhu (2006) Empirical likelihood for single-index models, *Journal of Multivariate Analysis*, 97, 1295-1312.

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

Yu, Y. and D. Ruppert (2002) Penalized spline estimation for partially linear single-index models, *Journal of the American Statistical Association*, 97, 1042-1054.

Yuan, A., Yin, A., & Tan, M. T. (2021). Enhanced Doubly Robust Procedure for Causal Inference. *Statistics in Biosciences,* 1-25.