Optimal Budget Allocation to Social Treatment Programs*

Ming He
Economics Discipline Group
University of Technology Sydney
Box 123, Sydney, NSW 2007

Bin Peng
Department of Economics
University of Bath
Bath BA2 7JP, UK

August 16, 2017

Abstract

In this paper, we study the problem of allocating a limited budget to competing social treatment programs. The optimal budget allocation scheme is defined by an optimal splitting point that maximizes a social value function, weighted across different treatment programs. We propose estimators for the social value function and the optimal splitting point, and establish their asymptotic properties. Monte Carlo experiments are conducted to evaluate the finite sample performances of our estimators. We illustrate the usefulness of our framework and method by studying the optimal budget allocation among the treatment programs of providing subsidized long-lasting insecticide-treatment bed net in Kenya to fight malaria.

Keywords: Budget Allocation, Nonparametric Regression, Treatment Effects.

*We thank Elie Tamer and participants in the 7th Shanghai Econometrics Workshop for helpful suggestions and comments. We are grateful to Pascaline Dupas for sharing her data with us.

1 Introduction

In the treatment effects literature, the emphasis has been on evaluating the effects of social treatment programs, such as average treatment effects, distributional treatment effects, quantile treatment effects, and so on, see Heckman and Vytlacil (2007a), Heckman and Vytlacil (2007b), and Imbens and Wooldridge (2009) for surveys on recent development.

In practice, there are usually parallel social treatment programs competing for limited social resources. In this situation, the central budget officer faces the problem of allocating a limited budget among these parallel programs. For example, a Department of Education officer needs to allocate a limited budget to different schools; a university officer faces the problem of allocating a limited budget to different departments; an officer of the Department of Labor needs to allocate a limited budget to different job training programs; and an officer of the World Health Organization faces the problem of allocating a limited medical resource to different countries or regions in Africa. Motivated by these practical questions, we consider the situation in which there is a fixed budget that can potentially be allocated among multiple treatment programs. Given an allocation of the total budget, each program will also be subject to a budget constraint, and the program coordinator maximizes a value function for that program by using some treatment assignment rule.

Our work is related to the literature on optimal treatment assignment rule, which aims to provide an individualized treatment recommendation based on observed characteristics. The relevant works include, but not limited to, Manski (2004), Dehejia (2005), Tetenov (2012), Hirano and Porter (2009), Bhattacharya and Dupas (2012), and so forth. To be more specific, Manski (2004) proposes to assess the welfare properties of treatment assignment rules by the maximum regret, and develops finite-sample bounds on the maximum regret of conditional empirical success rules. Dehejia (2005) studies the treatment assignment problem in a Bayesian framework. Tetenov (2012) extends Manski (2004) by differentiating between Type I and Type II regrets, and considers minimax regret treatment choice with asymmetric reference-dependent welfare functions exhibiting loss aversion. Hirano and Porter (2009) develop large-sample approximation to statistical treatment assignment problems and derive treatment assignment rules that are asymptotically optimal under average and minimax risk criteria. Bhattacharya and Dupas (2012) consider a practical situation with limited resource, and study the optimal treatment assignment under a budget constraint. More recently, Armstrong and Shen (2015) consider the inference problem for the optimal treatment rule. They use multiple hypothesis

testing methods to construct a random set for which there is overwhelming evidence that an individual with characteristics in this set should receive a treatment. Kitagawa and Tetenov (2015) develop an empirical welfare maximization method and estimate a treatment assignment policy by maximizing the sample analog of average social welfare over a class of feasible treatment policies.

We make the following contributions to the literature. First, we formulate the optimal budget allocation problem as one to maximize a social value function, weighted across competing treatment programs. The optimal budget allocation scheme is defined as an optimal splitting point for the proportion of money that is allocated to each program. Second, we consider estimation of the social value function and the optimal splitting point, and provide a formal statistical procedure to make an informed decision. Third, we apply our method to the optimal budget allocation in the treatment programs of providing subsidized long-lasting insecticide-treated bed net in Kenya to fight malaria, and find that a larger portion of the budget should be allocated to the program whose applicants have a child under ten years old.

Before proceeding to Section 2, it is convenient to introduce some notations that will be used throughout this paper. $\|A\|$ denotes the Euclidean norm of a vector or the Frobenius norm of a matrix; $\to_P$ denotes converging in probability; $\to_D$ denotes converging in distribution. Since we are interested in both parameters and functions in the paper, we define a norm in (1.1) below. Let θ represent a fixed dimension vector belonging to $\mathbb{R}^d$ with a norm $\|\cdot\|$, and let ϕ stand for a function belonging to a function space Φ with a norm $\|\phi\|_{L^2}$. Then for $\forall(\theta,\phi)\in\mathbb{R}^d\times\Phi$, we define the corresponding norm by

$$\|(\theta,\phi)\|_2=\left\{\|\theta\|^2+\|\phi\|_{L^2}^2\right\}^{1/2}. \tag{1.1}$$

Clearly, $\|\cdot\|_2$ satisfies the definition of a norm, and is in the same spirit as Newey and Powell (2003, p.1569). For square-integrable functions on a specific interval (say [0,1]), $\|\cdot\|_{L^2}$ can be $\|\phi\|_{L^2}=\left\{\int_{[0,1]}\phi^2(x)dx\right\}^{1/2}$. For non-integrable functions on the whole real line, one can use, for example, a function space $L^2(\mathbb{R},\exp(-x^2))$, so the corresponding norm will be $\|\phi\|_{L^2}=\left\{\int\phi^2(x)\exp(-x^2)dx\right\}^{1/2}$. More details, examples and discussions can be seen in Dong and Linton (2016).

The rest of this paper is organised as follows: Section 2 presents the theoretical framework, and defines the social value function as well as the optimal splitting point. Section 3 proposes estimators for the social value function and the optimal splitting point, and establishes their asymptotic properties. In Section 4, we conduct Monte Carlo simulations to evaluate the

finite sample performances of our estimators. In Section 5, we apply the proposed procedure to the budget allocation problem for treatment programs of providing subsidized long-lasting insecticide-treated bed net in Kenya. Section 6 concludes with discussions. All the proofs are given in Appendix A.

2 The Theoretical Framework

Suppose that there are L treatment programs, and without losing generality let $L = 2$. Define the primitive random vector for program ℓ by $W_\ell = (Y_{\ell_1}, Y_{\ell_0}, X_\ell, A_\ell)$ with $\ell = 1, 2$, where Y_{ℓ_a} with $a = 0, 1$ are the potential untreated and treated outcomes, X_ℓ is the vector of covariates for an individual in program ℓ, and A_ℓ is the treatment status with 1 as treated and 0 as untreated. Denote the population conditional expected potential outcome as $\mu^*_{\ell_a}(x) = \mathbb{E}[Y_{\ell_a}|X_\ell = x]$, and let the corresponding population conditional average treatment effect be $\delta^*_\ell(x) = \mu^*_{\ell_1}(x) - \mu^*_{\ell_0}(x)$.

With two treatment programs and a fixed amount of budget, the allocation problem is a two-stage one. The first-stage is for the budget officer to allocate the fixed amount of budget to the two programs. Let the total budget be b, and the proportion of budget for program ℓ be θ_ℓ with $\sum_{\ell=1}^{L} \theta_\ell = 1$, $\theta_\ell \in [0, 1]$. Given any pair (θ_1, θ_2) of budget allocation, the second-stage is for each program coordinator to design the optimal treatment allocation rule (i.e., the optimal propensity score) under the budget constraint for that program. For program ℓ, suppose there are N_ℓ applicants, and the treatment cost per person is c_ℓ. Thus, for $\forall \theta_\ell \in [0, 1]$, the proportion in population ℓ that can be treated is $b\theta_\ell / c_\ell N_\ell \equiv \nu_\ell \theta_\ell$.

According to the above setting, the optimization problem of the second-stage is formulated by

$$\begin{aligned} \rho_\ell(\theta_\ell) = \max_{p_\ell} \mathbb{E}\left\{\mu^*_{\ell_1}(X_\ell) p_\ell(X_\ell) + \mu^*_{\ell_0}(X_\ell)[1 - p_\ell(X_\ell)]\right\} \\ \text{s.t. } \mathbb{E}[p_\ell(X_\ell)] = \nu_\ell \theta_\ell \quad \text{for} \quad \ell = 1, 2, \end{aligned} \tag{2.1}$$

where $p_\ell(\cdot)$ is a candidate propensity score function. The constraint can be understood as follows: given the assignment rule $p_\ell(\cdot)$, the expected expense for one applicant of program ℓ is $\mathbb{E}[p_\ell(X_\ell)c_\ell + (1 - p_\ell(X_\ell))0]$, then the total expense for program ℓ under $p_\ell(\cdot)$ is $N_\ell \mathbb{E}[p_\ell(X_\ell)c_\ell] = \theta_\ell b$. The solution $\rho_\ell(\theta_\ell)$ represents the *ex ante* expected outcome for one applicant of program ℓ given the proportion of budget for program ℓ and its optimal assignment rule.

By Bhattacharya and Dupas (2012), the solution to (2.1) is the following optimal propensity

score function

$$p_\ell^*(x) = 1(\delta_\ell^*(x) \geq q_\ell^*(1 - \nu_\ell\theta_\ell)), \tag{2.2}$$

where $1(\cdot)$ is the indicator function, and $q_\ell^*(\cdot)$ is the quantile function of $\delta_\ell^*(X_\ell)$. Note that for program ℓ, $\nu_\ell\theta_\ell$ represents the proportion that can be treated. If this proportion decreases, then the threshold value $q_\ell^*(1-\nu_\ell\theta_\ell)$ will increase, and only those with very large positive treatment effect will be treated.

Given the optimal propensity score function at the second stage, the first-stage problem is formulated as

$$\max_{\theta_1+\theta_2=1} \rho(\theta_1, \theta_2) = \max_{\theta_1+\theta_2=1} w_1\rho_1(\theta_1) + w_2\rho_2(\theta_2), \tag{2.3}$$

where $w_\ell = \frac{N_\ell}{N_1+N_2}$ for $\ell = 1, 2$. Note that with the weights w_ℓ's, the social value function has already been normalized by the total number of applicants.

For notational simplicity, let $\phi_\ell^* = (\mu_{\ell_1}^*, \mu_{\ell_0}^*, q_\ell^*)$ be the vector of population functions associated with program ℓ, and $\phi_\ell = (\mu_{\ell_1}, \mu_{\ell_0}, q_\ell) \in \Phi_\ell$ be a generic element. Define

$$\rho(\theta, \phi) = \mathbb{E}[\rho(X; \theta, \phi)], \tag{2.4}$$

where $X = (X_1, X_2)$ with $X_1 \perp X_2$, $\rho(x; \theta, \phi) = \sum_{\ell=1,2} \rho_\ell(x_\ell; \theta, \phi_\ell)$, and

$$\rho_\ell(x_\ell; \theta, \phi_\ell) = \sum_{a=0,1} \rho_{\ell_a}\Big(x_\ell; 1(\ell = 1)\theta + 1(\ell = 2)(1-\theta), \phi_\ell\Big),$$
$$\rho_{\ell_a}(x_\ell; \theta, \phi_\ell) = w_\ell\mu_{\ell_a}(x_\ell)1\left((-1)^{a+1}\Big[\delta_\ell(x_\ell) - q_\ell(1-\nu_\ell\theta)\Big] \geq 0\right).$$

Further note that for two programs, the parameter space Θ of θ is not necessarily $[0, 1]$. By definition, we have $\nu_\ell\theta_\ell \in [0, 1]$. On one hand, $\nu_\ell\theta \geq 0$, because the money assigned to one program cannot be negative; on the other hand, $\nu_\ell\theta \leq 1$, as one program should not get more than what it needs for all its applicants. These restrictions together give $\Theta = \Big[\max(0, 1 - \frac{1}{\nu_2}), \min(1, \frac{1}{\nu_1})\Big]$.

Let's name $\rho(\theta, \phi^*)$ as the value frontier function, where ϕ^* is given but unknown. It represents the maximum weighted social value that the two treatment programs can achieve given a candidate budget allocation plan, and that the second-stage treatment is assigned according to the optimal rule. Then the optimal budget-splitting point can be obtained by

$$\theta^* = \arg\max_{\theta\in\Theta} \rho(\theta, \phi^*). \tag{2.5}$$

In the following, we are interested in recovering both θ^* and $\rho(\theta, \phi^*)$. While θ^* helps allocate the limited resource, the shape of $\rho(\theta, \phi^*)$ contains information on how much loss is associated with any suboptimal splitting point.

To better explain our motivation and the above settings, we consider the next simple example before proceeding further.

Example 1 *Let $X_\ell \sim U(0,1)$ denote the IQ of a student in population ℓ, and let A_ℓ be college education with $\ell = 1, 2$. For simplicity, we choose $b = 100$, $N_\ell = 100$ and $c_\ell = 1$, so it gives $\nu_\ell = 1$. Moreover, we let $\mu^*_{1_1}(x) = 3x, \mu^*_{1_0}(x) = x$, $\mu^*_{2_1}(x) = 2x$, $\mu^*_{2_0}(x) = x$, so simple algebra shows $\delta^*_1(x) = 2x$ and $\delta^*_2(x) = x$. Notice that $\delta^*_1(x) \geq \delta^*_2(x)$ is due to some unobserved factor, e.g., the students in population 1 are more perseverant than those in population 2.*

Then it can be shown that

$$\rho_1(\theta) = -\theta^2 + 2\theta + \frac{1}{2}, \ \rho_2(1-\theta) = -\frac{1}{2}\theta^2 + 1, \ \rho(\theta) = -\frac{3}{4}\theta^2 + \theta + \frac{3}{4},$$

which are respectively plotted in Figure 1 below. In this example, two programs compete for the limited budget. If only one program is considered, then θ should either be 1 or 0. However, in practice, both programs need to be considered, and the optimal splitting point is $\theta^* = \frac{2}{3} > 0.5$. Intuitively speaking, this means that we should allocate a larger portion of the budget to provide the students from population 1 college education in order to ensure a larger social value.

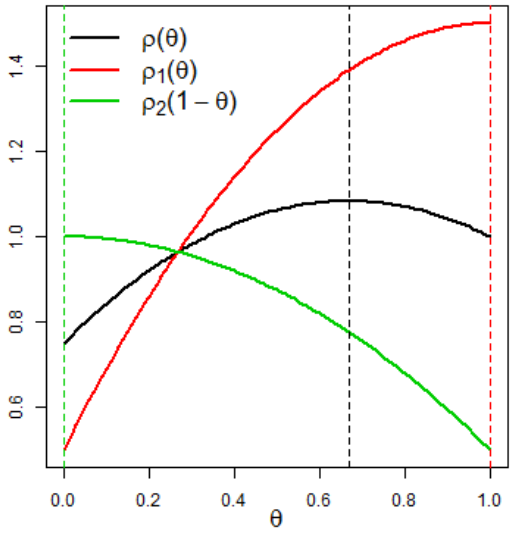


Figure 1: Competition between Programs and the Value Frontier Function

3 Estimation and Inference

Firstly, we provide the next lemma under a general framework, which can be applied to optimization problems similar to this study.

Lemma 3.1 *Let $\rho(\theta, \phi)$ define a population objective function mapping from $\Theta \times \Phi$ to $\mathbb{R}$, where Θ and Φ stand for a vector space and a function space with the norms $\|\cdot\|$ and $\|\cdot\|_{L^2}$, respectively. Suppose the following conditions hold:*

1. *$\Theta \times \Phi$ is compact;*
2. *For a given unknown ϕ^*, $\rho(\theta, \phi^*)$ has a unique maximum at $\theta = \theta^*$;*
3. *For $\forall(\theta_1, \phi_1),\ (\theta_2, \phi_2) \in \Theta \times \Phi$, suppose $|\rho(\theta_1, \phi_1) - \rho(\theta_2, \phi_2)| \leq C\|(\theta_1, \phi_1) - (\theta_2, \phi_2)\|_2$, where C is a positive constant and $\|\cdot\|_2$ is defined by (1.1);*
4. *For $\forall(\theta, \phi) \in \Theta \times \Phi$, $\widehat{\rho}(\theta, \phi) - \rho(\theta, \phi) = o_P(1)$, where $\widehat{\rho}(\theta, \phi)$ is the sample version of $\rho(\theta, \phi)$;*
5. *$\widehat{\phi}$ is a consistent nonparametric estimator of ϕ^* such that $\|\widehat{\phi} - \phi^*\|_{L^2} = o_P(1)$.*

Then the following two results hold.

1. *$\sup_{\theta \in \Theta} |\widehat{\rho}(\theta, \widehat{\phi}) - \rho(\theta, \phi^*)| = o_P(1)$;*
2. *$\widehat{\theta} - \theta^* = o_P(1)$, where $\widehat{\theta} = \arg\max_{\theta \in \Theta} \widehat{\rho}(\theta, \widehat{\phi})$.*

This lemma is in the same spirit as Theorem 1 of Chen et al. (2003), wherein a different set of assumptions and proofs are provided to solve another optimization problem. Condition 2 implies that (θ^*, ϕ^*) may not be a global minimum for $\rho(\cdot, \cdot)$. Condition 3 essentially defines the continuity of $\rho(\cdot, \cdot)$. Condition 4 can be justified easily for the independent and identically distributed (i.i.d.) cross-sectional data or the stationary time series. For the continuous dependent variable, one can follow for example Lemma A.1 of the Appendix A of this paper to justify Condition 5; for the binary dependent variables, one can follow Carroll et al. (1997) to obtain the consistent estimator of ϕ^* to ensure Condition 5 is satisfied. We refer interested readers to Chen (2007) for more complicated scenarios.

With Lemma 3.1 in hand, we are now ready to investigate the budget allocation problem given in (2.5).

3.1 The Value Frontier and Optimal Splitting Point

Note that the indicator function is involved in (2.2), however, the indicator function itself is not twice continuously differentiable which further causes troubles in establishing asymptotic properties. Therefore, we replace the indicator function with a cumulative distribution function (CDF) $K(\cdot)$ following the idea of kernel smoothing in the literature (e.g., Horowitz, 1992).

After the replacement, the objective function considered in (2.5) becomes $\rho_n(\theta, \phi^*)$, where $\rho_n(\theta,\phi) = \mathbb{E}[\rho_n(X;\theta,\phi)]$, $\rho_n(x;\theta,\phi) = \sum_{\ell=1,2} \rho_{\ell n_\ell}(x_\ell;\theta,\phi_\ell)$, and

$$\rho_{\ell n_\ell}(x_\ell;\theta,\phi_\ell) = \sum_{a=0,1} \rho_{\ell_a n_\ell}(x_\ell; 1(\ell=1)\theta + 1(\ell=2)(1-\theta), \phi_\ell),$$
$$\rho_{\ell_a n_\ell}(x_\ell;\theta,\phi_\ell) = w_\ell \mu_{\ell_a}(x_\ell) K\left(\frac{(-1)^{a+1}[\delta_\ell(x_\ell) - q_\ell(1-\nu_\ell\theta)]}{h_\ell}\right).$$

In the above, $h_\ell \to 0$ with $\ell = 1, 2$ are two bandwidth sequences. Finally, define

$$\widehat{\rho}(\theta,\phi) = \sum_{\ell=1,2} \mathbb{P}_{n_\ell}[\rho_{\ell n_\ell}(X_\ell;\theta,\phi_\ell)], \tag{3.1}$$

where $\mathbb{P}_{n_\ell}$ is the empirical measure based on the random sample from program ℓ. Then the estimator of θ^* is obtained as follows.

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \widehat{\rho}(\theta, \widehat{\phi}), \tag{3.2}$$

where $\widehat{\phi}(\cdot)$ is a consistent nonparametric estimator for ϕ^*.

For the choice of $\widehat{\phi}(\cdot)$, recall that $\phi^* = (\phi_1^*, \phi_2^*)$, where $\phi_\ell^* = (\mu_{\ell_1}^*, \mu_{\ell_0}^*, q_\ell^*)$. We estimate ϕ_ℓ^* as follows: under Assumption 1 (*iii*) below, $\mu_{\ell_a}^*(x)$ is estimated by a sieve estimator $\widehat{\mu}_{\ell_a}(x)$, where the Hermite polynomial sieve space is used. The proof for uniform consistency of $\widehat{\mu}_{\ell_a}(\cdot)$ is detailed in Lemma A.1. $q_\ell^*(\cdot)$ is estimated by the empirical quantile function $\widehat{q}_\ell(\cdot)$ based on the pseudo sample of $\{\widehat{\delta}_\ell(X_{\ell i})\}_{i=1}^{n_\ell}$, where $\widehat{\delta}_\ell(x) = \widehat{\mu}_{\ell_1}(x) - \widehat{\mu}_{\ell_0}(x)$. The uniform consistency of $\widehat{q}_\ell(\cdot)$ is shown in Lemma A.2.

Remark 1 *In practice, when there are more than two programs (i.e. $L > 2$), we can do a pairwise analysis for (θ_1, θ_ℓ), $\ell \neq 1$, under the restriction that $\theta_1 + \theta_\ell = 1$ to obtain the ratio $\widehat{\theta}_1/\widehat{\theta}_\ell = \widehat{r}_\ell$. After getting the point estimates $\widehat{r}_\ell$, then we can solve for the optimal allocation $(\widehat{\theta}_1, \cdots, \widehat{\theta}_L)$ by*

$$\widehat{\theta}_1\left(1 + \frac{1}{\widehat{r}_2} + \cdots + \frac{1}{\widehat{r}_L}\right) = 1,\ \widehat{\theta}_\ell = \frac{\widehat{\theta}_1}{\widehat{r}_\ell},\ \ell = 1, \cdots, L.$$

We further make the following assumptions to facilitate the development.

Assumption 1 (i). *Independence between programs:* $W_1 \perp W_2$. (ii). *Independence within programs:* $\{W_{\ell i}\}_{i=1}^{n_\ell}$ *are i.i.d. for* $\ell = 1, 2$. (iii). $(Y_{\ell_1}, Y_{\ell_0}) \perp A_\ell | X_\ell$ *for* $\ell = 1, 2$.

Assumption 2 *Let* $\rho(\theta, \phi)$ *be that defined in (2.4), and be defined on a compact set* $\Theta \times \Phi$. *Suppose that given* $\phi^* \in \Phi$, $\rho(\theta, \phi^*)$ *has a unique maximum at* $\theta = \theta^*$ *on* Θ. *Suppose further that* $\widehat{\phi}(\cdot)$ *is a consistent nonparametric estimator for* ϕ^*, *and* $\|\widehat{\phi} - \phi^*\|_{L^2} = o_P(1)$.

Assumption 3 *For* $\ell = 1, 2$, *let* G_ℓ^* *be the CDF of* $Z_\ell \equiv \delta_\ell^*(X_\ell)$. *Suppose that*

(i). Z_ℓ *belongs to a compact set* $\mathcal{Z}$;

(ii). G_ℓ^* *is strictly increasing and twice differentiable on* $\mathcal{Z}$, *and* $\sup_{z \in \mathcal{Z}} \left| \frac{\partial G_\ell^*(z)}{\partial z} \right| \geq c > 0$, *where* c *is a positive constant.*

Assumption 4

For $\ell = 1, 2$, *let* $f_\ell(u_0, u_1)$ *be the joint density of* $(\mu_{\ell_0}^*(X_\ell), \mu_{\ell_1}^*(X_\ell))$, *and* $f_{\ell u_a}^{(1)}(u_0, u_1) = \frac{\partial f_\ell(u_0, u_1)}{\partial u_a}$ *for* $a = 0, 1$. *Suppose that (i).* $\mu_{\ell_a}^*(X_\ell) \in [\underline{u}_{\ell_a}, \overline{u}_{\ell_a}]$; *(ii).* $\sup_{(u_0, u_1)} |f_{\ell u_a}^{(1)}(u_0, u_1)| < \infty$.

Assumption 5 (i). *Let* $K(\cdot)$ *be a CDF function such that* $K^{(1)}(\cdot)$ *(i.e., the corresponding probability density function (PDF)) is defined on* $[-1, 1]$; (ii). *For* $\ell = 1, 2$, $h_\ell \to 0$, $n_\ell h_\ell \to \infty$, $n_\ell h_\ell^3 \to 0$, $n_2 h_1^2 h_2 \to 0$, *and* $n_1 h_1 h_2^2 \to 0$.

Assumptions 1 (i) and (ii) impose independence between programs as well as independence across different individuals within a program, and (iii) imposes the standard selection on observables assumption widely used in the treatment effects literature. Under Assumption 1, $\mu_{\ell_a}^*$ is identified as $\mu_{\ell_a}^*(x) = \mathbb{E}[Y_\ell | X_\ell = x, A_\ell = a]$ for $\ell = 1, 2$ and $a = 0, 1$. Assumptions 2-4 impose some restrictions on the outcome functions, and are standard in the literature. Assumption 5 puts constraints on the kernel function and bandwidths.

Given these assumptions, we state the first main result of this study below.

Theorem 3.1 *Under Assumptions 1-5, as* $(n_1, n_2) \to (\infty, \infty)$,

1. $\sup_{\theta \in \Theta} |\widehat{\rho}(\theta, \widehat{\phi}) - \rho(\theta, \phi^*)| = o_P(1)$, *where* $\widehat{\rho}(\theta, \widehat{\phi})$ *and* $\rho(\theta, \phi^*)$ *are defined in (3.1) and (2.4) respectively,*

2. $\widehat{\theta} - \theta^* = o_P(1)$.

Before stating the asymptotic normality, we introduce some notations for better presentation. For $\ell = 1, 2$, let

$$
\begin{aligned}
&T_\ell^\dagger = T_{\ell_0}^\dagger + T_{\ell_1}^\dagger + 2T_{\ell_{01}}^\dagger, \\
&T_{\ell_{01}}^\dagger = (-1)^{1+2\ell} \int [K^{(1)}(v)]^2 dv \int u(u + e_\ell^*) f_\ell(u, u + e_\ell^*) du, \\
&T_{\ell_0}^\dagger = \int [K^{(1)}(v)]^2 dv \int u^2 f_\ell(u, u + e_\ell^*) du, \\
&T_{\ell_1}^\dagger = \int [K^{(1)}(v)]^2 dv \int u^2 f_\ell(u - e_\ell^*, u) du, \\
&e_\ell^* = e_\ell(\theta^*, \phi^*), \quad o_\ell^* = o_\ell(\theta^*, \phi^*), \\
&e_\ell(\theta, \phi) = q_\ell \left(1 - 1(\ell = 1)\nu_1\theta - 1(\ell = 2)\nu_2(1 - \theta)\right), \\
&o_\ell(\theta, \phi) = w_\ell \nu_\ell q_\ell' \left(1 - 1(\ell = 1)\nu_1\theta - 1(\ell = 2)\nu_2(1 - \theta)\right).
\end{aligned}
$$

The asymptotic normality is summarized in the next theorem.

Theorem 3.2 *Let Assumptions 1-5 hold. In addition, suppose that* $\frac{n_1 h_1}{n_2 h_2} \to \kappa$ *with* κ *being a positive constant. As* $(n_1, n_2) \to (\infty, \infty)$,

$$
\sqrt{\min(n_1 h_1, n_2 h_2)}(\widehat{\theta} - \theta^* - B) \to_D N\left(0, \sigma^2/\xi^2\right),
$$

where $B = \nabla_\theta \widehat{\rho}(\theta^*, \widehat{\phi}) - \nabla_\theta \widehat{\rho}(\theta^*, \phi^*)$ *is a bias term,* $\sigma^2 = \min(1, \frac{1}{\kappa}) o_1^{*2} T_1^\dagger + \min(1, \kappa) o_2^{*2} T_2^\dagger$ *and* $\xi = \lim_{h_1, h_2 \to 0} \nabla_{\theta\theta}^2 \rho_n(\theta^*, \phi^*)$.

Remark 2 *We would like to point out the bias term B can completely disappear from the system if we follow Yu and Ruppert (2002) to impose assumptions on the sample version of the objective function directly. The current form of B is in fact identical to the term* $\sqrt{n}\Gamma_2(\theta_o, h_o)(\widehat{h} - h_o)$ *given in Assumption 2.6 of Theorem 2 of Chen et al. (2003), where* $\widehat{h}$ *is their estimate on the unknown function* h_o *with i.i.d. data. Without imposing additional conditions or making additional assumptions regarding the sample version of the objective function, this term will not vanish.*

4 Simulation

In this section, we conduct a small-scale simulation experiment to evaluate the finite sample performances of our estimators. The simulation design is based on that in Example 1. That is, let $\mu_{1_1}^*(x) = 3x$, $\mu_{1_0}^*(x) = x$, $\mu_{2_1}^*(x) = 2x$, $\mu_{2_0}^*(x) = x$. For $\ell = 1, 2$ and $a = 0, 1$, we let

$X_\ell \sim U(0,1)$, $\epsilon_{\ell_a} \sim N(0,1)$, and $Y_{\ell_a} = \mu^*_{\ell_a}(X_\ell) + \epsilon_{\ell_a}$. Let $b = 1000$, $c_\ell = 1$. For the choice of (N_1, N_2), we experiment three combinations, $(600, 1000)$, $(1000, 600)$, and $(1000, 1000)$. A re-scaled Gaussian CDF is used for the kernel function $K(\cdot)$,[1] the bandwidth is set to be $h_\ell = n_\ell^{-2/5}$, and Hermite polynomials are used for the sieve approximation of the functions $\mu^*_{\ell_a}(\cdot)$,[2] with the truncation parameters set to $\lfloor n_\ell^{1/5} \rfloor + 1$.[3] For simplicity, we set $n_1 = n_2 = n$, where $n \in \{200, 500, 1000\}$ across all the experiments. The number of repetitions is 1000.

Experiment 1: The propensity scores are set to be $p_1(x) = p_2(x) = 0.5$.
Experiment 2: The propensity scores are set to be $p_1(x) = 0.3$, $p_2(x) = 0.7$.
Experiment 3: The propensity scores are set to be $p_1(x) = x$ and $p_2(x) = x^2$.

In Table 1, we report the result for bias and root mean squared error (RMSE) of $\widehat{\theta}$ under different combinations of (N_1, N_2) and sample sizes. On the one hand, the bias does not have a clear decreasing trend as the sample size increases due to the bias term in Theorem 3.2. On the other hand, the RMSE decreases in all cases as the sample size increases. For example, consider the $(N_1, N_2) = (1000, 600)$ combination in Experiment 2, the RMSEs are 0.0980, 0.0740, and 0.0598 when $n = 200, 500, 1000$, respectively.

In Figures 2, 3, and 4, we depict the estimated curve $\widehat{\rho}(\theta, \widehat{\phi})$ in each experiment for different combinations of (N_1, N_2), and the curves are obtained by averaging across all the repetitions. To be precise, the black solid curve represents the true function, while the coloured dash lines correspond to the averaged estimates for different sample sizes. Consider Figure 4 for example. All estimated curves are sufficiently close to the true one. Moreover, it is clear that as the sample size increases, the estimated curve gets closer to the true curve.

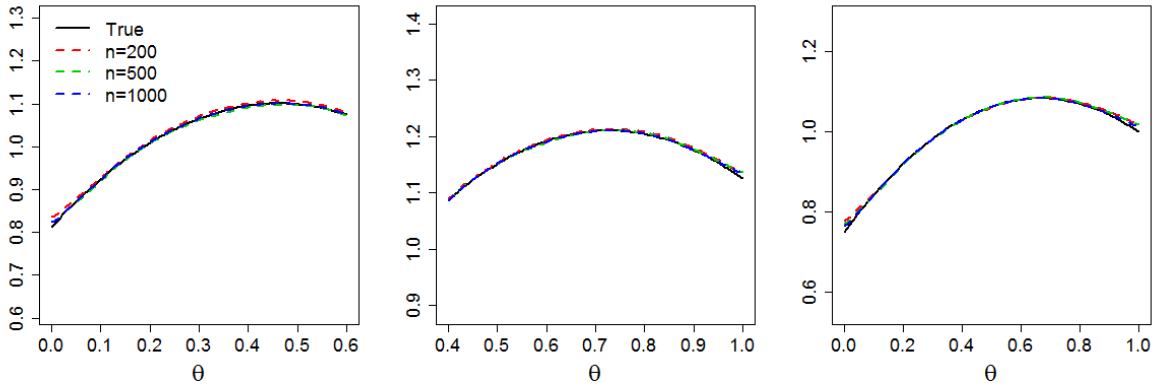
[1] The kernel function is defined as $K(x) = \frac{\Psi(x) - \Psi(-1)}{\Psi(1) - \Psi(-1)}$ for $x \in [-1, 1]$, where $\Psi(\cdot)$ is the CDF of the standard normal distribution.

[2] See Chen (2007) and Dong and Linton (2016) for detailed discussions on Hermite Polynomials.

[3] The above choices of bandwidths and truncation parameters may not be the optimal ones, but they satisfy all the requirements of our assumptions. Similar choices are also adopted by Su and Jin (2012) and Dong et al. (2016). While the choices of the bandwidths and truncations parameters have been well studied under some classic single equation models (e.g., Hall et al., 2007; Gao et al., 2002), how to choose the optimal ones together in the current setting of our paper is left to future research.

(N_1, N_2)	$(600, 1000)$		$(1000, 600)$		$(1000, 1000)$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Experiment 1						
$(n = 200)$	-0.0030	0.0745	0.0056	0.0968	0.0195	0.1348
$(n = 500)$	0.0077	0.0589	0.0100	0.0707	0.0145	0.0939
$(n = 1000)$	0.0038	0.0439	0.0052	0.0494	0.0015	0.0671
Experiment 2						
$(n = 200)$	-0.0009	0.0701	0.0048	0.0980	0.0130	0.1361
$(n = 500)$	0.0051	0.0599	0.0128	0.0740	0.0186	0.1012
$(n = 1000)$	0.0021	0.0496	0.0091	0.0598	0.0094	0.0757
Experiment 3						
$(n = 200)$	-0.0085	0.0843	-0.0023	0.1033	-0.0090	0.1397
$(n = 500)$	-0.0095	0.0651	-0.0048	0.0819	0.0080	0.1096
$(n = 1000)$	-0.0065	0.0530	0.0019	0.0668	0.0123	0.0840
θ^*	0.4615		0.7273		0.6667	

Table 1: Bias and RMSE Performance of the Estimator

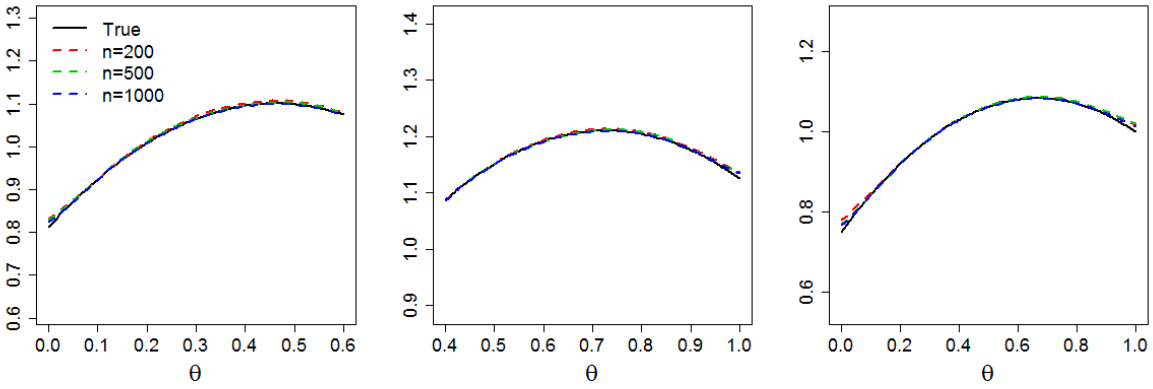


(a) $(N_1, N_2) = (600, 1000)$ (b) $(N_1, N_2) = (1000, 600)$ (c) $(N_1, N_2) = (1000, 1000)$

Figure 2: Averaged Estimates of $\rho(\theta, \phi^*)$ (Experiment 1)

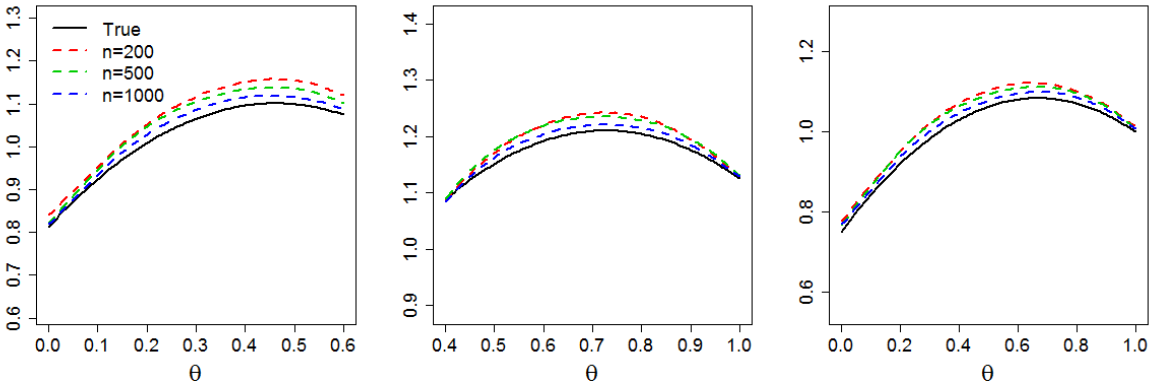
5 Empirical Application

In this section, we apply the proposed framework and estimation method to the data set on the provision of subsidized long-lasting insecticide-treated bed net (LL-ITN) in Kenya (see Dupas (2009) and Bhattacharya and Dupas (2012) for more details). Malaria causes around one million



(a) $(N_1, N_2) = (600, 1000)$ (b) $(N_1, N_2) = (1000, 600)$ (c) $(N_1, N_2) = (1000, 1000)$

Figure 3: Averaged Estimates of $\rho(\theta, \phi^*)$ (Experiment 2)



(a) $(N_1, N_2) = (600, 1000)$ (b) $(N_1, N_2) = (1000, 600)$ (c) $(N_1, N_2) = (1000, 1000)$

Figure 4: Averaged Estimates of $\rho(\theta, \phi^*)$ (Experiment 3)

deaths every year, of which an overwhelming proportion occurs in Africa. In regions of African, malaria has been one of the leading causes for the high childhood mortality. Moreover, it places an enormous economic burden on affected countries and has a highly detrimental effect on economic and social development.[4] LL-ITN is the most prominent malaria preventive measure for large-scale deployment in highly endemic areas. It has been documented in the literature that adoption of LL-ITN has reduced the child mortality by up to 38%, see Lengeler (2004) for a review. Moreover, adoption of LL-ITN can help reduce cost associated with medical treatment of malaria as well as loss in family income after being affected.

[4]Ettling et al. (1994) find that treating malaria episodes accounts for 28% of cash income for affected households in Malawi.

However, research has also suggested that the demand for malaria prevention is very price-sensitive, despite the large private return to preventing it. A LL-ITN costs between 5 and 7 U.S. dollars, which is not affordable for most of the African families (Cohen and Dupas, 2010). Subsidized program is one potential way of solving this problem. However, Teklehaimanot et al. (2007) calculate that providing one free LL-ITN for every two at-risk person in sub-Saharan Africa would cost 2.5 billion U.S. dollars, which is much higher than the available fund. In this case, the central budget officer faces the problem of allocating the total budget to different programs, and the coordinator of a specific program faces the problem of allocating limited money obtained from the central budget office to determine subsidies to different households based on their characteristics.

The data is from a randomized experiment conducted with rural households in Western Kenya in 2007 (Dupas, 2009; Bhattacharya and Dupas, 2012). Each household was given a random subsidized price at which they could purchase an LL-ITN, where the price takes values from \$0 to \$4, with increment of \$0.5. In our empirical application, we follow Bhattacharya and Dupas (2012) in defining the treatment and outcome variables. In particular, for the treatment variable, $A = 1$ if the household was assigned a low price (\$0 or \$0.5), and $A = 0$ if it was assigned a high price of \$2 or more. Therefore, the treatment is whether the household has access to a low price LL-ITN. For the outcome variable, $Y = 1$ if the household has redeemed the coupon and had started using the bed net at the time of a follow-up visit, and 0 otherwise. We divide our data set into two subsamples according to the criterion that whether there is a child under 10 in the household, and treat these two subsamples as from two different programs. The summary statistics for data from each program are provided in Table 2.

Given the sample information, we consider the budget allocation problem. Suppose that in the next wave of applications, the number of applicants for each program doubles, then $N_\ell = 2n_\ell$ for $\ell = 1, 2$, such that $N_1 + N_2 = 2016$. For the choice of cost parameters c_1, c_2, Bhattacharya and Dupas (2012) document that the net is roughly 5-7 U.S. dollars each, and we set $c_1 = c_2 = 6$. To subsidize free LL-ITN for all the applicants, the central budget officer would need $\bar{b} = (N_1 + N_2) \times 6$ dollars. Suppose that at most $n_1 + n_2$ households can be subsidized (one bed net for each household), and let $b = \alpha\bar{b}$, where $\alpha \in (0, 0.5)$ controls the actual fund available for both programs. In the empirical analysis, we set $\alpha = 0.25$ and 0.35 for the purpose of illustration.

In Figure 5, we present the estimated value frontier functions as well as the estimated

Program 1: with a child under 10	Min	Max	Mean	SD
Treatment	0	1	0.142	0.349
Outcome	0	1	0.164	0.370
Household Size	3	17	7.190	2.236
Wealth per Capita (US$)	8.564	158.195	45.018	28.037
Observations	543			
Program 2: without child under 10	Min	Max	Mean	SD
Treatment	0	1	0.183	0.387
Outcome	0	1	0.159	0.366
Household Size	1	31	6.647	3.044
Wealth per Capita (US$)	8.623	216.817	49.403	38.834
Observations	465			

Table 2: Summary Statistics

optimal budget splitting points and their confidence intervals when $\alpha = 0.25, 0.35$. In the left panel, only 25% of total needed fund is available. As expected, the estimated value frontier function is a humped curve. It peaks at 0.667, with the 90% and 95% confidence intervals to be $[0.541, 0.792]$ and $[0.517, 0.816]$, respectively.[5] This suggests to the budget officer that by taking into account information from previous treatment programs, and to maximize the social value, it is optimal to allocate 66.7% of available fund to the program in which each family has a child under 10. Similarly, in the right panel, the value frontier function peaks at 0.646, with the 90% and 95% confidence intervals to be $[0.548, 0.745]$ and $[0.529, 0.764]$, respectively. The estimated optimal budget splitting points in these two scenarios are both above 0.5, suggesting that families with a child under 10 are in more need of LL-ITN protection. Therefore, it is beneficial for the society to allocate more budget to the first program.

[5]The confidence interval is based on standard error estimated from the jackknife method. For instance, to estimate standard deviation (std) of the estimated θ^*, we just need to implement the leave-one-out estimate n times to obtain $\widehat{\theta}_{-i}$ with $i = 1, \ldots, n$, where $\widehat{\theta}_{-i}$ denotes the estimate that we obtain by maximizing (3.2) without the i^{th} individual. Then the estimated std is defined by $se = \left(\frac{n-1}{n}\sum_{i=1}^{n}(\widehat{\theta}_{-i} - \bar{\theta})^2\right)^{1/2}$, where $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{-i}$.

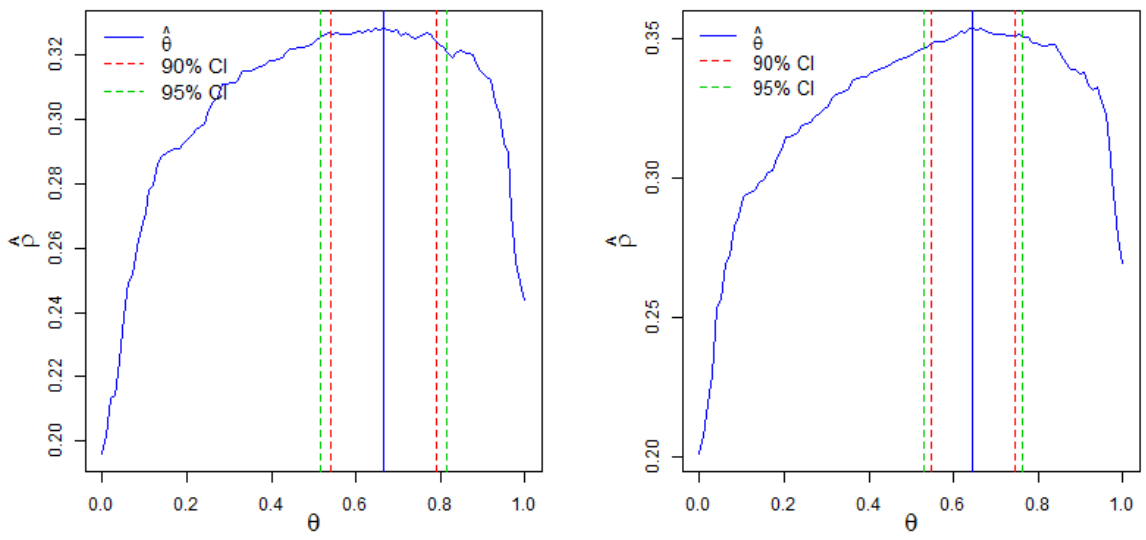


Figure 5: $\widehat{\rho}(\theta, \widehat{\phi})$, $\widehat{\theta}$ and Confidence Intervals (Left: $\alpha = 0.25$, Right: $\alpha = 0.35$)

6 Concluding Remarks

In this paper, we study the problem of allocating a limited budget to different social treatment programs. The optimal budget allocation is defined as an optimal splitting point that maximizes a social value function (called the value frontier function), weighted across different programs. We propose estimators for the value frontier function and the optimal splitting point. Monte Carlo experiments are conducted to evaluate the finite sample performances of our estimators.

We use our framework and econometric method to study the optimal budget allocation in the treatment programs of providing subsidized long-lasting insecticide-treated bed net in Kenya to fight malaria. Our estimated value for the optimal budget splitting point suggests that the central budget officer should allocate a larger percentage of the budget to the treatment program whose household applicants have a child under 10 than the program whose household applicants do not have. This finding is in line with the intuition that families with a child under 10 are in more need of LL-ITN protection, and we provide quantitative suggestions on the exact allocation scheme.

Beyond the LL-ITN treatment programs in this paper, our framework can be readily used in other empirical applications, such as a local education bureau's allocation of budget to different local schools, a university's budget allocation to different departments, the government's job training programs in different regions, and so on. An empirical work in either of these

applications would be of great interest for future research.

Appendix A

This appendix provides some preliminary lemmas and the proofs of the main results of this paper.

Proof of Lemma 3.1. *Step 1*: Firstly, we show that

$$\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\rho(\theta,\phi^*)\right|=\sup_{\theta\in\Theta}|\widehat{\rho}(\theta,\phi^*)-\rho(\theta,\phi^*)|+o_P(1). \tag{A.1}$$

Since

$$\begin{aligned}
&\sup_{\theta\in\Theta}|\widehat{\rho}(\theta,\phi^*)-\rho(\theta,\phi^*)|-\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\widehat{\rho}(\theta,\phi^*)\right| \\
\leq\ &\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\rho(\theta,\phi^*)\right| \\
\leq\ &\sup_{\theta\in\Theta}|\widehat{\rho}(\theta,\phi^*)-\rho(\theta,\phi^*)|+\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\widehat{\rho}(\theta,\phi^*)\right|,
\end{aligned} \tag{A.2}$$

it suffices to show that $\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\widehat{\rho}(\theta,\phi^*)\right|=o_P(1)$. Notice that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\widehat{\rho}(\theta,\phi^*)\right| \\
=\ &\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\rho(\theta,\widehat{\phi})+\rho(\theta,\widehat{\phi})-\rho(\theta,\phi^*)+\rho(\theta,\phi^*)-\widehat{\rho}(\theta,\phi^*)\right| \\
\leq\ &\sup_{\theta\in\Theta}\left|\widehat{\rho}(\theta,\widehat{\phi})-\rho(\theta,\widehat{\phi})\right|+\sup_{\theta\in\Theta}\left|\rho(\theta,\widehat{\phi})-\rho(\theta,\phi^*)\right|+\sup_{\theta\in\Theta}|\rho(\theta,\phi^*)-\widehat{\rho}(\theta,\phi^*)| \\
\equiv\ &A_1+A_2+A_3,
\end{aligned} \tag{A.3}$$

where the definitions of A_1 to A_3 should be obvious.

For A_1 and A_3, it suffices to show that $\sup_{(\theta,\phi)\in\Theta\times\Phi}|\widehat{\rho}(\theta,\phi)-\rho(\theta,\phi)|=o_P(1)$. We now use Lemma A2 of Newey and Powell (2003) to prove this argument.

Step 1.1: $\Theta\times\Phi$ is a compact set with respect to norm $\|\cdot\|_2$, which verifies the first condition of Lemma A2 of Newey and Powell (2003).

Step 1.2: By the fourth condition of this lemma, we have, for $\forall(\theta,\phi)\in\Theta\times\Phi$, $\widehat{\rho}(\theta,\phi)\to_P\rho(\theta,\phi)$, which verifies the second condition of Lemma A2 of Newey and Powell (2003).

Step 1.3: By the third condition of this lemma, we have

$$|\rho(\theta_1,\phi_1)-\rho(\theta_2,\phi_2)|\leq O(1)\|(\theta_1,\phi_1)-(\theta_2,\phi_2)\|_2.$$

In connection with *Step 1.2*, we have

$$\begin{aligned}
|\widehat{\rho}(\theta_1,\phi_1)-\widehat{\rho}(\theta_2,\phi_2)| &= (1+o_P(1))|\rho(\theta_1,\phi_1)-\rho(\theta_2,\phi_2)| \\
&\leq O_P(1)\|(\theta_1,\phi_1)-(\theta_2,\phi_2)\|_2,
\end{aligned}$$

which verifies the third condition of Lemma A2 of Newey and Powell (2003).

Thus, we have

$$\sup_{(\theta,\phi)\in\Theta\times\Phi} |\widehat{\rho}(\theta,\phi) - \rho(\theta,\phi)| = o_P(1), \tag{A.4}$$

which immediately implies that $A_1 = o_P(1)$ given that $\widehat{\theta}$ falls in Θ with probability approaching one and $A_3 = o_P(1)$.

We now deal with A_2. By the third condition of this lemma and the construction of $\|\cdot\|_2$,

$$A_2 = \sup_{\theta\in\Theta} \left|\rho(\theta,\widehat{\phi}) - \rho(\theta,\phi^*)\right| \leq O_P(1)\|\widehat{\phi} - \phi^*\|_{L^2} = o_P(1),$$

where the last equality follows from condition 5 of this lemma.

By analyses of A_1, A_2 and A_3, we can conclude that

$$\sup_{\theta\in\Theta} \left|\widehat{\rho}(\theta,\widehat{\phi}) - \widehat{\rho}(\theta,\phi^*)\right| = o_P(1),$$

which indicates (A.1) holds.

Then, we have completed the proof of *Step 1*.

Step 2:

Notice that

$$\widehat{\rho}(\theta,\widehat{\phi}) = \widehat{\rho}(\theta,\widehat{\phi}) - \rho(\theta,\phi^*) + \rho(\theta,\phi^*). \tag{A.5}$$

We now show that $\sup_{\theta\in\Theta} |\widehat{\rho}(\theta,\widehat{\phi}) - \rho(\theta,\phi^*)| = o_P(1)$. By *Step 1*, write

$$\sup_{\theta\in\Theta} |\widehat{\rho}(\theta,\widehat{\phi}) - \rho(\theta,\phi^*)| = \sup_{\theta\in\Theta} |\widehat{\rho}(\theta,\phi^*) - \rho(\theta,\phi^*)| + o_P(1) = o_P(1),$$

where the first equality follows from (A.1); and the second equality follows from (A.4).

Thus, we can further write (A.5) as

$$\widehat{\rho}(\theta,\widehat{\phi}) = \rho(\theta,\phi^*) + o_P(1). \tag{A.6}$$

By condition 2 of this lemma and (A.6), we know that if $\widehat{\theta} \not\to_P \theta^*$, there is a point θ^* satisfying that $\widehat{\rho}(\theta^*,\widehat{\phi}) < \widehat{\rho}(\widehat{\theta},\widehat{\phi})$ with probability approaching one, which violates the definition of $\widehat{\theta}$. Thus, we must have $\widehat{\theta} \to_P \theta^*$.

Based on *Step 1* and *Step 2*, the proof is now complete. ■

Lemma A.1 *Let $\{(y_i, x_i) \mid i = 1, \ldots, n\}$ be i.i.d. across i, and $\mathbb{E}[y_i|x_i] = \mu(x_i)$. Suppose the following conditions hold:*

1. *Let $\mu(x) \in L^2(\mathcal{X}, \varphi(x))$, where x is a $d \times 1$ vector, $\mathcal{X}$ is the support of x, and $\varphi(x)$ is the density of the Hilbert space $L^2(\mathcal{X}, \varphi(x))$.*

2. *Assume there exists an orthonormal function sequence $\{p_j(x)\}_{j=0}^{\infty}$ in the space $L^2(\mathcal{X}, \varphi(x))$ such that $\sup_{j\geq 0} \sup_{x\in\mathcal{X}} |p_j(x)\varphi^{1/2}(x)| < \infty$, so that*

$$\mu(x) = \sum_{j=0}^{\infty} c_j p_j(x) \equiv P_K(x)'C + \gamma_K(x),$$

where $P_K(x) = (p_0(x), p_1(x) \ldots, p_{K-1}(x))'$, $C = (c_0, \ldots, c_{K-1})'$ and $\gamma_K(x) = \sum_{j=K}^{\infty} c_j p_j(x)$. As $(K, n) \to (\infty, \infty)$, $K^2/n \to 0$. Moreover, assume that $\sum_{j=K}^{\infty} |c_j| = O(K^{-r/2})$, where r is a positive constant and $r \geq 2$.

3. *Let $f(x)$ be the density of $\{x_i\}$. Suppose that $\sup_{x\in\mathcal{X}} |f(x)/\varphi(x)| \leq \alpha_0 < \infty$.*

4. *Let $\Omega_K = \mathbb{E}[\varphi(x_1)P_K(x_1)P_K(x_1)']$. Assume $0 < \alpha_1 \leq \lambda_{min}(\Omega_K) \leq \lambda_{max}(\Omega_K) \leq \alpha_2 < \infty$ uniformly in K, where α_1 and α_2 are positive constants, and $\lambda_{min}(\Omega_K)$ and $\lambda_{max}(\Omega_K)$ stand for the minimum and maximum eigenvalues of Ω_K respectively.*

Define an objective function as

$$Q_n(C) = \frac{1}{n}\sum_{i=1}^{n}(y_i - P_K(x_i)'C)^2\varphi(x_i), \tag{A.7}$$

where $\varphi(\cdot)$ is defined in condition (1) and serves as a weight function. Then the corresponding estimator of C is given by

$$\widehat{C} = \arg\min\ Q_n(C) = \left(\sum_{i=1}^{n}\varphi(x_i)P_K(x_i)P_K(x_i)'\right)^{-1}\left(\sum_{i=1}^{n}\varphi(x_i)P_K(x_i)y_i\right). \tag{A.8}$$

Let $\widehat{\mu}(x) = P_K(x)'\widehat{C}$ and $\|\mu\|_{L^2} = \left\{\int_{\mathcal{X}} \mu^2(x)\varphi(x)dx\right\}^{1/2}$.

Based on the above conditions, we have

1. $\left\|\widehat{C} - C\right\| = O_P\left(\sqrt{\frac{K}{n}}\right) + O_P(K^{-r/2})$;

2. $\|\widehat{\mu} - \mu\|_{L^2} = O_P\left(\sqrt{\frac{K}{n}}\right) + O_P(K^{-r/2})$;

3. $\sup_{x\in\mathcal{X}} \varphi^{1/2}(x)\,|\widehat{\mu}(x) - \mu(x)| = O_P\left(\frac{K}{\sqrt{n}}\right) + O_P(K^{-r/2})$.

Remark A.3

1. *$\mathcal{X}$ can be $\mathbb{R}^d$, or a compact set of $\mathbb{R}^d$. It depends on the needs of the study. The current notation is in the same spirit of Dong and Linton (2016) and is designed to provide some generic results. For example, some widely used spaces (e.g., $L^2(\mathbb{R}^d)$, $L^2([0,1]^d)$, $L^2(\mathbb{R}, \exp(-w^2))$, etc.) are certainly captured under our setting. For $d > 1$, the curse of dimensionality kicks in through K (due to the use of tensor product usually when constructing $P_K(\cdot)$).*

2. *Note that $\varphi(\cdot)$ serves as a weight function. Without such a weighted least squares approach, more restrictive conditions have to be imposed on the density of x_i in order to ensure further development is possible for those basis like Hermite polynomials, which are non-integrable functions, and are defined on the whole real line. See Dong and Linton (2016) for more detailed discussions.*

3. *Condition 2 of this lemma is in the same spirit as assumptions of Section 6.1 of Hansen (2015).*

4. *The third result is consistent with result (2) of Theorem 1 of Newey (1997). Note that for $L^2(\mathbb{R})$ space, the third result will reduce to that $\sup_{\mathbb{R}} |\widehat{\mu}(x) - \mu(x)| = O_P\left(\frac{K}{\sqrt{n}}\right) + O_P(K^{-r/2})$.*

Proof of Lemma A.1.

(1). We point out two simple facts before investigating $\widehat{C}$. By condition (2), it is easy to know that

$$\|\gamma_K(x)\|_{L^2}^2 = \int_{\mathbb{R}^d} \gamma_K^2(x)\varphi(x)dx = \sum_{j=K}^{\infty} c_j^2 = O(K^{-r}) \tag{A.9}$$

and

$$\mathbb{E}\,\|\gamma_K(x_1)\|^2 = \int_{\mathbb{R}^d} \|\gamma_K(x)\|^2 \cdot \frac{f(x)}{\varphi(x)} \cdot \varphi(x)dx = O(K^{-r}). \tag{A.10}$$

We now start considering $\widehat{C}$, and simple algebra gives

$$\begin{aligned}
\widehat{C} - C &= \left(\sum_{i=1}^n \varphi(x_i)P_K(x_i)P_K(x_i)'\right)^{-1} \left(\sum_{i=1}^n \varphi(x_i)P_K(x_i)e_i\right) \\
&\quad + \left(\sum_{i=1}^n \varphi(x_i)P_K(x_i)P_K(x_i)'\right)^{-1} \left(\sum_{i=1}^n \varphi(x_i)P_K(x_i)\gamma_K(x_i)\right) \\
&\equiv A_1 + A_2,
\end{aligned}$$

where $e_i = y_i - \mu(x_i)$ stands for the error term, and the definitions of A_1 and A_2 should be obvious.

We start from considering A_1 now.

$$\begin{aligned} & \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)P_K(x_i)P_K(x_i)' - \Omega_K\right\|^2 \\ = & \frac{1}{n}\sum_{j_1=1}^{K}\sum_{j_2=1}^{K}\mathbb{E}|\varphi(x_1)p_{j_1}(x_1)p_{j_2}(x_1) - \Omega_{K,j_1j_2}|^2 \\ \leq & O(1)\frac{1}{n}\sum_{j_1=1}^{K}\sum_{j_2=1}^{K}\mathbb{E}|\varphi(x_1)p_{j_1}(x_1)p_{j_2}(x_1)|^2 = O\left(\frac{K^2}{n}\right) \end{aligned} \tag{A.11}$$

where Ω_{K,j_1j_2} stands for the $(j_1,j_2)^{th}$ element of Ω_K.

Secondly,

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)P_K(x_i)e_i\right\|^2 = O(1)\frac{1}{n^2}\sum_{j=1}^{K}\sum_{i=1}^{n}\mathbb{E}\left[\varphi^2(x_i)p_j^2(x_i)\right] = O\left(\frac{K}{n}\right),$$

which indicates that $\left\|\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)P_K(x_i)e_i\right\| = O_P\left(\sqrt{\frac{K}{n}}\right)$.

In connection with condition (4) of this lemma, we can conclude that $\|A_1\| = O_P\left(\sqrt{\frac{K}{n}}\right)$.

For A_2, we define the following variables for notational simplicity.

$$P_{nK} = (P_K(x_1),\ldots,P_K(x_n))' \quad \text{and} \quad \gamma_{nK} = (\gamma_K(x_1),\ldots,\gamma_K(x_n)).$$

Thus, write

$$\begin{aligned} \|A_2\|^2 &= \left\|\left(P_{nK}'P_{nK}\right)^{-1}P_{nK}'\gamma_{nK}\right\|^2 \\ &= \gamma_{nK}'P_{nK}\left(P_{nK}'P_{nK}/n\right)^{-1}\left(P_{nK}'P_{nK}\right)^{-1}P_{nK}'\gamma_{nK}/n \\ &\leq \left[\lambda_{min}\left(P_{nK}'P_{nK}/n\right)\right]^{-1}\cdot\gamma_{nK}'P_{nK}\left(P_{nK}'P_{nK}\right)^{-1}P_{nK}'\gamma_{nK}/n \\ &\leq \left[\lambda_{min}\left(P_{nK}'P_{nK}/n\right)\right]^{-1}\cdot\lambda_{max}(W)\cdot\left\|\gamma_{nK}\right\|^2/n, \end{aligned}$$

where the first inequality follows from the exercise 5 on page 267 of Magnus and Neudecker (2007). Note that $W = P_{nK}\left(P_{nK}'P_{nK}\right)^{-1}P_{nK}'$ is symmetric and idempotent, so $\lambda_{max}(W) = 1$. By (A.10), it is easy to know that $\mathbb{E}[\|\gamma_{nK}\|^2/n] = O\left(K^{-r}\right)$. In connection with (A.11) and condition (4) of this lemma, we immediately obtain that $\|A_2\| = O_P(K^{-r/2})$.

Based on the above analyses, the first result follows.

(2). The second statement of this lemma follows from (A.9) and the first result of this lemma immediately.

(3). Write

$$\begin{aligned}\sup_{x\in\mathcal{X}}\varphi^{1/2}(x)\,|\widehat{\mu}(x)-\mu(x)| &\leq \sup_{x\in\mathcal{X}}\varphi^{1/2}(x)\left|P_K(x)'(\widehat{C}-C)\right|+\sup_{x\in\mathcal{X}}\varphi^{1/2}(x)|\gamma_K(x)| \\ &\leq O(1)\sqrt{K}\|\widehat{C}-C\|+O(1)\sum_{j=K}^{\infty}|c_j| \\ &= O(1)\sqrt{K}\|\widehat{C}-C\|+O(1)K^{-r/2},\end{aligned}$$

where the first inequality follows from triangle inequality; the second inequality follows from the fact that $\varphi^{1/2}(x)p_j(x)$ is uniformly bounded on $\mathcal{X}$; and the first equality follows from the second condition of this lemma. In connection with the first result of this lemma, the third argument follows immediately.

■

Lemma A.2 *Let $\widehat{Z}_i=\widehat{\delta}(X_i)$, and $\widehat{q}(\tau)$ be the empirical quantile function based on $\{\widehat{Z}_i\}_{i=1}^n$. Let $Z_i=\delta(X_i)$, and $q(\tau)$ be the population quantile function of Z_i. Under Assumptions 1-5, $\sup_{\tau\in(0,1)}|\widehat{q}(\tau)-q(\tau)|=o_P(1)$.*

Proof of Lemma A.2. Denote that

$$F_{\widehat{z}}(z)=\frac{1}{n}\sum_{i=1}^{n}1(\widehat{Z}_i<z)\quad\text{and}\quad F_z(z)=\frac{1}{n}\sum_{i=1}^{n}1(Z_i<z).$$

Firstly, we show that $\sup_z|F_{\widehat{z}}(z)-F_z(z)|=o_P(1)$. Write

$$\begin{aligned}\sup_z|F_{\widehat{z}}(z)-F_z(z)| &\leq \frac{1}{n}\sum_{i=1}^{n}\sup_z\left|1(\widehat{Z}_i<z)-1(Z_i<z)\right| \\ &\leq \sup_{i,z}\left|1(\widehat{Z}_i-z<0)-1(Z_i-z<0)\right|.\end{aligned}$$

We then focus on $\sup_{i,z}\left|1(\widehat{Z}_i-z<0)-1(Z_i-z<0)\right|$ below.

Note that we have $\sup_x|\widehat{\delta}(x)-\delta(x)|=o_P(1)$ by Lemma A.1. It allows us to write

$$o_P(1)=\sup_i|\widehat{\delta}(X_i)-\delta(X_i)|=\sup_i|\widehat{Z}_i-Z_i|=\sup_{i,z}|(\widehat{Z}_i-z)-(Z_i-z)|,$$

which indicates that $\widehat{Z}_i-z$ and Z_i-z have the same sign with probability one uniformly in i and z. Thus, $\sup_{i,z}|1(\widehat{Z}_i-z<0)-1(Z_i-z<0)|=o_P(1)$. Then we can conclude that

$$\sup_z|F_{\widehat{z}}(z)-F_z(z)|\leq\sup_{i,z}\left|1(\widehat{Z}_i<z)-1(Z_i<z)\right|=o_P(1).$$

With the above uniform convergence in hand, we then can write $F_{\widehat{z}}(z)=F_z(z)\cdot(1+o_P(1))$. Thus, we just need to focus on the leading term while using Bahadur representation under Assumption 3 to

establish the convergence for quantile function. Further note that $\delta(X)$ belongs to a compact set of $\mathbb{R}$, so the PDF of Z_i is uniformly bounded away from 0. Thus, in view of Bahadur (1966, eq. 17 of p. 579), the result follows. ■

Proof of Theorem 3.1. Firstly, we point out some facts before we move on to the detailed proof. Note that the derivative of the Heaviside function $1(x \geq 0)$ is the Dirac delta function $\delta(x)$, i.e., $\frac{d}{dx}1(x \geq 0) = \delta(x)$, or equivalently, Heaviside function $1(x \geq 0)$ is the distribution function of the delta function. The property of the delta function gives that $\int \delta(x)dx = 1$ and $\int f(x)\delta(x)dx = f(0)$ for any continuous function $f(x)$. In the proof of this theorem, δ represents the Dirac delta function only.

In view of Assumptions 1-5 and Lemma 3.1, we now just need to verify the condition 3 of Lemma 3.1. By construction, it suffices to consider the continuity of the following function.

$$\rho(\theta, \mu_1, \mu_0, q) \quad = \quad \mathbb{E}\left[\mu_1(x) \cdot 1\left(\mu_1(x) - \mu_0(x) - q(1 - \nu\theta) \geq 0\right)\right],$$

where we have suppressed the some extra sub-indices.

In order to verify condition 3 of Lemma 3.1, it suffices to show that

$$\begin{aligned}
&(1). \quad |\rho(\theta_1, \mu_1, \mu_0, q) - \rho(\theta_2, \mu_1, \mu_0, q)| \leq O(1)|\theta_1 - \theta_2|,\\
&(2). \quad |\rho(\theta, \mu_{11}, \mu_0, q) - \rho(\theta, \mu_{12}, \mu_0, q)| \leq O(1)\|\mu_{11} - \mu_{12}\|_{L^2},\\
&(3). \quad |\rho(\theta, \mu_1, \mu_{01}, q) - \rho(\theta, \mu_1, \mu_{02}, q)| \leq O(1)\|\mu_{01} - \mu_{02}\|_{L^2},\\
&(4). \quad |\rho(\theta, \mu_1, \mu_{01}, q_1) - \rho(\theta, \mu_1, \mu_{02}, q_2)| \leq O(1)\|q_1 - q_2\|_{L^2}.
\end{aligned}$$

Start from (1), and write

$$\begin{aligned}
& |\rho(\theta_1, \mu_1, \mu_0, q) - \rho(\theta_2, \mu_1, \mu_0, q)|\\
\leq \;& \mathbb{E}\left[\mu_1(x) \cdot \{1\left(\mu_1(x) - \mu_0(x) - q(1 - \nu\theta_1) \geq 0\right) - 1\left(\mu_1(x) - \mu_0(x) - q(1 - \nu\theta_2) \geq 0\right)\}\right]\\
\leq \;& O(1)\mathbb{E}\left|1\left(\mu_1(x) - \mu_0(x) - q(1 - \nu\theta_1) \geq 0\right) - 1\left(\mu_1(x) - \mu_0(x) - q(1 - \nu\theta_2) \geq 0\right)\right|\\
\leq \;& O(1)|q(1 - \nu\theta_2) - q(1 - \nu\theta_1)| \cdot \mathbb{E}|\delta(\mu_1(x) - \mu_0(x) - q^*)|\\
= \;& O(1)|q(1 - \nu\theta_2) - q(1 - \nu\theta_1)| \cdot \int \delta(\mu_1(x) - \mu_0(x) - q^*)f(x)dx\\
\leq \;& O(1)|q(1 - \nu\theta_2) - q(1 - \nu\theta_1)| = O(1)|\theta_1 - \theta_2|,
\end{aligned}$$

where q^* lies between $\mu_1(x) - \mu_0(x) - q(1 - \nu\theta_1)$ and $\mu_1(x) - \mu_0(x) - q(1 - \nu\theta_2)$; $f(\cdot)$ denotes the PDF of x; the third inequality follows from Mean-Value theorem; and the last equality follows from the proof of Lemma A.2.

Similarly, for (3), we have

$$\begin{aligned}
& |\rho(\theta,\mu_1,\mu_{01},q)-\rho(\theta,\mu_1,\mu_{02},q)| \\
\leq\ & \mathbb{E}\left[\mu_1(x)\cdot\{1\left(\mu_1(x)-\mu_{01}(x)-q(1-\nu\theta)\geq 0\right)-1\left(\mu_1(x)-\mu_{02}(x)-q(1-\nu\theta)\geq 0\right)\}\right] \\
\leq\ & O(1)\mathbb{E}\left|1\left(\mu_1(x)-\mu_{01}(x)-q(1-\nu\theta)\geq 0\right)-1\left(\mu_1(x)-\mu_{02}(x)-q(1-\nu\theta)\geq 0\right)\right| \\
\leq\ & O(1)\mathbb{E}\left[|\mu_{01}(x)-\mu_{02}(x)|\cdot|\delta(\mu_1(x)-\mu_0^*-q(1-\nu\theta))|\right] \\
\leq\ & O(1)\sup_x|\mu_{01}(x)-\mu_{02}(x)|\cdot\mathbb{E}|\delta(\mu_1(x)-\mu_0^*-q(1-\nu\theta))| \\
\leq\ & O(1)\|\mu_{01}(x)-\mu_{02}(x)\|_{L^2}\cdot\mathbb{E}|\delta(\mu_1(x)-\mu_0^*-q(1-\nu\theta))| \\
=\ & O(1)\|\mu_{01}(x)-\mu_{02}(x)\|_{L^2},
\end{aligned}$$

where μ_0^* lies between $\mu_{01}(x)$ and $\mu_{02}(x)$; and the fifth inequality follows from Assumption 4.i.

The proofs of (2) and (4) follow in the similar fashion as the above, so omitted. With the continuity in hand, the proof follows from Lemma 3.1 immediately. ■

Proof of Theorem 3.2. By Taylor expansion, we have

$$0=\nabla_\theta\widehat{\rho}(\widehat{\theta},\widehat{\phi})=\nabla_\theta\widehat{\rho}(\theta^*,\widehat{\phi})+\nabla^2_{\theta\theta}\widehat{\rho}(\widetilde{\theta},\widehat{\phi})(\widehat{\theta}-\theta^*),$$

where $\widetilde{\theta}$ is between $\widehat{\theta}$ and θ^*. This implies that

$$\widehat{\theta}-\theta^*=-\left[\nabla^2_{\theta\theta}\widehat{\rho}(\widetilde{\theta},\widehat{\phi})\right]^{-1}\nabla_\theta\widehat{\rho}(\theta^*,\widehat{\phi}).$$

Write

$$\nabla_\theta\widehat{\rho}(\theta^*,\widehat{\phi})=[\nabla_\theta\widehat{\rho}(\theta^*,\phi^*)-\nabla_\theta\rho_n(\theta^*,\phi^*)]+\nabla_\theta\rho_n(\theta^*,\phi^*)+\left[\nabla_\theta\widehat{\rho}(\theta^*,\widehat{\phi})-\nabla_\theta\widehat{\rho}(\theta^*,\phi^*)\right],$$

where

$$\begin{aligned}
\nabla_\theta\widehat{\rho}(\theta,\phi)&=\sum_{\ell=1,2}o_\ell(\theta,\phi)\mathbb{P}_{n\ell}[T_{n\ell}(\theta,\phi)],\ \nabla_\theta\rho_n(\theta,\phi)=\sum_{\ell=1,2}o_\ell(\theta,\phi)\mathbb{E}[T_{n\ell}(\theta,\phi)],\\
o_\ell(\theta,\phi)&=w_\ell\nu_\ell q'_\ell\left(1-1(\ell=1)\nu_1\theta-1(\ell=2)\nu_2(1-\theta)\right),\\
T_{n\ell}(\theta,\phi)&=\sum_{a=0,1}\frac{(-1)^{a+\ell}}{h_\ell}\mu_{\ell_a}(X_\ell)k_{\ell_a}(X_\ell;\theta,\phi)\equiv\sum_{a=0,1}T_{n\ell_a}(\theta,\phi),\\
k_{\ell_a}(X_\ell;\theta,\phi)&=k\left(\frac{(-1)^{a+1}[\delta_\ell(X_\ell)-e_\ell(\theta,\phi)]}{h_\ell}\right),\\
e_\ell(\theta,\phi)&=q_\ell\left(1-1(\ell=1)\nu_1\theta-1(\ell=2)\nu_2(1-\theta)\right),
\end{aligned}$$

and $k(\cdot)=K^{(1)}(\cdot)$. For notational simplicity, in the following, we denote

$$\nabla_\theta\widehat{\rho}(\theta)=\nabla_\theta\widehat{\rho}(\theta,\widehat{\phi}),\quad\nabla_\theta\widetilde{\rho}(\theta)=\nabla_\theta\widehat{\rho}(\theta,\phi^*),\quad\nabla_\theta\rho_n(\theta)=\nabla_\theta\rho_n(\theta,\phi^*).$$

The claim follows from three steps: (1) $\sqrt{\min(n_1h_1, n_2h_2)}\,[\nabla_\theta\widetilde{\rho}(\theta^*) - \nabla_\theta\rho_n(\theta^*)] \to_D N(0,\sigma^2)$, where σ^2 is defined in (A.13); (2) $\sqrt{\min(n_1h_1, n_2h_2)}\nabla_\theta\rho_n(\theta^*) = o(1)$; (3) $\nabla^2_{\theta\theta}\widehat{\rho}(\widetilde{\theta}) - \nabla^2_{\theta\theta}\rho_\infty(\theta^*) = o_P(1)$.

Step (1): to show that

$$\sqrt{\min(n_1h_1, n_2h_2)}\,[\nabla_\theta\widetilde{\rho}(\theta^*) - \nabla_\theta\rho_n(\theta^*)] \to_D N(0,\sigma^2).$$

First, it is easy to see that $\mathbb{E}[\nabla_\theta\widetilde{\rho}(\theta^*) - \nabla_\theta\rho_n(\theta^*)] = 0$. Second, let $o^*_\ell = o_\ell(\theta^*,\phi^*)$ and write

$$\nabla_\theta\widetilde{\rho}(\theta^*) = \sum_{\ell=1,2} o^*_\ell \mathbb{P}_{n_\ell}[T^*_{n_\ell}],$$

where $T^*_{n_\ell} = T_{n_\ell}(\theta^*,\phi^*) \equiv \sum_{a=0,1} T^*_{n_{\ell_a}}$. Then

$$\mathrm{Var}[\nabla_\theta\widetilde{\rho}(\theta^*)] = \sum_{\ell=1,2} \frac{o^{*2}_\ell}{n_\ell}\mathrm{Var}[T^*_{n_\ell}]$$

under Assumption 1 (i).

Let $e^*_\ell = e_\ell(\theta^*,\phi^*)$, $(U_{\ell_1}, U_{\ell_0}) = (\mu^*_{\ell_1}(X_\ell), \mu^*_{\ell_0}(X_\ell))$ and $f_\ell(u_0,u_1)$ be its joint density, then

$$\begin{aligned}
\mathbb{E}[T^*_{n_{\ell_1}}] &= \frac{(-1)^{1+\ell}}{h_\ell}\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}}\int_{\underline{u}_{\ell_0}}^{\overline{u}_{\ell_0}} u_1 k\left(\frac{u_1-u_0-e^*_\ell}{h_\ell}\right) f_\ell(u_0,u_1)du_0du_1 \\
&\overset{(1)}{=}(-1)^{1+\ell}\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}}\int_{\underline{v}_{\ell_0}(u_1)}^{\overline{v}_{\ell_0}(u_1)} u_1 k(-v_0) f_\ell(u_1-e^*_\ell+h_\ell v_0,u_1)dv_0du_1 \\
&\overset{(2)}{=}(-1)^{1+\ell}\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}}\int_{\underline{v}_{\ell_0}(u_1)}^{\overline{v}_{\ell_0}(u_1)} u_1 f_\ell(u_1-e^*_\ell,u_1) k(v_0)dv_0du_1 \\
&+ (-1)^{1+\ell}h_\ell\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}}\int_{\underline{v}_{\ell_0}(u_1)}^{\overline{v}_{\ell_0}(u_1)} u_1 k(v_0) f^{(1)}_{\ell u_0}(\widetilde{u}_1,u_1)v_0dv_0du_1 \\
&= (-1)^{1+\ell}\int_{\mathcal{U}_{\ell_1}} u_1 f_\ell(u_1-e^*_\ell,u_1)\left[K(\overline{v}_{\ell_0}(u_1)) - K(\underline{v}_{\ell_0}(u_1))\right]du_1 \\
&+ (-1)^{1+\ell}h_\ell\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}}\int_{\underline{v}_{\ell_0}(u_1)}^{\overline{v}_{\ell_0}(u_1)} u_1 k(v_0) f^{(1)}_{\ell u_0}(\widetilde{u}_1,u_1)v_0dv_0du_1 \\
&\overset{(3)}{=}(-1)^{1+\ell}\int_{\mathcal{U}_{\ell_1}} u_1 f_\ell(u_1-e^*_\ell,u_1)du_1 + o(1) + O(h_\ell) \equiv A_{\ell_1} + o(1) + O(h_\ell), \qquad \text{(A.12)}
\end{aligned}$$

where $\widetilde{u}_1$ is between $u_1 - e^*_\ell + h_\ell v_0$ and $u_1 - e^*_\ell$, $\underline{v}_{\ell_0}(u_1) = \frac{\underline{u}_{\ell_0}+e^*_\ell-u_1}{h_\ell}$, $\overline{v}_{\ell_0}(u_1) = \frac{\overline{u}_{\ell_0}+e^*_\ell-u_1}{h_\ell}$, $\mathcal{U}_{\ell_1} = \{u_1 \in [\underline{u}_{\ell_1}, \overline{u}_{\ell_1}] : \overline{v}_{\ell_0}(u_1) \geq 0, \underline{v}_{\ell_0}(u_1) \leq 0\}$ since it must be that $u_1 - e^*_\ell \in [\underline{u}_{\ell_0}, \overline{u}_{\ell_0}]$ for $f_\ell(u_1 - e^*_\ell, u_1)$ to be nonzero under Assumption 4 (i). (1) follows from a change of variables $-v_0 = \frac{u_1-u_0-e^*_\ell}{h_\ell}$; (2) follows from a first-order Taylor expansion; (3) follows from the Dominated Convergence Theorem, and that

under Assumption 4 (ii),

$$\begin{aligned}
&\left|\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}}\int_{\underline{v}_{\ell_0}(u_1)}^{\overline{v}_{\ell_0}(u_1)} u_1 k(v_0) f^{(1)}_{\ell u_0}(\widetilde{u}_1,u_1) v_0 dv_0 du_1\right| \\
&\leq \int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}} |u_1| \int_{\underline{v}_{\ell_0}(u_1)}^{\overline{v}_{\ell_0}(u_1)} k(v_0)|v_0||f^{(1)}_{\ell u_0}(\widetilde{u}_1,u_1)| dv_0 du_1 \\
&\leq \int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}} |u_1| \int_{-\infty}^{\infty} k(v_0)|v_0| O(1) dv_0 du_1 = O(1).
\end{aligned}$$

Similar calculations yield

$$\begin{aligned}
\mathbb{E}[T^*_{n\ell_0}] &= (-1)^\ell \int_{\underline{u}_{\ell_0}}^{\overline{u}_{\ell_0}}\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}} u_0 k\left(\frac{-u_1+u_0+e^*_\ell}{h_\ell}\right) f_\ell(u_0,u_1) du_1 du_0 \\
&= (-1)^\ell \int_{\mathcal{U}_{\ell_0}} u_0 f_\ell(u_0, u_0+e^*_\ell) du_0 + o(1) + O(h_\ell) \equiv A_{\ell_0} + o(1) + O(h_\ell), \\
\mathbb{E}[T^*_{n\ell_0} T^*_{n\ell_1}] &= \frac{(-1)^{1+2\ell}}{h^2_\ell} \int_{\underline{u}_{\ell_0}}^{\overline{u}_{\ell_0}}\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}} u_0 u_1 k\left(\frac{-u_1+u_0+e^*_\ell}{h_\ell}\right) k\left(\frac{u_1-u_0-e^*_\ell}{h_\ell}\right) f_\ell(u_0,u_1) du_1 du_0 \\
&= \frac{1}{h_\ell}\left[(-1)^{1+2\ell}\kappa_2 \int_{\mathcal{U}_{\ell_0}} u_0(u_0+e^*_\ell) f_\ell(u_0, u_0+e^*_\ell) du_0 + o(1)\right] + O(1) \\
&\equiv \frac{1}{h_\ell}\left(T^\dagger_{\ell_{01}} + o(1)\right) + O(1),
\end{aligned}$$

where $\mathcal{U}_{\ell_0} = \{u_0 \in [\underline{u}_{\ell_0}, \overline{u}_{\ell_0}] : \overline{v}_{\ell_1}(u_0) \geq 0, \underline{v}_{\ell_1}(u_0) \leq 0\}$ with $\underline{v}_{\ell_1}(u_0) = \frac{\underline{u}_{\ell_1}-u_0-e^*_\ell}{h_\ell}$, $\overline{v}_{\ell_1}(u_0) = \frac{\overline{u}_{\ell_1}-u_0-e^*_\ell}{h_\ell}$, since it must be that $u_0 + e^*_\ell \in [\underline{u}_{\ell_1}, \overline{u}_{\ell_1}]$ in order for $f_\ell(u_0, u_0+e^*_\ell)$ to be nonzero under Assumption 4 (i), and $\kappa_2 = \int k^2(v)dv$. Also,

$$\begin{aligned}
\mathbb{E}[T^{*2}_{n\ell_1}] &= \frac{1}{h^2_\ell}\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}}\int_{\underline{u}_{\ell_0}}^{\overline{u}_{\ell_0}} u_1^2 k\left(\frac{u_1-u_0-e^*_\ell}{h_\ell}\right) f_\ell(u_0,u_1) du_0 du_1 \\
&= \frac{1}{h_\ell}\left[\kappa_2 \int_{\mathcal{U}_{\ell_1}} u_1^2 f_\ell(u_1-e^*_\ell, u_1) du_1 + o(1)\right] + O(1) \equiv \frac{1}{h_\ell}\left(T^\dagger_{\ell_1} + o(1)\right) + O(1), \\
\mathbb{E}[T^{*2}_{n\ell_0}] &= \frac{1}{h^2_\ell}\int_{\underline{u}_{\ell_0}}^{\overline{u}_{\ell_0}}\int_{\underline{u}_{\ell_1}}^{\overline{u}_{\ell_1}} u_0^2 k\left(\frac{-u_1+u_0+e^*_\ell}{h_\ell}\right) f_\ell(u_0,u_1) du_1 du_0 \\
&= \frac{1}{h_\ell}\left[\kappa_2 \int_{\mathcal{U}_{\ell_0}} u_0^2 f_\ell(u_0, u_0+e^*_\ell) du_0 + o(1)\right] + O(1) \equiv \frac{1}{h_\ell}\left(T^\dagger_{\ell_0} + o(1)\right) + O(1).
\end{aligned}$$

Then

$$\begin{aligned}
\mathrm{Var}[T^*_{n\ell}] &= \mathbb{E}[T^{*2}_{n\ell_0}] - \mathbb{E}^2[T^*_{n\ell_0}] + \mathbb{E}[T^{*2}_{n\ell_1}] - \mathbb{E}^2[T^*_{n\ell_1}] \\
&\quad + 2\Big(\mathbb{E}[T^*_{n\ell_0} T^*_{n\ell_1}] - \mathbb{E}[T^*_{n\ell_0}]\mathbb{E}[T^*_{n\ell_1}]\Big) \\
&= \frac{1}{h_\ell}\Big(T^\dagger_{\ell_0} + o(1)\Big) + O(1) - [A_{\ell_0} + o(1)]^2 + \frac{1}{h_\ell}\Big(T^\dagger_{\ell_1} + o(1)\Big) + O(1) \\
&\quad - [A_{\ell_1} + o(1)]^2 + 2\Big(\frac{1}{h_\ell}(T^\dagger_{\ell_{01}} + o(1)) - [A_{\ell_0} + o(1)][A_{\ell_1} + o(1)]\Big) \\
&= \frac{1}{h_\ell}\Big(T^\dagger_{\ell_0} + T^\dagger_{\ell_1} + 2T^\dagger_{\ell_{01}} + o(1)\Big) + O(1) \equiv \frac{1}{h_\ell}\Big(T^\dagger_\ell + o(1)\Big) + O(1),
\end{aligned}$$

and

$$\begin{aligned}
\sigma_n^2 &\equiv \mathrm{Var}\left[\sqrt{\min(n_1h_1, n_2h_2)}\nabla_\theta\widetilde{\rho}(\theta^*)\right] \\
&= \min(n_1h_1, n_2h_2)\left(\frac{o_1^{*2}}{n_1}\mathrm{Var}[T^*_{n_1}] + \frac{o_2^{*2}}{n_2}\mathrm{Var}[T^*_{n_2}]\right) \\
&= \min\left(1, \frac{n_2h_2}{n_1h_1}\right) o_1^{*2}T_1^\dagger + o(1) + \min\left(1, \frac{n_1h_1}{n_2h_2}\right) o_2^{*2}T_2^\dagger + o(1) \\
&\quad + O\Big(\min(h_1, \frac{n_2h_2}{n_1})\Big) + O\Big(\min(\frac{n_1h_1}{n_2}, h_2)\Big) \\
&\to \min(1, \frac{1}{\kappa})o_1^{*2}T_1^\dagger + \min(1, \kappa)o_2^{*2}T_2^\dagger \equiv \sigma^2, \qquad \text{(A.13)}
\end{aligned}$$

where $\frac{n_1h_1}{n_2h_2} \to \kappa \in (0, \infty)$. Then convergence result follows from the Liapunov Central Limit Theorem.

Step (2): Write

$$\begin{aligned}
\nabla_\theta\rho_n(\theta^*) &= \nabla_\theta\rho_n(\theta^*) - \nabla_\theta\rho(\theta^*) + \nabla_\theta\rho(\theta^*) \\
&\overset{(1)}{=} \nabla_\theta\rho_n(\theta^*) - \nabla_\theta\rho(\theta^*) \\
&= \sum_{\ell=1,2}\sum_{a=0,1}\Big\{o_\ell^*\mathbb{E}[T^*_{n\ell_a}] - \nabla_\theta\rho_{\ell_a}(\theta^*, \phi_\ell^*)\Big\}
\end{aligned}$$

where $\rho_{\ell_a}(\theta, \phi_\ell) \equiv \mathbb{E}[\rho_{\ell_a}(X_\ell; \theta, \phi_\ell)]$, and (1) follows from that $\nabla_\theta\rho(\theta^*) = 0$ by definition. It suffices to show that $\sqrt{\min(n_1h_1, n_2h_2)}\Big\{o_\ell^*\mathbb{E}[T^*_{n\ell_a}] - \nabla_\theta\rho_{\ell_a}(\theta^*, \phi_\ell^*)\Big\} = o(1)$ for $\ell = 1, 2$ and $a = 0, 1$. We show this for the term with $\ell = 1, a = 1$, other terms can be shown in exactly the same way.

Write

$$\begin{aligned}
\rho_{1_1}(\theta, \phi^*) &= w_1\mathbb{E}\left[\mu^*_{1_1}(X_1)1(\delta_1^*(X_1) - q_1^*(1 - \nu_1\theta) > 0)\right] \\
&= w_1\int_{\underline{u}_{1_1}}^{\overline{u}_{1_1}}\int_{\underline{u}_{1_0}}^{\overline{u}_{1_0}} u_1 1(u_1 - u_0 - q_1^*(1 - \nu_1\theta) > 0) f_1(u_0, u_1)du_0du_1 \\
&= w_1\int_{\underline{u}_{1_1}}^{\overline{u}_{1_1}} u_1\int_{\underline{u}_{1_0}}^{u_1 - q_1^*(1-\nu_1\theta)} f_1(u_0, u_1)du_0du_1 \\
&= w_1\int_{\underline{u}_{1_1}}^{\overline{u}_{1_1}} u_1\Big[\Delta f_1(u_1 - q_1^*(1 - \nu_1\theta), u_1) - \Delta f_1(\underline{u}_{1_0}, u_1)\Big]du_1,
\end{aligned}$$

where $\Delta f_1(u_0,u_1)=\int f_1(u_0,u_1)du_0$. Then

$$\nabla_\theta \rho_{1_1}(\theta,\phi^*)=w_1\nu_1 q_1^{*\prime}(1-\nu_1\theta)\int_{\underline{u}_{1_1}}^{\overline{u}_{1_1}} u_1 f_1(u_1-q_1^*(1-\nu_1\theta),u_1)du_1,$$

and $\nabla_\theta \rho_{1_1}(\theta^*,\phi^*)=o_1^*\int_{\underline{u}_{1_1}}^{\overline{u}_{1_1}} u_1 f_1(u_1-e_1^*,u_1)du_1$. It is shown in (A.12) that

$$\mathbb{E}[T^*_{n_{1_1}}]=\int_{\underline{u}_{1_1}}^{\overline{u}_{1_1}} u_1 f_1(u_1-e_1^*,u_1)\left[K(\overline{v}_{1_0})-K(\underline{v}_{1_0})\right]du_1+O(h_1).$$

Then

$$\begin{aligned}
&\sqrt{\min(n_1h_1,n_2h_2)}\Big(o_1^*\mathbb{E}[T^*_{n_{1_1}}]-\nabla_\theta\rho_{1_1}(\theta^*,\phi^*)\Big)\\
&=o_1^*\int_{\underline{u}_{1_1}}^{\overline{u}_{1_1}} u_1 f_1(u_1-e_1^*,u_1)\sqrt{\min(n_1h_1,n_2h_2)}\left[K(\frac{\overline{u}_{1_0}+e_1^*-u_1}{h_1})-K(\frac{\underline{u}_{1_0}+e_1^*-u_1}{h_1})-1\right]du_1\\
&+O\left(\sqrt{\min(n_1h_1^3,n_2h_1^2h_2)}\right)\overset{(1)}{=}o(1)+O\left(\sqrt{\min(n_1h_1^3,n_2h_1^2h_2)}\right)\overset{(2)}{=}o(1),
\end{aligned}$$

where (1) follows from that K has a bounded support $[-1,1]$, the fact that $\frac{\overline{u}_{1_0}+e_1^*-u_1}{h_1}\geq 0$ and $\frac{\underline{u}_{1_0}+e_1^*-u_1}{h_1}\leq 0$, and Lemma A.3 below; (2) follows from Assumption 5 (ii).

Step (3): To show that $\nabla^2_{\theta\theta}\widehat{\rho}(\widetilde{\theta})-\xi=o_P(1)$, where $\xi=\nabla^2_{\theta\theta}\rho_\infty(\theta^*)=\lim_{h_1,h_2\to 0}\nabla^2_{\theta\theta}\rho_n(\theta^*)$, and $\nabla^2_{\theta\theta}\rho_n(\theta)=\nabla^2_{\theta\theta}\rho_n(\theta,\phi^*)$. Write

$$\begin{aligned}
|\nabla^2_{\theta\theta}\widehat{\rho}(\widetilde{\theta})-\nabla^2_{\theta\theta}\rho_\infty(\theta^*)|&=\left|\nabla^2_{\theta\theta}\widehat{\rho}(\widetilde{\theta},\widehat{\phi})-\nabla^2_{\theta\theta}\rho_\infty(\widetilde{\theta},\phi^*)+\nabla^2_{\theta\theta}\rho_\infty(\widetilde{\theta},\phi^*)-\nabla^2_{\theta\theta}\rho_\infty(\theta^*,\phi^*)\right|\\
&\leq\sup_{\theta\in\Theta}\left|\nabla^2_{\theta\theta}\widehat{\rho}(\theta,\widehat{\phi})-\nabla^2_{\theta\theta}\rho_\infty(\theta,\phi^*)\right|+\left|\nabla^2_{\theta\theta}\rho_\infty(\widetilde{\theta},\phi^*)-\nabla^2_{\theta\theta}\rho_\infty(\theta^*,\phi^*)\right|\\
&=o_P(1)+o_P(1)=o_P(1),
\end{aligned}$$

where the first $o_P(1)$ term follows from similar assumptions as in Lemma 3.1, and the second $o_P(1)$ term follows from the continuity of $\nabla^2_{\theta\theta}\rho_\infty(\theta)$ and that $\widetilde{\theta}-\theta^*=o_P(1)$.

Based on the above analysis, the result follows immediately. ■

Lemma A.3 *Let $G(h)=\int_{\underline{u}}^{\bar{u}} g_0(u)\gamma_h\left[K\left(\frac{g_1(u)}{h}\right)-K\left(-\frac{g_2(u)}{h}\right)-1\right]du$, where $K(\cdot)$ is a CDF function such that $K^{(1)}(\cdot)$ (i.e., the corresponding PDF) is defined on a compact set $[-1,1]$, and $\gamma_h\to\infty$ as $h\to 0$. For $j\in\{1,2\}$, $0<\inf_{u\in[\underline{u},\bar{u}]}g_j(u)\leq\sup_{u\in[\underline{u},\bar{u}]}g_j(u)<\infty$ and $\sup_{u\in[\underline{u},\bar{u}]}|g_0(u)|<\infty$. Then as $h\to 0$, $G(h)\to 0$.*

Proof of Lemma A.3. We use Lemma A2 of Newey and Powell (2003) again to show

$$\sup_{u\in[\underline{u},\bar{u}]}\left|g_0(u)\gamma_h\left[K\left(\frac{g_1(u)}{h}\right)-K\left(-\frac{g_2(u)}{h}\right)-1\right]\right|=o(1).$$

Then the result of this lemma follows immediately due to the fact that

$$|G(h)| \leq (\bar{u} - \underline{u}) \sup_{u \in [\underline{u}, \bar{u}]} \left| g_0(u)\gamma_h \left[K\left(\frac{g_1(u)}{h}\right) - K\left(-\frac{g_2(u)}{h}\right) - 1 \right] \right|. \tag{A.14}$$

Step 1: $[\underline{u}, \bar{u}]$ is a compact set.

Step 2: For $\forall u \in [\underline{u}, \bar{u}]$, it is easy to obtain that as $h \to 0$,

$$\left| g_0(u)\gamma_h \left[K\left(\frac{g_1(u)}{h}\right) - K\left(-\frac{g_2(u)}{h}\right) - 1 \right] \right| = o(1),$$

as $K^{(1)}(\cdot)$ is defined on $[-1, 1]$.

Step 3: In this case, the continuity condition holds apparently.

Thus, we have verified all the conditions of Lemma A2 of Newey and Powell (2003), and can conclude that $\sup_{u \in [\underline{u}, \bar{u}]} \left| g_0(u)\gamma_h \left[K\left(\frac{g_1(u)}{h}\right) - K\left(-\frac{g_2(u)}{h}\right) - 1 \right] \right| = o(1)$.

In connection with (A.14), we can further obtain $|G(h)| = o(1)$. The proof is complete. ■

References

Armstrong, T. B. and Shen, S. (2015), Inference on optimal treatment assignments. Cowles Foundation Discussion Paper No. 1927RR.

Bahadur, R. R. (1966), 'A note on quantiles in large samples', *The Annals of Mathematical Statistics* **36**(3), 577–580.

Bhattacharya, D. and Dupas, P. (2012), 'Inferring welfare maximizing treatment assignment under budget constraints', *Journal of Econometrics* **167**(1), 168–196.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997), 'Generalized partially linear single-index models', *Journal of the American Statistical Association* **92**(438), 477–489.

Chen, X. (2007), 'Large sample sieve estimation of semi-nonparametric models', *Handbook of Econometrics* pp. 5549–5632.

Chen, X., Linton, O. and Keilegom, I. V. (2003), 'Estimation of semiparametric models when the criterion function is not smooth', *Econometrica* **71**(5), 1591–1608.

Cohen, J. and Dupas, P. (2010), 'Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment', *The Quarterly Journal of Economics* **125**(1), 1–45.

Dehejia, R. H. (2005), 'Program evaluation as a decision problem', *Journal of Econometrics* **125**(1), 141–173.

Dong, C., Gao, J. and Tjøstheim, D. (2016), 'Estimation for single-index and partially linear single-index nonstationary time series models', *Annals of Statistics* **44**(1), 425–453.

Dong, C. and Linton, O. (2016), Additive nonparametric models with time variable and both stationary and nonstationary regressors. https://ssrn.com/abstract=2847681.

Dupas, P. (2009), 'What matters (and what does not) in households' decision to invest in malaria prevention?', *The American Economic Review Papers and Proceedings* **99**(2), 224–230.

Ettling, M., McFarland, D. A., Schultz, L. J. and Chitsulo, L. (1994), 'Economic impact of malaria in malawian households.', *Tropical Medicine and Parasitology* **45**(1), 74–79.

Gao, J., Tong, H. and Wolff, R. (2002), 'Model specification tests in nonparametric stochastic regression models', *Journal of Multivariate Analysis* **83**, 324–359.

Hall, P., Li, Q. and Racine, J. S. (2007), 'Nonparametric estimation of regresssion functions in the presence of irrelevant regressors', *The Review of Economics and Statistics* **89**(4), 784–789.

Hansen, B. (2015), A unified asymptotic distribution theory for parametric and non-parametric least squares. Working Paper.

Heckman, J. J. and Vytlacil, E. J. (2007*a*), 'Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation', *Handbook of Econometrics* **6**, 4779–4874.

Heckman, J. J. and Vytlacil, E. J. (2007*b*), 'Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments', *Handbook of Econometrics* **6**, 4875–5143.

Hirano, K. and Porter, J. R. (2009), 'Asymptotics for statistical treatment rules', *Econometrica* **77**(5), 1683–1701.

Horowitz, J. L. (1992), 'A smoothed maximum score estimator for the binary response model', *Econometrica* **60**(3), 505–531.

Imbens, G. W. and Wooldridge, J. M. (2009), 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature* **47**(1), 5–86.

Kitagawa, T. and Tetenov, A. (2015), Who should be treated? empirical welfare maximization methods for treatment choice.

Lengeler, C. (2004), 'Insecticide-treated bed nets and curtains for preventing malaria', *Cochrane Database of Systematic Reviews* .

Magnus, J. R. and Neudecker, H. (2007), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, third edn, John Wiley & Sons Ltd.

Manski, C. F. (2004), 'Statistical treatment rules for heterogeneous populations', *Econometrica* **72**(4), 1221–1246.

Newey, W. K. (1997), 'Convergence rates and asymptotic normality for series estimators', *Journal of Econometrics* **79**(1), 147–168.

Newey, W. K. and Powell, J. L. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Su, L. and Jin, S. (2012), 'Sieve estimation of panel data models with cross section dependence', *Journal of Econometrics* **169**(1), 34–47.

Teklehaimanot, A., McCord, G. C. and Sachs, J. D. (2007), 'Scaling up malaria control in africa: an economic and epidemiological assessment', *The American journal of tropical medicine and hygiene* **77**(6), 138–144.

Tetenov, A. (2012), 'Statistical treatment choice based on asymmetric minimax regret criteria', *Journal of Econometrics* **166**(1), 157–165.

Yu, Y. and Ruppert, D. (2002), 'Penalized spline estimation for partially linear single-index models', *Journal of the American Statistical Association* **97**(460), 1042–1054.