Efficient Semiparametric Estimation of Dose-Response Functions[*]

Matias D. Cattaneo[†]
UC-Berkeley

PRELIMINARY AND INCOMPLETE DRAFT
COMMENTS WELCOME

April 2007

Abstract. A large fraction of the literature on program evaluation focuses on the identification and estimation of binary treatment effects under the assumption of unconfoundedness. However, in many empirical applications treatments are multi-valued, rendering available estimation techniques inappropriate. This paper proposes a simple two-step efficient semiparametric procedure to estimate a large class of population parameters when a finite multi-valued treatment assignment is ignorable. Estimation focuses on a general population parameter, the Dose-Response Function (DRF), that relates each treatment level to its corresponding outcome effect and that is defined as the solution of a moment equation. We provide a set of sufficient conditions that ensure root-N consistency, asymptotic normality and efficiency of this estimator, and we show that these conditions are satisfied for two particular cases, the Average Dose-Response Function and the Quantile Dose-Response Function, under mild assumptions. Using these large sample results, other important population parameters of interest may be efficiently estimated by means of continuous transformations of the estimator considered. Using this idea, previous estimators for average and quantile treatments effect are shown to be particular cases of the proposed estimation procedure when treatment is assumed to be dichotomous.

JEL Classification: C14, C21.

Keywords: Dose-response functions, multi-valued treatment effects, generalized propensity score, inverse probability weighting, semiparametric efficiency, efficient estimation, unconfoundness.

[*] I specially thank Guido Imbens, Michael Jansson and Jim Powell for helpful advice and support. I am grateful to David Brillinger, Richard Crump, and Paul Ruud for valuable comments and suggestions. The usual disclaimers apply.

[†] Department of Economics, University of California at Berkeley. Email: cattaneo@econ.berkeley.edu. Website: http:/socrates.berkeley.edu/~cattaneo/.

1. Introduction

A large literature on program evaluation concentrates on the estimation of treatment effects under the assumption of unconfoundedness or ignorability and focuses almost exclusively on the special case of binary treatments. However, in most empirical applications treatments are implicitly or explicitly multi-valued in nature. For example, in training programs, participants receive different hours of training or, in conditional cash transfer programs, households receive different amounts of money. Understanding the effect of multi-valued treatment assignments is crucial from a policy-making perspective because it provides additional information beyond the standard treatment effects considered in the classical dichotomous treatment literature. For instance, by looking at the effect of a multi-valued treatment it is possible to identify non-linearities or heterogeneous treatment impacts, a fact that may provide a better understanding of the specific policy.

This paper proposes a simple two-step efficient semiparametric procedure to estimate a large class of population parameters when a finite multi-valued treatment assignment is ignorable. We study the estimation of a general population parameter, labeled the Dose-Response Function (DRF), that relates each treatment level (dose) to its corresponding outcome effect (response), and that is defined as the solution of a moment condition. We provide a set of sufficient conditions for the estimation of the general DRF, and we show that these conditions are satisfied by two important examples: Average Dose-Response Function (ADRF) and Quantile Dose-Response Function (QDRF). The first population parameter, ADRF, relates each treatment level to its average effect on the outcome of interest, while the second population parameter, QDRF, estimates for a given quantile the effect of each treatment level on the outcome of interest. Together, these two examples alone provide a very rich set of population parameters, allowing not only for comparisons across and within treatment levels for both means and quantiles, but also for the construction of other quantities of interest such as pairwise differences, interquantile ranges, or incremental ratios. The latter measures may be more appropriate from a policy-making perspective because they capture better notions of inequality and differential treatment effects.

The results presented in this paper are closely related to both the program evaluation literature in econometrics and the missing data literature in statistics.[1] This literature was mainly motivated by the seminal work of Rubin (1974) and Rosenbaum and Rubin (1983), and often focuses on the construction of semiparametric (efficient) estimation procedures for different population parameters of interest. For the particular case of a binary treatment, great effort is devoted to the estimation of average treatments effects (ATE) and related quantities, using either regression methods (Hahn (1998), Heckman, Ichimura, and Todd (1998), Imbens, Newey, and Ridder (2006)), matching (Abadie and Imbens (2006)), procedures based on the propensity score (Hirano, Imbens, and Ridder (2003)), or methods involving both regression and the propensity score (Robins, Rotnitzky, and Zhao (1994), Bang and Robins (2005)). In a very recent contribution, Firpo (2007) has considered a different population parameter, focusing on the efficient estimation of quantile treatment effects for dichotomous treatment assignments using the propensity score.

Surprisingly less work is available in the literature for the case of multi-valued treatment assignments. In a recent paper, Imbens (2000) derives a generalization of the propensity score, termed the Generalized Propensity Score (GPS), for the context of finite treatments and shows that the results of Rosenbaum and Rubin (1983) continue to hold when the treatment is multi-valued. Further, Imai and Dyk (2004) obtain a similar result for a general framework that encompasses this and other extensions available in the literature. Concerning estimation, however, the work in this

[1] For recent surveys on these topics, usually with a particular emphasis on binary treatment assignments, see Rosenbaum (2002), Imbens (2004), Lee (2005), or Tsiatis (2006), among others.

framework is very limited. Recent contributions in the missing data literature in statistics apply to this problem when considering the special case of ADRF and doubly robust estimators (for a survey on these results see, e.g., Bang and Robins (2005) and the references therein).

This paper contributes to the literature of program evaluation in several ways. It provides a set of simple sufficient conditions that enable efficient estimation of the DRF and shows, in particular, how the ADRF and the QDRF can be estimated under mild regularity conditions. This result builds on the modern theory of empirical processes (see, e.g., van der Vaart and Wellner (1996)) and the modern theory of semiparametric efficiency (see, e.g., Newey (1990) and Bickel, Klaassen, Ritov, and Wellner (1993)), and proceeds by constructing an observable moment condition using an inverse probability weighting (IPW) scheme that involves both the population parameter of interest (DRF) and a nuisance parameter in the estimation (GPS). Interestingly, because of the way the theory is developed, the general large sample properties of the proposed estimator for the DRF are derived without formally specifying the structure of the nonparametric estimator of the GPS, but rather by assuming two well-known high-level conditions. This result not only provides a better understanding of the set of sufficient conditions required for the general procedure to work, but also allows for different choices of the nonparametric estimator of the GPS.

After establishing the general asymptotic results for the estimator of the DRF, we consider the nonparametric estimation of the GPS. Because in this case the infinite dimensional nuissance parameter is in fact a conditional probability, we propose a new nonparametric estimator labeled Multinomial Logistic Series Estimator (MLSE). This estimator has the key advantage of providing predicted positive probabilities that add up to one and is a generalization of the nonparametric estimator for the propensity score introduced by Hirano, Imbens, and Ridder (2003). Using this estimator for the GPS, we provide sufficient conditions that guarantees the efficient estimation of the DRF.

Using the efficient estimation procedure for the general DRF, we show how other important population parameters of interest can be efficiently estimated by means of continuous transformations of the estimator considered. It follows that this procedure not only enlarges the class of parameters covered by our estimation procedure, but it also allows for optimal hypothesis testing. Moreover, using this methodology it is shown how the results of Hahn (1998), Hirano, Imbens, and Ridder (2003), and Firpo (2007) may be seen as particular cases of our procedure when the treatment assignment is binary.

The rest of the paper is organized as follows. Section 2 presents the multi-valued treatment model, introduces the population parameter of interest, discusses identification and formalizes the estimation procedure proposed. Section 3 derives the large sample properties of the estimator. Section 4 discusses how efficient estimation and optimal testing of other interesting population parameters can be done using the estimator considered in this paper, and also presents straightforward extensions to our methodology that further enlarges the class of population parameters that may be covered. Section 5 presents an empirical application, and Section 6 concludes. We relegate all proofs to the Appendix, which includes some general results that may be of independent interest for other applications.

2. Statistical Model, Identification and Estimation Procedure

In this Section we introduce the statistical model, the class of population parameters of interest and the basic assumptions used for identification. Finally, we also present the two-step estimation procedure considered in this paper.

2.1. Multi-valued Treatment Model and Dose-Response Functions. The setup considered here is the natural extension of the well-known model used in the classical binary treatment literature (Rubin (1974)). We assume that a random sample of size N from a large population is observed where observational units are indexed by $n = 1, 2, ..., N$. Each unit receives a treatment level (dose) denoted by $T_n \in \mathcal{T}$ where $\mathcal{T}$ is assumed finite and, without loss of generality, of the form $\{0, 1, 2, \cdots, J\}$. Using the potential outcomes notation, let $Y_n(t) \in \mathcal{Y}$ be the potential outcome associated with each $J+1$ treatment level for unit n. We also assume that there exists a vector of pre-intervention covariates denoted $\mathbf{X}_n \in \mathcal{X}$ for each unit. Finally, we define the random variable $D_n(t) = \mathbf{1}\{T_n = t\}$, and we note that the observed outcome for each unit is given by $Y_n \equiv Y_n(T_n) = \sum_{t\in\mathcal{T}} D_n(t) \cdot Y_n(t)$. Notice that in this general setup, the fundamental problem of causal inference is exacerbated: for each unit n we only observe one of the $J+1$ potential outcomes.

The population parameter of interest in this paper is the Dose-Response Function (DRF), denoted $\boldsymbol{\beta}^* = [\beta_0^*, \beta_1^*, \cdots, \beta_J^*]' \in \mathcal{B}^{J+1}$, which is assumed to be implicitly defined by the collection of moment conditions

$$\mathbb{E}\left[m\left(Y(t), \mathbf{X}; \beta_t^*\right)\right] = 0 \text{ for all } t \in \mathcal{T}, \tag{1}$$

where $m(\cdot, \cdot; \cdot) : \mathcal{Y} \times \mathcal{X} \times \mathcal{B} \to \mathbb{R}$ is a known, possibly non-smooth function. This description of the model can be summarized by the following assumption:

Assumption 1. *(*Model Setup*)*

*(1.1) (*Sampling*) The (observed) random sample $\{[Y_n, T_n, \mathbf{X}_n'] : n = 1, 2, \cdots, N\}$ is i.i.d., where $D_n(t) = \mathbf{1}\{T_n = t\}$ and $Y_n = \sum_{t\in\mathcal{T}} D_n(t) \cdot Y_n(t)$.*

*(1.2) (*Identification*) $\mathbb{E}\left[m\left(Y(t), \mathbf{X}; \beta\right)\right] = 0$ if and only if $\beta = \beta_t^*$ for all $t \in \mathcal{T} = \{0, 1, 2, \cdots, J\}$.*

Assumption 1.1 is standard in the literature. It summarizes the cross-sectional random sample scheme considered in the paper and reflects the missing data problem underlying this model, i.e., that we only observe the potential outcome resulting from the corresponding treatment assignment for each unit. Assumption 1.2 imposes conventional identification conditions for M-estimation.

The conditions in Assumption 1 allow us to consider a large collection of population parameters of interest including those defined by non-smooth moment functions such as quantiles or other robust estimands. We provide a set of sufficient conditions that enable efficient estimation of the DRF for the multi-valued treatment model, and we also show that these conditions are satisfied by two leading examples that we develop throughout this paper:

Example 1: Average Dose-Response Function. The first leading example is the classical population parameter generally used in the literature of Biostatistics, Public Health or Medicine, among other fields. This population parameter captures the mean response for each treatment level and, in the context of program evaluation, it can be seen as an extension of the ATE as will be shown in Section 4. We denote $\boldsymbol{\mu}^* = [\mu_0^*, \mu_1^*, \cdots, \mu_J^*]'$ as the ADRF and by defining $m(Y(t), \mathbf{X}; \mu_t) = Y(t) - \mu_t$, for all $t \in T$, we obtain the ADRF whose t-th term is given by $\mu_t^* = \mathbb{E}[Y(t)]$. In this case Assumption 1.2 follows immediately after assuming finite first moments of the potential outcomes. □

Example 2: Quantile Dose-Response Function. Characterizing distributional impacts of a multi-valued treatment is crucial because these effects are closely related to usual inequality measures. The second leading example captures this idea by looking at the treatment effect at different quantiles of the outcome variable. We denote $\mathbf{q}^{*}(\tau) = [q_0^{*}(\tau), q_1^{*}(\tau), \cdots, q_J^{*}(\tau)]'$ as the QDRF for quantile $\tau \in (0,1)$, and by defining $m(Y(t), \mathbf{X}; q_t(\tau)) = \mathbf{1}\{Y(t) \leq q_t(\tau)\} - \tau$, for all $t \in \mathcal{T}$, we obtain the QDRF whose t-th term is given by $q_t^{*}(\tau) = \inf\{q : F_{Y(t)}(q) \geq \tau\}$, where $F_{Y(t)}$ is the c.d.f. of $Y(t)$. In this case, Assumption 1.2 is satisfied if we assume that $Y(t)$ is a continuous random variable with density $f_{Y(t)}(q_t^{*}(\tau)) > 0$, which we impose throughout this example. □

2.2. Identification.

The identification condition in Assumption 1.2 covers many cases of interest. However, it has the obvious drawback of being based on unobservable random variables, the potential outcomes, which makes estimation unfeasible. To make progress, we need to impose an additional identification restriction. Following the program evaluation literature, we make the "selection on observables" assumption which, combined with an inverse probability weighting scheme, recovers identification of the DRF from an observed moment condition. We summarize the key identifying assumption:

Assumption 2. *(*Ignorability*)*

*(2.1) (*Weak Unconfoundedness*)* $Y(t) \perp\!\!\!\perp T \mid \mathbf{X}$, *for all* $t \in \mathcal{T}$.

*(2.2) (*Common Support*)* $0 < p_{\min} \leq p_t^{*}(\mathbf{X}) \equiv \mathbb{P}[T = t \mid \mathbf{X}]$, *for all* $t \in \mathcal{T}$.

Assumption 2.1 assumes that the distribution of each potential outcome and the treatment status are conditionally independent for all treatment levels and consequently provides identification by imposing random assignment conditional on observables. Observe that 2.1 is weaker than the usual unconfoundedness assumption commonly used in the classical binary treatment literature, since it involves only the marginal distributions of the potential outcomes rather than their joint distribution. Assumption 2.2 requires that, conditional on the pre-intervention covariates, the probability of receiving any treatment level be strictly positive and ensures that, at least in large samples, there will be observations in each treatment category. Finally, note that the conditional probabilities $p_t^{*}(\mathbf{X}) = \mathbb{P}[T = t \mid \mathbf{X}] = \mathbb{E}[D(t) \mid \mathbf{X}]$, for all $t \in \mathcal{T}$, correspond to the generalized propensity score in the context of the multi-valued treatment model. See Imbens (2000) for more details on these assumptions and results regarding the GPS.

Since we are interested not only in mean effects but also in other population parameters, the general DRF is defined as the solution to an implicit moment equation. For this reason, conditioning on observed covariates $\mathbf{X}$ to remove bias and then averaging out will not, in general, recover the parameter of interest. An alternative procedure is to use the GPS in an Inverse Probability Weighting scheme that, under Assumption 1 and Assumption 2, provides identification by transforming the infeasible moment equation (1) into a moment condition that depends only on observed random variables. It is easy to verify that

$$\mathbb{E}\left[\frac{D(t) \cdot m(Y, \mathbf{X}; \beta)}{p_t^{*}(\mathbf{X})}\right] = \mathbb{E}[m(Y(t), \mathbf{X}; \beta)] = 0 \text{ if and only if } \beta = \beta_t^{*}, \tag{2}$$

for all $t \in \mathcal{T}$.

Inverse probability weighting schemes have been considered by many authors in different contexts at least since the work of Horvitz and Thompson (1952). This procedure achieves identification by reweighting the observations to make them representative of the population of interest. This idea has been exploited in the context of program evaluation by Imbens (2000), Hirano, Imbens, and Ridder (2003) and Firpo (2007), and in the context of missing data by Robins, Rotnitzky, and Zhao (1994) and Robins, Rotnitzky, and Zhao (1995), among others.

Identification of the DRF follows directly from equation (2) together with Assumption 1 and Assumption 2. This result applies directly to our leading examples:

Example 1 (Continued): ADRF. In this case we obtain

$$\mathbb{E}\left[\frac{D(t)\cdot(Y-\mu_t^*)}{p_t^*(\mathbf{X})}\right]=0,\qquad \text{for all } t\in\mathcal{T}.\ \square$$

Example 2 (Continued): QDRF. Let $\tau\in(0,1)$ and we obtain

$$\mathbb{E}\left[\frac{D(t)\cdot(\mathbf{1}\{Y\leq q_t^*(\tau)\}-\tau)}{p_t^*(\mathbf{X})}\right]=0,\qquad \text{for all } t\in\mathcal{T}.\ \square$$

2.3. Two-Step Estimation Procedure. The (feasible) identification condition discussed in the previous section suggests a simple semi-parametric two-step minimum distance estimator for the class of DRF considered in this paper. To motivate the procedure we use the analogy principle (Manski (1988)).

Recall that our goal is to estimate the parameters implicitly defined by the moment conditions $\mathbb{E}[m(Y(t),\mathbf{X};\beta_t^*)]=0$ for all $t\in\mathcal{T}$. Had we observed the random variables $(Y(0),\cdots,Y(J))$, a natural estimator would simply solve the sample analog counterpart of each of these $(J+1)$ equations leading to an standard M-estimation procedure. However, due to the fundamental problem of causal inference, we cannot perform such estimation. Instead, we can use the result in Equation (2) to obtain $(J+1)$ equations based only on observed random variables. This alternative, however, has the drawback that now the estimating equations involve both the finite dimensional parameter of interest (DRF) and an infinite dimensional nuisance parameter (GPS). This line of reasoning suggests that if we could construct a preliminary estimator for the GPS that converges to the true GPS sufficiently fast, we would still be able to estimate the finite dimensional parameter of interest.

Using this idea, we consider a simple semi-parametric two-step estimation procedure where each dose-effect is estimated separately once the nonparametric nuisance parameter using the full data has been estimated. This procedure involves two steps:

Step 1. Construct a nonparametric estimator of the GPS, denoted $\hat{\mathbf{p}}(\cdot)=[\hat{p}_0(\cdot),\cdots,\hat{p}_J(\cdot)]'$.

Step 2. Obtain the efficient estimate of the DRF, $\hat{\boldsymbol{\beta}}$, by minimizing the sample analogue of the observed identification condition; that is, the components of $\hat{\boldsymbol{\beta}}$ are given by:

$$\hat{\beta}_t=\arg\min_{\beta\in\mathcal{B}}\left|\frac{1}{N}\sum\nolimits_{n=1}^{N}\frac{D_n(t)\cdot m(Y_n,\mathbf{X}_n;\beta)}{\hat{p}_t(\mathbf{X}_n)}\right|,\qquad \text{for all } t\in\mathcal{T}.$$

To fix ideas, we show how this estimation procedure applies to our leading examples:

Example 1 (Continued): ADRF. In this case, the procedure leads to a closed solution given by

$$\hat{\mu}_t = \frac{1}{N}\sum\nolimits_{n=1}^{N} \frac{D_n(t)\cdot Y_n}{\hat{p}_t(\mathbf{X}_n)}, \qquad \text{for all } t \in \mathcal{T}. \ \square$$

Example 2 (Continued): QDRF. In this case, for fixed $\tau \in (0,1)$, the estimator is given by

$$\hat{q}_t(\tau) = \arg\min_{q\in\mathcal{B}} \left| \frac{1}{N}\sum\nolimits_{n=1}^{N} \frac{D_n(t)\cdot(\mathbf{1}\{Y_n \le q\} - \tau)}{\hat{p}_t(\mathbf{X}_n)} \right|, \qquad \text{for all } t \in \mathcal{T}. \ \square$$

3. Large Sample Properties

In this Section we present the main large sample results of the paper in five stages. First, we establish consistency of the estimator for the DRF. Second, we compute the Efficient Influence Function and corresponding Semiparamentric Efficiency Bound. Third, we present a set of sufficient conditions to obtain asymptotic normality and efficiency of the estimator. Fourth, we introduce a nonparametric estimator appropriate for the estimation of the GPS. Finally, we construct consistent uncertainty estimates. The results presented in this Section build on more general results included in Appendix A and Appendix B, which may be of independent interest.

To reduce the notational burden, we define the augmented moment equation and its sample analogue, given respectively by

$$M(\beta, p_t(\cdot)) = \mathbb{E}\left[\frac{D(t)\cdot m(Y,\mathbf{X};\beta)}{p_t(\mathbf{X})}\right], \text{ and } M_N(\beta, p_t(\cdot)) = \frac{1}{N}\sum\nolimits_{n=1}^{N} \frac{D_n(t)\cdot m(Y_n,\mathbf{X}_n;\beta)}{p_t(\mathbf{X}_n)},$$

for $t \in \mathcal{T}$.

3.1. Consistency.

Consistency of the proposed two-step estimator will follow from two mild conditions imposed on the underlying unfeasible moment identification function $m(\cdot;\beta)$. Interestingly, to verify consistency we do not need to impose any particular structure on the nonparametric component or its estimator beyond the model assumptions and a very basic condition on the nonparametric estimator.

Assumption 3. *(*Consistency*) $\mathcal{B}$ is compact, and*

*(3.1) (*Glivenko-Cantelli Property*) $\mathcal{M} = \{m(\cdot;\beta) : \beta \in \mathcal{B}\}$ is Glivenko-Cantelli.*

*(3.2) (*Integrable Envelope*) $\mathbb{E}\left[\sup_{\beta\in\mathcal{B}} |m(Y(t),\mathbf{X};\beta)|\right] < \infty$, for $t \in \mathcal{T}$.*

Assumption 3.1 builds on the modern theory of Empirical Processes (van der Vaart and Wellner (1996)) and restricts the class of functions $m(\cdot;\beta)$ characterizing the DRF that we may consider. This assumption is slightly stronger than required. In fact, inspection of the proof of Theorem 1 reveals that a weaker sufficient condition is $M_N(\beta, p_t^*(\cdot)) \xrightarrow{p} M(\beta, p_t^*(\cdot))$ uniformly in $\beta \in \mathcal{B}$. It is well-known that uniform laws of large numbers are available when underlying functions are (close to) continuous (see, e.g., Newey and McFadden (1994)). However, to cover interesting nonsmooth cases (such as quantiles) we need to rely on slightly stronger results such as those covered by the empirical process literature. A more primitive condition for Assumption 3.1 would involve a standard covering number argument such as those employed in Ai and Chen (2003) in the context of GMM estimation. Assumption 3.2 is close to a regularity condition.

Theorem 1. *(Consistency) Suppose Assumption 1, Assumption 2, Assumption 3 and the following condition hold:*

(C.1) (Consistency of Nonparametric Estimator) $\|\hat{\mathbf{p}}(\cdot) - \mathbf{p}^*(\cdot)\|_\infty = o_p(1)$.

Then, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$.

Condition (C.1) is weak, requiring only that the nonparametric estimator is uniformly consistent. Moreover, below we will require a stronger assumption to obtain asymptotic normality of the estimator. Similarly, Assumption 3.1 will be automatically implied by the stronger requirement that the class of functions $m(\cdot;\beta)$ is Donsker, which will also be imposed later along with other conditions when deriving the asymptotic normality of the estimator. It is interesting to note that the result in Theorem 1 implies that for any consistent nonparametric estimator of the GPS, the two-step estimator proposed in this paper is consistent for the DRF.

For most applications, the key Assumption 3.1 is either readily available from the literature or can be verified directly by well-known results (see, e.g., Andrews (1994) or van der Vaart and Wellner (1996)). For example, it is well-known that continuous functions, Lipschitz functions or indicator functions enjoy this property, provided some envelop condition holds. For our examples, given the model assumptions, Assumption 3 is automatically satisfied:

Example 1 (Continued): ADRF. Assume $\mathcal{B}$ is compact and $\mathbb{E}[|Y(t)|] < \infty$, for $t \in \mathcal{T}$. Assumption 3 follows directly because the class of functions $\mathcal{M} = \{(y-\mu) : \mu \in \mathcal{B}\}$ is Glivenko-Cantelli. Therefore, using Theorem 1 we conclude that $\hat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}^*$. □

Example 2 (Continued): QDRF. Assume $\mathcal{B}$ is compact. Assumption 3 follows directly because the class of functions $\mathcal{M} = \{(\mathbf{1}\{y \leq q\} - \tau) : q \in \mathcal{B}\}$ is Glivenko-Cantelli. Therefore, using Theorem 1 we conclude that $\hat{\mathbf{q}}(\tau) \xrightarrow{p} \mathbf{q}^*(\tau)$. □

3.2. Efficient Influence Function and Semi-Parametric Efficiency Bound.

Semiparemetric efficiency theory has received considerable attention in econometrics since the seminal works of Newey (1990) and Bickel, Klaassen, Ritov, and Wellner (1993). This general theory provides the necessary ingredients for the construction of efficient estimators of finite dimensional parameters in the context of semiparametric models. First, it provides the analogue concept of the Cramer-Rao Lower Bound for semiparametric models, that is, an efficiency benchmark for regular estimators of the population parameter of interest. Second, and more importantly, it provides a way of constructing efficient estimators using the efficient influence function or efficient score of the model. In the simplest possible case, the construction of an efficient estimator starts by deriving the efficient influence function of the statistical model and then verifying that the proposed estimator admits an asymptotic linear representation based on this efficient influence function. In the next section we use this idea to derive asymptotic normality and efficiency of the estimator.

In the context of program evaluation with binary treatments, efficient influence functions and efficiency bounds have been computed by Hahn (1998), Hirano, Imbens, and Ridder (2003), and Firpo (2007) for different treatment effect parameters using the methodology outlined in Bickel, Klaassen, Ritov, and Wellner (1993). In the closely related framework of missing data, and under the assumption of "missing at random", Robins, Rotnitzky, and Zhao (1994), Robins, Rotnitzky, and Zhao (1995), and Robins and Rotnitzky (1995) have developed a general methodology to construct efficient scores and compute the corresponding efficiency bounds.

To compute the efficient influence function of the DRF we impose the following additional assumption:

Assumption 4. *(*Semi-parametric Efficiency*)*

*(4.1) (*Finite Second Moment*)* $\mathbb{E}\left[m\left(Y\left(t\right),\mathbf{X}_n;\beta_t^*\right)^2\right]<\infty$ *for all* $t\in\mathcal{T}$.

*(4.2) (*Implicit Function Theorem*)* $v_t\left(\beta_t^*\right)\equiv\partial\mathbb{E}\left[m\left(Y\left(t\right),\mathbf{X};\beta\right)\right]/\partial\beta|_{\beta_t^*}\neq 0$ *for all* $t\in\mathcal{T}$.

The main role of Assumption 4 (together with Assumption 1.2) is to ensure that the bound is finite. Notice that semiparametric efficiency computations require some additional mild regularity conditions on the underlying statistical model, which we are not explicitly including in this set of assumptions. We refer to Newey (1990), Bickel, Klaassen, Ritov, and Wellner (1993) and Newey (1994) for a discussion of such regularity conditions. A key necessary requirement, however, is that the population parameter of interest be pathwise differentiable. This result is established in the Appendix under these assumptions.

Theorem 2. *(*Efficient Influence Function and SPEB*) Suppose that Assumption 1, Assumption 2, and Assumption 4 hold, then the efficient influence function associated with the DRF is given by* $\boldsymbol{\psi}\left(\mathbf{Z};\boldsymbol{\beta}^*,\mathbf{p}^*\left(\cdot\right)\right)\in\mathbb{R}^{J+1}$*, with typical element* $t\in\mathcal{T}$

$$\psi_t\left(\mathbf{Z};\boldsymbol{\beta}^*,\mathbf{p}^*\left(\cdot\right)\right)=\frac{1}{v_t\left(\beta_t^*\right)}\cdot\left(\frac{D\left(t\right)\cdot\left(m\left(Y,\mathbf{X};\beta_t^*\right)-\mathcal{E}_t\left(\mathbf{X};\beta_i^*\right)\right)}{p_t^*\left(\mathbf{X}\right)}+\mathcal{E}_t\left(\mathbf{X};\beta_i^*\right)\right),$$

where $\mathcal{E}_t\left(\mathbf{X};\beta\right)\equiv\mathbb{E}\left[m\left(Y\left(t\right),\mathbf{X};\beta\right)\mid\mathbf{X}\right]$ *for all* $t\in\mathcal{T}$*. Consequently, the Semiparametric Efficiency Bound is given by the matrix* $SPEB\left(\boldsymbol{\beta}^*\right)\in\mathbb{R}^{(J+1)\times(J+1)}$ *with typical* (i,j)*-th element*

$$SPEB_{i,j}\left(\boldsymbol{\beta}^*\right)=\mathbb{E}\left[\mathbf{1}\left\{i=j\right\}\cdot\frac{\mathbb{V}ar\left[m\left(Y\left(i\right),\mathbf{X};\beta_i^*\right)\middle|\,\mathbf{X}\right]}{v_i\left(\beta_i^*\right)^2\cdot p_i^*\left(\mathbf{X}\right)}+\frac{\mathcal{E}_i\left(\mathbf{X};\beta_i^*\right)\cdot\mathcal{E}_j\left(\mathbf{X};\beta_j^*\right)}{v_i\left(\beta_i^*\right)\cdot v_j\left(\beta_j^*\right)}\right].$$

Observe that Theorem 2 provides the general form of the efficient influence function and the SPEB for any DRF covered by the model considered in this paper. Our derivation follows the work of Newey (1990), Bickel, Klaassen, Ritov, and Wellner (1993) and Newey (1994). Alternatively, this result can be obtained by means of the high-level methodology introduced in the context of missing data by Robins, Rotnitzky, and Zhao (1994) as mentioned before.

Following Newey (1994), we may provide additional intuition for the influence function derived in Theorem 2 by considering the typical t-th element of $\boldsymbol{\psi}\left(\mathbf{Z};\boldsymbol{\beta}^*,\mathbf{p}^*\left(\cdot\right)\right)$ and rearranging terms to obtain

$$\psi_t\left(\mathbf{Z};\boldsymbol{\beta}^*,\mathbf{p}^*\left(\cdot\right)\right)=\frac{1}{v_t\left(\beta_t^*\right)}\cdot\left(\frac{D\left(t\right)\cdot m\left(Y,\mathbf{X};\beta_t^*\right)}{p_t^*\left(\mathbf{X}\right)}+\alpha_t\left(T,X;p_t^*\left(\cdot\right)\right)\right),$$

where

$$\alpha_t\left(T,X;p_t^*\left(\cdot\right)\right)=-\frac{\mathcal{E}_t\left(\mathbf{X}_n\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(D_n\left(t\right)-p_t^*\left(\mathbf{X}_n\right)\right).$$

Thus, the vector-valued function $\boldsymbol{\alpha}\left(\cdot\right)=\left[\alpha_0\left(\cdot\right),\alpha_1\left(\cdot\right),\cdots,\alpha_J\left(\cdot\right)\right]'$ corresponds to the adjustment term due to the fact that we need to estimate the nuisance parameter (GPS). This decomposition will prove useful when deriving the asymptotic normality and constructing a consistent estimator for the asymptotic matrix of variances and covariances.

Finally, we apply the results of Theorem 2 to the examples under study:

Example 1 (Continued): ADRF. Assume $\mathbb{E}\left[Y(t)^2\right] < \infty$ and note that $v_t(\mu_t^*) = 1$ for all $t \in \mathcal{T}$ in this case. Thus, Assumption 4 is satisfied and Theorem 2 implies that the SPEB for the ADRF is given by $SPEB(\boldsymbol{\mu}^*)$ with typical (i,j)-th element

$$SPEB_{i,j}(\boldsymbol{\mu}^*) = \mathbb{E}\left[\mathbf{1}\{i=j\} \cdot \frac{\sigma_i^2(\mathbf{X})}{p_i^*(\mathbf{X})} + (\mu_i(\mathbf{X}) - \mu_i^*) \cdot \left(\mu_j(\mathbf{X}) - \mu_j^*\right)\right],$$

where $\sigma_i^2(\mathbf{X}) = \mathbb{V}ar[Y(i)|\,\mathbf{X}]$, $\mu_i(\mathbf{X}) = \mathbb{E}[Y(i)|\,\mathbf{X}]$, for all $i \in \mathcal{T}$. □

Example 2 (Continued): QDRF. Using Leibniz's rule we have $v_t(\mu_t^*(\tau)) = f_{Y(t)}(\mu_t^*(\tau))$ for $t \in \mathcal{T}$, which was assumed strictly positive. Thus, Assumption 4 is satisfied and Theorem 2 implies that the SPEB for the QDRF is given by $SPEB(\mathbf{q}^*(\tau))$ with typical (i,j)-th element

$$SPEB_{i,j}(\mathbf{q}^*(\tau)) = \mathbb{E}\left[\mathbf{1}\{i=j\} \cdot \frac{\sigma_t^2(\mathbf{X};\tau)}{f_{Y(i)}(\mu_i^*(\tau))^2 \cdot p_i^*(\mathbf{X})} + \frac{\mu_t(\mathbf{X};\tau) \cdot \mu_t(\mathbf{X};\tau)}{f_{Y(i)}(\mu_i^*(\tau)) \cdot f_{Y(j)}\left(\mu_j^*(\tau)\right)}\right],$$

where $\sigma_i^2(\mathbf{X};\tau) = \mathbb{V}ar[\mathbf{1}\{Y(i) \le \mu_i^*(\tau)\}|\,\mathbf{X}]$, $\mu_i(\mathbf{X};\tau) = \mathbb{E}[\mathbf{1}\{Y(i) \le \mu_i^*(\tau)\} - \tau|\,\mathbf{X}]$, for all $i \in \mathcal{T}$. □

3.3. Asymptotic Normality and Efficiency.

We may now derive an asymptotic linear representation based on the efficient influence function that will not only provide asymptotic normality of the estimator, but will also establish its asymptotic efficiency. The following conditions are needed:

Assumption 5. (Asymptotic Normality) *Suppose $\mathcal{B}$ is compact, $\boldsymbol{\beta}^* \in \operatorname{int}(\mathcal{B})$, and*

(5.1) (Donsker Property) *$\mathcal{M} = \{m(\cdot;\beta) : |\beta - \beta_t^*| < \delta\}$ is Donsker, for some $\delta > 0$.*

(5.2) (Square-Integrable Envelope) $\mathbb{E}\left[\sup_{\beta\in\mathcal{B}} |m(Y,\mathbf{X};\beta)|^2\right] < \infty$

(5.3) (Smoothness) *$\mathbb{E}[|m(Y(t),\mathbf{X};\beta) - m(Y(t),\mathbf{X};\beta_t^*)|] \le C \cdot |\beta - \beta_t^*|$ for all $t \in \mathcal{T}$, for some positive constant C, and for all β such that $|\beta - \beta_t^*| < \delta$, for some $\delta > 0$.*

(5.4) (L2 Continuity) *$\mathbb{E}\left[|m(Y(t),\mathbf{X};\beta) - m(Y(t),\mathbf{X};\beta_t^*)|^2\right] \to 0$ as $|\beta - \beta_t^*| \to 0$.*

Assumption 5.1 and Assumption 5.2 are standard sufficient conditions for weak convergence in the literature of empirical processes. These conditions can be relaxed either by directly proving weak convergence or by restricting the class of functions using a covering number argument. For most applications, Assumption 5.1 is already established or can be easily established by some "permanence theorem" (see, e.g. Andrews (1994) or van der Vaart and Wellner (1996)). Assumption 5.3 and Assumption 5.4 are key assumptions that allow us to derive the asymptotic normality result without specifying a nonparametric estimator for the GPS.

Theorem 3. (Asymptotic Normality and Efficiency) *Suppose Assumption 1, Assumption 2, and Assumption 4 hold. Further, assume the following conditions hold:*

(AN.1) (Nonparametric Estimator Rate) $N^{\frac{1}{4}} \cdot \left\|\hat{\mathbf{p}}(\cdot) - \mathbf{p}^*(\cdot)\right\|_\infty = o_p(1)$.

(AN.2) (Asymptotic Linear Expansion) *For all* $t \in \mathcal{T}$,

$$\sqrt{N} \cdot M_N\left(\beta_t^*, \hat{p}_t(\cdot)\right) = \sqrt{N} \cdot M_N\left(\beta_t^*, p_t^*(\cdot)\right) + \frac{1}{\sqrt{N}} \sum\nolimits_{n=1}^{N} \alpha_t\left(T_n, X_n; p_t^*(\cdot)\right) + o_p(1).$$

Then the asymptotic linear representation of the DRF is given by

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) = \frac{1}{\sqrt{N}} \sum\nolimits_{n=1}^{N} \boldsymbol{\psi}\left(Y_n, T_n, \mathbf{X}_n; \boldsymbol{\beta}^*, \mathbf{p}^*(\cdot)\right) + o_p(1),$$

and consequently $\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \xrightarrow{d} \mathcal{N}\left(0, SPEB\left(\boldsymbol{\beta}^*\right)\right)$.

The first condition of the theorem is standard in the literature and, of course, implies condition (C.1) in Theorem 1, and will be satisfied for most reasonable nonparametric estimators. On the other hand, condition (AN.2) turns out to be crucial. This condition involves only the nonparametric estimator (at the true DRF) and requires a linear expansion to hold. Newey (1994) provides a general discussion of high-level conditions, involving stochastic equicontinuity and mean-square continuity, that ensure that this condition holds. In the next section we verify these two conditions directly for the particular nonparametric estimator considered in this paper. For other standard nonparametric estimators, conditions in Newey (1994) or Newey and McFadden (1994) apply directly provided some additional regularity conditions are assumed.

Now we consider our leading examples:

Example 1 (Continued): ADRF. The class of functions $\mathcal{M} = \{(y - \mu) : \mu \in \mathcal{B}\}$ is Donsker and

$$\mathbb{E}\left[\left|m\left(Y(t), \mathbf{X}; \mu_t\right) - m\left(Y(t), \mathbf{X}; \mu_t^*\right)\right|\right] = \int \left|(y - \mu_t) - (y - \mu_t^*)\right| \cdot dF_{Y(t)}(y) = \left|\mu_t - \mu_t^*\right|,$$

giving Assumption 5.3 and Assumption 5.4. Thus, using Theorem 3 we conclude that $\sqrt{N}\left(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\right) \xrightarrow{d} \mathcal{N}\left(0, SPEB\left(\boldsymbol{\mu}^*\right)\right)$. □

Example 2 (Continued): QDRF. For fixed $\tau \in (0,1)$, the class of functions $\mathcal{M} = \{(\mathbf{1}\{y \leq \mu(\tau)\} - \tau) : \mu(\tau) \in \mathcal{B}\}$ is Donsker and

$$\begin{aligned} &\mathbb{E}\left[\left|m\left(Y(t), \mathbf{X}; q_t(\tau)\right) - m\left(Y(t), \mathbf{X}; q_t^*(\tau)\right)\right|\right] \\ &= \int \left|\mathbf{1}\{y \leq q_t(\tau)\} - \mathbf{1}\{y \leq q_t^*(\tau)\}\right| \cdot dF_{Y(t)}(y) \\ &= F_{Y(t)}\left(\max\{q_t(\tau), q_t^*(\tau)\}\right) - F_{Y(t)}\left(\min\{q_t(\tau), q_t^*(\tau)\}\right) \\ &\leq C \cdot \left|q_t(\tau) - q_t^*(\tau)\right|, \end{aligned}$$

for all $q_t(\tau)$ such that $\left|q_t(\tau) - q_t^*(\tau)\right| < \delta$, for some $\delta > 0$, under the assumptions imposed at the beginning of this example. This verifies Assumption 5.3, while Assumption 5.4 follows by using the same argument. Thus, using Theorem 3 we conclude that $\sqrt{N}\left(\hat{\mathbf{q}}(\tau) - \mathbf{q}^*(\tau)\right) \xrightarrow{d} \mathcal{N}\left(0, SPEB\left(\mathbf{q}^*(\tau)\right)\right)$. □

3.4. Nonparametric Estimator. Theorem 1 and Theorem 3 establish consistency, asymptotic normality and efficiency of the estimator considered in this paper. These results have been obtained without formally specifying the nonparametric estimator used for the GPS, but rather by imposing high-level assumptions concerning the behavior of such estimator. In this Section we present a new nonparametric estimator appropriate for the estimation of the GPS and we show that the conditions required in these theorems are met.

Recall we only need to verify conditions (AN.1) and (AN.2) in Theorem 3. Since the GPS is a conditional expectation, many standard nonparametric estimators are available. In particular, the arguments in Newey (1994) or Newey and McFadden (1994) can be used to verify that Conditions (AN.1) and (AN.2) are met when series or kernels are used to estimate nonparametrically the nuisance parameter.

However, the GPS is not only a conditional expectation but also a conditional probability (i.e., all elements are positive and add up to one), which imposes additional restrictions that cannot be captured by standard nonparametric estimators. Thus, in this section we present a new nonparametric estimator consistent with this additional requirements. In particular, we consider a generalization of the estimator introduced by Hirano, Imbens, and Ridder (2003) for the particular context of binary treatment, labeled Multinomial Logistic Series Estimator. This estimator can be interpreted as a non-linear sieve (see Chen (2005)) and works as follows.

Recall that our goal is to nonparametrically estimate the GPS, that is, the $(J+1)$ vector-valued function given by $\mathbf{p}^*(\cdot) = [p_0^*(\cdot), p_1^*(\cdot), \cdots, p_J^*(\cdot)]'$ with $p_t^*(\cdot) : \mathcal{X} \to (0,1)$ for all $t \in \mathcal{T}$. We consider a sequence of approximating functions of the form $\mathbf{r}_K(\mathbf{x}) = [r_{1K}(\mathbf{x}), r_{2K}(\mathbf{x}), \cdots, r_{KK}(\mathbf{x})]'$, where for simplicity the basis is restricted to be either power series or splines and are assumed to be the same for all $(J+1)$ probabilities. See Appendix B for details and more general results. The construction of this K-dimensional vector of functions is standard and a full description can be found in, for example, Newey (1997). Let $\boldsymbol{\gamma}_K = \left[\boldsymbol{\gamma}_{K,0}', \boldsymbol{\gamma}_{K,1}', \cdots, \boldsymbol{\gamma}_{K,J}'\right]'$ be a vector of approximating coefficients for each of the $(J+1)$ conditional expectations. To construct the MLSE we consider the (MLE) problem

$$\hat{\boldsymbol{\gamma}}_K = \arg \max_{\boldsymbol{\gamma}_K : \boldsymbol{\gamma}_{K,0}' = \mathbf{0}} \ell_N(\boldsymbol{\gamma}_K) = \frac{1}{N} \sum\nolimits_{n=1}^{N} \sum\nolimits_{t=0}^{J} D_n(t) \cdot \log\left(\frac{\exp\left\{\mathbf{r}_K(\mathbf{X}_n)' \boldsymbol{\gamma}_{K,t}\right\}}{\sum_{j=0}^{J} \exp\left\{\mathbf{r}_K(\mathbf{X}_n)' \boldsymbol{\gamma}_{K,j}\right\}} \right),$$

where we have imposed the usual normalization $\boldsymbol{\gamma}_{K,0}' = \mathbf{0}$. Then the MLSE is defined as $\hat{\mathbf{p}}(\mathbf{X}_n)$ with typical element given by

$$\hat{p}_t(\mathbf{X}_n) = \frac{\exp\left\{\mathbf{r}_K(\mathbf{X}_n)' \hat{\boldsymbol{\gamma}}_{K,t}\right\}}{\sum_{j=0}^{J} \exp\left\{\mathbf{r}_K(\mathbf{X}_n)' \hat{\boldsymbol{\gamma}}_{K,j}\right\}}.$$

It is straightforward to verify that this nonparametric estimator satisfies the additional restrictions underlying the GPS. To derive the asymptotic properties of this estimator, the following assumptions are required:

Assumption 6. *(*MLSE with Power Series and Splines*)*

(6.1) (Covariates Distribution) *$\mathbf{X}_n \in \mathcal{X} \subset \mathbb{R}^r$ compact and its density is bounded and bounded away from zero on $\mathcal{X}$.*

(6.2) (Smoothness) *For all $t \in \mathcal{T}$, the functions $p_t^*(\cdot)$ and $\mathcal{E}_t(\cdot;\beta_i^*)$ are s times differentiable with $s/r > 3.5$.*

Theorem 4. (Conditions (AN.1) and (AN.2)) *Suppose Assumption 1, Assumption 2, Assumption 4, and Assumption 6 hold. Then, Conditions (AN.1) and (AN.2) are satisfied by the MLSE if $K = N^{\nu}$ with*

$$\frac{1}{4\left(s/r-2\eta\right)} < \nu < \frac{1}{2\left(2\eta+1\right)}$$

where $\eta = 1$ or $\eta = 1/2$ depending on whether power series or splines are used as basis functions, respectively.

3.5. Uncertainty Estimation. TO BE COMPLETED.

The theorem is:

Theorem 5. (Consistent Estimation of Covariance Matrix) *Under assumptions,*

$$\hat{\mathbf{V}} = \frac{1}{N}\sum\nolimits_{n=1}^{N} \boldsymbol{\psi}\left(\mathbf{Z}_n, \hat{\boldsymbol{\beta}}, \hat{\mathbf{p}}\left(\mathbf{X}_n\right)\right)\boldsymbol{\psi}\left(\mathbf{Z}_n, \hat{\boldsymbol{\beta}}, \hat{\mathbf{p}}\left(\mathbf{X}_n\right)\right)' \xrightarrow{p} \mathbf{V} = SPEB\left(\boldsymbol{\mu}\right).$$

4. Recovering Other Population Parameters, Optimal Testing and Extensions

In this Section we discuss how other population parameters of interest based on DRF can be estimated efficiently, which leads to optimal hypothesis testing. Then we introduce other population parameters already covered by our methodology as well as some natural extensions.

4.1. Efficient Estimation and Optimal Testing of Continuous Functions of DRF.

Continuous functions of Euclidean efficient estimators are efficient. Thus, any population parameter of interest that can be written as a function of the DRF can be estimated efficiently by a standard delta-method argument. For example, pairwise comparisons (in the spirit of ATE), differences between pairwise comparisons or incremental ratios can be estimated efficiently. Furthermore, because tests based on asymptotically efficient estimators are asymptotically optimal, the usual testing strategies apply directly to this problem and deliver asymptotically optimal tests. Thus, we can test for differential effects along treatment levels in an straightforward manner.

We exploit these ideas further in Section 5 when we present the empirical application. Finally, to fix ideas, we show how the papers of Hahn (1998), Hirano, Imbens, and Ridder (2003) and Firpo (2007) can be thought as particular cases of the method discuss here:

Example 1 (Continued): ADRF. Let $\mathcal{T} = \{0,1\}$ and observe that the ATE can be written as $\Delta^{ATE} \equiv \mathbb{E}\left[Y\left(1\right)\right] - \mathbb{E}\left[Y\left(0\right)\right] = \mathbf{v}'\boldsymbol{\mu}^*$, where $\mathbf{v} = [-1,1]'$. Using Theorem 2, we conclude that

$$SPEB\left(\boldsymbol{\mu}^*\right) = \mathbb{E}\left[\begin{array}{cc} \frac{\sigma_0^2(\mathbf{X})}{p(0,\mathbf{X})} + \left(\mu_0\left(\mathbf{X}\right) - \mu_0^*\right)^2 & \left(\mu_0\left(\mathbf{X}\right) - \mu_0^*\right)\cdot\left(\mu_1\left(\mathbf{X}\right) - \mu_1^*\right) \\ \left(\mu_0\left(\mathbf{X}\right) - \mu_0^*\right)\cdot\left(\mu_1\left(\mathbf{X}\right) - \mu_1^*\right) & \frac{\sigma_1^2(\mathbf{X})}{p(1,\mathbf{X})} + \left(\mu_1\left(\mathbf{X}\right) - \mu_1^*\right)^2 \end{array}\right],$$

where $\sigma_t^2\left(\mathbf{X}\right) = \mathbb{V}ar\left[Y\left(t\right)\right|\mathbf{X}]$, $\mu_t\left(\mathbf{X}\right) = \mathbb{E}\left[Y\left(t\right)\right|\mathbf{X}]$, for $t \in \mathcal{T}$. Using Theorem 3 and the transformation $g\left(\mathbf{z}\right) = \mathbf{v}'\mathbf{z}$, we conclude that

$$\sqrt{N}\cdot\left(\hat{\Delta}^{ATE} - \Delta^{ATE}\right) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \mathbf{v}'SPEB\left(\boldsymbol{\mu}^*\right)\mathbf{v}\right],$$

where

$$\mathbf{v}'SPEB\left(\boldsymbol{\mu}^*\right)\mathbf{v} = \mathbb{E}\left[\frac{\sigma_0^2\left(\mathbf{X}\right)}{p\left(0,\mathbf{X}\right)} + \frac{\sigma_1^2\left(\mathbf{X}\right)}{p\left(1,\mathbf{X}\right)} + \left(\Delta^{ATE}\left(\mathbf{X}\right) - \Delta^{ATE}\right)^2\right].$$

Observe that the asymptotic variance is the SPEB found by Hahn (1998) and the resulting estimator is the one considered in Hirano, Imbens, and Ridder (2003). □

Example 2 (Continued): QDRF. Let $\mathcal{T} = \{0, 1\}$ and observe that the QTE can be written as $\Delta^{QTE} \equiv q_1^*(\tau) - q_0^*(\tau) = \mathbf{v}'\mathbf{q}^*(\tau)$, where $\mathbf{v} = [-1, 1]'$ and $q_1^*(\tau) = \inf\{q : \mathbb{E}[\mathbf{1}\{Y(t) \leq q\} - \tau]\}$ for $t \in \mathcal{T}$. Using Theorem 3 and the transformation $g(\mathbf{z}) = \mathbf{v}'\mathbf{z}$, we conclude that

$$\sqrt{N} \cdot \left(\hat{\Delta}^{QTE} - \Delta^{QTE}\right) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \mathbf{v}'SPEB\left(\mathbf{q}^*(\tau)\right)\mathbf{v}\right].$$

Observe that the asymptotic variance is the SPEB found by Firpo (2007) and the resulting estimator coincides with Firpo's estimator for the QTE. □

4.2. Other Population Parameters and Extensions. Other population parameters that may be easily included in our methodology are: Robust ADRF (RADRF) in the spirit of Huber's robust location estimator, regression based estimators widely considered in the literature of missing data, and weighted average treatment effects (such as treatment effects conditional on a subpopulation) as discussed in Hirano, Imbens, and Ridder (2003).

TO BE COMPLETED.

5. Empirical Application.

TO BE COMPLETED.

6. Conclusions

TO BE COMPLETED.

References

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Andrews, D. W. K. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics, Volumen IV*, ed. by R. F. Engle, and D. L. McFadden, chap. 37, pp. 2247–2294. Elsevier Science B. V.

Bang, H., and J. M. Robins (2005): "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer.

Chen, X. (2005): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volumne VI*, ed. by J. Heckman, and E. Leamer. North-Holland Publishers.

Chen, X., O. Linton, and I. V. Keilegom (2003): "Estimation of Semiparametric Models When The Criterion Function Is Not Smooth," *dsdds*, 71, 1591–1608.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.

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

Heckman, J. J., H. Ichimura, and P. E. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 1017–1098.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using The Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Horvitz, D. G., and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Population," *Journal of the American Statistical Association*, 47, 663–685.

Imai, K., and D. A. V. Dyk (2004): "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99, 854–866.

Imbens, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710.

——— (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W., W. Newey, and G. Ridder (2006): "Mean-Squared-Error Calculations for Average Treatment Effects," Working Paper.

Lee, M. J. (2005): *Micro-Econometrics for Policy, Program and Treatment Effects*. Oxford University Press, Oxford.

Manski, C. (1988): *Analog Estimation Methods in Econometrics.* Chapman and Hall.

Newey, W. K. (1990): "Semiparemetric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Thesting," in *Handbook of Econometrics, Volume IV*, ed. by R. F. Engle, and D. L. McFadden, chap. 36, pp. 2112–2245. Elsevier Science.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

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

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

——— (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Rosenbaum, P. R. (2002): *Observational Studies.* Springer, New York.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688–701.

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

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

A. Appendix A: M-Estimation

This Appendix develops two general M-estimation results appropriate for the problem studied in this paper. Recall that the model discussed here involves both a parametric (finite dimensional) and a nonparametric (infinite dimensional) component that enter a criterion (loss) function, where the former is the population parameter of interest and the latter is simply a nuisance parameter in the estimation procedure. We propose an estimation procedure very popular for this class of models: a two-step procedure where first we construct a nonparametric estimator of the infinite dimensional component, and then we plug in this preliminary estimate in the criterion function and optimize to obtain an estimate of the finite dimensional parameter of interest.

Originally, this class of models and their asymptotic properties have been studied by Andrews (1994), Bickel, Klaassen, Ritov, and Wellner (1993), Newey (1994), and Newey and McFadden (1994), among others, under the assumption that the criterion function is smooth in both the finite and the infinite dimensional parameters. Recently, Chen, Linton, and Keilegom (2003) have relaxed the smoothness assumption and have provided quite general sufficient conditions under which the large sample properties of this class of models can be derived when neither the parametric nor the nonparametric components enter in an smooth fashion in the criterion function.

The model studied in this paper implies that the criterion function is smooth in the nonparametric component, while it imposes no assumptions on the finite dimensional parameter of interest. Consequently, although the general M-estimation framework of Chen, Linton, and Keilegom (2003) can be applied to our problem, relatively simpler conditions that exploit the additional smoothness assumption can be derived. The large sample theory presented in this Appendix follows the original style of proof from Pakes and Pollard (1989) and the recent extension due to Chen, Linton, and Keilegom (2003). The key difference in the results presented here is that by imposing stronger assumptions (i.e., smoothness in the non-parametric component of the criterion function) we derive easy to check sufficient conditions as well as some stronger results than those available in the literature.

We briefly outline the general setup (for a more complete description see Chen, Linton, and Keilegom (2003)). Let $\{Z_n\}_{n=1}^N$ be a random sample of size N from a distribution $\mathbb{P}$ with support $\mathcal{Z} \subseteq \mathbb{R}^d$. The statistical model of interest is a two-step semiparametric minimum distance procedure, which is based on a preliminary nonparametric estimator of the infinite dimensional component. The parameters of the model are given by $(\theta, h(\cdot;\theta)) \in \Theta \times \mathcal{H}$, where $\Theta \subset \mathbb{R}^k$ is a compact finite dimensional parameter set and $\mathcal{H}$ is an infinite dimensional parameter set, which can in principle depend on θ. We assume that Θ and $\mathcal{H}$ are normed vector spaces and we denote their norms $\|\cdot\|$ and $\|\cdot\|_{\mathcal{H}}$, respectively. The norm $\|\cdot\|$ is set to $\|A\| = \sqrt{\text{trace}(A'WA)}$ for any matrix A. The norm $\|\cdot\|_{\mathcal{H}}$ is set to be a sup-norm metric with respect to the θ-argument and any pseudo-metric with respect to the remaining arguments. We assume there exists a nonrandom vector-valued measurable function $M : \Theta \times \mathcal{H} \to \mathbb{R}^l$, $k \leq l$, with the property that $M(\theta, h_0(\cdot,\theta)) = 0$ if and only if $\theta = \theta_0$, where $(\theta_0, h_0(\cdot;\theta_0)) \in \Theta \times \mathcal{H}$ are the true finite and infinite dimensional parameters. Observe that we suppress the arguments of the nonparametric component to notational ease.

Given a preliminary nonparametric estimator, we are interested in the estimator of the finite dimensional parameter defined as the (approximate) solution of the distance minimization problem:

$$\min_{\theta\in\Theta} \left\| M_N\left(\theta, \hat{h}(\theta)\right)\right\| = \min_{\theta\in\Theta} \left[M_N\left(\theta, \hat{h}(\theta)\right)\right]' W \left[M_N\left(\theta, \hat{h}(\theta)\right)\right],$$

where $M_N : \Theta \times \mathcal{H} \to \mathbb{R}^l$ is a random vector-valued measurable function that depends on the random sample $\{Z_n\}_{n=1}^N$ and that is assumed to be close in norm to the function $M(\theta, h_0(\cdot,\theta))$ at $\theta = \theta_0$. To simplify the notation, in the sequel we assume $k = l$ and $W = I$, the identity matrix, since for our purpose this assumption covers our problem. The results below extend naturally to the GMM context.

The next Theorem provides a set of sufficient conditions for consistency of the finite dimensional estimator.

Theorem 6. (Consistency) *Suppose that $\theta_0 \in \Theta$ satisfies $M(\theta_0, h_0) = 0$, and the following conditions hold:*

(C1) (Estimator) $\left\|M_N\left(\hat{\theta}, \hat{h}\right)\right\| \leq \inf_{\theta\in\Theta}\left\|M_N\left(\theta, \hat{h}\right)\right\| + o_p(1)$,

(C2) (Identification) for all $\varepsilon > 0$, there exists $\delta(\varepsilon) > 0$ such that $\inf_{\|\theta-\theta_0\|>\varepsilon} \|M(\theta, h_0)\| \geq \delta(\varepsilon) > 0$,

(C3) (Smoothness of Nonparametric Component)

$$\sup_{\theta\in\Theta} \left\|M_N\left(\theta, \hat{h}\right) - M_N(\theta, h_0)\right\| = O_p\left(\left\|\hat{h} - h_0\right\|_{\mathcal{H}}\right),$$

(C4) (Consistency of Nonparametric Estimator) $\left\|\hat{h} - h_0\right\|_{\mathcal{H}} = o_p(1)$, and

(C5) (Stochastic Equicontinuity of Parametric Component at $h = h_0$)

$$\sup_{\theta\in\Theta} \frac{\|M_N(\theta, h_0) - M(\theta, h_0)\|}{1 + \|M_N(\theta, h_0)\| + \|M(\theta, h_0)\|} = o_p(1).$$

Then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof of Theorem 6:
By condition (C2) we have $\mathbb{P}\left[\left\|\hat{\theta}-\theta_0\right\|>\delta\right] \leq \mathbb{P}\left[\left\|M\left(\hat{\theta},h_0\right)\right\| \geq \varepsilon(\delta)\right]$ and hence it is sufficient to show that $\left\|M\left(\hat{\theta},h_0\right)\right\| = o_p(1)$. Next, observe that by conditions (C3), (C4) and (C5) we have

$$\begin{aligned}\left\|M\left(\hat{\theta},h_0\right)\right\| &\leq \left\|M\left(\hat{\theta},h_0\right)-M_N\left(\hat{\theta},h_0\right)\right\|+\left\|M_N\left(\hat{\theta},h_0\right)-M_N\left(\hat{\theta},\hat{h}\right)\right\|+\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\| \\ &\leq o_p(1)\cdot\left[1+\left\|M_N\left(\hat{\theta},h_0\right)\right\|+\left\|M\left(\hat{\theta},h_0\right)\right\|\right]+o_p(1)+\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\|.\end{aligned}$$

Now, rearranging and using conditions (C1), (C3) and (C4) we have

$$\begin{aligned}\left\|M\left(\hat{\theta},h_0\right)\right\|\cdot[1-o_p(1)] &\leq o_p(1)\cdot\left\|M_N\left(\hat{\theta},h_0\right)-M_N\left(\hat{\theta},\hat{h}\right)\right\|+[1+o_p(1)]\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\|+o_p(1) \\ &\leq [1+o_p(1)]\cdot\left\|M_N\left(\theta_0,\hat{h}\right)\right\|+o_p(1) \\ &\leq [1+o_p(1)]\cdot\left\|M_N\left(\theta_0,\hat{h}\right)-M_N(\theta_0,h_0)\right\|+[1+o_p(1)]\cdot\|M_N(\theta_0,h_0)\|+o_p(1) \\ &\leq [1+o_p(1)]\cdot\|M_N(\theta_0,h_0)\|+o_p(1),\end{aligned}$$

and finally by (C5) and the fact that $M(\theta_0,h_0)=0$, we obtain $\|M_N(\theta_0,h_0)\|=o_p(1)$ and the result follows. **Q.E.D.**

Remark 1. (Sufficient Condition of (C5)) *Similarly as in Pakes and Pollard (1989) and Chen, Linton, and Keilegom (2003), condition C5 is implied by*

(C5') *(Uniform Consistency at $h=h_0$)*

$$\sup_{\theta\in\Theta}\|M_N(\theta,h_0)-M(\theta,h_0)\|=o_p(1).$$

Let $\Theta_\delta=\{\theta\in\Theta:\|\theta-\theta_0\|\leq\delta\}$ and $\mathcal{H}_\delta=\{h\in\mathcal{H}:\|h-h_0\|\leq\delta\}$ for some $\delta>0$. The next theorem derives an asymptotic linear representation and provides a set of sufficient conditions for asymptotic normality of the finite dimensional estimator.

Theorem 7. (Asymptotic Normality) *Suppose there exists a unique $\theta_0\in\operatorname{int}(\Theta)$ that satisfies $M(\theta_0,h_0)=0$ and $\hat{\theta}-\theta_0=o_p(1)$, and the following conditions hold:*

(AN1) *(Estimator)* $\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\|=\inf_{\theta\in\Theta_\delta}\left\|M_N\left(\theta,\hat{h}\right)\right\|+o_p\left(N^{-1/2}\right)$,

(AN2) *(Differentiability of Parametric Component) the ordinary derivative w.r.t. θ of $M(\theta,h_0)$, denoted $\Gamma(\theta,h_0)$, exists for $\theta\in\Theta_\delta$ and the matrix $\Gamma_0\equiv\Gamma(\theta_0,h_0)$ is full rank,*

(AN3) *(Smooth Linearization of Nonparametric Component) there exists a function $\Delta_N(\theta,h)$ linear in h such that (i)*

$$\sup_{\theta\in\Theta_\delta}\left\|M_N\left(\theta,\hat{h}\right)-M_N(\theta,h_0)-\Delta_N\left(\theta,\hat{h}-h_0\right)\right\|=O_p\left(\left\|\hat{h}-h_0\right\|_{\mathcal{H}}^2\right),$$

and (ii) $\left\|\Delta_N\left(\hat{\theta},\hat{h}-h_0\right)-\Delta_N\left(\theta_0,\hat{h}-h_0\right)\right\|=O_p\left(\left\|\hat{\theta}-\theta_0\right\|\cdot\left\|\hat{h}-h_0\right\|_{\mathcal{H}}\right)$,

(AN4) *(Nonparametric Estimator Rate)* $\mathbb{P}\left[\hat{h}\in\mathcal{H}\right]\longrightarrow 1$ *and* $\left\|\hat{h}-h_0\right\|_{\mathcal{H}}=o_p\left(N^{-1/4}\right)$,

(AN5) *(Stochastic Equicontinuity of Parametric Component) for all sequences of positive numbers $\{\delta_n\}$ with $\delta_n=o(1)$,*

$$\sup_{\|\theta-\theta_0\|<\delta_n}\frac{\sqrt{N}\cdot\|M_N(\theta,h_0)-M(\theta,h_0)-M_N(\theta_0,h_0)\|}{1+\sqrt{N}\cdot\|M_N(\theta,h_0)\|+\sqrt{N}\cdot\|M(\theta,h_0)\|}=o_p(1),$$

and

(AN6) *(Asymptotic Normality at $\theta=\theta_0$)* $\sqrt{N}\cdot M_N\left(\theta_0,\hat{h}\right)\xrightarrow{d}\mathcal{N}[0,\Omega]$.

Then $\sqrt{N}\cdot\left(\hat{\theta}-\theta_0\right)=\left[-\left(\Gamma_0'\Gamma_0\right)^{-1}\Gamma_0'\right]\cdot\sqrt{N}\cdot M_N\left(\theta_0,\hat{h}\right)+o_p\left(1\right)\xrightarrow{d}\mathcal{N}\left[0,V\right]$, *where* $V=\left(\Gamma_0'\Gamma_0\right)^{-1}\Gamma_0'\Omega\Gamma_0\left(\Gamma_0'\Gamma_0\right)^{-1}$.

Proof of Theorem 7:
The proof of this theorem is done in two parts. First we establish $\sqrt{N}$-consistency of the estimator $\hat{\theta}$ and then we use this result to derive an asymptotic linear representation to obtain the desired result.

$\sqrt{N}$-consistency. Because both the parametric and nonparametric estimators are consistent, we can choose a positive sequence $\delta_N=o\left(1\right)$ slow enough such that $\mathbb{P}\left[\left\|\hat{\theta}-\theta_0\right\|>\delta_N,\ \left\|\hat{h}-h_0\right\|>\delta_N\right]\to 0$, which implies that we only need to work with $(\theta,h)\in\Theta_\delta\times\mathcal{H}_\delta$. By condition (AN3) (i), (AN4), and (AN6) we have

$$\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)\right\|=\sqrt{N}\cdot\left\|M_N\left(\theta_0,h_0\right)+\Delta_N\left(\theta_0,\hat{h}-h_0\right)\right\|+o_p\left(1\right)=O_p\left(1\right).$$

By condition (AN1), and the previous result we have

$$\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\|\leq\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)\right\|+o_p\left(1\right)=O_p\left(1\right).$$

By the previous results and condition (AN5)

$$\begin{aligned}&\sqrt{N}\cdot\left\|M\left(\hat{\theta},h_0\right)\right\|\\&\leq\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)-M\left(\hat{\theta},h_0\right)-M_N\left(\theta_0,h_0\right)\right\|+\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|+\sqrt{N}\cdot\left\|M_N\left(\theta_0,h_0\right)\right\|\\&=o_p\left(1\right)\cdot\left[1+\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|+\sqrt{N}\cdot\left\|M\left(\hat{\theta},h_0\right)\right\|\right]+\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|+O_p\left(1\right)\\&=\left[1+o_p\left(1\right)\right]\cdot\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|+o_p\left(1\right)\cdot\sqrt{N}\cdot\left\|M\left(\hat{\theta},h_0\right)\right\|+O_p\left(1\right),\end{aligned}$$

and therefore

$$\sqrt{N}\cdot\left\|M\left(\hat{\theta},h_0\right)\right\|=O_p\left(\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|\right)+O_p\left(1\right).$$

By the previous results and conditions (AN3), (AN4) and (AN5)

$$\begin{aligned}&\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|\\&\leq\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)-M_N\left(\hat{\theta},h_0\right)-\Delta_N\left(\hat{\theta},\hat{h}-h_0\right)\right\|+\sqrt{N}\cdot\left\|\Delta_N\left(\hat{\theta},\hat{h}-h_0\right)-\Delta_N\left(\theta_0,\hat{h}-h_0\right)\right\|\\&\qquad+\sqrt{N}\cdot\left\|\Delta_N\left(\theta_0,\hat{h}-h_0\right)\right\|+\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\|\\&=o_p\left(1\right)\cdot\left[1+\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|+\sqrt{N}\cdot\left\|M\left(\hat{\theta},h_0\right)\right\|\right]+O_p\left(\left\|\hat{h}-h_0\right\|_{\mathcal{H}}\right)+O_p\left(1\right),\end{aligned}$$

and therefore

$$\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)\right\|=O_p\left(\sqrt{N}\cdot\left\|\hat{\theta}-\theta_0\right\|\cdot\left\|\hat{h}-h_0\right\|_{\mathcal{H}}\right)+O_p\left(1\right).$$

Finally, by condition (AN2) and the previous results,

$$\sqrt{N}\cdot\left\|\Gamma_0\right\|\cdot\left\|\hat{\theta}-\theta_0\right\|=\sqrt{N}\cdot\left\|M\left(\hat{\theta},h_0\right)\right\|+o_p\left(\sqrt{N}\cdot\left\|\hat{\theta}-\theta_0\right\|\right)=O_p\left(1\right),$$

i.e., $\hat{\theta}$ is $\sqrt{N}$-consistent.

Asymptotic Normality. Let $\tilde{\theta}=\theta_0-\left[\left(\Gamma_0'\Gamma_0\right)^{-1}\Gamma_0'\right]M_N\left(\theta_0,\hat{h}\right)$ and observe that $\tilde{\theta}$ is also $\sqrt{N}$-consistent by construction. Then, by condition (AN2) wpa1

$$\sqrt{N}\cdot\left\|M\left(\tilde{\theta},h_0\right)\right\|=\sqrt{N}\cdot\left\|\Gamma\left(\tilde{\theta}-\theta_0\right)\right\|+o_p\left(\sqrt{N}\cdot\left\|\tilde{\theta}-\theta_0\right\|\right)=O_p\left(1\right).$$

and also by triangular inequality, the previous results and condition (AN5)

$$\begin{aligned}&\sqrt{N}\cdot\left\|M_N\left(\tilde{\theta},h_0\right)\right\|-\sqrt{N}\cdot\left\|M\left(\tilde{\theta},h_0\right)\right\|-\sqrt{N}\cdot\left\|M_N\left(\theta_0,h_0\right)\right\|\\&\leq\sqrt{N}\cdot\left\|M_N\left(\tilde{\theta},h_0\right)-M\left(\tilde{\theta},h_0\right)-M_N\left(\theta_0,h_0\right)\right\|\\&=o_p\left(1\right)+o_p\left(\sqrt{N}\cdot\left\|M\left(\tilde{\theta},h_0\right)\right\|\right)+o_p\left(\sqrt{N}\cdot\left\|M_N\left(\theta_0,h_0\right)\right\|\right),\end{aligned}$$

which implies that

$$\sqrt{N}\cdot\left\|M_N\left(\tilde{\theta},h_0\right)\right\|=O_p\left(1\right).$$

Next, by $\sqrt{N}$-consistency, conditions (AN2), (AN3), (AN4), (AN5) and (AN6), and the previous results we have

$$\begin{aligned}
&\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\|-\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma\left(\hat{\theta}-\theta_0\right)\right\| \\
&\leq\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)-M_N\left(\theta_0,\hat{h}\right)-\Gamma\left(\hat{\theta}-\theta_0\right)\right\| \\
&\leq\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)-M_N\left(\hat{\theta},h_0\right)-\Delta_N\left(\hat{\theta},\hat{h}-h_0\right)\right\| \\
&\qquad+\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)-M_N\left(\theta_0,\hat{h}\right)-D_N\left(\theta_0,\hat{h}-h_0\right)\right\| \\
&\qquad+\sqrt{N}\cdot\left\|\Delta_N\left(\hat{\theta},\hat{h}-h_0\right)-\Delta_N\left(\theta_0,\hat{h}-h_0\right)\right\| \\
&\qquad+\sqrt{N}\cdot\left\|M_N\left(\hat{\theta},h_0\right)-M\left(\hat{\theta},h_0\right)-M_N\left(\theta_0,\hat{h}\right)\right\| \\
&\qquad+\sqrt{N}\cdot\left\|M\left(\hat{\theta},h_0\right)-\Gamma_0\left(\hat{\theta}-\theta_0\right)\right\| \\
&=o_p\left(1\right),
\end{aligned}$$

and observe that the same result holds for $\hat{\theta}$ replaced by $\tilde{\theta}$ wpa1. Putting these results together and using condition (AN1) we have wpa1

$$\begin{aligned}
\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\hat{\theta}-\theta_0\right)\right\| &= \sqrt{N}\cdot\left\|M_N\left(\hat{\theta},\hat{h}\right)\right\|+o_p\left(1\right) \\
&\leq \sqrt{N}\cdot\left\|M_N\left(\tilde{\theta},\hat{h}\right)\right\|+o_p\left(1\right) \\
&= \sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\tilde{\theta}-\theta_0\right)\right\|+o_p\left(1\right),
\end{aligned}$$

and this result implies that

$$\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\hat{\theta}-\theta_0\right)\right\|-\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\tilde{\theta}-\theta_0\right)\right\|=o_p\left(1\right).$$

Finally, observe that by the definition of $\tilde{\theta}-\theta_0$ (i.e., a projection onto the column space of Γ') we have

$$\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\hat{\theta}-\theta_0\right)\right\|^2=\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\tilde{\theta}-\theta_0\right)\right\|^2+\left\|\Gamma_0\left(\hat{\theta}-\tilde{\theta}\right)\right\|^2$$

and using the previous results and the fact that $(a-b)^2=(a-b)\cdot(a+b)$,

$$\begin{aligned}
N\cdot\left\|\hat{\theta}-\tilde{\theta}\right\|^2 &\leq \left\|\Gamma_0^{-1}\right\|^2\cdot N\cdot\left\|\Gamma_0\left(\hat{\theta}-\tilde{\theta}\right)\right\|^2 \\
&\leq N\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\hat{\theta}-\theta_0\right)\right\|^2-N\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\tilde{\theta}-\theta_0\right)\right\|^2 \\
&=\left(\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\hat{\theta}-\theta_0\right)\right\|-\sqrt{N}\cdot\left\|M_N\left(\theta_0,\hat{h}\right)+\Gamma_0\left(\tilde{\theta}-\theta_0\right)\right\|\right)\cdot O_p\left(1\right) \\
&=o_p\left(1\right),
\end{aligned}$$

which in turn implies

$$\sqrt{N}\cdot\left(\hat{\theta}-\tilde{\theta}\right)=\sqrt{N}\cdot\left(\hat{\theta}-\theta_0\right)-\left[-\left(\Gamma_0'\Gamma_0\right)^{-1}\Gamma_0'\right]\cdot\sqrt{N}\cdot M_N\left(\theta_0,\hat{h}\right)=o_p\left(1\right),$$

since the matrix Γ_0 is full rank. **Q.E.D.**

B. Appendix B: Multinomial Logistic Series Estimator

In this appendix we derive uniform rates of convergence for the non-linear sieve estimator proposed in Section 3.4, labeled Multinomial Logistic Series Estimator (MLSE). Recall that our goal is to estimate nonparametrically the GPS, which has the additional property that the sum of its positive elements should add up to one. These restrictions imply that standard nonparametric procedures may seem less appropriate for this case. Recently, Hirano, Imbens, and Ridder (2003) proposed a solution to this problem for the particular case $J = 1$ by considering a non-linear sieve estimation procedure where the conditional probability is estimated within a Logistic model, labeled Logistic Series Estimation. In this appendix we generalized this procedure in three ways: (i) we allow for arbitrary number of outcomes, (ii) we allow for an arbitrary choice of approximating basis, and (iii) we derive the approximation rates using only the approximating sequence. The results presented here not only encompass those in Hirano, Imbens, and Ridder (2003), but also allow for different nonparametric procedures as well as a wider class of problems since we reduce the requirement on smoothness of the underlying conditional expectation by speeding up the rates of convergence.

We consider the nonparametric estimation of the generalized propensity score, that is, the vector-valued function given by $\mathbf{p}^*(\cdot) = [p_0^*(\cdot), p_1^*(\cdot), \cdots, p_J^*(\cdot)]'$ with $p_t^*(\cdot) : \mathcal{X}_t \to \mathcal{P}_t$ for all $t \in \mathcal{T}$. For simplicity we assume that $\mathcal{X}_t = \mathcal{X}$ and $\mathcal{P}_t = \mathcal{P}$ for all $t \in \mathcal{T}$, and that the true functions $p_t^*(\cdot)$ enjoy the same degree of smoothness. These restrictions can be weaken without altering the conclusion in this Appendix, provided the notation is modified accordingly. The proposed estimator works as follows. Let K denote the number of basis functions in the series and define vector $\mathbf{r}_K(\mathbf{x}) = [r_{1K}(\mathbf{x}), \cdots, r_{KK}(\mathbf{x})]'$ of approximating functions. It is customary to use the matrix norm $\|A\| = \sqrt{\mathsf{trace}(A'A)}$, whose properties are well-known. Under some conditions imposed below and by choosing an appropriate non-singular linear transformation we can assumed without loss of generality that $\mathbb{E}[\mathbf{r}_K(\mathbf{X})\,\mathbf{r}_K'(\mathbf{X})] = \mathbf{I}_K$, where $\mathbf{I}_K$ is the $(K \times K)$ identity matrix (see Newey (1997) for details). Let $\zeta(K) = \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{r}_K(\mathbf{x})\|$, and observe that in general this bound will depend on the approximating functions chosen. To reduce notational burden we use the same number of approximating functions for each conditional probability, a feature that can be easily relaxed.

To deal with all the relevant probabilities simultaneously we define $\mathbf{p}_{-0}(\mathbf{X}_n) = [p_1(\mathbf{X}_n), \cdots, p_J(\mathbf{X}_n)]' \in \mathbb{R}^J$, $\gamma_{-0,K} = \left[\gamma_{K,1}', \cdots, \gamma_{K,J}'\right]' \in \mathbb{R}^{JK}$, and $\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K) = \left[\mathbf{r}_K(\mathbf{X}_n)'\gamma_{K,1}, \cdots, \mathbf{r}_K(\mathbf{X}_n)'\gamma_{K,J}\right]' \in \mathbb{R}^J$. Recall that $p_0^*(\mathbf{X}_n) = 1 - \sum_{j=1}^J p_j^*(\mathbf{X}_n)$.

Define for a vector $\mathbf{z} \in \mathbb{R}^J$, $\mathbf{z} = [z_1, \cdots, z_J]'$, the functions $L_t : \mathbb{R}^J \to \mathbb{R}$ and $L_t^{-1} : \mathbb{R}^J \to \mathbb{R}$, for all $t = 1, 2, \cdots, J$,

$$L_t(\mathbf{z}) = \frac{\exp\{z_t\}}{1 + \sum_{j=1}^J \exp\{z_j\}}, \quad \text{and} \quad L_t^{-1}(\mathbf{z}) = \log\left\{\frac{z_t}{1 - \sum_{j=1}^J z_t}\right\}.$$

and set $L_0(\mathbf{z}) = 1 - \sum_{j=1}^J L_t(\mathbf{z})$. The gradient of $L_t : \mathbb{R}^J \to \mathbb{R}$ is given by

$$\dot{L}_t(\mathbf{z}) = [-L_t(\mathbf{z})\cdot L_1(\mathbf{z}),\ \cdots,\ -L_t(\mathbf{z})\cdot L_{t-1}(\mathbf{z}),\ L_t(\mathbf{z})\cdot(1 - L_t(\mathbf{z})),\ -L_t(\mathbf{z})\cdot L_{t+1}(\mathbf{z}),\ -L_t(\mathbf{z})\cdot L_J(\mathbf{z})]'$$

and observe that $\sup_{\mathbf{z}}\left|\dot{L}_t(\mathbf{z})\right| < C(J)$ since $|L_t(\mathbf{z})\cdot L_1(\mathbf{z})| < 1$ and $L_t(\mathbf{z})\cdot(1 - L_t(\mathbf{z})) < 1/4$. Also define the vector valued functions $\mathbf{L}(\mathbf{z}) = [L_1(\mathbf{z}), \cdots, L_J(\mathbf{z})]'$ and $\mathbf{L}^{-1}(\mathbf{z}) = \left[L_1^{-1}(\mathbf{z}),\ \cdots,\ L_J^{-1}(\mathbf{z})\right]'$ and observe that the function $\mathbf{L}(\cdot)$ is differentiable with gradient (matrix) $\dot{\mathbf{L}}(\mathbf{z}) = \left[\dot{L}_1(\mathbf{z}),\ \cdots,\ \dot{L}_J(\mathbf{z})\right] \in \mathbb{R}^{J\times J}$ and notice that $\sup_{\mathbf{z}}\left|\dot{\mathbf{L}}(\mathbf{z})\right| < C(J)$, for some constant $C(J)$ that only depends on J.

We consider the model

$$p_t(\mathbf{X}_n) = \mathbb{P}[T = t|\,\mathbf{X}_n] = \frac{\exp\left\{\mathbf{r}_K(\mathbf{X}_n)'\gamma_{K,t}\right\}}{1 + \sum_{j=1}^J \exp\left\{\mathbf{r}_K(\mathbf{X}_n)'\gamma_{K,j}\right\}} = L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))$$

for $t = 1, 2, \cdots, J$ and we set $p_0(\mathbf{X}_n) = \mathbb{P}[T = 0|\,\mathbf{X}_n] = L_0(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))$ to gain identification (i.e., we set $\gamma_{K,0} = \mathbf{0}$).

The multinomial logistic log-likelihood is given by

$$\ell_N(\gamma_K) = \frac{1}{N}\sum_{n=1}^N \sum_{t=0}^J D_n(t)\cdot\log(L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))),$$

with solution $\hat{\gamma}_K = \arg\max_{\gamma_K} \ell_N(\gamma_K)$ and estimated probabilities given by $\hat{p}_t(\mathbf{X}_n) = L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \hat{\gamma}_K))$ for all $t \in \mathcal{T}$. Verify that

$$\begin{aligned}
\frac{\partial}{\partial\gamma_{K,t}}\ell_N(\gamma_K) &= \frac{1}{N}\sum_{n=1}^N [D_n(t) - L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))]\cdot\mathbf{r}_K(\mathbf{X}_n), \qquad \text{for } t = 1, 2, ..., J, \\
\frac{\partial^2}{\partial\gamma_{K,i}\partial\gamma_{K,j}'}\ell_N(\gamma_K) &= -\frac{1}{N}\sum_{n=1}^N L_j(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))\cdot[\mathbf{1}\{i = j\} - L_i(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))]\cdot\mathbf{r}_K(\mathbf{X}_n)\,\mathbf{r}_K(\mathbf{X}_n)',
\end{aligned}$$

and in matrix notation we have

$$\begin{aligned}
\frac{\partial}{\partial\gamma_K}\ell_N(\gamma_K) &= \frac{1}{N}\sum_{n=1}^N [\mathbf{D}_n - \mathbf{L}(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))]\otimes\mathbf{r}_K(\mathbf{X}_n), \\
\frac{\partial^2}{\partial\gamma_K\partial\gamma_K'}\ell_N(\gamma_K) &= -\frac{1}{N}\sum_{n=1}^N \mathbf{H}(\mathbf{X}_n, \gamma_K)\otimes\mathbf{r}_K(\mathbf{X}_n)\,\mathbf{r}_K(\mathbf{X}_n)',
\end{aligned}$$

where $\mathbf{D}_n = [D_n(t) : t = 1, 2, \cdots J]'$ and

$$\begin{aligned}\mathbf{H}(\mathbf{X}_n, \gamma_K) &= [L_i(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K)) \cdot (\mathbf{1}\{i = j\} - L_j(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))) : i = 1, 2, \cdots J,\ j = 1, 2, \cdots J] \\ &= \operatorname{diag}(\mathbf{L}(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))) - \mathbf{L}(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))\mathbf{L}(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma_K))'.\end{aligned}$$

To derive the uniform rates of convergence, we impose the followings conditions:

Assumption 7. *(7.1) (Random Variables) $\{(T_n, \mathbf{X}_n)\}_{n=1}^N$ are i.i.d. with $T_n \in \mathcal{T}$, $\mathbf{X}_n \in \mathcal{X}^r$ and $p_t^* \in \mathcal{P} = [\underline{p},\ \bar{p}]$ $(0 <\underline{p}< \bar{p} < 1)$.*

(7.2) (Nonsingular Second Moment Matrix) The smallest eigenvalue of $\mathbb{E}[\mathbf{r}_K(\mathbf{X})\mathbf{r}_K'(\mathbf{X})]$ is bounded away from zero uniformly in K.

(7.3) (Series Bound) There is a sequence of constants $\zeta(K)$ satisfying $\sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{r}_K(\mathbf{x})\| \le \zeta(K)$, for $K = K(N) \to \infty$ and $\zeta^2(K)KN^{-1} \to 0$, as $N \to \infty$.

(7.4) (Log-Odds Smoothness) For all $t \in \mathcal{T}$ there exists $\gamma_{K,t}^0 \in \mathbb{R}^K$ and $\alpha > 0$ such that

$$\sup_{\mathbf{x}\in\mathcal{X}} \left| \log\left(\frac{p_t^*}{p_0^*}\right)(\mathbf{x}) - \mathbf{r}_K(\mathbf{x})'\gamma_{K,t}^0 \right| = O\left(K^{-\alpha}\right), \quad \textit{and} \quad \zeta(K)^2 K^{-\alpha} \to 0.$$

Before stating the main result, we need the following Lemma that allows us to control the matrix $\mathbf{H}(\mathbf{X}, \gamma_K)$.

Lemma 1. *(Uniform Lower Bound for $\mathbf{H}(\mathbf{X}, \gamma_K)$)*

$$\inf_{\mathbf{x}\in\mathcal{X}} \mathbf{H}(\mathbf{x}, \gamma) \ge \inf_{\mathbf{x}\in\mathcal{X}} \prod_{t=0}^{J} L_t(\mathbf{R}_{-0}(\mathbf{x}, \gamma)) \cdot \mathbf{I}_J.$$

Proof of Lemma 1 (Uniform Lower Bound for $\mathbf{H}(\mathbf{X}, \gamma_K)$):

Recall that $\mathbf{L}(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma)) = [L_1(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma)), \cdots, L_J(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma))]'$, with $L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma)) > 0$, for all $t = 1, 2, \cdots, J$, and $\sum_{t=1}^J L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma)) < 1$. Using the same reasoning as in Watson (1996), it is seen that $0 < \lambda_{\min}(\mathbf{H}(\mathbf{x}, \gamma)) \le \lambda_{\max}(\mathbf{H}(\mathbf{x}, \gamma)) \le \max_t L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma)) < 1$, which implies that $\lambda_{\min}(\mathbf{H}(\mathbf{x}, \gamma)) \ge \det(\mathbf{H}(\mathbf{X}_n, \gamma))$. Next, using the exact choleskey decomposition derived by Tanabe and Sagae (1992) and the properties of the determinant we conclude that

$$\lambda_{\min}(\mathbf{H}(\mathbf{x}, \gamma)) \ge \det(\mathbf{H}(\mathbf{X}_n, \gamma)) = \prod_{t=0}^{J} L_t(\mathbf{R}_{-0}(\mathbf{X}_n, \gamma)).$$

Finally, since $\mathbf{H}(\mathbf{X}_n, \gamma)$ is a symmetric positive definite matrix, an orthogonal decomposition gives $\mathbf{H}(\mathbf{X}_n, \gamma) = \mathbf{O\Lambda O'} \ge \lambda_{\min}(\mathbf{H}(\mathbf{X}_n, \gamma)) \cdot \mathbf{I}_J$ and the result follows. **Q.E.D.**

The following theorem provides the uniform rate of convergence for the MLSE.

Theorem 8. *(Uniform Rate of Convergence) Under Assumption 7,*

$$\begin{aligned}(i) &\quad : \quad \sup_{\mathbf{x}\in\mathcal{X}} \left|\mathbf{p}_K^0(\mathbf{x}) - \mathbf{p}^*(\mathbf{x})\right| = O\left(K^{-\alpha}\right). \\ (ii) &\quad : \quad \left|\hat{\gamma}_K - \gamma_K^0\right| = O_p\left(K^{1/2}N^{-1/2} + \zeta(K)K^{-\alpha}\right).\end{aligned}$$

Conclude that

$$\sup_{\mathbf{x}\in\mathcal{X}} |\hat{\mathbf{p}}(\mathbf{x}) - \mathbf{p}^*(\mathbf{x})| = O_p\left(\zeta(K)K^{1/2}N^{-1/2} + \zeta(K)^2 K^{-\alpha}\right)$$

Proof of Lemma 1 (Uniform Rate of Convergence): Assumption (7.4) implies that $\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{L}^{-1}\left(\mathbf{p}^*_{-0}\right)(\mathbf{x})-\mathbf{r}\left(\mathbf{x},\gamma_K^0\right)\right|=O\left(K^{-\alpha}\right)$. Since the mapping $\mathbf{L}(\cdot)$ is differentiable with $\sup_{\mathbf{z}}\left|\dot{\mathbf{L}}(\mathbf{z})\right|<C(J)$, an application of the mean value theorem gives

$$\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{p}^*_{-0}(\mathbf{x})-\mathbf{L}\left(\mathbf{R}_{-0}\left(\mathbf{x},\gamma_K^0\right)\right)\right|=\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{L}\left(\mathbf{L}^{-1}\left(\mathbf{p}^*_{-0}\right)\right)(\mathbf{x})-\mathbf{L}\left(\mathbf{R}_{-0}\left(\mathbf{x},\gamma_K^0\right)\right)\right|\leq C(J)\cdot\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{L}^{-1}\left(\mathbf{p}^*_{-0}\right)(\mathbf{x})-\mathbf{R}_{-0}\left(\mathbf{x},\gamma_K^0\right)\right|,$$

giving (i). (Extending the previous result to all $(J+1)$ probabilities follows directly since $p_0^*(\mathbf{x})=1-\sum_{j=1}^J p_j^*(\mathbf{X}_n)$ and $L_0\left(\mathbf{R}_{-0}\left(\mathbf{x},\gamma_K^0\right)\right)=1-\sum_{j=1}^J L_j\left(\mathbf{R}_{-0}\left(\mathbf{x},\gamma_K^0\right)\right)$.)

Next we establish (ii). Let $\hat{\Omega}_K=N^{-1}\sum_{n=1}^N\mathbf{r}_K(\mathbf{X}_n)\,\mathbf{r}_K(\mathbf{X}_n)'$, and observed that according to Newey (1997) $\left|\hat{\Omega}_K-\mathbf{I}_K\right|=O_p\left(\zeta(K)\,K^{1/2}N^{-1/2}\right)$. define the event $\mathbb{A}_N=\left\{\lambda_{\min}\left(\hat{\Omega}_K\right)>1/2\right\}$ and note that under Assumption 7.3 $O_p\left(\zeta(K)\,K^{1/2}N^{-1/2}\right)=o_p(1)$, which implies $\mathbb{P}[\mathbb{A}_N]\to1$.

Now, note that

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{\partial}{\partial\gamma}\ell_N\left(\gamma_K^0\right)\right|\right]\\
&=\mathbb{E}\left[\left|\frac{1}{N}\sum\nolimits_{n=1}^N\left[\mathbf{D}_n-\mathbf{L}\left(\mathbf{R}_{-0}\left(\mathbf{X}_n,\gamma_K^0\right)\right)\right]\otimes\mathbf{r}_K(\mathbf{X}_n)\right|\right]\\
&\leq\left(\mathbb{E}\left[\left|\frac{1}{N}\sum\nolimits_{n=1}^N\left[\mathbf{D}_n-\mathbf{p}^*(\mathbf{X}_n)\right]\otimes\mathbf{r}_K(\mathbf{X}_n)\right|^2\right]\right)^{1/2}+\mathbb{E}\left[\left|\frac{1}{N}\sum\nolimits_{n=1}^N\left[\mathbf{p}^*(\mathbf{X}_n)-\mathbf{L}\left(\mathbf{R}_{-0}\left(\mathbf{X}_n,\gamma_K^0\right)\right)\right]\otimes\mathbf{r}_K(\mathbf{X}_n)\right|\right]\\
&\leq C\cdot\left(\frac{1}{N}\cdot\mathbb{E}\left[\left|\left[\mathbf{D}_n-\mathbf{p}^*(\mathbf{X}_n)\right]\otimes\mathbf{r}_K(\mathbf{X}_n)\right|^2\right]\right)^{1/2}+C\cdot\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{p}^*(\mathbf{x})-\mathbf{L}\left(\mathbf{R}_{-0}\left(\mathbf{x},\gamma_K^0\right)\right)\right|\cdot\sup_{\mathbf{x}\in\mathcal{X}}\left|\mathbf{r}_K(\mathbf{X}_n)\right|\\
&=O\left(K^{1/2}N^{-1/2}+\zeta(K)\,K^{-\alpha}\right),
\end{aligned}$$

under Assumption 7.4 and by the Markov's Inequality we conclude

$$\left|\frac{\partial}{\partial\gamma}\ell_N\left(\gamma_K^0\right)\right|=O_p\left(K^{1/2}N^{-1/2}+\zeta(K)\,K^{-\alpha}\right),$$

which implies that for any fixed constant $\varsigma>0$ the probability of the event

$$\mathbb{B}_N(\varsigma)=\left\{\left|\frac{\partial}{\partial\gamma}\ell_N\left(\gamma_K^0\right)\right|<\varsigma\cdot\left(K^{1/2}N^{-1/2}+\zeta(K)\,K^{-\alpha}\right)\right\}$$

goes to one, i.e., $\mathbb{P}[\mathbb{B}_N(\varsigma)]\to1$.

Next, let $\delta=\inf_{\mathbf{x}\in\mathcal{X}}\prod_{t=0}^J L_t\left(\mathbf{R}_{-0}\left(\mathbf{x},\gamma_K^0\right)\right)$ and observe that for K large enough $\delta>0$ by (i) and the assumption that the true probabilities are exactly between zero and one. Define the sets $\Gamma_K^\delta=\left\{\gamma_K\in\mathbb{R}^{JK}:\inf_{\mathbf{x}\in\mathcal{X}}\prod_{t=0}^J L_t\left(\mathbf{R}_{-0}(\mathbf{x},\gamma)\right)>\delta/2\right\}$ and $\Gamma_K^0(\varrho)=\left\{\gamma_K\in\mathbb{R}^{JK}:\left|\gamma_K-\gamma_K^0\right|\leq\varrho\cdot\left(K^{1/2}N^{-1/2}+\zeta(K)\,K^{-\alpha}\right)\right\}$ for any $\varrho>0$, and because

$$\begin{aligned}
\sup_{\mathbf{x}\in\mathcal{X},\boldsymbol{\gamma}\in\Gamma_K^0(\varrho)}\left|\mathbf{L}\left(\mathbf{r}(\mathbf{x},\gamma_K)\right)-\mathbf{L}\left(\mathbf{r}\left(\mathbf{x},\gamma_K^0\right)\right)\right| &\leq \sup_{\mathbf{x}\in\mathcal{X},\boldsymbol{\gamma}\in\Gamma_K^0(\varrho),\tilde{\boldsymbol{\gamma}}_K}\left|\dot{\mathbf{L}}\left(\mathbf{r}(\mathbf{x},\tilde{\gamma}_K)\right)\otimes\mathbf{r}_K(\mathbf{x})'\right|\cdot\left|\gamma_K-\gamma_K^0\right|\\
&\leq C\cdot\zeta(K)\cdot\sup_{\boldsymbol{\gamma}\in\Gamma_K^0(\varrho)}\left|\gamma_K-\gamma_K^0\right|=O\left(\zeta(K)\,K^{1/2}N^{-1/2}+\zeta(K)^2K^{-\alpha}\right)
\end{aligned}$$

and $O\left(\zeta(K)\,K^{1/2}N^{-1/2}+\zeta(K)^2K^{-\alpha}\right)=o(1)$ by Assumptions 7.3 and 7.4, we conclude that for K for large enough $\Gamma_K^\delta\subset\Gamma_K^0(\varrho)$.

To finish the argument, choose N large enough so that $\Gamma_K^\delta\subset\Gamma_K^0(C)$, $\mathbb{P}[\mathbb{A}_N]\geq1-\varepsilon/2$ and $\mathbb{P}[\mathbb{B}_N(\delta C/8)]\geq1-\varepsilon/2$, for any $C>0$. Then for any $\gamma_K\in\Gamma_K^0$ we have

$$\begin{aligned}
-\frac{\partial}{\partial\gamma\partial\gamma'}\ell_N(\gamma_K) &= \frac{1}{N}\sum\nolimits_{n=1}^N\mathbf{H}(\mathbf{X}_n,\gamma_K)\otimes\mathbf{r}_K(\mathbf{X}_n)\,\mathbf{r}_K(\mathbf{X}_n)'\\
&\geq \frac{1}{N}\sum\nolimits_{n=1}^N\left[\inf_{\mathbf{x}\in\mathcal{X}}\prod\nolimits_{t=0}^J L_t\left(\mathbf{R}_{-0}(\mathbf{x},\gamma_K)\right)\cdot\mathbf{I}_J\right]\otimes\mathbf{r}_K(\mathbf{X}_n)\,\mathbf{r}_K(\mathbf{X}_n)'\\
&\geq \frac{\delta}{2}\cdot\left[\mathbf{I}_J\otimes\hat{\Omega}_K\right],
\end{aligned}$$

which implies that with probability at least $(1-\varepsilon)$,

$$\lambda_{\min}\left(-\frac{\partial}{\partial\gamma\partial\gamma'}\ell_N\left(\gamma_K\right)\right)\geq\frac{\delta}{4}.$$

Moreover, under the same conditions (i.e., also with probability at least $(1-\varepsilon)$) we verify that for any $\gamma_K\in\Gamma_K^0-\left\{\gamma_K^0\right\}$ we have

$$\begin{aligned}\ell_N\left(\gamma_K\right)-\ell_N\left(\gamma_K^0\right) &= \frac{\partial}{\partial\gamma}\ell_N\left(\gamma_K^0\right)\cdot\left(\gamma_K-\gamma_K^0\right)-\frac{1}{2}\left(\gamma_K-\gamma_K^0\right)'\left[-\frac{\partial}{\partial\gamma\partial\gamma'}\ell_N\left(\tilde{\gamma}_K\right)\right]\left(\gamma_K-\gamma_K^0\right)\\ &\leq \left|\frac{\partial}{\partial\gamma}\ell_N\left(\gamma_K^0\right)\right|\cdot\left|\gamma_K-\gamma_K^0\right|-\frac{\delta}{8}\cdot\left|\left(\gamma_K-\gamma_K^0\right)\right|^2\\ &\leq \left(\left|\frac{\partial}{\partial\gamma}\ell_N\left(\gamma_K^0\right)\right|-\frac{\delta}{8}\cdot C\cdot\left(K^{1/2}N^{-1/2}+\zeta\left(K\right)K^{-\alpha}\right)\right)\cdot\left|\gamma_K-\gamma_K^0\right|<0,\end{aligned}$$

for some $\tilde{\gamma}_K$ such that $\left|\tilde{\gamma}_K-\gamma_K^0\right|\leq\left|\gamma_K-\gamma_K^0\right|$. Since $\ell_N\left(\gamma_K\right)$ is continuous and concave, it follows that $\hat{\gamma}_K$ maximizes $\ell_N\left(\gamma_K\right)$ and $\hat{\gamma}_K$ satisfies the first order condition wpa1. Now the result follows directly. **Q.E.D.**

C. Appendix C: Proofs of Theorems

This appendix provides proofs for the theorems in the paper and uses the results in Appendix A and Appendix B. Let C denote a generic constant that may be different depending on the context.

Proof of Theorem 1 (Consistency):

This result follows directly by an application of Theorem 6 in Appendix A, after we verify the required sufficient conditions. First, observe that Condition (C1) is automatically verified by the (two-step) estimator considered. Condition (C2) follows directly from the identification Condition (**??**). Next, note that for N large enough,

$$\begin{aligned}\sup_{\beta\in\mathcal{B}}\left\|M_N\left(\beta_t,\hat{p}_t\left(\cdot\right)\right)-M_N\left(\beta_t,p_t^*\left(\cdot\right)\right)\right\| &= \sup_{\beta\in\mathcal{B}}\left\|\frac{1}{N}\sum\nolimits_{n=1}^N\frac{D_n\left(t\right)\cdot m\left(Y_n,\mathbf{X}_n;\beta_t\right)}{\hat{p}_t\left(\mathbf{X}_n\right)\cdot p_t^*\left(\mathbf{X}_n\right)}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right)-p_t^*\left(\mathbf{X}_n\right)\right)\right\|\\ &\leq C\cdot\left\|p_t\left(\cdot\right)-p_t^*\left(\cdot\right)\right\|_\infty\cdot\frac{1}{N}\sum\nolimits_{n=1}^N\sup_{\beta\in\mathcal{B}}\left|m\left(Y_n\left(t\right),\mathbf{X}_n;\beta\right)\right|\\ &= O_p\left(\left\|p_t\left(\cdot\right)-p_t^*\left(\cdot\right)\right\|_\infty\right),\end{aligned}$$

where the second line uses Assumption 2.2 and the third line uses Assumption 3.2, establishing condition (C3). Condition (C4) is assumed by the theorem. Finally, to verify Condition (C5) simply note that by Assumption 2.2 and an application of Theorem 2.10.6 of van der Vaart and Wellner (1996) we conclude that the class of functions (for any fixed $j\in\mathcal{T}$) $\mathcal{F}_j=\left\{\mathbf{1}\left\{\cdot=j\right\}\cdot m\left(\cdot;\beta\right)/p_t^*\left(\cdot\right):m\left(\cdot;\beta\right)\in\mathcal{M}\right\}$ is Glivenko-Cantelli with finite integrable envelop by Assumptions 3.1 and 3.2. **Q.E.D.**

Proof of Theorem 2 (Efficient Influence Function and SPEB):

The proof given here follows the theoretical approach in Bickel, Klaassen, Ritov, and Wellner (1993), and Newey (1990). The derivation involves three main steps: tangent space characterization, pathwise differentiability of the parameter of interest, and SPEB computation. Let $L_0^2\left(F_W\right)$ be the usual Hilbert space of zero mean squared integrable functions with respect to the distribution function F_W.

Step 1: Tangent Space Characterization. For a (regular) parametric submodel (see, e.g., Appendix A in Newey (1990) for definitions and regularity conditions) of the distribution of $\mathbf{Z}=\left(Y,T,\mathbf{X}\right)$, the observed data model, the log-likelihood is given by is given by

$$\log f\left(y,\tau,\mathbf{x};\theta\right)=\sum\nolimits_{t\in\mathcal{T}}\mathbf{1}\left\{t=\tau\right\}\cdot\left[\log f_t\left(y\mid\mathbf{x};\theta\right)+\log p_t\left(\mathbf{x};\theta\right)\right]+\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right),$$

which equals $\log f\left(y,\tau,\mathbf{x}\right)$ when $\theta=\theta_0$, and where we have used the definition $f_\tau\left(y\mid\mathbf{x}\right)\equiv\int\cdots\int f_{\mathbf{Y}^*|\mathbf{X}}\left(\mathbf{y}\mid\mathbf{x}\right)\cdot d\mathbf{y}_{-\tau}$ with $\mathbf{y}_{-\tau}\equiv\left[y\left(t\right):t\in\mathcal{T}-\left\{\tau\right\}\right]'$ and Assumption 2.2. The corresponding score is given by

$$S\left(y,\tau,\mathbf{x};\theta_0\right)=\frac{d}{d\theta}\log f\left(y,\tau,\mathbf{x};\theta\right)\bigg|_{\theta_0}=\sum\nolimits_{t\in\mathcal{T}}\mathbf{1}\left\{t=\tau\right\}\cdot s_t\left(y\mid\mathbf{x}\right)+\gamma\left(\tau,\mathbf{x}\right)+s_x\left(\mathbf{x}\right),$$

where $s_t\left(y \mid \mathbf{x}\right) \equiv \frac{d}{d\theta}\log f_t\left(y \mid \mathbf{x};\theta\right)\big|_{\theta_0}$, $\gamma\left(\tau,\mathbf{x}\right) \equiv \sum_{t\in\mathcal{T}} \mathbf{1}\left\{t=\tau\right\}\cdot\frac{\dot{p}_t(\mathbf{x};\theta_0)}{p_t(\mathbf{x})}$ where $\dot{p}_t\left(\mathbf{x};\theta_0\right) = \frac{d}{d\theta}p_t\left(\mathbf{x};\theta\right)\big|_{\theta_0}$, and $s_x\left(\mathbf{x}\right) \equiv \frac{d}{d\theta}\log f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\big|_{\theta_0}$. Therefore, the tangent space of this statistical model is characterized by the set of functions $\mathbb{T} \equiv \mathbb{T}_y + \mathbb{T}_p + \mathbb{T}_x$, where

$$\begin{aligned}
\mathbb{T}_y &\equiv \left\{\left\{s_t\left(Y\left(t\right) \mid \mathbf{X}\right)\right\}_{t\in\mathcal{T}} : s_t\left(Y\left(t\right) \mid \mathbf{X}\right) \in L_0^2\left(F_{Y(t)|\mathbf{X}}\right), \forall t\right\}, \\
\mathbb{T}_p &\equiv \left\{\gamma\left(T,\mathbf{X}\right) : \gamma\left(T,\mathbf{X}\right) \in L_0^2\left(F_{T|\mathbf{X}}\right)\right\}, \text{ and} \\
\mathbb{T}_x &\equiv \left\{s_x\left(\mathbf{X}\right) : s_x\left(\mathbf{X}\right) \in L_0^2\left(F_{\mathbf{X}}\right)\right\}.
\end{aligned}$$

In particular, observe that

$$\mathbb{E}\left[\gamma\left(T,\mathbf{X}\right)\middle|\,\mathbf{X}\right] = \mathbb{E}\left[\sum\nolimits_{t\in\mathcal{T}} D\left(t\right)\cdot\frac{\dot{p}_t\left(\mathbf{X};\theta_0\right)}{p_t\left(\mathbf{X}\right)}\,\middle|\,\mathbf{X}\right] = \sum\nolimits_{t\in\mathcal{T}} \dot{p}_t\left(\mathbf{X};\theta_0\right),$$

and

$$\mathbb{E}\left[\left(\gamma\left(T,\mathbf{X}\right)\right)^2\middle|\,\mathbf{X}\right] = \mathbb{E}\left[\sum\nolimits_{i\in\mathcal{T}}\sum\nolimits_{j\in\mathcal{T}} D\left(i\right)\cdot\frac{\dot{p}_i\left(\mathbf{X};\theta_0\right)}{p_i\left(\mathbf{X}\right)}\cdot D\left(j\right)\cdot\frac{\dot{p}_j\left(\mathbf{X};\theta_0\right)}{p_j\left(\mathbf{X}\right)}\,\middle|\,\mathbf{X}\right] = \sum\nolimits_{t\in\mathcal{T}}\frac{\left(\dot{p}_t\left(\mathbf{X};\theta_0\right)\right)^2}{p_t\left(\mathbf{X}\right)},$$

and hence it is required that $p_t\left(\mathbf{x}\right)$ and $\dot{p}_t\left(\mathbf{x};\theta_0\right)$ are any measurable functions such that $\sum_{t\in\mathcal{T}}\dot{p}_t\left(\mathbf{X};\theta_0\right) = 0$ and $\sum_{t\in\mathcal{T}}\frac{(\dot{p}_t(\mathbf{X};\theta_0))^2}{p_t(\mathbf{X})} < \infty$, almost surely. Notice that the first condition implies that by varying the model, the probabilities should change in such a way that they still add up to one. This is guaranteed in the model since, by assumption, $\sum_{t\in\mathcal{T}} p_t\left(\mathbf{x};\theta_0\right) = 1$. The second condition is automatically satisfied by Assumption 2.2 and the fact that T is finite.

Step 2: Pathwise Differentiability of the Parameter of Interest. Observe that using the implicit function theorem,

$$\left[\frac{\partial}{\partial\theta}\beta_t\left(\theta\right)\bigg|_{\theta=\theta_0} : t\in\mathcal{T}\right]' = -\left[\frac{\partial}{\partial\beta_t}\mathbb{E}\left[\left[m\left(Y\left(t\right),\mathbf{X};\beta_t\right) : t\in\mathcal{T}\right]'\right]\right]^{-1}\left[\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[\left[m\left(Y\left(t\right),\mathbf{X};\beta_t\right) : t\in\mathcal{T}\right]'\right]\bigg|_{\theta=\theta_0}\right].$$

In this case, we have

$$\left[\frac{\partial}{\partial\beta_t}\mathbb{E}\left[\left[m\left(Y\left(t\right),\mathbf{X};\beta_t\right) : t\in\mathcal{T}\right]'\right]\right] = \operatorname{diag}\left[v_t\left(\beta_t\right) : t\in\mathcal{T}\right]$$

and hence for the t-th coordinate

$$\begin{aligned}
\frac{\partial}{\partial\theta}\beta_t\left(\theta\right)\bigg|_{\theta=\theta_0} &= \frac{1}{v_t\left(\beta_t\right)}\cdot\left(\frac{\partial}{\partial\theta}\iint m\left(y,\mathbf{x};\beta_t\right)\cdot f_t\left(y\middle|\,\mathbf{x};\theta\right)\cdot f_{\mathbf{X}}\left(\mathbf{x};\theta\right)\cdot dy\cdot d\mathbf{x}\bigg|_{\theta=\theta_0}\right) \\
&= \frac{1}{v_t\left(\beta_t\right)}\cdot\left(\mathbb{E}\left[m\left(Y\left(t\right),\mathbf{X};\beta_t\right)\cdot s_t\left(Y\left(t\right)\middle|\,\mathbf{X}\right)\right] + \mathbb{E}\left[\mathcal{E}_t\left(\mathbf{X};\beta_t\right)\cdot s_x\left(\mathbf{X}\right)\right]\right),
\end{aligned}$$

where $\mathcal{E}_t\left(\mathbf{X};\beta_t\right) = \mathbb{E}\left[m\left(Y\left(t\right),\mathbf{X};\beta_t\right)\middle|\,\mathbf{X}\right]$, for all $t\in\mathcal{T}$. To show that the parameter is pathwise differentiable, a function $\mathbf{d}_{\boldsymbol{\beta}}\left(y,t,\mathbf{x}\right)\in\mathbb{R}^{J+1}$ is needed such that for all regular parametric submodels

$$\left[\frac{\partial}{\partial\theta}\beta_t\left(\theta\right)\bigg|_{\theta=\theta_0} : t\in\mathcal{T}\right]' = \mathbb{E}\left[\mathbf{d}_{\boldsymbol{\beta}}\left(Y,T,\mathbf{X}\right)\cdot S\left(Y,T,\mathbf{X};\theta_0\right)\right].$$

It is an standard exercise to verify that the function $\mathbf{d}_{\boldsymbol{\beta}}\left(Y,T,\mathbf{X}\right)$ is given by

$$\mathbf{d}_{\boldsymbol{\beta}}\left(Y,T,\mathbf{X}\right) \equiv \left[d_{\beta_t}\left(Y,T,\mathbf{X}\right) : t\in\mathcal{T}\right]' \equiv \left[\frac{1}{v_t\left(\beta_t\right)}\cdot\left(\frac{D\left(t\right)}{p_t\left(\mathbf{X}\right)}\cdot\left(m\left(Y,\mathbf{X};\beta_t\right) - \mathcal{E}_t\left(\mathbf{X};\beta_t\right)\right) + \mathcal{E}_t\left(\mathbf{X};\beta_t\right)\right) : t\in\mathcal{T}\right]',$$

and therefore the population parameter of interest is pathwise differentiable.

Step 3: SPEB Computation. Since $d_{\beta_t}\left(Y,T,\mathbf{X}\right)\in\mathbb{T}$, it follows that $\mathit{Proj}\left(d_{\beta_t}\left(Y,T,\mathbf{X}\right)\middle|\,\mathbb{T}\right) = d_{\beta_t}\left(Y,T,\mathbf{X}\right)$, for all $t\in\mathcal{T}$, and therefore the variance bound is given by $\mathit{SPEB}\left(\boldsymbol{\beta}\right) = \mathbb{E}\left[\mathbf{d}_{\boldsymbol{\beta}}\left(Y,T,\mathbf{X}\right)\mathbf{d}_{\boldsymbol{\beta}}'\left(Y,T,\mathbf{X}\right)\right]$. **Q.E.D.**

Proof of Theorem 3 (Asymptotic Linear Representation and Asymptotic Normality):

This result follows directly by an application of Theorem 7 in Appendix A, after we verify the required sufficient conditions. First, set the sequence $\delta_N = o(1)$ accordingly. Condition (AN1) follows directly from the definition of the estimator and the identifying condition, while Condition (AN2) holds by Assumption 4 since $M\left(\beta, p_t^*(\cdot)\right) = \mathbb{E}\left[m\left(Y(t), \mathbf{X}; \beta\right)\right]$.

Next, define

$$\Delta_N\left(\beta_t, p_t(\cdot) - p_t^*(\cdot)\right) = \frac{1}{N}\sum\nolimits_{n=1}^{N} \frac{D_n(t)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t\right)}{p_t^*\left(\mathbf{X}_n\right)^2} \cdot \left(p_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right),$$

and observe that for N large enough

$$\begin{aligned}&\sup_{\beta\in\mathcal{B}_{\delta_N}} \left|M_N\left(\beta, \hat{p}_t(\cdot)\right) - M_N\left(\beta, p_t^*(\cdot)\right) - \Delta_N\left(\beta, \hat{p}_t(\cdot) - p_t^*(\cdot)\right)\right| \\ &\leq C\cdot\left\|\hat{p}_t(\cdot) - p_t^*(\cdot)\right\|_\infty^2 \cdot \frac{1}{N}\sum\nolimits_{n=1}^{N} \sup_{\beta\in\mathcal{B}} \left|m\left(Y_n(t), \mathbf{X}_n; \beta\right)\right| = O_p\left(\left\|\hat{p}_t(\cdot) - p_t^*(\cdot)\right\|_\infty^2\right),\end{aligned}$$

where the result uses Assumption 2.2 and Assumption 5.2. This gives the first part of condition (AN3). To verify the second part, define the empirical process

$$\upsilon_N(\beta) = \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\frac{D_n(t)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left|m\left(Y_n, \mathbf{X}_n; \beta\right) - m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)\right| - \mathbb{E}\left[\left|m\left(Y(t), \mathbf{X}; \beta\right) - m\left(Y(t), \mathbf{X}; \beta_t^*\right)\right|\right]\right\},$$

and observe that $\upsilon_N\left(\beta_t^*\right) = 0$. Also, notice that the parametrization is L^2-continuous by Assumption 5.4 and by an application of Theorem 2.10.6 of van der Vaart and Wellner (1996) we conclude that the class of functions (for any fixed $j \in \mathcal{T}$) $\mathcal{F}_j = \left\{\mathbf{1}\{\cdot = j\}\cdot\left|m(\cdot;\beta) - m(\cdot;\beta_t^*)\right|/p_t^*(\cdot) : m(\cdot;\beta)\in\mathcal{M}\right\}$ is Donsker with finite square-integrable envelop by Assumptions 5.1 and 5.2. Hence, Lemma 3.3.5 of van der Vaart and Wellner (1996) gives $\sup_{\beta\in\mathcal{B}_{\delta_N}}\left|\upsilon_N(\beta)\right| = o_p(1)$. Using this result, letting $\Upsilon(\beta) = \mathbb{E}\left[\left|m\left(Y(t), \mathbf{X}; \beta\right) - m\left(Y(t), \mathbf{X}; \beta_t^*\right)\right|\right]$ and by Assumption 5.3 we obtain for N large enough

$$\begin{aligned}&\left|D_N\left(\hat{\beta}_t, p_t(\cdot) - p_t^*(\cdot)\right) - D_N\left(\beta_t^*, p_t(\cdot) - p_t^*(\cdot)\right)\right| \\ &= \left|\frac{1}{N}\sum\nolimits_{n=1}^{N}\frac{D_n(t)}{p_t^*\left(\mathbf{X}_n\right)^2}\cdot\left(m\left(Y_n, \mathbf{X}_n; \hat{\beta}_t\right) - m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)\right)\cdot\left(p_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right)\right| \\ &\leq C_1\cdot\left\|p_t(\cdot) - p_t^*(\cdot)\right\|_\infty\cdot N^{-1/2}\cdot\sup_{\beta\in\mathcal{B}_{\delta_N}}\left|\upsilon_n(\beta)\right| + C_2\cdot\left\|p_t(\cdot) - p_t^*(\cdot)\right\|_\infty\cdot\Upsilon\left(\hat{\beta}_t\right) \\ &= O_p\left(\left\|p_t(\cdot) - p_t^*(\cdot)\right\|_\infty\cdot\left|\hat{\beta}_t - \beta_t^*\right|\right),\end{aligned}$$

which verifies condition (AN3). Finally, Condition (AN4) is assumed, Condition (AN5) follows directly by an application of Theorem 2.10.6 of van der Vaart and Wellner (1996) to conclude that the class of functions (for all fixed $j \in \mathcal{T}$) $\mathcal{F}_j = \left\{\mathbf{1}\{\cdot = j\}\cdot m(\cdot;\beta)/p_t^*(\cdot) : m(\cdot;\beta)\in\mathcal{M}\right\}$ is Donsker with finite integrable envelop by Assumption 4.2, and Condition (AN6) is also assumed directly in the theorem at this point.

As a consequence we conclude that $\sqrt{N}\cdot\left(\hat{\beta}_t - \beta_t^*\right) = -\sqrt{N}\cdot M_N\left(\beta^*, \hat{p}(\cdot)\right)/\upsilon_t\left(\beta_t^*\right) + o_p(1)$, giving the result of the theorem after applying the second condition. The asymptotic normality and efficiency of the estimator follows directly. **Q.E.D.**

Proof of Theorem 4 (Conditions (AN.1) and (AN.2)):

First, observe that for power series and splines, we have $\zeta(K) = K^\eta$, with $\eta = 1$ and $\eta = 1/2$ respectively, and using Assumption 6 (which implies Assumption 7 in this cases), have $\alpha = s/r$ (see, e.g., Newey (1997)). Now Theorem 8 implies

$$N^{1/4}\cdot\sup_{\mathbf{x}\in\mathcal{X}}\left|\hat{\mathbf{p}}(\mathbf{x}) - \mathbf{p}^*(\mathbf{x})\right| = N^{1/4}\cdot O_p\left(K^\eta K^{1/2} N^{-1/2} + K^{2\eta}K^{-s/r}\right) = o_p(1),$$

under the assumptions of the Theorem and therefore Condition (AN.1) holds.

Next, we consider condition (AN.2). Observe that it is enough to show the result for a typical t-th component of the vector. Thus,

$$\begin{aligned}
&\left| M_N\left(\beta_t^*, \hat{p}_t\left(\cdot\right)\right) - M_N\left(\beta_t^*, p_t^*\left(\cdot\right)\right) + \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N} \alpha_t\left(T_n, \mathbf{X}_n; p_t^*\left(\cdot\right)\right)\right| \\
&= \left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N} \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{\hat{p}_t\left(\mathbf{X}_n\right)} - \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{ \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} - \frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(D_n\left(t\right) - p_t^*\left(\mathbf{X}_n\right)\right)\right\}\right| \\
&\leq \left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{ \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{\hat{p}_t\left(\mathbf{X}_n\right)} - \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} + \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}\right)\right)\right\}\right|
\end{aligned}\tag{3}$$

$$+ \left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{ -\frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right) + \frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}\right)\right)\right\}\right| \tag{4}$$

$$+ \left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{ -\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right) + \frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(D_n\left(t\right) - p_t^*\left(\mathbf{X}_n\right)\right)\right\}\right|. \tag{5}$$

The bound of term (3) is given by

$$\begin{aligned}
&\left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{ \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{\hat{p}_t\left(\mathbf{X}_n\right)} - \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} + \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right)\right\}\right| \\
&\leq \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left| \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{\hat{p}_t\left(\mathbf{X}_n\right)\cdot p_t^*\left(\mathbf{X}_n\right)^2}\right|\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right)^2\right\} \\
&\leq \sqrt{N}\cdot\left(\sup_{\mathbf{x}\in\mathcal{X}}\left|\hat{p}_t\left(\mathbf{x}\right) - p_t^*\left(\mathbf{x}\right)\right|\right)^2\cdot\frac{1}{N}\sum\nolimits_{n=1}^{N}\frac{D_n\left(t\right)\cdot\left|m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)\right|}{\left|\hat{p}_t\left(\mathbf{X}_n\right)\right|\cdot p_t^*\left(\mathbf{X}_n\right)^2} \\
&= \sqrt{N}\cdot O_p\left(\left(\zeta\left(K\right)K^{1/2}N^{-1/2} + \zeta\left(K\right)^2 K^{-\alpha}\right)^2\right),
\end{aligned}$$

for N large enough.

The bound of term (4) is given by

$$\begin{aligned}
&\left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{ -\frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right) + \frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right)\right\}\right| \\
&\leq \left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} - \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\right)\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_{K,t}^0\left(\mathbf{X}_n\right)\right)\right\}\right|
\end{aligned}\tag{6}$$

$$+ \left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} - \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\right)\cdot\left(p_{K,t}^0\left(\mathbf{X}_n\right) - p_t^*\left(\mathbf{X}_n\right)\right)\right\}\right|. \tag{7}$$

Now, to obtain a bound on the term (6), first notice that by a second order Taylor expansion and using the results in Appendix B we obtain

$$\begin{aligned}
\hat{p}_t\left(\mathbf{X}_n\right) - p_{K,t}^0\left(\mathbf{X}_n\right) &= \left[\dot{\mathbf{L}}_t\left(\mathbf{R}_{-0}\left(\mathbf{X}_n, \gamma_K^0\right)\right)\otimes\mathbf{r}_K\left(\mathbf{x}\right)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right) \\
&\quad + \frac{1}{2}\left(\hat{\gamma}_K - \gamma_K^0\right)'\left[\mathbf{H}\left(\mathbf{X}_n, \tilde{\gamma}_K\right)\otimes\mathbf{r}_K\left(\mathbf{X}_n\right)\mathbf{r}_K\left(\mathbf{X}_n\right)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right) \\
&\leq \left[\dot{\mathbf{L}}_t\left(\mathbf{R}_{-0}\left(\mathbf{X}_n, \gamma_K^0\right)\right)\otimes\mathbf{r}_K\left(\mathbf{x}\right)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right) + C\cdot\left(\hat{\gamma}_K - \gamma_K^0\right)'\left[\mathbf{I}_J\otimes\mathbf{r}_K\left(\mathbf{X}_n\right)\mathbf{r}_K\left(\mathbf{X}_n\right)'\right]\left(\hat{\gamma}_K - \gamma_K^0\right),
\end{aligned}$$

for some $\tilde{\gamma}_K$ such that $\left|\tilde{\gamma}_K - \gamma_K^0\right| \leq \left|\hat{\gamma}_K - \gamma_K^0\right|$ and K large enough. This implies that

$$\begin{aligned}
&\left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} - \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\right)\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right) - p_{K,t}^0\left(\mathbf{X}_n\right)\right)\right\}\right| \\
&\leq \left|\hat{\gamma}_K - \gamma_K^0\right|\cdot\left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} - \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\right)\cdot\left[\dot{\mathbf{L}}_t\left(\mathbf{R}_{-0}\left(\mathbf{X}_n, \gamma_K^0\right)\right)\otimes\mathbf{r}_K\left(\mathbf{X}_n\right)'\right]\right\}\right| \\
&\quad + \left|\hat{\gamma}_K - \gamma_K^0\right|^2\cdot\left| \frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)} - \frac{D_n\left(t\right)\cdot m\left(Y_n, \mathbf{X}_n; \beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\right)\cdot\left[\mathbf{I}_J\otimes\mathbf{r}_K\left(\mathbf{X}_n\right)\mathbf{r}_K\left(\mathbf{X}_n\right)'\right]\right\}\right| \\
&= O_p\left(K^{1/2}N^{-1/2} + \zeta\left(K\right)K^{-\alpha}\right)\cdot O\left(K\right) + o_p\left(1\right),
\end{aligned}$$

where the bound follows because the random variables inside the sums are mean zero and variance bounded by K.

Now, to obtain a bound on the term (7), observe that using a similar reasoning, we obtain

$$\left|\frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}-\frac{D_n\left(t\right)\cdot m\left(Y_n,\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)^2}\right)\cdot\left(p_{K,t}^0\left(\mathbf{X}_n\right)-p_t^*\left(\mathbf{X}_n\right)\right)\right\}\right|=\sqrt{N}\cdot O_p\left(K^{-\alpha}\right).$$

Finally, the bound of term (5) is given by

$$\begin{aligned}&\left|\frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{-\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(\hat{p}_t\left(\mathbf{X}_n\right)-p_t^*\left(\mathbf{X}_n\right)\right)+\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(D_n\left(t\right)-p_t^*\left(\mathbf{X}_n\right)\right)\right\}\right|\\&=\left|\frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}\cdot\left(D_n\left(t\right)-\hat{p}_t\left(\mathbf{X}_n\right)\right)\right\}\right|\\&=\left|\frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}-\mathbf{r}_K'\left(\mathbf{X}_n\right)\boldsymbol{\theta}\right)\cdot\left(D_n\left(t\right)-\hat{p}_t\left(\mathbf{X}_n\right)\right)\right\}\right|,\end{aligned}$$

using the first order condition for MLSE, which implies that $\sum_{n=1}^{N}\left(D_n\left(t\right)-\hat{p}_t\left(\mathbf{X}_n\right)\right)\mathbf{r}_K\left(\mathbf{X}_n\right)=\mathbf{0}$, and where $\boldsymbol{\theta}\in\mathbb{R}^K$ is any vector. Now, by choosing $\boldsymbol{\theta}$ appropriately, we conclude by standard series estimation results (see, e.g., Newey (1997) or Imbens, Newey, and Ridder (2006)) that

$$\begin{aligned}&\left|\frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\left\{\left(\frac{\mathcal{E}_t\left(\mathbf{X}_n;\beta_t^*\right)}{p_t^*\left(\mathbf{X}_n\right)}-\mathbf{r}_K'\left(\mathbf{X}_n\right)\boldsymbol{\theta}\right)\cdot\left(D_n\left(t\right)-\hat{p}_t\left(\mathbf{X}_n\right)\right)\right\}\right|\\&\leq\sqrt{N}\cdot\sup_{\mathbf{x}\in\mathcal{X}}\left|\frac{\mathcal{E}_t\left(\mathbf{x};\beta_t^*\right)}{p_t^*\left(\mathbf{x}\right)}-\mathbf{r}_K'\left(\mathbf{x}\right)\boldsymbol{\theta}_{K,t}\right|\cdot\sup_{\mathbf{x}\in\mathcal{X}}\left|D_n\left(t\right)-\hat{p}_t\left(\mathbf{x}\right)\right|\\&\leq\sqrt{N}\cdot O\left(K^{-s/r}\right)\cdot\left(\sup_{\mathbf{x}\in\mathcal{X}}\left|D_n\left(t\right)-p_t^*\left(\mathbf{x}\right)\right|+\sup_{\mathbf{x}\in\mathcal{X}}\left|p_t^*\left(\mathbf{x}\right)-\hat{p}_t\left(\mathbf{x}\right)\right|\right)\\&\leq\sqrt{N}\cdot O\left(K^{-s/r}\right)\cdot O_p\left(K^{\eta}K^{1/2}N^{-1/2}+K^{2\eta}K^{-s/r}\right)+o_p\left(1\right).\end{aligned}$$

Using the bounds derived, it is easily verified that under the assumptions of Theorem 4, we obtain

$$\left|M_N\left(\beta_t^*,\hat{p}_t\left(\cdot\right)\right)-M_N\left(\beta_t^*,p_t^*\left(\cdot\right)\right)+\frac{1}{\sqrt{N}}\sum\nolimits_{n=1}^{N}\alpha_t\left(T_n,\mathbf{X}_n;p_t^*\left(\cdot\right)\right)\right|=o_p\left(1\right),$$

which verifies condition (AN.2) as desired. **Q.E.D.**