Efficient Semiparametric Estimation of Multi-valued Treatment Effects*

Matias D. Cattaneo†
UC-Berkeley

Job Market Paper

November 1, 2007

Abstract. A large fraction of the literature on program evaluation focuses on efficient estimation of binary treatment effects under the assumption of unconfoundedness. In practice, however, treatments are frequently multi-valued and available econometric techniques in this literature cannot be applied directly. This paper studies the efficient estimation of a large class of multi-valued treatment effects as implicitly defined by a collection of possibly over-identified non-smooth moment conditions when treatment assignment is assumed to be ignorable. We propose two estimators, one based on an inverse probability weighting scheme and the other based on the efficient influence function of the model, and provide a set of sufficient conditions that ensure root-N consistency, asymptotic normality and efficiency of these estimators. Under mild assumptions, these conditions are satisfied for the marginal mean treatment effect and marginal quantile treatment effect, two estimands of particular importance for empirical applications. Furthermore, based on these large sample results, other important population parameters of interest may be efficiently estimated by means of transformations of the two estimators considered. Using this idea, previous results for average and quantile treatments effects may be seen as particular cases of the methods proposed here when treatment is assumed to be dichotomous. We illustrate the procedures presented in this paper by studying the effect of maternal smoking intensity during pregnancy on birth weight. Our main findings suggest the presence of approximately homogeneous, non-linear treatment effects concentrated on the first 10 cigarettes-per-day smoked during pregnancy.

Keywords: multi-valued treatment effects, semiparametric efficiency, efficient estimation, inverse probability weighting, unconfoundedness, sieve estimation.

*I especially would like to thank Guido Imbens, Michael Jansson and Jim Powell for advice and support. I am grateful to David Brillinger, Richard Crump, Paul Ruud and Rocio Titiunik for valuable comments and suggestions. I thank Douglas Almond, Ken Chay and David Lee for generously providing the data used in the empirical illustration of this paper. 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 fraction of the literature on program evaluation focuses on the efficient estimation of treatment effects under the assumption of unconfoundedness. This literature concentrates almost exclusively on the special case of binary treatment assignments, despite the fact that in many empirical applications treatments are implicitly or explicitly multi-valued in nature. For example, in training programs participants receive different hours of training, in conditional cash transfer programs households receive different levels of transfers, and in educational interventions individuals are assigned to different classroom sizes. In cases such as these, a common empirical practice is to collapse the multi-valued treatment status into a binary indicator for eligibility or participation, a procedure that allows for the application of available semiparametric econometric techniques at the expense of a considerable loss of information. Important phenomena such as non-linearities and differential effects across treatment levels cannot be captured by the classical dichotomous treatment literature. This is especially important in a policy-making context where this additional information may provide a better understanding of the policy under consideration.

This paper is concerned with the efficient estimation of a general class of finite multi-valued treatment effects when treatment assignment is assumed to be ignorable. We study two estimation procedures for a population parameter implicitly defined by a possibly over-identified non-smooth collection of moment restrictions and we provide a set of sufficient conditions that guarantees that these estimators be efficient in large samples. This general model covers important estimands for applied work such as marginal mean treatment effects and marginal quantile treatment effects, and provides the basis for the analysis of a rich set of population parameters by allowing not only for comparisons across and within treatment levels, but also for the construction of other quantities of interest. For example, the researcher may easily construct measures of inequality, differential treatment effects, and heterogeneous treatment effects by considering different functions of means and quantiles such as pairwise differences, interquantile ranges and incremental ratios. Moreover, the model considered in this paper may provide further efficiency gains in the estimation of treatment effects by allowing for over-identification. For instance, if the underlying distributions of the potential outcomes are assumed to be symmetric, we may incorporate this information to obtain more efficient treatment effect estimators.

The results presented in this paper are closely related to the program evaluation literature, the missing data literature and the measurement error literature in both econometrics and statistics.[1] Most of these works may be traced back to the seminal papers of Rubin (1974) and Rosenbaum and Rubin (1983), and often focus on the identification and semiparametric (efficient) estimation of different population parameters of interest. In the context of program evaluation and for the particular case of binary treatments, great effort has been devoted to the efficient estimation of the average treatment effect (ATE) and average treatment effect on the treated (ATT) using either nonparametric regression methods (Hahn (1998), Heckman, Ichimura, and Todd (1998), Imbens, Newey, and Ridder (2006)), matching techniques (Abadie and Imbens (2006)), or procedures based on the nonparametric estimation of the propensity score (Hirano, Imbens, and Ridder (2003)). Recently, Firpo (2007) considered a

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

different population parameter by studying the efficient estimation of quantile treatment effects for dichotomous treatment assignments using a nonparametrically estimated propensity score. In the closely related context of missing data, Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995) and Robins, Rotnitzky, and Zhao (1995) develop a general (locally) efficient estimation strategy for models where the missingness indicator is binary involving the parametric estimation of both a regression function and the propensity score. Finally, two recent contributions by Chen, Hong, and Tamer (2005) and Chen, Hong, and Tarozzi (2007) study efficient GMM estimation in the context of measurement error models under a set of assumptions similar to ignorability with a binary missingness indicator.

Considerably less work is available in the literature for the case of multiple treatment assignments. In the context of program evaluation, Imbens (2000) derives a generalization of the propensity score, termed the Generalized Propensity Score (GPS), and shows that the results of Rosenbaum and Rubin (1983) continue to hold when the treatment status is multi-valued. Concerning identification and estimation, Imbens (2000) and Lechner (2001) discuss marginal mean treatment effects but do not assess the asymptotic properties of their estimators, while Abadie (2005) studies the large sample properties of an estimator for the marginal mean treatment effect conditional on a treatment level in the context of a difference-in-differences model. In the framework of missing data and under the assumption of missing at random there are further results in terms of limiting distributions and (local) efficiency properties for estimators of the marginal means; for a survey on these results see the recent paper of Bang and Robins (2005) and the references therein. Finally, in the context of missing data but without the assumption of missing at random, Horowitz and Manski (2000) develop sharp bounds for different multi-valued marginal mean treatment effects.

This paper contributes to the literature of program evaluation by developing a unified framework for the efficient estimation of a large class of multi-valued treatment effects. This general framework not only includes as particular cases important results from the program evaluation literature when treatment is binary, but also allows for the efficient estimation of other estimands of interest. We begin by computing the Efficient Influence Function (EIF) and Semiparametric Efficiency Bound (SPEB) for the general population parameter of interest using the methodology outlined in Newey (1990) and Bickel, Klaasen, Ritov, and Wellner (1993). We then propose two estimators of multi-valued treatment effects which are the solution to a general GMM model. To circumvent the fundamental problem of causal inference, we construct both estimators by forming sample analogues of two (possibly over-identified) moment conditions that depend only on observed data. For the first estimator, the observed moment condition is obtained by an Inverse Probability Weighting (IPW) scheme based on the GPS which may be interpreted as a moment condition exploiting a portion of the EIF. For the second estimator, the observed moment condition is obtained by using the complete form of the EIF and involves both the GPS and another conditional expectation. Because the observed moment conditions include not only the treatment effects of interest but also some infinite dimensional nuisance parameters, both estimators are of the two-step variety. In the first step, the infinite dimensional nuisance parameters are estimated and, in the second step, the corresponding GMM problem is solved.

The large sample results presented in this paper are derived in two basic stages. In the first stage, we establish consistency, asymptotic normality and efficiency of both estimators for any given nonparametric estimator of the infinite dimensional nuisance parameters.

These results are obtained by imposing a set of mild sufficient conditions concerning the underlying moment identification functions as well as two well-known high-level conditions involving the nonparametric estimators. This strategy provides a better understanding of the set of sufficient conditions required for the general procedure to work and allows for different choices of the nonparametric estimator of the nuisance parameters. The mild conditions imposed for the underlying moment identification functions are easily verified in applications, as shown in the examples discussed below, while the two-high level conditions generally require additional work. Thus, in the second stage, we discuss the nonparametric estimation of the two nuisance parameters for the particular case of series estimation. Since both nuisance parameters are conditional expectations, results from the nonparametric series (or sieve) estimation literature may be applied directly. However, since the GPS is a conditional probability we propose a new nonparametric estimator, labeled Multinomial Logistic Series Estimator, which is based on series estimation and captures the specific features of this nuisance parameter. This estimator generalizes the nonparametric estimator for the propensity score introduced by Hirano, Imbens, and Ridder (2003) and may be interpreted as a nonlinear sieve procedure (Chen (2007)) having the key advantage of providing predicted positive probabilities that add up to one. Using these nonparametric estimators, we provide simple primitive conditions that guarantee the efficient estimation of general multi-valued treatment effects.

Once an efficient estimation procedure is available, we discuss how other important population parameters of interest may be efficiently estimated by means of transformations. Intuitively, because semiparametric efficiency is preserved by a standard delta-method argument, other treatment effects that may be written as functions of the general population parameter of interest are also efficiently estimated. For the case of binary treatments, this implies that the results of Hahn (1998), Hirano, Imbens, and Ridder (2003), and Firpo (2007) may be seen as particular cases of our procedure. Furthermore, our general procedure allows for the efficient estimation of restricted treatment effects by means of a simple minimum distance estimator based on the efficiently estimated, unrestricted treatment effects. In addition to enlarging the class of treatment effects covered by our results, these ideas also allow for "optimal" testing of many hypotheses of interest.

Finally, to illustrate the results discussed in this paper we report a brief empirical study of the effect of maternal smoking intensity on birth weight that extends the analysis of Almond, Chay, and Lee (2005). These authors study the costs of low birth weight using different nonexperimental techniques and find an important negative effect of maternal smoking on birth weight defining maternal smoking as a binary treatment. Exploiting the fact that their rich database includes the number of cigarettes-per-day smoked by the mother, we extend the analysis to a multi-valued treatment setup and study the effect of maternal smoking intensity on birth weight. Our main findings suggest the presence of a nonlinear negative effect where two thirds of the full impact of smoking on birth weight are due to the first 5 cigarettes, while the remaining third is explained by the next 5 cigarettes with no important effects beyond the tenth cigarette-per-day smoked. Moreover, these effects appear to be additive, shifting parallelly the entire distribution of birth weight along smoking intensity.

The rest of the paper is organized as follows. Section 2 introduces the multi-valued treatment effect model and discusses identification of the general population parameter of interest. Section 3 includes the semiparametric efficiency calculations for the model con-

sidered and presents the general form of the EIF and SPEB. Section 4 describes the two proposed estimators and Section 5 presents the large sample results. Section 6 discusses efficient estimation of other interesting population parameters and optimal hypothesis testing. Section 7 presents the empirical illustration and Section 8 concludes. All proofs are collected in the Appendix.

2. STATISTICAL MODEL AND IDENTIFICATION

In this section we describe the multi-valued treatment effect model and discuss identification of the general population parameter of interest.

2.1. The Model. We study a multi-valued treatment effect model that is the natural extension of the well-known model used in the classical binary treatment effect literature.[2] Assume there exists a finite collection of multiple treatment status (categorical or ordinal) indexed by $t \in \mathcal{T}$, where without loss of generality $\mathcal{T} = \{0, 1, 2, \cdots, J\}$ with $J \in \mathbb{N}$ fixed. The random variables $\{Y(t),\ t \in \mathcal{T}\}$, with $Y(t) \in \mathcal{Y} \subset \mathbb{R}$, denote the collection of potential outcomes under treatment $t \in \mathcal{T}$, while the random variable $T \in \mathcal{T}$ indicates which of the $J+1$ potential outcomes is observed. Thus, the observed outcome is the random variable $Y = \sum_{t \in \mathcal{T}} D_t \cdot Y(t)$, where $D_t = \mathbf{1}\{T = t\}$ for all $t \in \mathcal{T}$ and $\mathbf{1}\{\cdot\}$ is the indicator function. We also assume there exists a real-valued random vector $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$, $d_x \in \mathbb{N}$, which is always observed.

The population parameter of interest is the vector $\beta^* = [\beta_0^{*\prime}, \beta_1^{*\prime}, \cdots, \beta_J^{*\prime}]'$, where $\beta_t^* \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ for all $t \in \mathcal{T}$ and $d_\beta \in \mathbb{N}$. We assume that this parameter solves a collection of $J+1$ (possibly over-identified) moment conditions given by

$$\mathbb{E}\left[m\left(Y(t); \beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*,\ \forall t \in \mathcal{T}, \tag{1}$$

where the function $m : \mathcal{Y} \times \mathcal{B} \to \mathbb{R}^{d_m}$ is known (possibly non-smooth) with $d_m \geq d_\beta$.[3] The maintained assumption in equation (1) imposes a conventional high-level identification condition for GMM estimation as defined by the collection of moment conditions. This model allows for a large class of population parameters of interest including those defined by non-smooth moment functions such as quantiles or other robust estimands.

We assume a random sample of size n from (Y, T, X) is observed, which is denoted by (Y_i, T_i, X_i), $i = 1, 2, \cdots, n$. This leads to a cross-sectional random sample scheme where only the potential outcome corresponding to $T = t$ is observed, which implies that we effectively sample from the conditional distribution of $Y(t)$ given $T = t$ rather than from the marginal distribution of $Y(t)$, a fact that will in general induce a bias in the estimation. Notice that

[2] For a review on the binary treatment effect literature see Imbens (2004), and for a review on the multi-valued treatment effect literature see Frölich (2004).

[3] The model considered in this paper corresponds to a slightly specialized case of a general GMM model with multi-level missing data. This may be verified by a simple change in notation: let $Y(t) \in \mathbb{R}^{d_y}$ with $d_y \geq 1$ and (abusing notation) redefine $Y(t) = (Y(t), X)$ for all $t \in \mathcal{T}$. Although all the results in this paper apply to this more general model without changes, for simplicity we restrict our attention to the multi-valued treatment effect model. Furthermore, observe that we have set all dimensions and moment conditions equal across treatment levels $t \in \mathcal{T}$. This is done only to simplify notation since all the results presented continue to hold in the more general case where different dimensions and/or moment conditions depending on t are considered.

in this model the fundamental problem of causal inference is exacerbated: for each unit we only observe one of the $J+1$ potential outcomes.

Of particular relevance for applied work are the following particular forms of this model:

EXAMPLE 1: MARGINAL MEAN TREATMENT EFFECT (MMTE). The first leading example is a classical population parameter of interest in the literature of Biostatistics, Public Health and Medicine, among other fields. This population parameter, sometimes called the Dose-Response Function, captures the mean response for each treatment level and, in the context of program evaluation, can be seen as an extension of the ATE. The MMTE is denoted by $\mu^* = [\mu_0^*, \mu_1^*, \cdots, \mu_J^*]'$ and solves equation (1) with $m(Y(t), X; \mu_t) = Y(t) - \mu_t$, for all $t \in \mathcal{T}$, which leads to $\mu_t^* = \mathbb{E}[Y(t)]$. In this case identification follows immediately after assuming a finite first moment of the potential outcomes. □

EXAMPLE 2: MARGINAL QUANTILE TREATMENT EFFECT (MQTE). Characterizing distributional impacts of a multi-valued treatment is crucial because these effects are closely related to usual inequality and heterogeneity measures. The second leading example captures this idea by looking at the treatment effect at different quantiles of the outcome variable. For some $\tau \in (0,1)$, the MQTE is denoted by $q^*(\tau) = [q_0^*(\tau), q_1^*(\tau), \cdots, q_J^*(\tau)]'$ and it is assumed to solve equation (1) with $m(Y(t); q_t(\tau)) = \mathbf{1}\{Y(t) \leq q_t(\tau)\} - \tau$, for all $t \in \mathcal{T}$, which leads to $q_t^*(\tau) \in \inf\{q : F_{Y(t)}(q) \geq \tau\}$, where $F_{Y(t)}$ is the c.d.f. of $Y(t)$. In this case, a simple sufficient condition for identification is that $Y(t)$ be a continuous random variable with density $f_{Y(t)}(q_t^*(\tau)) > 0$. □

EXAMPLE 3: MMTE WITH SYMMETRY. A very simple example where further efficiency gains may be obtained in the estimation of treatment effects is when the distribution of $Y(t)$ is assumed to be symmetric for location. In this case, mean and median coincide and hence we obtain two moment conditions for the same parameter of interest. Thus, the population parameter of interest solves the following (over-identified) moment condition: $m(Y(t), X; \vartheta_t) = (Y(t) - \vartheta_t, \mathbf{1}\{Y(t) \leq \vartheta_t\} - 1/2)$, for all $t \in \mathcal{T}$. Since this moment condition collects the moment conditions used in Example 1 and Example 2, the results for this case will follow from the conditions and results discussed for the first two examples. □

2.2. Identification. The identification condition in equation (1) covers many cases of interest. However, it has the obvious drawback of being based on unobservable random variables, the potential outcomes, which makes estimation infeasible. To proceed, we need to impose an additional identification restriction. Following the program evaluation literature, we make a "selection on observables" assumption based on the always observed random vector X:

Assumption 1. *For all $t \in \mathcal{T}$: (a) $Y(t) \perp\!\!\!\perp D_t \mid X$; and (b) $0 < p_{\min} \leq p_t^*(X) \equiv \mathbb{P}[T = t \mid X]$.*

In the context of multi-valued treatment effects, Assumption 1 is sometimes referred to as Ignorability while the conditional probabilities $p_t^*(X)$, $t \in \mathcal{T}$, are known as the Generalized Propensity Score. Imbens (2000) and Lechner (2001) provide a detailed discussion of this

assumption and discuss the role of the GPS in the estimation of the particular population parameter covered by Example 1.

Part (a) of Assumption 1 has been widely used in the program evaluation, missing data and measurement error literatures. This condition, sometimes called Unconfoundedness or Missing at Random, ensures that the distribution of each potential outcome and the treatment level indicator are conditionally independent and consequently provides identification by imposing "random assignment" conditional on observables. Intuitively, this assumption guarantees that, after conditioning on X, the conditional distribution of $Y(t)$ given $T = t$ and the marginal distribution of $Y(t)$ be identical. This assumption turns out to be sufficient for identification of β^* because it leads to

$$\mathbb{E}\left[\mathbb{E}\left[m\left(Y;\beta_t\right) \mid T = t, X\right]\right] = \mathbb{E}\left[m\left(Y(t);\beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*, \forall t \in \mathcal{T}. \quad (2)$$

Part (b) of Assumption 1 is important for at least two reasons. First, it is a necessary condition for finiteness of the semiparametric efficiency bound for regular estimators of β^* as discussed in the next section. Second, together with part (a), it provides the opportunity to consider alternative identification conditions based on the observed random variables. For example, we may easily verify that

$$\mathbb{E}\left[\frac{D_t \cdot m\left(Y;\beta_t\right)}{p_t^*\left(X\right)}\right] = \mathbb{E}\left[m\left(Y(t);\beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*, \forall t \in \mathcal{T}, \quad (3)$$

and

$$\mathbb{E}\left[\frac{D_t \cdot \mathbb{E}\left[m\left(Y;\beta_t\right) \mid X\right]}{p_t^*\left(X\right)}\right] = \mathbb{E}\left[m\left(Y(t);\beta_t\right)\right] = 0 \text{ if and only if } \beta_t = \beta_t^*, \forall t \in \mathcal{T}, \quad (4)$$

which leads to two additional observed moment conditions.[4]

Using equations (2), (3) and (4) as a starting point, several estimation procedures and their corresponding efficiency properties have been considered in the literature for the particular case of binary treatment effects (or binary missingness indicator). Estimators that exploit moment conditions (2) or (4) are usually known as "imputation" or "projection" estimators because first a conditional expectation function is (nonparametrically) estimated and then missing outcomes are imputed for all (or some subset of) the observations and averaged out. Recent examples of papers studying these kind of estimators are Hahn (1998) and Imbens, Newey, and Ridder (2006) in the context of program evaluation with binary treatments, and Chen, Hong, and Tamer (2005) and Chen, Hong, and Tarozzi (2007) in the context of nonclassical measurement error. In the framework of missing data, there is a vast literature known as Doubly Robust Estimation that is based on moment conditions such as equation (4) and uses parametric specifications of the unknown functions. Bang and Robins (2005) present a comprehensive review on this topic.

Estimators constructed from the moment condition (3) lead naturally to an Inverse Probability Weighting (IPW) scheme and have been considered by many authors in different contexts at least since the work of Horvitz and Thompson (1952). Intuitively, this procedure

[4]Other identification conditions are also available in the literature based on this idea. For example, see Hahn (1998).

achieves identification by reweighting the observations to make them representative of the population of interest. This idea has been exploited in the program evaluation literature by Imbens (2000), Hirano, Imbens, and Ridder (2003) and Firpo (2007), in the missing data literature by Robins, Rotnitzky, and Zhao (1994) and Robins, Rotnitzky, and Zhao (1995), and in the measurement error literature by Chen, Hong, and Tarozzi (2007), among others. Wooldridge (2007) provides a very interesting discussion of this estimation strategy.

Assumption 1 leads to an important collection of alternative asymptotically equivalent efficient estimators in the context of program evaluation. In this paper we study two efficient estimators for the case of multi-valued treatment effects. The first estimator is based on equation (3), while the second estimator is based on a different moment condition that may be constructed as a linear combination of equations (2), (3) and (4). These estimators are also asymptotically equivalent to those available in the literature in the special case of binary treatment effects. It remains as an important open research question to rank the large class of available semiparametric efficient estimators.

2.3. Notation. Before turning to the discussion of efficient estimation in the context of multi-valued treatment effects, it is convenient to introduce some notation that will simplify the presentation. We work with two important functions: the $J+1$ vector-valued function representing the GPS, denoted by $p^*(\cdot) = [p_0^*(\cdot), \cdots, p_J^*(\cdot)]'$, and the $(J+1)\cdot d_m$ vector-valued function of conditional expectations denoted by $e^*(\cdot;\beta) = \left[e_0^*(\cdot;\beta_0)', \cdots, e_J^*(\cdot;\beta_J)'\right]'$, where $e_t^*(X;\beta_t) = \mathbb{E}[m(Y(t);\beta_t) \mid X]$. We assume that $p_t^*(\cdot) \in \mathcal{P}$ and $e_t^*(\cdot;\beta_t) \in \mathcal{E}$ for all $\beta_t \in \mathcal{B}$ and $t \in \mathcal{T}$, where $\mathcal{P}$ and $\mathcal{E}$ represent some (smooth) space of functions. For simplicity, in the remaining of the paper we will drop the arguments of the functions considered whenever it is clear from the context.

We let $|\cdot|$ denote the matrix norm given by $|A| = \sqrt{\text{trace}(A'A)}$ for any matrix A. As for functions, we work with the sup-norm in all arguments denoted by $\|\cdot\|_\infty$. In particular, for all $t \in \mathcal{T}$, we have $\|p_t\|_\infty = \sup_{x\in\mathcal{X}} |p_t(x)|$ for some $p_t \in \mathcal{P}$, $\|e_t(\beta_t)\|_\infty = \sup_{x\in\mathcal{X}} |e_t(x;\beta_t)|$ and $\|e_t\|_\infty = \sup_{\beta\in\mathcal{B},x\in\mathcal{X}} |e_t(x;\beta_t)|$ for some $e_t(\beta_t) \in \mathcal{E}$, and similarly for the vector-valued functions p and e. Later in the paper we will restrict the class of functions considered to enable the nonparametric estimation of these nuisance parameters.

Finally, to reduce the notational burden we introduce the following vector-valued functions

$$m(Y,T,X;\beta,p) = \left[\frac{D_0}{p_0(X)}\cdot m(Y;\beta_0)', \cdots, \frac{D_J}{p_J(X)}\cdot m(Y;\beta_J)'\right]',$$

and

$$\alpha(T,X;p,e(\beta)) = \left[e_0(X;\beta_0)'\cdot\frac{D_0-p_0(X)}{p_0(X)}, \cdots, e_J(X;\beta_J)'\cdot\frac{D_J-p_J(X)}{p_J(X)}\right]',$$

for some $p \in \mathcal{P}^{J+1}$ and $e(\beta) \in \mathcal{E}^{J+1}$ for all $\beta \in \mathcal{B}^{J+1}$.

3. SEMIPARAMETRIC EFFICIENCY CALCULATIONS

In this section we provide basic semiparametric efficiency calculations essential for the construction of efficient estimators of β^*. Semiparametric efficiency theory has received considerable attention in econometrics at least since the seminal work of Bickel, Klaasen, Ritov, and

Wellner (1993) (see also Newey (1990) for an excellent survey). This general theory provides the necessary ingredients for the construction of efficient estimators of finite dimensional parameters in the context of semiparametric models under some mild regularity conditions. 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 the efficient score of the model. In the simplest possible case, the construction of an efficient estimator starts by deriving the EIF in the statistical model and then verifying that the proposed estimator admits an asymptotic linear representation based on this function. In this paper we use these ideas to verify that our estimators are in fact efficient.

Several semiparametric efficiency calculations are available in the literature when some form of Assumption 1 holds. 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 average and quantile treatment effect parameters using the methodology outlined in Bickel, Klaasen, Ritov, and Wellner (1993). In models of missing data, Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995) develop a general methodology to construct efficient scores and compute the corresponding efficiency bounds when the missingness indicator is binary. In a recent contribution, Chen, Hong, and Tarozzi (2007) provide semiparametric efficiency calculations for GMM models in the context of nonclassical measurement error with one auxiliary sample. The results presented in this section cover all these cases by considering a multi-level missing mechanism in a GMM model. In Section 6 we show how the efficiency bounds derived in the program evaluation literature may be recovered from the calculations presented here.

Assumption 2. *(a) For all* $t \in \mathcal{T}$, $\mathbb{E}[|m(Y(t);\beta_t)|^2] < \infty$ *and* $\mathbb{E}[m(Y(t);\beta_t)]$ *is differentiable in* $\beta_t \in \mathcal{B}$ *at* β_t^**; and (b)* $\operatorname{rank}(\Gamma_*) = (J+1)\cdot d_\beta$*, where*

$$\Gamma_* = \begin{bmatrix} \Gamma_0^* & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Gamma_1^* & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Gamma_J^* \end{bmatrix},$$

where $\mathbf{0}$ *is a* $(d_m \times d_\beta)$ *matrix of zeros and*

$$\Gamma_t^* = \frac{\partial}{\partial \beta_t'} \mathbb{E}[m(Y(t);\beta_t)]\Big|_{\beta_t=\beta_t^*}, \forall t \in \mathcal{T}.$$

The main role of Assumption 2 (together with part (b) of Assumption 1) is to ensure that the bound is finite. The full column rank assumption on the gradient matrix Γ_* ensures a local identification condition necessary for the semiparametric calculations. A key necessary requirement to provide semiparametric calculations is to establish the pathwise differentiability of the population parameter of interest, which is done in the Appendix under this Assumption (and Assumption 1).

The following theorem provides the general form of the EIF and SPEB for the model considered in this paper.

Theorem 1. *(*EIF AND SPEB*) Let Assumptions 1 and 2 hold. Then the EIF for any regular estimator of β^* is given by*

$$\Psi\left(y,t,x;\beta^*,p^*,e^*\left(\beta^*\right)\right) = -\left(\Gamma_*'V_*^{-1}\Gamma_*\right)^{-1}\Gamma_*'V_*^{-1}\psi\left(y,t,x;\beta^*,p^*,e^*\left(\beta^*\right)\right),$$

where $\psi\left(y,t,x;\beta^*,p^*,e^*\left(\beta^*\right)\right) = m\left(y,t,x;\beta^*,p^*\right) - \alpha\left(t,x;\beta^*,p^*,e^*\left(\beta^*\right)\right)$ *and*

$$V_* = \mathbb{V}\left[\psi\left(Y,T,X;\beta^*,p^*,e^*\left(\beta^*\right)\right)\right].$$

Consequently, the SPEB for any regular estimator of β^* is given by $V^* = \left(\Gamma_*'V_*^{-1}\Gamma_*\right)^{-1}$.

The results in Theorem 1 may be directly compared to those presented in Newey (1994). This leads to a natural interpretation for the EIF, where the vector-valued function $\alpha(\cdot)$ corresponds to the "adjustment term" in the influence function due to the presence of the unknown nuisance parameter (GPS) when the estimator is constructed from the sample analogue of the moment condition given by equation (3). In the next section we use this interpretation to compare the two estimators considered in this paper.

It is possible to provide additional intuition for the structure of the SPEB after noting that

$$V_* = \mathbb{E}\left[\mathbb{V}\left[m\left(Y,T,X;\beta^*,p^*\right) \mid X\right]\right] + \mathbb{E}\left[e^*\left(X;\beta^*\right)e^*\left(X;\beta^*\right)'\right]. \tag{5}$$

Using this decomposition, we see that the results in Theorem 1 may be interpreted as the multi-level generalization of the SPEB in Theorem 1 of Chen, Hong, and Tarozzi (2007) in the context of measurement error with "verify-in-sample" auxiliary data. Extending the results of Hahn (1998) and Chen, Hong, and Tarozzi (2007) to the context of multi-valued treatments, we verify that (i) the GPS is ancillary for the estimation of β^* (i.e., the SPEB does not change whether we assume the GPS to be known), and (ii) if the distribution of X is known or correctly specified the SPEB is reduced (in particular, if the distribution of X is assumed to be known, then the second term in equation (5) drops out). We do not provide details for these results to conserve space.

It is important to note that we have explicitly allowed for the components $\beta_0^*, \cdots, \beta_J^*$ of the population parameter β^* to be different. Under this assumption, the SPEB obtained in Theorem 1 will be in general larger than the one we would obtain had we imposed $\beta_0^* = \cdots = \beta_J^*$. Since our main goal is to estimate efficiently the components of β^* (i.e., treatment effects), the result presented in Theorem 1 seems to be the most appropriate. The SPEB for the "restricted" case may be easily obtained by similar derivations to those presented in the appendix.

One important simplification in Theorem 1 is achieved in the important case of exact identification:

Corollary 1. *If $d_m = d_\beta$, then Theorem 1 implies that the EIF for any regular estimator of β^* is given by*

$$\Psi\left(y,t,x;\beta^*,p^*,e^*\left(\beta^*\right)\right) = \Gamma_*^{-1}\psi\left(y,t,x;\beta^*,p^*,e^*\left(\beta^*\right)\right).$$

Consequently, the SPEB for any regular estimator of β^* is given by $V^* = \Gamma_*^{-1}V_*\Gamma_*'^{-1}$.

Notice further that in the just-identified case, $\Gamma_* = \text{diag}\left(\Gamma_0^*, \cdots, \Gamma_J^*\right)$.

The result in Corollary 1 is important because it shows that in the just-identified case the EIF may be constructed by collecting in a single vector the EIF's corresponding to each $\beta_0^*, \cdots, \beta_J^*$. Moreover, using this result, it will follow that in the just-identified case we may estimate efficiently β^* by estimating each $\beta_0^*, \cdots, \beta_J^*$ independently. We discuss this result further in the following sections.

Finally, we apply the results of Theorem 1 to the examples under study:

EXAMPLE 1 (CONTINUED): MMTE. Assume $\mathbb{E}[Y(t)^2] < \infty$ and note that $\Gamma_t^* = 1$ for all $t \in \mathcal{T}$ in this case. Thus, Assumption 2 is satisfied and Theorem 1 implies that the SPEB for the MMTE is given by V^* with typical (i,j)-th element

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

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

EXAMPLE 2 (CONTINUED): MQTE. Using Leibniz's rule we have $\Gamma_t^* = f_{Y(t)}^*\left(q_t^*\left(\tau\right)\right)$ for $t \in \mathcal{T}$, which was assumed strictly positive. Thus, Assumption 2 is satisfied and Theorem 1 implies that the SPEB for the MQTE is given by V^* with typical (i,j)-th element

$$V_{[i,j]}^* = \mathbb{E}\left[\mathbf{1}\left\{i=j\right\} \cdot \frac{\sigma_i^2\left(X;\tau\right)}{f_{Y(i)}^*\left(q_i^*\left(\tau\right)\right)^2 \cdot p_i^*\left(X\right)} + \frac{q_i\left(X;\tau\right) \cdot q_j\left(X;\tau\right)}{f_{Y(i)}^*\left(q_i^*\left(\tau\right)\right) \cdot f_{Y(j)}\left(q_j^*\left(\tau\right)\right)}\right],$$

where $\sigma_i^2\left(X;\tau\right) = \mathbb{V}\left[\mathbf{1}\left\{Y\left(i\right) \leq q_i^*\left(\tau\right)\right\} \mid X\right]$, $q_i\left(X;\tau\right) = \mathbb{E}\left[\mathbf{1}\left\{Y\left(i\right) \leq q_i^*\left(\tau\right)\right\} - \tau \mid X\right]$, for all $i \in \mathcal{T}$. □

4. ESTIMATION PROCEDURES

In this paper we consider two estimators for the multi-valued treatment effects. The first estimation procedure uses an IPW approach and is based on equation (3), while the second estimation procedure combines the IPW and imputation approaches and is based on the EIF derived in Theorem 1. For simplicity, in the over-identified case, the construction does not use continuously updated GMM but rather uses a consistent estimator of the corresponding weighting matrix.[5] In particular, we assume that A_n is a $\left((J+1)\cdot d_\beta\right) \times \left((J+1)\cdot d_m\right)$ (random) matrix such that $A_n = A + o_p\left(1\right)$ for some positive semidefinite matrix $W = A'A$.

4.1. Inverse Probability Weighting Estimator (IPWE). We may motivate this procedure by a simple sample analog principle. Recall that our goal is to estimate the parameters implicitly defined by the moment conditions $\mathbb{E}\left[m\left(Y(t);\beta_t^*\right)\right] = 0$ for all $t \in \mathcal{T}$. Had we observed the random variables $\left(Y(0), \cdots, Y(J)\right)$, a natural estimator would simply solve the sample analog counterpart of the $J+1$ moment conditions leading to a standard GMM estimation procedure. Unfortunately, due to the presence of the missingness mechanism, we cannot perform such estimation since we only observe Y. Instead, we may use the result in Equation (3) to obtain a moment condition based only on observed random

[5] A generalization to a continuously updated GMM model is straightforward provided the corresponding additional regularity conditions are imposed.

variables. This alternative has the drawback that now the feasible moment conditions involve both the finite dimensional parameter of interest, β^*, and an infinite dimensional nuisance parameter (GPS). This 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 consistently estimate the finite dimensional parameter of interest.

Using these ideas, we may consider a simple semiparametric two-step GMM estimation procedure where the parameter β^* is estimated after a preliminary nonparametric estimator for the GPS has been constructed. In particular, to save on notation, define the moment condition

$$M^{IPW}(\beta, p) = \mathbb{E}\left[m\left(Y, T, X; \beta, p\right)\right],$$

and its sample analogue

$$M_n^{IPW}(\beta, p) = \frac{1}{n}\sum_{i=1}^{n} m\left(Y_i, T_i, X_i; \beta, p\right).$$

Formally the IPWE may be described by the following steps. First, construct a nonparametric estimator of the GPS based on the full sample, denoted $\hat{p} = [\hat{p}_0, \cdots, \hat{p}_J]'$. Second, the IPWE for β^* is given by

$$\hat{\beta}^{IPW} = \arg\min_{\beta \in \mathcal{B}^{J+1}} \left| A_n M_n^{IPW}(\beta, \hat{p}) \right| + o_p(n^{-1/2}).$$

This estimation procedure has the important advantage of being based only on the nonparametric estimator of the GPS. Note that the infinite dimensional component does not depend on β and therefore we only need to estimate it once to form the GMM problem, leading to a very simple two-step procedure. On the other hand, this estimation procedure has an important drawback based on its construction. Because it only involves the first part of the EIF derived in the previous section, to ensure its semiparametric efficiency the nonparametric estimator $\hat{p}$ will have to play two roles simultaneously: not only does it have to approximate p^* fast enough, but it also has to do it in such a way that the limiting GMM problem becomes a GMM problem based on the EIF. For example, as pointed out by Hirano, Imbens, and Ridder (2003) in the model of binary treatment effects, the extreme case where $\hat{p} = p^*$ will not lead in general to an efficient estimator because this procedure will be solving the incorrect GMM problem. We will make explicit the requirements on $\hat{p}$ in the next section when we study the large sample properties of this estimator.[6]

For the just-identified case, the procedure leading to the IPWE is equivalent to solving

$$\hat{\beta}_t^{IPW} = \arg\min_{\beta \in \mathcal{B}} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{D_{t,i} \cdot m\left(Y_i; \beta_t\right)}{\hat{p}_t\left(X_i\right)} \right| + o_p(n^{-1/2}),\ \forall t \in \mathcal{T},$$

which leads to a very simple estimator.

[6] The role of the propensity score and how information about it may be efficiently incorporated in semiparametric models have received considerable attention in the literature of program evaluation and related areas of study. See, e.g., Hahn (1998), Heckman, Ichimura, and Todd (1998), Hirano, Imbens, and Ridder (2003), and Chen, Hong, and Tarozzi (2007), among others, for a discussion on this topic.

This applies directly to our examples:

EXAMPLE 1 (CONTINUED): MMTE. In this case, we obtain a closed-form solution given by

$$\hat{\mu}_t^{IPW} = \left(\sum\nolimits_{i=1}^{n} \frac{D_{t,i}}{\hat{p}_t(X_i)}\right)^{-1} \sum\nolimits_{i=1}^{n} \frac{D_{t,i} \cdot Y_i}{\hat{p}_t(X_i)},$$

which corresponds to a properly re-weighted average for each $t \in \mathcal{T}$. □

EXAMPLE 2 (CONTINUED): MQTE. In this case we cannot obtain a closed-form solution to the minimization problem. Instead, we have for fixed $\tau \in (0,1)$,

$$\hat{q}_t^{IPW}(\tau) = \arg\min_{q \in \mathcal{B}} \left| \frac{1}{n} \sum\nolimits_{i=1}^{n} \frac{D_{t,i} \cdot (\mathbf{1}\{Y_i \leq q\} - \tau)}{\hat{p}_t(X_i)} \right|$$

for all $t \in \mathcal{T}$. □

4.2. Efficient Influence Function Estimator (EIFE). This estimator is based on the EIF derived in Theorem 1. This procedure can also be motivated by the analogue principle after observing that $\mathbb{E}[\psi(Y,T,X;\beta,p,e(\beta))] = 0$ if and only if $\beta = \beta^*$, $p = p^*$ and $e = e^*$. In words, the EIF provides another collection of moment conditions that can be exploited to obtain a GMM estimator. Inspection of $\mathbb{E}[\psi(Y,T,X;\beta,p,e(\beta))]$ shows that its sample analogue corresponds to a linear combination of three sample analogues already discussed in the literature for the special case of binary treatment effects. In particular, this moment condition includes (i) the moment condition leading to an IPW estimator, (ii) the moment condition leading to a nonparametric version of the doubly robust estimator, and (iii) the moment condition leading to an imputation estimator.

To describe the estimator, define the moment condition

$$M^{EIF}(\beta,p,e(\beta)) = \mathbb{E}[\psi(Y,T,X;\beta,p,e(\beta))],$$

and its sample analogue

$$M_n^{EIF}(\beta,p,e(\beta)) = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i,T_i,X_i;\beta,p,e(\beta)).$$

Formally the EIFE may be described by the following steps. First, construct a nonparametric estimator of the GPS, denoted $\hat{p} = [\hat{p}_0, \cdots, \hat{p}_J]'$, and for each $\beta \in \mathcal{B}$ construct a nonparametric estimator of $e(\beta)$, denoted $\hat{e}(\beta) = [\hat{e}_0(\beta)', \cdots, \hat{e}_J(\beta)']'$. Second, the EIFE for β^* is given by

$$\hat{\beta}^{EIF} = \arg\min_{\beta \in \mathcal{B}^{J+1}} \left| A_n M_n^{EIF}(\beta, \hat{p}, \hat{e}(\beta)) \right| + o_p(n^{-1/2}).$$

This estimator appears to be in general more complicated than the IPWE because it requires the nonparametric estimation of two infinite dimensional parameters, one of which is a function of β itself. On the other hand, it has the attractive feature of being based

on the EIF and therefore each nonparametric estimator would only be required to have the intuitive role of approximating well its own population counterpart. For example, it is now possible to consider the extreme case of $\hat{p} = p^*$ and still obtain an efficient estimator, as we discuss below.[7]

As for the IPWE, in the just-identified case this procedure is equivalent to solve for all $t \in \mathcal{T}$,

$$\hat{\beta}_t^{EIF} = \arg\min_{\beta_t \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} \cdot m\left(Y_i; \beta_t\right) - \hat{e}_t\left(X_i; \beta_t\right) \cdot \left(D_{t,i} - \hat{p}_t\left(X_i\right)\right)}{\hat{p}_t\left(X_i\right)} \right| + o_p(n^{-1/2}).$$

In the case of our leading example, this estimation procedure gives:

EXAMPLE 1 (CONTINUED): MMTE. In this case, for $t \in \mathcal{T}$ we have $e_t^*\left(X; \mu_t\right) = \mu_t^*\left(X\right) - \mu_t$ and therefore we may obtain a close form solution,

$$\hat{\mu}_t^{EIF} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} \cdot Y_i - \hat{\mu}_t\left(X_i\right) \cdot \left(D_{t,i} - \hat{p}_t\left(X_i\right)\right)}{\hat{p}_t\left(X_i\right)},$$

where $\hat{\mu}_t\left(x\right)$ represents some nonparametric estimator of $\mu_t^*\left(x\right)$. □

EXAMPLE 2 (CONTINUED): MQTE. In this example, $e_t^*\left(X; \beta_t\right) = F_{Y(t)}^*\left(q_t\left(\tau\right) \mid X\right) - \tau$, for $t \in \mathcal{T}$, and the minimization problem becomes for fixed $\tau \in (0,1)$,

$$\hat{q}_t^{EIF}\left(\tau\right) = \arg\min_{q_t \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} \cdot \left(\mathbf{1}\left\{Y_i \leq q_t\right\} - \tau\right) - \left(\hat{F}_{Y(t)}\left(q_t \mid X_i\right) - \tau\right) \cdot \left(D_{t,i} - \hat{p}_t\left(X_i\right)\right)}{\hat{p}_t\left(X_i\right)} \right|,$$

where $\hat{F}_{Y(t)}\left(y \mid x\right)$ represents some nonparametric estimator of $F_{Y(t)}^*\left(y \mid x\right)$. □

5. LARGE SAMPLE PROPERTIES

This section presents the main large sample results of the paper in four stages. First, we establish consistency of both the IPWE and EIFE under mild regularity conditions. Second, we provide sufficient conditions for asymptotic normality and efficiency of the IPWE and EIFE for any nonparametric estimators of the infinite dimensional nuisance parameters based on a set of high-level conditions. Third, we construct estimators for the different components of the SPEB derived in Theorem 1. Finally, we discuss nonparametric estimation of the infinite dimensional nuisance parameters and thereby provide a full data-driven procedure for the efficient estimation of β^*.

The large sample theory presented in this paper is based on the work of Pakes and Pollard (1989).[8] In the following discussion, we will repeatedly employ terminology and results from the modern theory of empirical processes. For consistency and to simplify the exposition, all references to this literature are based on van der Vaart and Wellner (1996) (see also Andrews (1994) and van der Vaart (1998) for excellent reviews on this topic).

[7] It is important to note that this is not the only way in which information about the (generalized) propensity score may be incorporated in semiparametric efficient estimators. For two other examples, see the recent work of Chen, Hong, and Tarozzi (2007) in the context of measurement error models.

[8] Alternatively, we may apply the general large sample theory of Chen, Linton, and van Keilegom (2003). However, because in our case the criterion function is smooth in the infinite dimensional nuisance parameters, the results from Pakes and Pollard (1989) turn out to be sufficient.

5.1. Consistency. Consistency of the IPW estimator will follow from two mild conditions imposed on the underlying identifying function $m(\cdot;\beta)$:

Assumption 3. *For all $t \in \mathcal{T}$, (a) the class of functions $\{m(\cdot;\beta_t) : \beta_t \in \mathcal{B}\}$ is Glivenko-Cantelli, and (b) $\mathbb{E}[\sup_{\beta_t \in \mathcal{B}} |m(Y(t);\beta_t)|] < \infty$.*

Part (a) of Assumption 3 restricts the class of functions that may be considered to implicitly define the population parameter of interest. Functions in this class enjoy an important property: sample averages of these functions are uniform consistent in β for their population mean. Although consistency may be established by other means, requiring an uniform consistency property of the underlying sample moment conditions is standard in the GMM literature. Newey and McFadden (1994) discuss this and other related conditions. A simple set of sufficient conditions for Assumption 3(a) are $\mathcal{B}$ compact, $m(\cdot;\beta_t)$ continuous in β_t, and Assumption 3(b). Although this set of conditions is reasonably weak, it is still stronger than necessary. In fact, to cover interesting nonsmooth cases (such as quantiles) it is necessary to rely on slightly stronger results such as those presented in the empirical process literature. From this literature, many classes of functions are known to be Glivenko-Cantelli and many other classes may be formed by some "permanence" theorem.[9]

Part (b) of Assumption 3 is a usual dominance condition.

Theorem 2. *(CONSISTENCY OF IPWE) Let Assumptions 1 and 3 hold. Assume the following additional condition holds:*

(2.1) $\|\hat{p} - p^*\|_\infty = o_p(1)$.

Then, $\hat{\beta}^{IPW} = \beta^* + o_p(1)$.

The additional condition (2.1) in Theorem 2 is very weak, requiring only that the nonparametric estimator of the GPS is uniformly consistent.

Next, we consider the EIFE. For this estimator, we additionally assume:

Assumption 4. *For all $t \in \mathcal{T}$, the class of functions $\{e_t^*(\cdot;\beta_t) : \beta_t \in \mathcal{B}\}$ is Glivenko-Cantelli.*

Assumption 4 captures the ideas implied by Assumption 3(a). In this case, however, this assumption may be easier to verify because the functions $e_t^*(\cdot;\beta_t)$ are conditional expectations and therefore it is natural to assume they are smooth in β_t. Thus, verifying the underlying uniform consistency requirement should be straightforward in this case, possibly after imposing some additional mild regularity conditions.

Theorem 3. *(CONSISTENCY OF EIFE) Let Assumptions 1, 3, and 4 hold. Assume the following additional condition holds:*

(3.1) $\|\hat{p} - p^*\|_\infty = o_p(1)$ *and* $\|\hat{e} - e^*\|_\infty = o_p(1)$.

[9] Primitive conditions that ensure a given class of functions to be Glivenko-Cantelli (or Donsker) usually involve some explicit assumption concerning the "size" of the class as measured by some version of the entropy numbers. For a recent example in the context of GMM estimation see Ai and Chen (2003).

Then, $\hat{\beta}^{EIF} = \beta^* + o_p(1)$.

Since we are now using the full EIF to construct the estimator, it is natural to observe that Theorem 3 also requires the nonparametric estimator $\hat{e}$ to be uniformly consistent for e^* in both arguments (the covariates X and the parameter β). This condition is still weak and reasonable for most nonparametric estimators.

For our examples, Assumptions 3 and 4 may be easily verified:

EXAMPLE 1 (CONTINUED): MMTE. Assume $\mathcal{B}$ is compact and $\mathbb{E}[|Y(t)|] < \infty$ for all $t \in \mathcal{T}$. Assumption 3 follows directly because the class of functions $\{(\cdot - \mu_t) : \mu_t \in \mathcal{B}\}$ is Glivenko-Cantelli. Therefore, Theorem 2 implies $\hat{\mu}^{IPW} \xrightarrow{p} \mu^*$. Moreover, the class of $\{(\mu_t^*(\cdot) - \mu_t) : \mu_t \in \mathcal{B}\}$ is also Glivenko-Cantelli and Theorem 3 implies $\hat{\mu}^{EIF} \xrightarrow{p} \mu^*$. □

EXAMPLE 2 (CONTINUED): MQTE. Assumption 3 follows immediately because the class of functions $\{(\mathbf{1}\{\cdot \leq q_t\} - \tau) : q_t \in \mathcal{B}\}$ is Glivenko-Cantelli and Theorem 2 gives $\hat{q}^{IPW}(\tau) \xrightarrow{p} q^*(\tau)$. Furthermore, if the class of functions $\{F^*_{Y(t)}(q_t \mid \cdot) - \tau : q_t \in \mathcal{B}\}$ is Glivenko-Cantelli, Theorem 3 gives $\hat{q}^{EIF}(\tau) \xrightarrow{p} q^*(\tau)$. The last requirement may be verified if, for example, we have $\mathcal{B}$ compact and $F^*_{Y(t)}(y \mid x)$ continuous in y for every x. □

5.2. Asymptotic Normality and Efficiency. We are now ready to discuss the conditions needed to establish the limiting distribution and efficiency of the two estimators considered in this paper. We begin by stating a set of sufficient conditions for the IPWE:

Assumption 5. *For all $t \in \mathcal{T}$ and some $\delta > 0$: (a) $\{m(\cdot;\beta_t) : |\beta_t - \beta_t^*| < \delta\}$ is a Donsker class; (b) $\mathbb{E}[\,|m(Y(t);\beta_t) - m(Y(t);\beta_t^*)|^2] \to 0$ as $\beta_t \to \beta_t^*$; (c) there exists a constant $C > 0$ such that $\mathbb{E}[|m(Y(t);\beta_t) - m(Y(t);\beta_t^*)|] \leq C \cdot |\beta_t - \beta_t^*|$ for all β_t with $|\beta_t - \beta_t^*| < \delta$; and (d) $\mathbb{E}[\sup_{|\beta_t - \beta_t^*| < \delta} |m(Y(t);\beta_t)|^2] < \infty$.*

Similar to the requirement for consistency, Part (a) of Assumption 5 restricts the class of functions defining the population parameter of interest that may be considered. This assumption is standard from the empirical process literature and ensures that an uniform (in β_t) central limit theorem holds. In turn, this result together with part (b) and part (c) will ensure that a certain stochastic equicontinuity condition applies, which allows us to obtain an asymptotic linear representation for the estimator. For most applications, Assumption 5(a) is already established or can be easily verified by some "permanence theorem". Assumptions 5(b) and 5(c) are standard in the literature and may be verified directly, while Assumption 5(d) is a usual dominance condition.

Theorem 4. *(*ASYMPTOTIC LINEAR REPRESENTATION OF IPWE*) Let $\beta^* \in \operatorname{int}(\mathcal{B}^{J+1})$, $\hat{\beta}^{IPW} = \beta^* + o_p(1)$, and Assumptions 1, 2, and 5 hold. Assume the following additional conditions hold:*

(4.1) $\|\hat{p} - p^*\|_\infty = o_p\left(n^{-1/4}\right)$.

(4.2) $M_n^{IPW}(\beta^*, \hat{p}) = M_n^{EIF}(\beta^*, p^*, e^*(\beta^*)) + o_p\left(n^{-1/2}\right)$.

Then,

$$\hat{\beta}^{IPW} - \beta^* = -\left(\Gamma_*' W \Gamma_*\right)^{-1} \Gamma_*' W M_n^{EIF}\left(\beta^*, p^*, e^*\left(\beta^*\right)\right) + o_p(n^{-1/2}).$$

Asymptotic normality of $\hat{\beta}^{IPW}$ follows directly from Theorem 4 while efficiency is easily obtained by an appropriate choice of the limiting weighting matrix W. This theorem requires two important additional conditions involving the estimator of the GPS. These conditions imply certain restrictions in terms of smoothness for the class of functions $\mathcal{P}$ and $\mathcal{E}$, depending on the nonparametric estimator chosen and the dimension of $\mathcal{X}$.

Condition (4.1) is standard in the literature and imposes a lower bound in the uniform rate of convergence of $\hat{p}$. Condition (4.2) is crucial. This condition involves the sample moment condition (at $\beta = \beta^*$) and the nonparametric estimator, and requires a particular linear expansion based on the EIF to hold. Newey (1994) provides an in-depth general discussion of this particular condition and outlines high-level assumptions that ensure this condition holds. This assumption is very important because it employs the exact form of the EIF to guarantee that the resulting estimator is efficient (provided the weighting matrix is chosen appropriately). If condition (4.2) holds for a function different than $M_n^{EIF}\left(\beta^*, p^*, e^*\left(\beta^*\right)\right)$, then the estimator cannot be efficient. For example, if the GPS is known and we replace $\hat{p} = p^*$ in $M_n^{IPW}\left(\beta^*, \hat{p}\right)$ when constructing the estimation procedure, then the resulting estimator will not be efficient as mentioned before. In this sense, Condition (4.2) imposes an upper bound on the uniform rate of convergence of $\hat{p}$. Intuitively, this is due to the fact that $\hat{p}$ plays two roles simultaneously: it estimates nonparametrically p^*, and it also nonparametrically approximates the correction term $\alpha\left(\cdot; p, e\left(\beta\right)\right)$ present in the EIF. Consequently, even if the GPS is known, one may obtain an efficient estimator only if the GPS is nonparametrically estimated.

One way to avoid requiring $\hat{p}$ to play this dual role is to consider the full EIF, which leads to the EIFE. This estimator will be asymptotically normal if the following additional assumption holds:

Assumption 6. *For all $t \in \mathcal{T}$, some $\delta > 0$, and for all $x \in \mathcal{X}$ and all β_t such that $|\beta_t - \beta_t^*| < \delta$: (a) $e_t^*\left(x; \beta_t\right)$ is continuously differentiable with derivative given by $\partial_{\beta_t} e_t^*\left(x; \beta_t\right) \equiv \frac{\partial}{\partial \beta_t} e_t^*\left(x; \beta_t\right)$ with $\mathbb{E}[\sup_{|\beta_t - \beta_t^*| < \delta} \left|\partial_{\beta_t} e_t^*\left(X; \beta_t\right)\right|] < \infty$; and (b) there exists $\epsilon > 0$ and a measurable function $b\left(x\right)$, with $\mathbb{E}\left[\left|b\left(X\right)\right|\right] < \infty$, such that $\left|\partial_{\beta_t} e_t\left(x; \beta_t\right) - \partial_{\beta_t} e_t^*\left(x; \beta_t\right)\right| \leq b\left(x\right) \cdot \left\|e_t - e_t^*\right\|_\infty^\epsilon$ for all functions $e_t\left(\beta_t\right) \in \mathcal{E}$ such that $\left\|e_t - e_t^*\right\|_\infty < \delta$.*

Assumption 6 basically restricts the class of functions $\mathcal{G} = \{e_t : e_t\left(\beta\right) \in \mathcal{E}, \left\|e_t - e_t^*\right\|_\infty < \delta$ and $|\beta_t - \beta_t^*| < \delta\}$, where $e_t^* \in \mathcal{G}$ by construction. Part (a) of this assumption is simple and natural, requiring only mild smoothness conditions of the conditional expectation $e_t\left(\beta_t\right)$ in β_t as well as a usual dominance condition. Note that this part of the assumption will imply the smoothness requirement in Assumption 2 whenever integration and differentiation can be interchanged. Part (b) of Assumption 6 further restricts the possible class of functions by requiring that functions that are uniformly close also have their derivatives close. This special technical requirement has also been used by Chen, Hong, and Tamer (2005) and Chen, Hong, and Tarozzi (2007) in the context of nonclassical measurement error. Assumption 6(b) is imposed because uniform convergence is not enough to ensure uniform convergence of derivatives, a result needed in the proof of the following theorem.

Theorem 5. *(ASYMPTOTIC LINEAR REPRESENTATION OF EIFE) Let* $\beta^* \in \operatorname{int}\left(\mathcal{B}^{J+1}\right)$, $\hat{\beta}^{EIF} = \beta^* + o_p(1)$ *and Assumptions 1, 2, 5 and 6 hold. Assume the following additional conditions hold:*

(5.1) $\|\hat{p} - p^*\|_\infty = o_p\left(n^{-1/4}\right)$.

(5.2) $\sup_{|\beta-\beta^*|<\delta} \|\hat{e}(\beta) - e^*(\beta)\|_\infty = o_p(1)$, *for some* $\delta > 0$.

(5.3) $M_n^{EIF}(\beta^*, \hat{p}, \hat{e}(\beta^*)) = M_n^{EIF}(\beta^*, p^*, e^*(\beta^*)) + o_p(n^{-1/2})$.

Then,

$$\hat{\beta}^{EIF} - \beta^* = -\left(\Gamma_*' W \Gamma_*\right)^{-1} \Gamma_*' W M_n^{EIF}(\beta^*, p^*, e^*(\beta^*)) + o_p\left(n^{-1/2}\right).$$

Asymptotic normality of $\hat{\beta}^{EIF}$ also follows directly from Theorem 5. This time, three additional conditions involving the nonparametric estimators are imposed. Condition (5.1) is the same as Condition (4.1) in Theorem 4. Condition (5.2) further requires uniform consistency of the nonparametric estimator of e^* in both arguments, although in this case no particular rate is required. This result follows from the additional smoothness assumptions imposed in this theorem. Finally, Condition (5.3) is the analogue of Condition (4.2) in Theorem 4, although much easier to verify in general. In this case, additional knowledge about the GPS may be easily incorporated in the estimation without affecting the asymptotic variance, provided the asymptotic linear representation continues to hold.

Efficiency of the estimators follow directly from Theorems 4 and 5:

Corollary 2. *If* $d_m = d_\beta$ *(just-identified case) or* $W = V_*^{-1}$ *(as given in Theorem 1), then the IPWE and EIFE are efficient for* β^*.

This corollary distinguishes two cases. First, if the problem is exactly identified (as in our Examples 1 and 2), then the estimators are efficient without further work. Second, if the problem is over-identified (as in our Example 3), then a consistent estimator of the matrix V_*^{-1} is needed, generating an intermediate step in the construction of the GMM problems for the IPWE and the EIFE. A consistent estimator for V_*^{-1} is easy to construct without further assumptions, as we show below after we consider our leading examples.

EXAMPLE 1 (CONTINUED): MMTE. The class of functions $\{(\cdot - \mu_t) : |\mu_t - \mu_t^*| < \delta\}$ is Donsker and $\mathbb{E}[|m(Y(t); \mu_t) - m(Y(t); \mu_t^*)|] = |\mu_t - \mu_t^*|$, giving Assumption 5. Thus, under the conditions of Theorem 4 and Corollary 2 we conclude that $\sqrt{n}(\hat{\mu}^{IPW} - \mu^*) \xrightarrow{d} \mathcal{N}(0, V^*)$, and the estimator $\hat{\mu}^{IPW}$ is efficient. Further, in this case Assumption 6 is trivially satisfied and therefore under the conditions of Theorem 5 and Corollary 2 we obtain $\sqrt{n}(\hat{\mu}^{EIF} - \mu^*) \xrightarrow{d} \mathcal{N}(0, V^*)$, and the estimator $\hat{\mu}^{EIF}$ is also efficient. □

EXAMPLE 2 (CONTINUED): MQTE. The class of functions $\{(\mathbf{1}\{y \leq q_t(\tau)\} - \tau) : |q_t(\tau) - q_t^*(\tau)| < \delta\}$ is Donsker and

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

for all $q_t(\tau)$ such that $|q_t(\tau) - q_t^*(\tau)| < \delta$, for some $\delta > 0$, under regularity conditions. It follows from this calculation that Assumption 5 is satisfied in this case and under the conditions of Theorem 4 and Corollary 2 we conclude $\sqrt{n}(\hat{q}^{IPW}(\tau) - q^*(\tau)) \xrightarrow{d} \mathcal{N}(0, V^*)$, and $\hat{q}_t^{IPW}(\tau)$ is efficient. Turning to Assumption 6, part (a) may be easily verified under mild regularity conditions because $e_t^*(X;\beta_t) = F_{Y(t)}^*(q_t(\tau) \mid X) - \tau$, while part (b) requires further restrictions on the class of distribution functions allowed for in this case. Thus, under regularity conditions, we obtain $\sqrt{n}(\hat{q}^{EIF}(\tau) - q^*(\tau)) \xrightarrow{d} \mathcal{N}(0, V^*)$ with $\hat{q}^{EIF}(\tau)$ efficient. □

5.3. Optimal Weighting Matrix and Uncertainty Estimation. Now we turn to the estimation of V_* and Γ_*, the variance of the EIF and the "sandwich" matrix appearing in the SPEB, respectively. For the over-identified case, estimation of V_* is crucial since the square-root of this matrix is the optimal weighting matrix of both GMM problems.

The natural plug-in estimator of V_* is given by

$$V_n = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, T_i, X_i, \hat{\beta}, \hat{p}, \hat{e}(\hat{\beta}))\psi(Y_i, T_i, X_i, \hat{\beta}, \hat{p}, \hat{e}(\hat{\beta}))',$$

for some consistent estimator $\hat{\beta}$ of β^*.

Theorem 6 gives a set of simple sufficient conditions that ensure $\hat{V}_n$ is consistent for V_*.

Theorem 6. (CONSISTENT ESTIMATOR OF V^*) *Let Assumptions 1, 2, 5, and 6(a) with* $\mathbb{E}[\sup_{|\beta_t - \beta_t^*|<\delta} \left|\partial_{\beta_t} e_t^*(X;\beta_t)\right|^2] < \infty$ *hold. If* $\hat{\beta} = \beta^* + o_p(1)$, $\|\hat{p} - p^*\|_\infty = o_p(1)$ *and* $\sup_{|\beta-\beta^*|<\delta} \|\hat{e}(\beta) - e^*(\beta)\|_\infty = o_p(1)$, *for some* $\delta > 0$, *then* $V_n = V_* + o_p(1)$.

Observe that the conditions imposed in Theorem 6 are the same as those assumed in Theorem 4 plus the mild smoothness and dominance condition on e^*. Next, consider the estimation of Γ_*. Because this matrix has a very particular structure there are several simple alternative approaches to construct a consistent estimator. For example, it is possible to consider a numerical derivative approach directly applied to the sample analogue (e.g., Pakes and Pollard (1989)) or, in some cases, the estimator may be constructed by taking into consideration the explicit form of the matrix (for instance, in Example 2 we have $\Gamma_t(\beta_t^*) = f_{Y(t)}^*(q_t^*)$). As a third alternative, it is also possible to construct a generic estimator, under the assumptions we have already imposed if integration and differentiation can be interchanged. In this case, we see that for all $t \in \mathcal{T}$ we have

$$\Gamma_t^* = \frac{\partial}{\partial \beta_t}\mathbb{E}\left[m(Y(t);\beta_t)\right]\Big|_{\beta_t=\beta_t^*} = \mathbb{E}\left[\frac{\partial}{\partial \beta_t} e_t(X;\beta_t)\Big|_{\beta_t=\beta_t^*}\right],$$

which suggests the plug-in estimator given by

$$\hat{\Gamma}_{t,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta_t}\hat{e}_t(X;\beta_t)\Big|_{\beta_t=\hat{\beta}_t}.$$

We verify the consistency of this plug-in estimator in the following theorem.

Theorem 7. (CONSISTENT ESTIMATOR OF Γ_*) *Let Assumptions 1, 2, 6 hold. If* $\hat{\beta} = \beta^* + o_p(1)$ *and* $\sup_{|\beta-\beta^*|<\delta} \|\hat{e}(\beta) - e^*(\beta)\|_\infty = o_p(1)$, *for some* $\delta > 0$, *then* $\hat{\Gamma}_{t,n} = \Gamma_t^* + o_p(1)$.

From Theorem 7 it is straightforward to form a consistent estimator of the gradient matrix Γ_*.

5.4. Nonparametric Estimation of Nuisance Parameters. We have established asymptotic normality and efficiency of the estimators considered in this paper. These results have been obtained by imposing high-level assumptions concerning the behavior of the nonparametric estimators used for the estimation of the infinite dimensional nuisance parameters rather than by specifying a particular form of such estimators. In this section we discuss explicitly the nonparametric estimation of p^* and e^* and verify the additional high-level conditions imposed in Theorems 4 and 5.

Since both p^* and e^* are (possibly high-dimensional) conditional expectations, a nonparametric series estimator seems an appropriate choice. These estimators are attractive because they are computational convenient and they can incorporate dimension reduction restrictions easily. This nonparametric estimation procedure has been studied in detailed by Newey (1997) and may be interpreted as a linear sieve estimator as discussed in Chen (2007). To briefly describe the estimator, let $g(X) = \mathbb{E}[Z \mid X]$ for some random variable Z and random vector $X \in \mathcal{X}$, and let $\{r_k(x)\}_{k=1}^{\infty}$ be a sequence of known approximating functions with the property that a linear combination of $R_K(x) = (r_1(x), \cdots, r_K(x))'$ can approximate $g(x)$ for $K = 1, 2, \cdots$. An approximating function is formed by $g(X; \gamma_K) = R_K(X)' \gamma_K$ and the series estimator based on an i.i.d. random sample (Z_i, X_i), $i = 1, 2, \cdots, n$, is given by $\hat{g}(X) = g(X; \hat{\gamma}_K)$, with

$$\hat{\gamma}_K = \arg\min_{\gamma_K} \sum_{i=1}^{n} (Z_i - g(X_i; \gamma_K))^2,$$

where, in this case, the closed-form solution is given by

$$\hat{\gamma}_K = \left(\sum_{i=1}^{n} R_K(X_i) R_K(X_i)' \right)^{-} \sum_{i=1}^{n} R_K(X_i) Z_i \tag{6}$$

with B^- denoting the generalized inverse of the matrix B.

By choosing the approximating basis appropriately and under suitable conditions on the function $g(\cdot)$ and growth rate of K it is possible to establish the consistency and rate of convergence (in both L_2 and uniform sense) of this nonparametric estimator. Two common choices for an approximating basis are power series and splines, leading to polynomial regression and spline regression, respectively. See Newey (1997) for further details.

This nonparametric estimator may be used directly to estimate the vector valued function e^*. For all $t \in \mathcal{T}$, let $Z(\beta_t) = m(Y; \beta_t)'$ and let $\hat{\gamma}_{t,K}(\beta_t)$ to be defined as in equation (6) but when only the data for $T = t$ is used. Then, for all $t \in \mathcal{T}$, the series nonparametric estimator of $e_t^*(X; \beta_t)$, $\beta_t \in \mathcal{B}$, is given by $\hat{e}_t(X; \beta_t)' = R_K(X)' \hat{\gamma}_{t,K}(\beta_t)$ where

$$\hat{\gamma}_{t,K}(\beta_t) = \left(\sum_{i=1}^{n} D_{t,i} \cdot R_K(X_i) R_K(X_i)' \right)^{-} \sum_{i=1}^{n} D_{t,i} \cdot R_K(X_i) m(Y_i; \beta_t)'.$$

We may construct similarly a series estimator for p^*. 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 this standard nonparametric estimator. Thus, in this case we consider a nonparametric estimator consistent with this additional requirements. We study a generalization of the estimator introduced by Hirano, Imbens, and Ridder (2003) for the particular context of binary treatments, labeled Multinomial Logistic Series Estimator (MLSE), which may be interpreted as a non-linear sieve (Chen (2007)) estimation procedure.

Intuitively, since we are estimating nonparametrically $J+1$ conditional probabilities it is reasonable to embed them within a multinomial logistic model. Using the notation introduced for series estimation, for all $t \in \mathcal{T}$, let $g\left(X;\gamma_{t,K}\right) = R_K\left(X\right)'\gamma_{t,K}$ be the approximating function and for notational simplicity let $\gamma_K = (\gamma_{0,K}', \gamma_{1,K}', \cdots, \gamma_{J,K}')'$. When the coefficients $\gamma_{t,K}$, $t \in \mathcal{T}$, are chosen as in equation (6) with $Z = D_t$ we obtain the usual series estimator for the components of p^*. Alternatively, the MLSE chooses simultaneously all the vectors in γ_K by solving the maximum likelihood multinomial logistic problem

$$\hat{\gamma}_K = \arg \max_{\gamma_K | \gamma_{0,K}' = 0_K} \sum_{i=1}^{n}\sum_{t=0}^{J} D_{t,i} \cdot \log\left(\frac{\exp\left\{g\left(X_i;\gamma_{t,K}\right)\right\}}{\sum_{j=0}^{J}\exp\left\{g\left(X_i;\gamma_{t,K}\right)\right\}}\right),$$

where 0_K represents a $K \times 1$ vector of zeros used to impose the usual normalization $\gamma_{K,0} = 0_K$ needed to achieve identification in this model. In this case, the nonparametric estimator $\hat{p}\left(\cdot\right)$ has typical t-th element given by

$$\hat{p}_t\left(X\right) = \frac{\exp\left\{R_K\left(X\right)'\hat{\gamma}_{t,K}\right\}}{1+\sum_{j=1}^{J}\exp\left\{R_K\left(X\right)'\hat{\gamma}_{t,K}\right\}}.$$

It is straightforward to verify that this nonparametric estimator satisfies the additional restrictions underlying the GPS. The rates of convergence of this non-linear sieve estimator are established in Appendix B.

For simplicity and to reduce the notational burden, we restrict attention to power series and splines as possible approximation basis and we assume that the same bases is used for all the nonparametric estimators. The following simple assumption is enough to establish the appropriate large sample results for both the linear series estimator and the MLSE.

Assumption 7. *(a) For all $t \in \mathcal{T}$, $p_t^*\left(\cdot\right)$ and $e_t^*\left(\cdot,\beta_i^*\right)$ are s times differentiable with $s/d_x > 5\eta/2 + 1/2$, where $\eta = 1$ or $\eta = 1/2$ depending on whether power series or splines are used as basis functions, respectively; (b) X is continuously distributed with density bounded and bounded away from zero on its compact support $\mathcal{X}$; and (c) for all $t \in \mathcal{T}$ and some $\delta > 0$, $\mathbb{V}\left[m\left(Y\left(t\right);\beta_t\right) \mid X = x\right]$ is uniformly bounded for all $x \in \mathcal{X}$ and all β_t such that $|\beta_t - \beta_t^*| < \delta$.*

Part (a) of Assumption 7 provides the exact restrictions needed on the spaces $\mathcal{P}$ and $\mathcal{E}$, describing the minimum smoothness required as a function of the dimension of X and the choice of basis of approximation. Part (b) of Assumption 7 restricts X to be continuous on a compact support with "well-behaved" density. These assumptions may be relaxed

considerably at the expense of some additional notation. For example, it is possible to allow for some components of X to be discretely distributed and to permit $\mathcal{X}$ to be unbounded by restricting the tail-behavior of the density of X (see Chen, Hong, and Tamer (2005) for an example). Part (c) of Assumption 7 is standard from the series (or sieve) nonparametric estimation literature.

Theorem 8. *(NONPARAMETRIC ESTIMATION) Let Assumptions 1(b) and 7 hold. Then, conditions (4.1) and (4.2) in Theorem 4, and conditions (5.1), (5.2) and (5.3) in Theorem 5 are satisfied by the nonparametric estimators introduced in this section if $K = n^{\nu}$ with*

$$\frac{1}{4s/d_x - 6\eta} < \nu < \frac{1}{4\eta + 2}$$

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

6. OTHER POPULATION PARAMETERS AND HYPOTHESIS TESTING

The results presented in the paper so far allow for the joint efficient estimation of several multi-valued treatment effects. For instance, using the procedures discussed we may easily estimate jointly (and efficiently) several marginal quantiles as well as the marginal mean of all potential outcomes. However, in many applications the population parameters of interest may be not only the marginal treatment effects but also other quantities involving possibly more than one marginal treatment effect. Because differentiable transformations of efficient estimators of Euclidean parameters lead to efficient estimators for the corresponding population parameters, a simple delta-method argument allows us to easily recover any collection of treatment effects that can be written as (or approximated by) a differentiable function of the marginal treatment effects.

Using this idea and Examples 1 and 2, we may efficiently estimate many other treatment effects such as pairwise comparisons (in the spirit of ATE), differences between pairwise comparisons, incremental ratios, interquantile ranges, quantile ratios or other measures of differential and heterogeneous treatments effects. Moreover, by a straightforward extension of Example 1, we can consider the efficient estimation of the effect of different treatments on dispersion as measured by the standard deviation of the potential outcome distribution. We exploit these ideas further in the next section when we present the empirical illustration.

In particular, because the ATE and QTE are continuous transformations of the treatment effects studied in Example 1 and Example 2, we may also obtain the important results of Hahn (1998), Hirano, Imbens, and Ridder (2003) and Firpo (2007) from the binary treatment effect literature as particular cases of our examples:

EXAMPLE 1 (CONTINUED): MMTE. If $\mathcal{T} = \{0, 1\}$ and because the ATE can be written as $\Delta^{ATE} \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = v'\mu^*$, where $v = (-1, 1)'$, using Theorem 1 we conclude that

$$V^* = \mathbb{E}\left[\begin{array}{cc} \frac{\sigma_0^2(X)}{p_0(X)} + (\mu_0(X) - \mu_0^*)^2 & (\mu_0(X) - \mu_0^*) \cdot (\mu_1(X) - \mu_1^*) \\ (\mu_0(X) - \mu_0^*) \cdot (\mu_1(X) - \mu_1^*) & \frac{\sigma_1^2(X)}{p_1(X)} + (\mu_1(X) - \mu_1^*)^2 \end{array}\right].$$

Then, either Theorem 4 or Theorem 5 and the transformation $g(z) = v'z$ gives

$$\sqrt{n}\left(\hat{\Delta}^{ATE} - \Delta^{ATE}\right) \xrightarrow{d} \mathcal{N}\left[0, v'V^*v\right],$$

where

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

and $\Delta^{ATE}(X) = \mu_1(X) - \mu_0(X)$. In this case, the asymptotic variance is the SPEB found by Hahn (1998) and the resulting estimator in the case of Theorem 4 is essentially the same as the one considered in Hirano, Imbens, and Ridder (2003) (see also Imbens, Newey, and Ridder (2006) for another similar modification of this estimator). □

EXAMPLE 2 (CONTINUED): MQTE. If $\mathcal{T} = \{0, 1\}$ and because the QTE may also be written as $\Delta^{QTE} \equiv q_1^*(\tau) - q_0^*(\tau) = v'q^*(\tau)$, where $v = (-1, 1)'$, either Theorem 4 or Theorem 5 gives

$$\sqrt{n}\left(\hat{\Delta}^{QTE} - \Delta^{QTE}\right) \xrightarrow{d} \mathcal{N}\left[0, v'V^*v\right].$$

In this case, the asymptotic variance coincides with the SPEB derived in Firpo (2007) and the resulting estimator in the case of Theorem 4 corresponds to the Z-estimator version of Firpo's estimator for the QTE. □

Furthermore, because in some applications incorporating additional information about the treatment effects in a general over-identified model may be challenging, we can consider an alternative approach to the efficient estimation of multiple restricted treatment effects. In particular, suppose that the restrictions of interests can be imposed by writing the marginal treatment effects as a function of the parameters π^*, and denote this function by $\beta(\pi^*)$. Then, it can be verified that, under mild regularity conditions, an efficient estimator of π^* is given by

$$\hat{\pi} = \arg\min_{\pi} [\hat{\beta} - \beta(\pi)]' \left(\Gamma_n' V_n^{-1} \Gamma_n\right) [\hat{\beta} - \beta(\pi)],$$

where $\hat{\beta}$ is an efficient estimator of β^*, Γ_n is a consistent estimator of Γ_*, and V_n is a consistent estimator of V_*. In this case, we obtain

$$\sqrt{n}(\hat{\pi} - \pi^*) \xrightarrow{d} \mathcal{N}\left[0, \left(\partial\beta(\pi^*)' \Gamma_*' V_*^{-1} \Gamma_* \partial\beta(\pi^*)\right)^{-1}\right],$$

where $\partial\beta(\pi^*) = \frac{\partial}{\partial\pi}\beta(\pi)\big|_{\pi=\pi^*}$. From this result, a consistent estimator of the covariance matrix of $\hat{\pi}$ may be constructed using a plug-in approach.

To fix ideas, consider the case of Example 3. As discussed before, under the assumption of symmetry we may incorporate this information to form a over-identified GMM problem. Alternatively, we could first estimate jointly $(\mu^*, q^*(.5))$ using either the IPWE or the EIFE and then solve the following problem:

$$\hat{\pi} = \arg\min_{\pi} \begin{bmatrix} \hat{\mu} - \pi \\ \hat{q}(.5) - \pi \end{bmatrix}' \left(\Gamma_n' V_n^{-1} \Gamma_n\right) \begin{bmatrix} \hat{\mu} - \pi \\ \hat{q}(.5) - \pi \end{bmatrix},$$

which leads to an efficient estimator of the multi-valued treatment effect for location under symmetry. Similarly, using this idea we may also incorporate additional restrictions on different quantiles and other estimands of interest.

Finally, because testing procedures based on efficient estimators are optimal (possibly after restricting the class of allowed tests), it is straightforward to perform optimal testing of different hypotheses concerning multi-valued treatment effects. This can be done within and across treatment levels for marginal treatment effects, for treatment effects obtained by means of some (differentiable) transformation of these parameters, and for restricted treatment effects by relying on standard testing strategies.

7. EMPIRICAL ILLUSTRATION

To show how our procedures work in practice, we report a brief empirical exercise that studies the effect of maternal smoking during pregnancy on birth weight. In a recent paper, Almond, Chay, and Lee (2005) (ACL hereafter) present detailed empirical evidence on the economic costs of low birth weight (LBW). In their paper, the authors estimate the direct economic costs imposed by LBW on society and also study the possible causes of LBW using different nonexperimental techniques. In particular, ACL present empirical evidence on the effect of maternal smoking on birth weight for a rich database of singletons in Pennsylvania and find a strong effect of about 200-250 gram reduction in birth weight using both subclassification on the propensity score and regression adjusted methods.

In our application, we extend the results of ACL by considering the effect of maternal smoking *intensity* during pregnancy on birth weight. The database used by ACL not only includes almost half a million singleton births and many pre-intervention covariates, but also records the mother's declared number of cigarettes-per-day smoked during pregnancy. This additional information allows us to consider multi-valued treatment effects and address several interesting questions, particularly relevant from a policy-making perspective. For example, we assess whether the effect of smoking is constant across levels of smoking, whether there exist differential and/or heterogeneous treatment effects, and whether the variability in birth weight is affected by smoking intensity.

The empirical illustration uses the same database, response variable and pre-intervention variables as ACL. In this sample, approximately 80% of mothers did not smoke during pregnancy, while for the remaining 20% inspection of the empirical distribution of smoked cigarettes reveals important mass points approximately every 5 cigarettes ranging from 1 to 25. This feature suggests considering 5-cigarette bins as a starting point for the empirical analysis. We collapse the number of smoked cigarettes into 6 categories ($J = 5$) {0, 1-5, 6-10, 11-15, 16-20, 21+} and we consider the joint estimation of five quantiles (.9,.75,.5,.25,.1), the mean and standard deviation for each potential outcome, leading to 42 treatment effects. For $t \in \mathcal{T}$, the identifying moment function in this case is given by the vector-valued function $m(y;\beta_t) = ((\mathbf{1}\{y \leq \beta_{1t}\} - 0.95),\ (\mathbf{1}\{y \leq \beta_{2t}\} - 0.75),\ (\mathbf{1}\{y \leq \beta_{3t}\} - 0.5),\ (y - \beta_{4t}),\ (\mathbf{1}\{y \leq \beta_{5t}\} - 0.25),\ (\mathbf{1}\{y \leq \beta_{6t}\} - 0.1),\ (y^2 - \beta_{7t}))'$ for $\beta_t = (\beta_{1t},\ \beta_{2t},\ \beta_{3t},\ \beta_{4t},\ \beta_{5t},\ \beta_{6t})'$. We first jointly estimate β^* using both the IPWE and EIFE and then we recover the marginal population parameters of interest by means of the delta method.

To ensure comparability we use the same pre-intervention covariates as in ACL. In particular, we include 43 dummy variables (mother's demographics, father's demographics, prenatal care, alcohol use, pregnancy history, month of birth and county of residency) and 6

"continuous" covariates (mother's age and education, father's age and education, number of prenatal visits, months since last birth and order of birth).[10] For the estimation of both nonparametric nuisance parameters, we use cubic B-splines with knots ranging from 1 to 3 depending on the continuous covariate, and to reduce the computational burden we impose an additive separability assumption on the approximating functions. We experimented with different choices of smoothing parameters for the splines as well as with different interactions between the dummies and the smoothed covariates. In all the cases considered, the results appear to be robust to the particular specification of the nonparametric estimators.[11]

Because in this case the model is exactly identified, we may estimate each treatment effect separately and then form the full EIF to estimate the SPEB. Table 1 presents the point and uncertainty estimates for the 42 treatment effects using three estimators: a simple dummy regression estimator (DRE), the IPWE and the EIFE. In this sample, estimates from the (inefficient, possibly inconsistent) DRE appear to be very similar to those obtained from the (consistent and efficient) IPWE and EIFE. This result is consistent with the findings in ACL. The standard errors of our estimators appear to be very similar to each other and considerably lower than those of the DRE in the case of the mean, while for the quantiles the standard errors are slightly higher.[12]

A simple way to present the information in Table 1 is by means of Figure 1, which gives important qualitative information about the treatment effects. This figure shows the point estimates and their 95% (marginal) confidence intervals for the case of the MMTE and MQTE when estimated using the IPWE. Interestingly, we observe a parallel shift in the entire distribution of birth weight along smoking intensity. In particular, there is a large reduction of about 150 grams when the mother starts to smoke (1-5 cigarettes), an additional reduction of approximately 70 grams when changing from 1-5 to 6-10 cigarettes-per-day, and no additional effects once the mother smokes at least 11 cigarettes. These findings provide qualitative evidence that differential treatment effects are non-linear and approximately homogeneous along the distribution of the potential outcomes. In particular, we observe a close to symmetric distribution with approximately constant dispersion (as measured by both interquartile ranges and standard deviation).

The qualitative results summarized in Figure 1 may be formally tested. Since we have jointly estimated the 42 marginal treatment effects, it is straightforward to test the hypotheses suggested by Figure 1 as well as other hypotheses of interest. Table 2 presents a collection of hypothesis tests regarding pairwise differences and difference-in-differences of marginal mean treatment effects. On the diagonal, we report pairwise differences across

[10] A full description of the variables used is given in footnote 36 of ACL. We do not include maternal medical risk factors in the analysis; see also footnote 39 of ACL.

[11] This is consistent with the available literature on semiparametric estimation suggesting that the choice of basis or smoothing parameters are relatively unimportant (see for example Newey (1994), Ai and Chen (2003), Chen, Hong, and Tamer (2005), or Chen, Hong, and Tarozzi (2007)). Based on these results, and for computational simplicity, we did not consider data-driven procedures such as cross-validation for the selection of the smoothing parameters.

[12] In the quantile dummy regression case the standard errors were calculated using a kernel density estimator with bandwidth set by Silverman's rule-of-thumb. In the case of IPWE and EIFE, we estimate the gradient matrix Γ_* using its exact form (implemented by a weighted kernel density estimator with bandwidth set by Silverman's rule-of-thumb as in Firpo (2007)). We also experimented with the general numerical derivative approach (implemented by a simple numerical difference), which led to very similar estimates.

treatment levels. For example, the reduction in birth weight induced by increasing maternal smoking from 0 to 1-5 cigarettes is 146 grams (statistically significant), while the corresponding reduction induced by increasing maternal smoking from 6-10 to 11-15 cigarettes is 37 grams (not statistically significant). This table also reports the difference-in-differences comparisons which may be used to test for non-linearities. For example, increasing maternal smoking from 0 to 1-5 cigarettes induces an additional 75 gram reduction in birth weight when compared to the corresponding reduction induced by increasing maternal smoking from 1-5 to 6-10 cigarettes. This differential effect is statistically significant and provides formal evidence of non-linear treatment effects. Importantly, non-linearities disappear beyond the tenth cigarette smoked during pregnancy. Similar results are obtained when analyzing the MQTE.

Table 3 illustrates additional multiple-hypotheses tests of interest. In the first row, we jointly test the hypothesis of no treatment effect (as measured by mean, quantile and spread) for the highest three treatment levels, while in the second and third rows we present the analogous tests considering the highest four and highest five treatment levels, respectively. As shown in this table, increasing smoking intensity beyond 10 cigarettes per day has no further effect on birth weight. The remaining rows in Table 3 test for different hypotheses involving possible distributional effects across and within treatment levels. We find small but statistically significant differences on the interquantile ranges.

Finally, based on our main finding that most of the effect of smoking on birth weight appears to be concentrated on the first 10 cigarettes-per-day smoked, we replicate our analysis for the subpopulation of mothers who smoked between 0 and 10 cigarettes-per-day breaking up the treatment variable into 2-cigarette bins.[13] To conserve space, we only present qualitative results in Figure 2. According to this figure, the treatment effects continue to be non-linear and approximately homogenous at all quantile levels. Interestingly, the main reduction in birth weight appears to be caused by increasing the number of cigarettes smoked from 0 to 1-2. This effect appears constant until the fourth cigarette. Increasing smoking beyond the fourth cigarette has an additional negative effect on birth weight, although this effect is smaller than the effect from 0 to 1-2.

8. Final Remarks

We study the efficient estimation of a large class of multi-valued treatment effects implicitly defined by a possibly over-identified non-smooth collection of moment conditions. We propose two alternative estimators based on standard GMM arguments combined with the corresponding modifications needed to circumvent the fundamental problem of causal inference. Under regularity conditions, these estimators are shown to be root-N consistent, asymptotically normal and efficient for the general population parameter of interest. Using these estimators we show how other estimands of interest may also be efficiently estimated, allowing the researcher to recover a rich class of population parameters. We verify that important results in the literature of program evaluation with binary treatment assignments may be seen as particular cases of our procedure when the treatment is dichotomous.

Considering multi-valued treatment assignments provides the opportunity for a better characterization of the program under study. As illustrated in the empirical application,

[13] Unfortunately, 1-cigarette bins could not be used due to sample size restrictions.

collapsing a multiple treatment into a binary indicator may prevent the researcher from detecting the presence of important non-linear effects. More generally, in many applications we may expect to have multiple differential impacts within and across treatments, which highlights the relevance of considering multi-valued treatments, when possible, for making informed policy decisions. A possible extension of our results would focus on studying the effect of policies that change the distribution of multiple treatments and how these alternative configurations may affect the population (or subpopulations) under consideration.

Our results have been obtained under the assumption of finite multi-valued treatments, which leads to a statistical model where many estimands of interest are regular, this is, they can be estimated at the parametric rate. A natural extension would be to relax this assumption to continuous treatment assignments. This may be appealing from an empirical perspective, but would make many population parameters of interest irregular. Nonetheless, when treatments are continuous, it may be possible to consider relevant regular estimands such as specific functionals of the treatment effect process or, more interestingly, alternative restrictions on the underlying statistical model that may deliver regular population parameters.

The results presented in this paper could also be extended based on the developments available in the literature of binary treatment effects. For example, in applications it may be of interest to consider the multi-valued analogue of weighted treatment effects (Hirano, Imbens, and Ridder (2003)), including average and quantile treatment effects for a given treatment level as particular cases. Efficiency calculations and the corresponding efficient estimation procedures for these estimands may be derived by following and extending the work discussed here. For an alternative extension, consider the important concern in empirical work about the lack of common support in the estimated propensity score, a pathology likely to be exacerbated in the context of multi-valued treatments. Using the results presented here, it may be possible to consider a systematic approach to deal with limited overlap by extending the recent work of Crump, Hotz, Imbens, and Mitnik (2007) to the context of multiple treatments.

Finally, in this paper we proposed two estimators that are first-order efficient. However, as in the binary treatment case, other efficient estimators may also be considered, which implies that an important open question for future research is how to rank the large class of first-order efficient estimators available. Although it seems unclear how to rank these estimators, the results of this paper justify focusing on the marginal treatment effects as the target estimand when ranking the competing first-order efficient estimators.

References

Abadie, A. (2005): "Semiparametric Difference-in-Differences Estimators," *Review of Economic Studies*, 72(1), 1–19.

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 7(1), 235–267.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71(6), 1795–1843.

Almond, D., K. Y. Chay, and D. S. Lee (2005): "The Costs of Low Birth Weight," *Quarterly Journal of Economics*, 120(3), 1031–1083.

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

——— (2002): "Generalized Method of Moments Estimation When a Parameter Is on a Boundary," *Journal of Business and Economic Statistics*, 20(4), 530–544.

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. Klaasen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volumn VI*, ed. by J. Heckman, and E. Leamer. Elsevier Science B.V.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343–366.

Chen, X., H. Hong, and A. Tarozzi (2007): "Semiparametric Efficiency in GMM Models With Auxiliary Data," *The Annals of Statistics*, forthcoming.

Chen, X., O. Linton, and van Keilegom (2003): "Estimation of Semiparametric Models when The Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Crump, R. K., V. Hotz, G. W. Imbens, and O. A. Mitnik (2007): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," Working Paper.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

Frölich, M. (2004): "Programme Evaluation With Multiple Treatments," *Journal of Economic Surveys*, 18(2), 181–224.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *The Review of Economic Studies*, 65(2), 261–294.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

HOROWITZ, J. L., AND C. F. MANSKI (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Jounal of the American Statistical Association*, 95(449), 77–84.

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(260), 663–685.

IMBENS, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87(3), 706–710.

——— (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, 86(1), 4–29.

IMBENS, G. W., W. K. NEWEY, AND G. RIDDER (2006): "Mean-Squared-Error Calculations for Average Treatment Effects," Working Paper.

LECHNER, M. (2001): "Identification and Estimation of Causal Effects of Multiple Treatments Under The Conditional Independence Assumption," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 43–58. Physica/Springer, Heidelberg.

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

NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 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 Testing," in *Handbook of Econometrics, Volumn IV*, ed. by R. F. Engle, and D. L. McFadden, pp. 2112–2245. Elsevier Science B.V.

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 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(429), 122–129.

ROBINS, J. M., A. ROTNITZKY, AND L. ZHAO (1994): "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89(427), 846–866.

——— (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90(429), 846–866.

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(1), 41–55.

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

TANABE, K., AND M. SAGAE (1992): "An Exact Cholesky Decomposition and the Generalized Inverse of the Variance-Covariance Matrix of the Multinomial Distribution, with Applications," *Journal of the Royal Statistical Society. Series B Methodological*, 54(1), 211–219.

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

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, New York.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer, New York.

——— (2000): "Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes," in *High Dimensional Probability II*, ed. by E. Giné, D. Mason, and J. A. Wellner, pp. 115–134. Bikhäuser, Boston.

WOOLDRIDGE, J. M. (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, forthcoming.

APPENDIX A. PROOFS OF THEOREMS

In this appendix we let C denote a generic positive constant which may vary depending on the context. Also, for any vector v we denote its t-th element by $v_{[t]}$, and for any matrix A we denote its (i,j)-th element by $A_{[i,j]}$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalue of the matrix A, respectively.

Proof of Theorem 1 (EIF AND SPEB)**:** the proof given is based on the theoretical approach described in Bickel, Klaasen, Ritov, and Wellner (1993) and Newey (1990), and follows the results presented in Hahn (1998) and Chen, Hong, and Tarozzi (2007). The derivation is completed in three steps: characterization of the tangent space, verification of pathwise differentiability of the parameter of interest, and SPEB computation. Let $L_0^2(F_W)$ be the usual Hilbert space of zero-mean, square-integrable functions with respect to the distribution function F_W.

First, consider a (regular) parametric submodel of the joint distribution of (Y,T,X), the observed data model, with c.d.f. $F(y,t,x;\theta)$ and log-likelihood given by

$$\log f(y,t,x;\theta) = \sum\nolimits_{j\in\mathcal{T}} \mathbf{1}\{t=j\} \cdot \Big[\log f_j(y \mid x;\theta) + \log p_j(x;\theta)\Big] + \log f_X(x;\theta),$$

which equals $\log f(y,t,x)$ when $\theta=\theta_0$, and where $f_j(y \mid x;\theta)$ corresponds to the density of $Y(j)|X$, $p_j(x;\theta) = \mathbb{P}[D_j=1|x;\theta]$ and $p_j(x;\theta_0) = p_j^*(x)$ for all $j\in\mathcal{T}$. The corresponding score is given by

$$S(y,t,x;\theta_0) = \frac{d}{d\theta}\log f(y,t,x;\theta)\Big|_{\theta_0} = S_y(y,t,x) + S_p(t,x) + S_x(x),$$

where

$$\begin{aligned}
S_y(y,t,x) &= \sum\nolimits_{j\in\mathcal{T}} \mathbf{1}\{t=j\}\cdot s_j(y,x), & s_j(y,x) &= \frac{d}{d\theta}\log f_j(y \mid x;\theta)\Big|_{\theta_0},\\
S_p(t,x) &= \sum\nolimits_{j\in\mathcal{T}} \mathbf{1}\{t=j\}\cdot \frac{\dot p_j^*(x)}{p_j^*(x)}, & \dot p_j^*(x) &= \frac{d}{d\theta}p_j(x;\theta)\Big|_{\theta_0},\\
S_x(x) &= \frac{d}{d\theta}\log f_X(x;\theta)\Big|_{\theta_0}.
\end{aligned}$$

Therefore, the tangent space of this statistical model is characterized by the set of functions $\mathcal{T} \equiv \mathcal{T}_y + \mathcal{T}_p + \mathcal{T}_x$, where

$$\begin{aligned}
\mathcal{T}_y &= \Big\{S_y(Y,T,X) : s_j(Y(t),X) \in L_0^2\left(F_{Y(t)|X}\right), \forall j\in\mathcal{T}\Big\},\\
\mathcal{T}_p &= \Big\{S_p(T,X) : S_p(T,X) \in L_0^2\left(F_{T|X}\right)\Big\},\\
\mathcal{T}_x &= \Big\{S_x(X) : S_x(X) \in L_0^2(F_X)\Big\}.
\end{aligned}$$

In particular, observe that

$$\mathbb{E}[S_p(T,X) \mid X] = \mathbb{E}\left[\sum\nolimits_{t\in\mathcal{T}} D_j \cdot \frac{\dot p_t^*(X)}{p_t^*(X)} \;\middle|\; X\right] = \sum\nolimits_{t\in\mathcal{T}} \dot p_t(X;\theta_0),$$

and

$$\mathbb{E}\left[S_p(T,X)^2 \;\middle|\; X\right] = \mathbb{E}\left[\sum\nolimits_{i\in\mathcal{T}}\sum\nolimits_{j\in\mathcal{T}} D_i\cdot\frac{\dot p_i^*(X)}{p_i^*(X)}\cdot D_j\cdot\frac{\dot p_j^*(X)}{p_j^*(X)} \;\middle|\; X\right] = \sum\nolimits_{t\in\mathcal{T}}\frac{\dot p_t^*(X)^2}{p_t^*(X)},$$

and hence it is required that $p_t^*(x)$ and $\dot p_t(x;\theta_0)$ are measurable functions such that $\sum_{t\in\mathcal{T}}\dot p_t^*(X) = 0$ and $\sum_{t\in\mathcal{T}}\dot p_t^*(X)^2/p_t^*(X) < \infty$, 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. The second condition is verified by Assumption 1(b) and the fact that T is finite.

Next, define $\underline{\mathrm{m}}(\beta) = [m(Y(0);\beta_0)', \cdots, m(Y(J);\beta_J)']$ and let A be any $(d_\beta \cdot (J+1) \times d_m \cdot (J+1))$ positive semi-definite matrix. Then the population parameter of interest satisfies $A\mathbb{E}[\underline{\mathrm{m}}(\beta)] = 0$ if and only if $\beta = \beta^*$, and using the implicit function theorem we obtain

$$\frac{\partial}{\partial\theta}\beta^*(\theta) = -(A\Gamma_*)^{-1} A\Upsilon(\theta_0),$$

where

$$\Gamma_* = \frac{\partial}{\partial\beta}\mathbb{E}[\underline{\mathrm{m}}(\beta)]\Big|_{\beta=\beta^*}, \qquad \Upsilon(\theta_0) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[\underline{\mathrm{m}}(\beta^*)]\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\int \underline{\mathrm{m}}(\beta^*)\, dF(y,t,x;\theta)\Big|_{\theta=\theta_0},$$

and observe that

$$\Upsilon(\theta_0) = \left[\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[m(Y(0);\beta_0)'\right]\Big|_{\theta=\theta_0}, \cdots, \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[m(Y(J);\beta_J)'\right]\Big|_{\theta=\theta_0}\right]$$

with typical element $j \in \mathcal{T}$,

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta\left[m\left(Y(j);\beta_j^*\right)'\right]\Big|_{\theta=\theta_0} = \mathbb{E}\left[m\left(Y(j);\beta_j^*\right)\cdot s_j(Y(j) \mid X)\right] + \mathbb{E}\left[e_j^*\left(X;\beta_j^*\right)\cdot S_x(X)\right].$$

Now, to show that the parameter is pathwise differentiable we need to find a $d_\beta\cdot(J+1)$-valued function $\Psi_\beta(y,t,x;A) \in \mathcal{T}$ such that for all regular parametric submodels

$$\frac{\partial}{\partial\theta}\beta^*(\theta) = \mathbb{E}[\Psi_\beta(Y,T,X;A)\cdot S(Y,T,X;\theta_0)].$$

It is not difficult to verify that the function satisfying such condition is given by

$$\Psi_\beta(Y,T,X;A) = -(A\Gamma_*)^{-1} A\psi(Y,T,X;\beta^*,p^*,e^*(\beta^*)),$$

for a fixed choice of the matrix A.

Finally, it follows from semiparametric efficiency theory and standard GMM arguments that the EIF is obtained when $A = \Gamma_*'V_*^{-1}$, which leads to the SPEB given by $V^* = (\Gamma_* V_*^{-1}\Gamma_*')^{-1}$. ■

Proof of Theorem 2 (CONSISTENCY OF IPWE): we apply Corollary 3.2 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\beta) = A_n M_n^{IPW}(\beta,\hat{p})$, $G(\beta) = AM^{IPW}(\beta,p^*)$, and verifying their three sufficient conditions (i), (ii), and (iii). First observe that conditions (i) and (ii) are satisfied by construction of the estimator and the model considered. Next, because $A_n - A = o_p(1)$, to verify condition (iii) it is enough to show

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left|M_{[t],n}^{IPW}(\beta,\hat{p}_t) - M_{[t]}^{IPW}(\beta,p_t^*)\right| \\
&\leq \sup_{\beta\in\mathcal{B}}\left|M_{[t],n}^{IPW}(\beta,\hat{p}_t) - M_{[t],n}^{IPW}(\beta,p_t^*)\right| + \sup_{\beta\in\mathcal{B}}\left|M_{[t],n}^{IPW}(\beta,p_t^*) - M_{[t]}^{IPW}(\beta,p_t^*)\right| = o_p(1),
\end{aligned}$$

for all $t \in \mathcal{T}$. Now the result follows because for n large enough we have

$$\sup_{\beta\in\mathcal{B}}\left|M_{[t],n}^{IPW}(\beta,\hat{p}_t) - M_{[t],n}^{IPW}(\beta,p_t^*)\right| \leq C\cdot\|\hat{p}_t - p_t^*\|_\infty \cdot \frac{1}{n}\sum_{i=1}^n \frac{D_{t,i}}{p_t^*(X_i)} \cdot \sup_{\beta_t\in\mathcal{B}}|m(Y_i;\beta_t)| = o_p(1),$$

by Assumption 3(b), and

$$\sup_{\beta\in\mathcal{B}}\left|M_{[t],n}^{IPW}(\beta,p_t^*) - M_{[t]}^{IPW}(\beta,p_t^*)\right| = \sup_{\beta_t\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^n \frac{D_{t,i}\cdot m(Y_i;\beta_t)}{p_t^*(X_i)} - \mathbb{E}\left[\frac{D_t\cdot m(Y;\beta_t)}{p_t^*(X)}\right]\right| = o_p(1)$$

because (assuming $d_m = 1$ or applying the following argument element by element) the class of functions $\mathcal{F}_t = \{\mathbf{1}\{\cdot = t\} \cdot m(\cdot;\beta)/p_t^*(\cdot) : \beta \in \mathcal{B}\}$ is Glivenko-Cantelli by Assumptions 1(b) and 3 (van der Vaart and Wellner (2000)). ■

Proof of Theorem 3 (CONSISTENCY OF EIFE): the proof of this theorem follows the same logic as the proof of Theorem 2. We apply Corollary 3.2 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\theta) = A_n M_n^{EIF}(\beta, \hat{p}, \hat{e})$, $G(\theta) = AM^{EIF}(\beta, p^*, e^*)$, and verifying their three sufficient conditions (i), (ii), and (iii). Using the same arguments in the proof and the conclusion of Theorem 2, it is sufficient to show

$$\sup_{\beta\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{e}_t(X_i;\beta)\cdot\frac{D_{t,i}-\hat{p}_t(X_i)}{\hat{p}_t(X_i)}\right| = o_p(1),$$

for all $t \in \mathcal{T}$. To establish this result, first notice that by Assumption 3(b) we have $\mathbb{E}\left[\sup_{\beta\in\mathcal{B}}|e_t^*(X;\beta)|\right] < \infty$ for all $t \in \mathcal{T}$. Now, for n large enough we have

$$\begin{aligned}
&\sup_{\beta\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{e}_t(X_i;\beta)\cdot\frac{D_{t,i}-\hat{p}_t(X_i)}{\hat{p}_t(X_i)}\right| \\
&\leq C\cdot\sup_{\beta\in\mathcal{B}}\left\|\hat{e}_t(\beta)-e_t^*(\beta)\right\|_\infty + \sup_{\beta_t\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^{n}e_t^*(X_i;\beta)\cdot\frac{D_{t,i}-p_t^*(X_i)}{p_t^*(X_i)}\right| + o_p(1) = o_p(1),
\end{aligned}$$

because (assuming $d_m = 1$ or applying the argument element by element) the class of functions $\mathcal{F}_t = \{e_t^*(\cdot;\beta)\cdot(\mathbf{1}\{\cdot = t\} - p_t^*(\cdot))/p_t^*(\cdot) : \beta \in \mathcal{B}\}$ is Glivenko-Cantelli by Assumptions 1(b) and 3 (van der Vaart and Wellner (2000)). ■

Proof of Theorem 4 (ASYMPTOTIC LINEAR REPRESENTATION OF IPWE): we apply Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\beta) = A_n M_n^{IPW}(\beta, \hat{p})$, $G(\beta) = AM^{IPW}(\beta, p^*)$, and verifying their five sufficient conditions (i)-(v). First, observe that conditions (i), (ii), (iv) and (v) hold in our case by the construction of the estimator, Assumptions 2 and 5, and condition (2.1). Thus it only remains to show the stochastic equicontinuity condition (iii). To establish this condition, it suffices to show (see, e.g., Lemma 3.5 in Pakes and Pollard (1989) and Lemma 1 in Andrews (2002)) for all sequences $\delta_n = o(1)$ that

$$\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\frac{n^{1/2}\cdot\left|M_{[t],n}^{IPW}(\beta,\hat{p})-M_{[t]}^{IPW}(\beta^*,p^*)-M_{[t],n}^{IPW}(\beta^*,\hat{p})\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|} = o_p(1),$$

for all $t \in \mathcal{T}$. Now, to verify this final condition define

$$\Delta_{[t],n}(\beta, p-p^*) = -\frac{1}{n}\sum_{i=1}^{n}\frac{D_{t,i}\cdot m(Y_i;\beta_t)}{p_t^*(X_i)^2}\cdot(p_t(X_i)-p_t^*(X_i)),$$

and consider the following decomposition

$$\begin{aligned}
&\left|M_{[t],n}^{IPW}(\beta,\hat{p})-M_{[t]}^{IPW}(\beta^*,p^*)-M_{[t],n}^{IPW}(\beta^*,\hat{p})\right| \\
&\leq \left|M_{[t],n}^{IPW}(\beta,p^*)-M_{[t]}^{IPW}(\beta^*,p^*)-M_{[t],n}^{IPW}(\beta^*,p^*)\right| && \text{(A-1)}\\
&\quad+\left|M_{[t],n}^{IPW}(\beta,\hat{p})-M_{[t],n}^{IPW}(\beta,p^*)-\Delta_{[t],n}(\beta,\hat{p}-p^*)\right| && \text{(A-2)}\\
&\quad+\left|M_{[t],n}^{IPW}(\beta^*,\hat{p})+M_{[t],n}^{IPW}(\beta^*,p^*)-\Delta_{[t],n}(\beta^*,\hat{p}-p^*)\right| && \text{(A-3)}\\
&\quad+\left|\Delta_{[t],n}(\beta,\hat{p}-p^*)-\Delta_{[t],n}(\beta^*,\hat{p}-p^*)\right|. && \text{(A-4)}
\end{aligned}$$

Now, for n large enough and using the first term (A-1) we have

$$\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\frac{n^{1/2}\cdot\left|M_{[t],n}^{IPW}(\beta,p^*)-M_{[t]}^{IPW}(\beta^*,p^*)-M_{[t],n}^{IPW}(\beta^*,p^*)\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|} = o_p(1)$$

because (assuming $d_m = 1$ or applying the following argument element by element) the class of functions $\mathcal{F}_t = \{\mathbf{1}\{\cdot = t\} \cdot m(\cdot;\beta)/p_t^*(\cdot) : |\beta - \beta_t^*| \leq \delta\}$ is Donsker with finite integrable envelope by Assumption 5 (Theorem 2.10.6 of van der Vaart and Wellner (1996)) and L_2 continuous by Assumptions 2 and 5 (compare to Lemma 2.17 in Pakes and Pollard (1989)).

For the second term (A-2) we have

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \cdot \left| M_{[t],n}^{IPW}(\beta, \hat{p}) - M_{[t],n}^{IPW}(\beta, p^*) - \Delta_{[t],n}(\beta, \hat{p} - p^*) \right|}{1 + C \cdot n^{1/2} \cdot |\beta_t - \beta_t^*|}$$
$$\leq C \cdot n^{1/2} \cdot \|\hat{p}_t - p_t^*\|_\infty^2 \cdot \frac{1}{n} \sum_{i=1}^n \frac{D_i(t) \cdot \sup_{|\beta_t - \beta_t^*| \leq \delta_n} |m(Y_i;\beta_t)|}{p_t^*(X_i)} = o_p(1),$$

by condition (2.1) and Assumption 2.

For the third term (A-3) we have

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \cdot \left| M_{[t],n}^{IPW}(\beta^*, \hat{p}) + M_{[t],n}^{IPW}(\beta^*, p^*) - \Delta_{[t],n}(\beta^*, \hat{p} - p^*) \right|}{1 + C \cdot n^{1/2} \cdot |\beta_t - \beta_t^*|}$$
$$\leq C \cdot n^{1/2} \cdot \|\hat{p}_t - p_t^*\|_\infty^2 \cdot \frac{1}{n} \sum_{i=1}^n \frac{D_i(t) \cdot |m(Y_i;\beta_t^*)|}{p_t^*(X_i)} = o_p(1),$$

by condition (2.1) and Assumption 2.

Finally, for the last term (A-4) consider the following decomposition

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \cdot \left| \Delta_{[t],n}(\beta, \hat{p} - p^*) - \Delta_{[t],n}(\beta^*, \hat{p} - p^*) \right|}{1 + C \cdot n^{1/2} \cdot |\beta_t - \beta_t^*|}$$
$$\leq C \cdot \|\hat{p}_t - p_t^*\|_\infty \cdot n^{1/2} \cdot \sup_{|\beta_t - \beta_t^*| \leq \delta_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{D_{t,i} \cdot |m(Y_i;\beta_t) - m(Y_i;\beta_t^*)|}{p_t^*(X_i)} - \mathbb{E}\left[\frac{D_{t,i} \cdot |m(Y_i;\beta_t) - m(Y_i;\beta_t^*)|}{p_t^*(X_i)} \right] \right|$$
$$+ C \cdot \|\hat{p}_t - p_t^*\|_\infty \cdot \sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \cdot \mathbb{E}[|m(Y(t), X; \beta_t) - m(Y(t), X; \beta_t^*)|]}{1 + C \cdot n^{1/2} \cdot |\beta_t - \beta_t^*|}$$
$$= o_p(1),$$

because (assuming $d_m = 1$ or applying the following argument element by element) the class of functions $\mathcal{F}_t = \{\mathbf{1}\{\cdot = t\} \cdot |m(\cdot;\beta) - m(\cdot;\beta_t^*)|/p_t^*(\cdot) : |\beta - \beta^*| \leq \delta\}$ is Donsker with finite integrable envelop by Assumption 5 (Theorem 2.10.6 of van der Vaart and Wellner (1996)) and L_2 continuous by Assumptions 2 and 5.

This establishes condition (iii) of Theorem 3.3 in Pakes and Pollard (1989). ■

Proof of Theorem 5 (ASYMPTOTIC LINEAR REPRESENTATION OF EIFE): the proof of this theorem follows the same logic as the proof of Theorem 4. We apply Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\theta) = A_n M_n^{EIF}(\beta, \hat{p}, \hat{e})$, $G(\theta) = AM^{EIF}(\beta, p^*, e^*)$, and verifying their five sufficient conditions (i)-(v). Like in the proof of Theorem 4, conditions (i), (ii), (iv) and (v) are already satisfied in our case, thus it only remains to establish the stochastic equicontinuity condition (iii), which is implied by the following condition: for all sequences $\delta_n = o(1)$,

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \cdot \left| M_{[t],n}^{EIF}(\beta, \hat{p}, \hat{e}) - M_{[t]}^{EIF}(\beta, p^*, e^*(\beta)) - M_{[t],n}^{EIF}(\beta^*, \hat{p}, \hat{e}) \right|}{1 + C \cdot n^{1/2} \cdot |\beta_t - \beta_t^*|} = o_p(1),$$

for all $t \in \mathcal{T}$. Now, using the results in Theorem 4, it only remains to show that

$$\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{n^{1/2} \cdot \left| \frac{1}{n} \sum_{i=1}^n (\hat{e}_t(X_i;\beta_t) - \hat{e}_t(X_i;\beta_t^*)) \cdot (D_{t,i} - \hat{p}_t(X_i))/\hat{p}_t(X_i) \right|}{1 + C \cdot n^{1/2} \cdot |\beta_t - \beta_t^*|} = o_p(1).$$

Now, for n large enough we obtain

$$\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\frac{n^{1/2}\cdot\left|\frac{1}{n}\sum_{i=1}^n\left(\hat{e}_t\left(X_i;\beta_t\right)-\hat{e}_t\left(X_i;\beta_t^*\right)\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|}$$

$$\leq\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\|e_t-e_t^*\|_\infty\leq\delta_n}\frac{n^{1/2}\cdot\left|\frac{1}{n}\sum_{i=1}^n\left(e_t\left(X_i;\beta\right)-e_t\left(X_i;\beta_t^*\right)\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|}$$

$$\leq\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\|e_t-e_t^*\|_\infty\leq\delta_n}\frac{n^{1/2}\cdot\left|\frac{1}{n}\sum_{i=1}^n\left(\frac{\partial}{\partial\beta}e_t(X_i;\tilde{\beta})-\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\right)\cdot\left(\beta_t-\beta_t^*\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|} \tag{A-5}$$

$$+\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\frac{n^{1/2}\cdot\left|\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\cdot\left(\beta_t-\beta_t^*\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|}, \tag{A-6}$$

for some convex linear combination $\tilde{\beta}$ (between β_t and β_t^*).

Next, for the first term (A-5) we obtain for n large enough,

$$\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\|e_t-e_t^*\|_\infty\leq\delta_n}\frac{n^{1/2}\cdot\left|\frac{1}{n}\sum_{i=1}^n\left(\frac{\partial}{\partial\beta}e_t(X_i;\tilde{\beta})-\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\right)\cdot\left(\beta_t-\beta_t^*\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|}$$

$$\leq C\cdot\sup_{|\beta_t-\beta_t^*|\leq\delta_n,\|e_t-e_t^*\|_\infty\leq\delta_n}\frac{1}{n}\sum_{i=1}^n\left|\frac{\partial}{\partial\beta}e_t\left(X_i;\beta_t\right)-\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta\right)\right|$$

$$+C\cdot\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\left|\frac{1}{n}\sum_{i=1}^n\left(\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t\right)-\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\right)\cdot\frac{D_{t,i}-p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\right|$$

$$+C\cdot\frac{1}{n}\sum_{i=1}^n\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\left|\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t\right)\right|\cdot\left|\frac{D_{t,i}-\hat{p}_t\left(X_i\right)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}-p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\right|$$

$$=o_p\left(1\right),$$

because the first term is $o_p\left(1\right)$ by Assumption 6(b), the second term is $o_p\left(1\right)$ because (assuming $d_m=1$ or applying the argument element by element) the class of functions $\mathcal{F}_t=\{(\partial_\beta e_t^*\left(\cdot;\beta\right)-\partial_\beta e_t^*\left(\cdot;\beta_t^*\right))\cdot(\mathbf{1}\left\{\cdot=t\right\}-p_t^*\left(\cdot\right))/p_t^*\left(\cdot\right):|\beta-\beta_t^*|\leq\delta\}$ is Glivenko-Cantelli for some $\delta>0$ by Assumption 6(a) (van der Vaart and Wellner (2000)), and the third term is $o_p\left(1\right)$ by Assumption 6(a).

The second term (A-6) is

$$\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\frac{n^{1/2}\cdot\left|\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\cdot\left(\beta_t-\beta_t^*\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)/\hat{p}_t\left(X_i\right)\right|}{1+C\cdot n^{1/2}\cdot|\beta_t-\beta_t^*|}$$

$$\leq\left|\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\cdot\frac{D_{t,i}-p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\right|+\frac{1}{n}\sum_{i=1}^n\left|\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t^*\right)\right|\cdot\left|\frac{D_{t,i}-\hat{p}_t\left(X_i\right)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}-p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\right|$$

$$=o_p\left(1\right),$$

by Assumption 6(a).

This establishes condition (iii) of Theorem 3.3 in Pakes and Pollard (1989). ■

Proof of Theorem 6 (CONSISTENT ESTIMATOR OF V^*): first we establish the following two results: for all sequences $\delta_n=o\left(1\right)$ and for all $t\in\mathcal{T}$,

$$\frac{1}{n}\sum_{i=1}^n\left|m(Y_i,T_i,X_i;\hat{\beta},\hat{p})-m\left(Y_i,T_i,X_i;\beta^*,p^*\right)\right|^2=o_p\left(1\right) \tag{A-7}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left|\alpha(T_i,X_i;\hat{p},\hat{e}(\hat{\beta}))-\alpha\left(T_i,X_i;p^*,e^*\left(\beta^*\right)\right)\right|^2=o_p\left(1\right). \tag{A-8}$$

The first result (A-7) follows because for n large enough and for all $t\in\mathcal{T}$ we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_{t,i}\cdot m(Y_i;\hat{\beta}_t)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\right|^2\\&\leq\left\|\hat{p}_t-p_t^*\right\|_\infty^2\cdot\frac{C}{n}\sum_{i=1}^{n}\frac{D_{t,i}\cdot\sup_{|\beta-\beta_t^*|\leq\delta_n}\left|m\left(Y_i;\beta\right)\right|^2}{p_t^*\left(X_i\right)}+\frac{C}{n}\sum_{i=1}^{n}\frac{D_{t,i}}{p_t^*\left(X_i\right)}\cdot\left|m(Y_i;\hat{\beta}_t)-m\left(Y_i;\beta_t^*\right)\right|^2=o_p\left(1\right),\end{aligned}$$

by the same arguments and assumptions used in Theorem 4 and an application of Theorem 2.10.14 of van der Vaart and Wellner (1996). The second result (A-8) follows because for n large enough and for all $t\in\mathcal{T}$ we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{e}_t(X_i;\hat{\beta}_t)}{\hat{p}_t\left(X_i\right)}\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)-\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\cdot\left(D_{t,i}-p_t^*\left(X_i\right)\right)\right|^2\\&\leq\frac{C}{n}\sum_{i=1}^{n}\left|e_t^*(X_i;\hat{\beta}_t)-e_t^*\left(X_i;\beta_t^*\right)\right|^2+o_p\left(1\right)\\&\leq\frac{C}{n}\sum_{i=1}^{n}\sup_{|\beta_t-\beta_t^*|\leq\delta_n}\left|\frac{\partial}{\partial\beta}e_t^*\left(X_i;\beta_t\right)\right|^2\cdot\left|\hat{\beta}_t-\beta_t^*\right|+o_p\left(1\right)=o_p\left(1\right).\end{aligned}$$

Now, define

$$V_n=\frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)',$$

and notice that $V_n-V_*=o_p\left(1\right)$. Next, using Holder's Inequality we have

$$\left|\hat{V}_n-V_*\right|\leq\left|\hat{V}_n-V_n\right|+\left|V_n-V_*\right|\leq R_{1,n}+R_{2,n}+R_{3,n}+R_{4,n}+R_{5,n}+o_p\left(1\right),$$

where

$$\begin{aligned}R_{1,n}&=\frac{1}{n}\sum_{i=1}^{n}\left|m(Y_i,T_i,X_i;\hat{\beta},\hat{p})-m\left(Y_i,T_i;\beta^*,p^*\right)\right|^2,\\R_{2,n}&=\frac{1}{n}\sum_{i=1}^{n}\left|\alpha(T_i,X_i;\hat{p},\hat{e}(\hat{\beta}))-\alpha\left(Y_i,T_i;p^*,e^*\left(\beta^*\right)\right)\right|^2,\\R_{3,n}&=2\cdot R_{1,n}^{1/2}\cdot R_{2,n}^{1/2},\\R_{4,n}&=2\cdot R_{1,n}^{1/2}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left|\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)\right|^2\right)^{1/2},\\R_{5,n}&=2\cdot R_{2,n}^{1/2}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left|\psi\left(Y_i,T_i,\beta^*,p^*,e^*\left(\beta^*\right)\right)\right|^2\right)^{1/2},\end{aligned}$$

and using (A-7) and (A-8) the result follows. ■

Proof of Theorem 7 (CONSISTENT ESTIMATOR OF Γ^*): follows directly by the same arguments given in the proof of Theorem 5. ■

Proof of Theorem 8 (NONPARAMETRIC ESTIMATION): first, for power series and splines, we have $\zeta(K) = K^{\eta}$, with $\eta = 1$ and $\eta = 1/2$, respectively, and using Assumption 7 (which for these cases implies Assumption B-1 in Appendix B), we have $\alpha = s/d_x$ (Newey (1997)). Now Theorem B-1 in Appendix B implies

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

under the assumptions of the theorem and therefore condition (4.1) in Theorem 4 holds.

Next, we consider condition (4.2) in Theorem 4. It is enough to show the result for a typical t-th component of the vector. Thus,

$$\begin{aligned}
&\sqrt{n} \cdot \left| M^{IPW}_{[t],n}(\beta^*_t, \hat{p}_t) - M^{EIF}_{[t],n}(\beta^*_t, p^*_t, e^*_t(\beta^*_t)) \right| \\
&\leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{\hat{p}_t(X_i)} - \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)} + \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)^2} \cdot (\hat{p}_t(X_i) - p^*_t(X_i)) \right\} \right| \qquad \text{(A-9)} \\
&\quad + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)^2} \cdot (\hat{p}_t(X_i) - p^*_t(X_i)) + \frac{e^*_t(X_i; \beta^*_t)}{p^*_t(X_i)} \cdot (\hat{p}_t(X_i) - p^*_t(X_i)) \right\} \right| \qquad \text{(A-10)} \\
&\quad + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\frac{e^*_t(X_i; \beta^*_t)}{p^*_t(X_i)} \cdot (\hat{p}_t(X_i) - p^*_t(X_i)) + \frac{e^*_t(X_i; \beta^*_t)}{p^*_t(X_i)} \cdot (D_{t,i} - p^*_t(X_i)) \right\} \right|. \qquad \text{(A-11)}
\end{aligned}$$

The bound of term (A-9) is given by (for n large enough)

$$\begin{aligned}
&\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{\hat{p}_t(X_i)} - \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)} + \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)^2} \cdot (\hat{p}_t(X_i) - p^*_t(X_i)) \right\} \right| \\
&\leq C \cdot \sqrt{n} \cdot \|\hat{p}_t - p^*_t\|^2_{\infty} \cdot \frac{1}{n} \sum_{i=1}^{n} \frac{D_{t,i} \cdot |m(Y_i; \beta^*_t)|}{p^*_t(X_i)} = \sqrt{n} \cdot O_p\left((K^{\eta} K^{1/2} n^{-1/2} + K^{\eta} K^{1/2} K^{-s/d_x})^2 \right).
\end{aligned}$$

The bound of term (A-10) is given by

$$\begin{aligned}
&\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)^2} \cdot (\hat{p}_t(X_i) - p^*_t(X_i)) + \frac{e^*_t(X_i; \beta^*_t)}{p^*_t(X_i)} \cdot (\hat{p}_t(X_i) - p^*_t(X_i)) \right\} \right| \\
&\leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t(X_i; \beta^*_t)}{p^*_t(X_i)} - \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)^2} \right) \cdot \left(\hat{p}_t(X_i) - p^0_{K,t}(X_i) \right) \right| \qquad \text{(A-12)} \\
&\quad + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t(X_i; \beta^*_t)}{p^*_t(X_i)} - \frac{D_{t,i} \cdot m(Y_i; \beta^*_t)}{p^*_t(X_i)^2} \right) \cdot \left(p^0_{K,t}(X_i) - p^*_t(X_i) \right) \right|, \qquad \text{(A-13)}
\end{aligned}$$

using the notation introduced in Appendix B. Now, to obtain a bound on the term (A-12), first notice that by a second order Taylor expansion and using the results in Appendix B we obtain for some $\tilde{\gamma}_K$ such that $|\tilde{\gamma}_K - \gamma^0_K| \leq |\hat{\gamma}_K - \gamma^0_K|$ and n large enough,

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

which implies that

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\right)\cdot\left(\hat{p}_t\left(X_i\right)-p_{K,t}^0\left(X_i\right)\right)\right| \\
&\leq\left|\hat{\gamma}_K-\gamma_K^0\right|\cdot\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\right)\cdot\left[\dot{\mathbf{L}}_t\left(g_{-0}\left(X_i,\gamma_K^0\right)\right)\otimes R_K\left(X_i\right)'\right]\right| \\
&\qquad+\left|\hat{\gamma}_K-\gamma_K^0\right|^2\cdot\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\right)\cdot\left[\mathbf{I}_J\otimes R_K\left(X_i\right)R_K\left(X_i\right)'\right]\right| \\
&=O_p\left(K^{1/2}n^{-1/2}+K^{1/2}K^{-s/d_x}\right)\cdot O\left(K^{1/2}\right),
\end{aligned}$$

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

Now, for the term (A-13) we have

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\right)\cdot\left(p_{K,t}^0\left(X_i\right)-p_t^*\left(X_i\right)\right)\right|=O_p\left(K^{-s/d_x}\right)=o_p\left(1\right).$$

Finally, the bound of term (A-11) is given by

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\cdot\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)+\frac{e_t\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\cdot\left(D_{t,i}-p_t^*\left(X_i\right)\right)\right\} \\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-R_K\left(X_i\right)'\theta\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right),
\end{aligned}$$

using the first order condition for MLSE, which implies that $\sum_{i=1}^{n}(D_{t,i}-\hat{p}_t\left(X_i\right))\cdot R_K\left(X_i\right)=\mathbf{0}$, and where $\theta\in\mathbb{R}^K$ is any vector. Now, by choosing θ appropriately, we conclude for n large enough that

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}-R_K\left(X_i\right)'\theta\right)\cdot\left(D_{t,i}-\hat{p}_t\left(X_i\right)\right)\right| \\
&\leq C\cdot\sqrt{n}\cdot\sup_{x\in\mathcal{X}}\left|\frac{e_t^*\left(x;\beta_t^*\right)}{p_t^*\left(x\right)}-R_K\left(x\right)'\theta\right|\cdot\sup_{x\in\mathcal{X}}\left|D_{t,i}-\hat{p}_t\left(x\right)\right| \\
&=\sqrt{n}\cdot O\left(K^{-s/d_x}\right)\cdot\left(\sup_{x\in\mathcal{X}}\left|D_n\left(t\right)-p_t^*\left(x\right)\right|+\sup_{x\in\mathcal{X}}\left|p_t^*\left(x\right)-\hat{p}_t\left(x\right)\right|\right) \\
&=\sqrt{n}\cdot O\left(K^{-s/d_x}\right)\cdot O_p\left(K^{\eta}K^{1/2}n^{-1/2}+K^{\eta}K^{1/2}K^{-s/d_x}\right).
\end{aligned}$$

Using the bounds derived and under the assumptions of Theorem 8, we obtain

$$\left|M_n^{IPW}\left(\beta^*,\hat{p}\right)-M_n^{EIF}\left(\beta^*,p^*,e^*\left(\beta^*\right)\right)\right|=o_p\left(n^{-1/2}\right),$$

which verifies condition (4.2) in Theorem 4 as desired.

Next, consider Theorem 5. Conditions (5.1) and (5.2) follow directly from the previous calculations and the first part of Proposition A1 in Chen, Hong, and Tamer (2005), respectively. It remains only to show condition (5.3) in Theorem 5. From Newey (1997) it follows immediately that

$$n^{1/4}\cdot\sup_{x\in\mathcal{X}}\left|\hat{e}\left(x;\beta^*\right)-e^*\left(x;\beta^*\right)\right|=n^{1/4}\cdot O_p\left(K^{\eta}K^{1/2}n^{-1/2}+K^{\eta}K^{-s/r}\right)=o_p\left(1\right).$$

Now, to establish the final condition is enough to show the result for the typical t-th component. From the previous calculations we have

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left\{\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{\hat{p}_t\left(X_i\right)}-\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\right\}=-\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\cdot m\left(Y_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\cdot\left(\hat{p}_t\left(X_i\right)-p_t^*\left(X_i\right)\right)+o_p\left(1\right),$$

and using the identity

$$\frac{\hat{a}}{\hat{b}} = \frac{a}{b} + \frac{1}{b}(\hat{a}-a) - \frac{a}{b^2}\left(\hat{b}-b\right) + \frac{a}{b^2\widehat{b}}\left(\hat{b}-b\right)^2 - \frac{1}{b\hat{b}}(\hat{a}-a)\left(\hat{b}-b\right)$$

we also obtain

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left\{\frac{D_{t,i}\cdot\hat{e}_t\left(X_i;\beta_t^*\right)}{\hat{p}_t\left(X_i\right)} - \frac{D_{t,i}\cdot e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)}\right\}\\ &= \frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\cdot\left(\hat{e}_t\left(X_i;\beta_t^*\right) - e_t^*\left(X_i;\beta_t^*\right)\right)}{p_t^*\left(X_i\right)} - \frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\cdot e_t^*\left(X_i;\beta_t^*\right)}{p_t^*\left(X_i\right)^2}\cdot\left(\hat{p}_t\left(X_i\right) - p_t^*\left(X_i\right)\right) + o_p\left(1\right).\end{aligned}$$

Putting these results together, we see that

$$\begin{aligned}&\sqrt{n}\cdot\left|M_{[t],n}^{EIF}\left(\beta_t^*,\hat{p}_t,\hat{e}_t\left(\beta_t^*\right)\right) - M_{[t],n}^{EIF}\left(\beta_t^*,p_t^*,e_t^*\left(\beta_t^*\right)\right)\right|\\ &\leq \left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\cdot\left(m\left(Y_i;\beta_t^*\right) - e_t^*\left(X_i;\beta_t^*\right)\right)}{p_t^*\left(X_i\right)^2}\cdot\left(\hat{p}_t\left(X_i\right) - p_t^*\left(X_i\right)\right)\right|\\ &\qquad + \left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i} - p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\cdot\left(\hat{e}_t\left(X_i;\beta_t^*\right) - e_t^*\left(X_i;\beta_t^*\right)\right)\right| + \left|\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left(\hat{e}_t\left(X_i;\beta_t^*\right) - e_t^*\left(X_i;\beta_t^*\right)\right)\right|.\end{aligned}$$

Finally, observe that by the same arguments as those used for term (A-10) above, we obtain

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i}\cdot\left(m\left(Y_i;\beta_t^*\right) - e_t^*\left(X_i;\beta_t^*\right)\right)}{p_t^*\left(X_i\right)^2}\cdot\left(\hat{p}_t\left(X_i\right) - p_t^*\left(X_i\right)\right) = o_p\left(1\right),$$

and by analogous arguments, but for the case of series (linear sieves) we may verify that

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\frac{D_{t,i} - p_t^*\left(X_i\right)}{p_t^*\left(X_i\right)}\cdot\left(\hat{e}_t\left(X_i;\beta_t^*\right) - e_t^*\left(X_i;\beta_t^*\right)\right) = o_p\left(1\right),$$

and

$$\frac{1}{\sqrt{n}}\sum\nolimits_{i=1}^{n}\left(\hat{e}_t\left(X_i;\beta_t^*\right) - e_t^*\left(X_i;\beta_t^*\right)\right) = o_p\left(1\right),$$

under the assumptions of this theorem. Therefore we conclude that

$$\sqrt{n}\cdot\left|M_n^{EIF}\left(\beta^*,\hat{p},\hat{e}\left(\beta^*\right)\right) - M_n^{EIF}\left(\beta^*,p^*,e^*\left(\beta^*\right)\right)\right| = o_p\left(1\right),$$

which gives condition (5.3) in Theorem 5 as needed. ■

APPENDIX B. MULTINOMIAL LOGISTIC SERIES ESTIMATOR

In this appendix we derive uniform rates of convergence for the non-linear sieve estimator proposed for the estimation of the GPS. The results presented here generalize those in Hirano, Imbens, and Ridder (2003) by allowing for arbitrary number of outcomes, arbitrary choice of approximating basis, and less stringent requirements in terms of smoothness of the underlying conditional expectation.

We begin by introducing some normalizations and notation. Under some conditions imposed below and by choosing an appropriate non-singular linear transformation we can assumed without loss of generality that $\mathbb{E}\left[R_K\left(X\right)R_K\left(X\right)'\right] = \mathbf{I}_K$, where $\mathbf{I}_K$ is the $(K\times K)$ identity matrix (see Newey (1997) for details). Let $\zeta\left(K\right) = \sup_{x\in\mathcal{X}}\left|R_K\left(x\right)\right|$, 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 may be relaxed at the expense of only additional notation. To deal with all the relevant probabilities simultaneously we define $p_{-0}\left(X\right) = \left(p_1\left(X\right),\cdots,p_J\left(X\right)\right)' \in \mathbb{R}^J$, $\gamma_{-0,K} = \left(\gamma_{K,1}',\cdots,\gamma_{K,J}'\right)' \in \mathbb{R}^{JK}$, and $g_{-0}\left(X,\gamma_K\right) = \left[R_K\left(X\right)'\gamma_{K,1},\cdots,R_K\left(X\right)'\gamma_{K,J}\right]' \in \mathbb{R}^J$. Recall that $p_0^*\left(X\right) = 1 - \sum_{j=1}^{J}p_j^*\left(X\right)$.

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

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

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

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

and observe that $\sup_z |\dot{L}_t(z)| < C$ since $|L_t(z) \cdot L_j(z)| < 1$ and $L_t(z) \cdot (1 - L_t(z)) < 1/4$. Also define the vector-valued functions $\mathbf{L}(z) = [L_1(z), \cdots, L_J(z)]'$ and $\mathbf{L}^{-1}(z) = [L_1^{-1}(z), \ \cdots, \ L_J^{-1}(z)]'$ and observe that the function $\mathbf{L}(\cdot)$ is differentiable with gradient (matrix) $\dot{\mathbf{L}}(z) = [\dot{L}_1(z), \ \cdots, \ \dot{L}_J(z)] \in \mathbb{R}^{J \times J}$ and notice that $\sup_z |\dot{\mathbf{L}}(z)| < C$, for some constant C that only depends on J. With this notation, we obtain $p\left(X; \gamma_{t,K}\right) = L_t\left(g_{-0}\left(X, \gamma_K\right)\right)$ for $t \in \mathcal{T}$ (recall $\gamma_{K,0} = \mathbf{0}_K$ for identification purposes).

The multinomial logistic log-likelihood is given by

$$\ell_n(\gamma_K) = \sum_{i=1}^n \sum_{t=0}^J D_{t,i} \cdot \log\left(L_t\left(g_{-0}\left(X_i, \gamma_K\right)\right)\right),$$

with solution $\hat{\gamma}_K = \arg\max_{\gamma_K} \ell_n(\gamma_K)$ and estimated probabilities given by $\hat{p}_t(X) = L_t\left(g_{-0}\left(X_i, \hat{\gamma}_K\right)\right)$ for all $t \in \mathcal{T}$. Verify that

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{K,t}} \ell_n(\gamma_K) &= \sum_{i=1}^n \left[D_{t,i} - L_t\left(g_{-0}\left(X_i, \gamma_K\right)\right)\right] \cdot R_K(X_i), \\
\frac{\partial^2}{\partial \gamma_{K,t} \partial \gamma_{K,l}'} \ell_n(\gamma_K) &= -\sum_{i=1}^n L_l\left(g_{-0}\left(X_i, \gamma_K\right)\right) \cdot \left[\mathbf{1}\{t = l\} - L_t\left(g_{-0}\left(X_i, \gamma_K\right)\right)\right] \cdot R_K(X_i) R_K(X_i)',
\end{aligned}$$

for $t = 1, 2, \cdots, J$, $l = 1, 2, \cdots, J$, and in matrix notation we have

$$\begin{aligned}
\frac{\partial}{\partial \gamma_K} \ell_n(\gamma_K) &= \sum_{i=1}^n \left[\mathbf{D}_i - \mathbf{L}\left(g_{-0}\left(X_i, \gamma_K\right)\right)\right] \otimes R_K(X_i), \\
\frac{\partial^2}{\partial \gamma_K \partial \gamma_K'} \ell_n(\gamma_K) &= -\sum_{i=1}^n \mathbf{H}\left(X_i, \gamma_K\right) \otimes R_K(X_i) R_K(X_i)',
\end{aligned}$$

where $\mathbf{D}_i = (D_{1,i}, D_{2,i} \cdots, D_{J,i})'$ and $\mathbf{H}\left(X_i, \gamma_K\right) = \operatorname{diag}\left(\mathbf{L}\left(g_{-0}\left(X_i, \gamma_K\right)\right)\right) - \mathbf{L}\left(g_{-0}\left(X_i, \gamma_K\right)\right) \mathbf{L}\left(g_{-0}\left(X_i, \gamma_K\right)\right)'$.

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

Assumption B-1. (a) The smallest eigenvalue of $\mathbb{E}\left[R_K(X) R_K(X)'\right]$ is bounded away from zero uniformly in K; (b) there is a sequence of constants $\zeta(K)$ satisfying $\sup_{x \in \mathcal{X}} |R_K(x)| \leq \zeta(K)$, for $K = K(n) \to \infty$ and $\zeta(K) K^{1/2} n^{-1/2} \to 0$, as $n \to \infty$; and (c) for all $t \in \mathcal{T}$ there exists $\gamma_{t,K}^0 \in \mathbb{R}^K$ and $\alpha > 0$ such that

$$\sup_{x \in \mathcal{X}} \left| \log\left(\frac{p_t^*(x)}{p_0^*(x)}\right) - R_K(x)' \gamma_{t,K}^0 \right| = O\left(K^{-\alpha}\right),$$

and $\zeta(K) K^{1/2} K^{-\alpha} \to 0$.

Assumption B-1 is automatically satisfied in the case of power series or splines if the GPS is smooth enough. Parts (a) and (b) are standard in the literature (Newey (1997)), while Part (c) is slightly stronger than its counterpart for linear series because it imposes a lower bound in $\alpha > 0$. Part (c) guarantees the existence of an approximating sequence that can approximate the function uniformly well. For notational simplicity, we denote such sequence by $p_{t,K}^0(X) = L_t\left(g_{-0}\left(X, \gamma_K^0\right)\right)$, for all $t \in \mathcal{T}$, and define $p_K^0 = [p_{0,K}^0, \cdots, p_{J,K}^0]'$.

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

Theorem B-1. (UNIFORM RATE OF CONVERGENCE OF MLSE) If Assumptions 1(b) and B-1 hold, then

(i) $\left\|p_K^0 - p^*\right\|_\infty = O\left(K^{-\alpha}\right)$,

(ii) $\left|\hat{\gamma}_K - \gamma_K^0\right| = O_p\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right)$,

and consequently $\left\|\hat{p} - p^*\right\|_\infty = O_p(\zeta\left(K\right)K^{1/2}n^{-1/2} + \zeta\left(K\right)K^{1/2}K^{-\alpha})$.

Proof of Theorem B-1 (UNIFORM RATE OF CONVERGENCE OF MLSE):

First, Assumption B-1(c) implies that $\sup_{x\in\mathcal{X}}\left|\mathbf{L}^{-1}\left(p_{-0}^*\left(x\right)\right) - g_{-0}\left(x,\gamma_K^0\right)\right| = O\left(K^{-\alpha}\right)$. Since the mapping $\mathbf{L}\left(\cdot\right)$ is differentiable with $\sup_z\left|\dot{\mathbf{L}}\left(z\right)\right| < C$, an application of the mean value theorem gives

$$\sup_{x\in\mathcal{X}}\left|p_{-0}^*\left(x\right) - \mathbf{L}\left(g_{-0}\left(x,\gamma_K^0\right)\right)\right| \leq C\cdot\sup_{x\in\mathcal{X}}\left|\mathbf{L}^{-1}\left(p_{-0}^*\left(x\right)\right) - g_{-0}\left(x,\gamma_K^0\right)\right|,$$

and since $p_0^*\left(x\right) = 1 - \sum_{j=1}^J p_j^*\left(x\right)$ and $L_0\left(g_{-0}\left(x,\gamma_K^0\right)\right) = 1 - \sum_{j=1}^J L_j\left(g_{-0}\left(x,\gamma_K^0\right)\right)$ part (i) follows directly.

For part (ii), first recall that $L_t\left(g_{-0}\left(x,\gamma\right)\right) > 0$, for all $t = 1,2,\cdots,J$, and $\sum_{t=1}^J L_t\left(g_{-0}\left(x,\gamma\right)\right) < 1$. The special structure of the matrix $\mathbf{H}\left(x,\gamma\right)$ and Theorem 1 in Tanabe and Sagae (1992) shows that $\mathbf{H}\left(x,\gamma\right)$ is symmetric positive definite with $0 < \lambda_{\min}\left(\mathbf{H}\left(x,\gamma\right)\right) \leq \lambda_{\max}\left(\mathbf{H}\left(x,\gamma\right)\right) < 1$, which implies that $\mathbf{H}\left(x,\gamma\right) \geq \lambda_{\min}\left(\mathbf{H}\left(x,\gamma\right)\right)\cdot\mathbf{I}_J$ and $\lambda_{\min}\left(\mathbf{H}\left(x,\gamma\right)\right) \geq \det\left(\mathbf{H}\left(x,\gamma\right)\right)$. These results and the exact Cholesky decomposition of $\mathbf{H}\left(x,\gamma\right)$ gives

$$\inf_{x\in\mathcal{X}}\mathbf{H}\left(x,\gamma\right) \geq \inf_{x\in\mathcal{X}}\prod_{t=0}^J L_t\left(g_{-0}\left(x,\gamma\right)\right)\cdot\mathbf{I}_J,$$

in a positive semidefinite sense.

Now, let $\hat{\Omega}_K = n^{-1}\sum_{i=1}^n R_K\left(X_i\right)R_K\left(X_i\right)'$, and observe that (Newey (1997)) $|\hat{\Omega}_K - \mathbf{I}_K| = O_p(\zeta\left(K\right)K^{1/2}n^{-1/2})$. Define the event $\mathcal{A}_n = \{\lambda_{\min}(\hat{\Omega}_K) > 1/2\}$ and by Assumption B-1(b) we have $O_p\left(\zeta\left(K\right)K^{1/2}n^{-1/2}\right) = o_p\left(1\right)$, which implies $\mathbb{P}\left[\mathcal{A}_n\right] \to 1$.

Next, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\frac{\partial}{\partial\gamma}\ell_n\left(\gamma_K^0\right)\right|\right] \\
&= \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n\left[\mathbf{D}_i - \mathbf{L}\left(g_{-0}\left(X_i,\gamma_K^0\right)\right)\right]\otimes R_K\left(X_i\right)\right|\right] \\
&\leq \left(\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n\left[\mathbf{D}_i - p_{-0}^*\left(X_i\right)\right]\otimes R_K\left(X_i\right)\right|^2\right]\right)^{1/2} + \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n\left[p_{-0}^*\left(X_i\right) - \mathbf{L}\left(g_{-0}\left(X_i,\gamma_K^0\right)\right)\right]\otimes R_K\left(X_i\right)\right|\right] \\
&\leq C\cdot\left(\frac{1}{n}\cdot\mathbb{E}\left[\left|\left[\mathbf{D}_i - p_{-0}^*\left(X_i\right)\right]\otimes R_K\left(X_i\right)\right|^2\right]\right)^{1/2} + C\cdot\sup_{x\in\mathcal{X}}\left|p_{-0}^*\left(x\right) - \mathbf{L}\left(g_{-0}\left(x,\gamma_K^0\right)\right)\right|\cdot\mathbb{E}\left[\left|R_K\left(X\right)\right|\right] \\
&= O\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right),
\end{aligned}$$

and by Markov's Inequality we conclude

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

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

$$\mathcal{B}_n\left(\varsigma\right) = \left\{\left|\frac{1}{n}\frac{\partial}{\partial\gamma}\ell_n\left(\gamma_K^0\right)\right| < \varsigma\cdot\left(K^{1/2}n^{-1/2} + K^{-\alpha+1/2}\right)\right\}$$

approaches one, i.e., $\mathbb{P}\left[\mathcal{B}_n\left(\varsigma\right)\right] \to 1$.

Let $\delta = \inf_{x\in\mathcal{X}} \prod_{t=0}^{J} L_t\left(g_{-0}\left(x,\gamma_K^0\right)\right)$ and observe that for K large enough $\delta > 0$ by part (i) and the assumption that the true probabilities are strictly between zero and one. Define the sets

$$\Gamma_K^\delta = \left\{\gamma \in \mathbb{R}^{JK} : \inf_{x\in\mathcal{X}} \prod_{t=0}^{J} L_t\left(g_{-0}\left(x,\gamma\right)\right) > \frac{\delta}{2}\right\},$$

and $\Gamma_K^0\left(\varrho\right) = \left\{\gamma \in \mathbb{R}^{JK} : \left|\gamma - \gamma_K^0\right| \leq \varrho \cdot \left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right)\right\}$ for any $\varrho > 0$, and because (for some intermediate point $\tilde{\gamma}_K$),

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

by Assumptions B-1(b) and B-1(c), we conclude that for n for large enough $\Gamma_K^\delta \subset \Gamma_K^0\left(\varrho\right)$.

To finish the argument, choose n large enough so that $\Gamma_K^\delta \subset \Gamma_K^0\left(C\right)$, $\mathbb{P}\left[\mathcal{A}_n\right] \geq 1-\varepsilon/2$ and $\mathbb{P}\left[\mathcal{B}_n\left(\delta C/8\right)\right] \geq 1-\varepsilon/2$, for some $C > 0$. Then for any $\gamma_K \in \Gamma_K^0$ we have

$$\begin{aligned}
-\frac{\partial}{\partial\gamma\partial\gamma'}\ell_n\left(\gamma_K\right) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{H}\left(X_i,\gamma_K\right) \otimes R_K\left(X_i\right) R_K\left(X_i\right)' \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \left[\inf_{x\in\mathcal{X}} \prod_{t=0}^{J} L_t\left(g_{-0}\left(x,\gamma_K\right)\right) \cdot \mathbf{I}_J\right] \otimes R_K\left(X_i\right) R_K\left(X_i\right)' \\
&\geq \frac{\delta}{2} \cdot \left[\mathbf{I}_J \otimes \hat{\Omega}_K\right],
\end{aligned}$$

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 $\gamma_K \in \Gamma_K^0 \setminus \left\{\gamma_K^0\right\}$ we have

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

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 with probability approaching one.

Now the result follows directly. ■

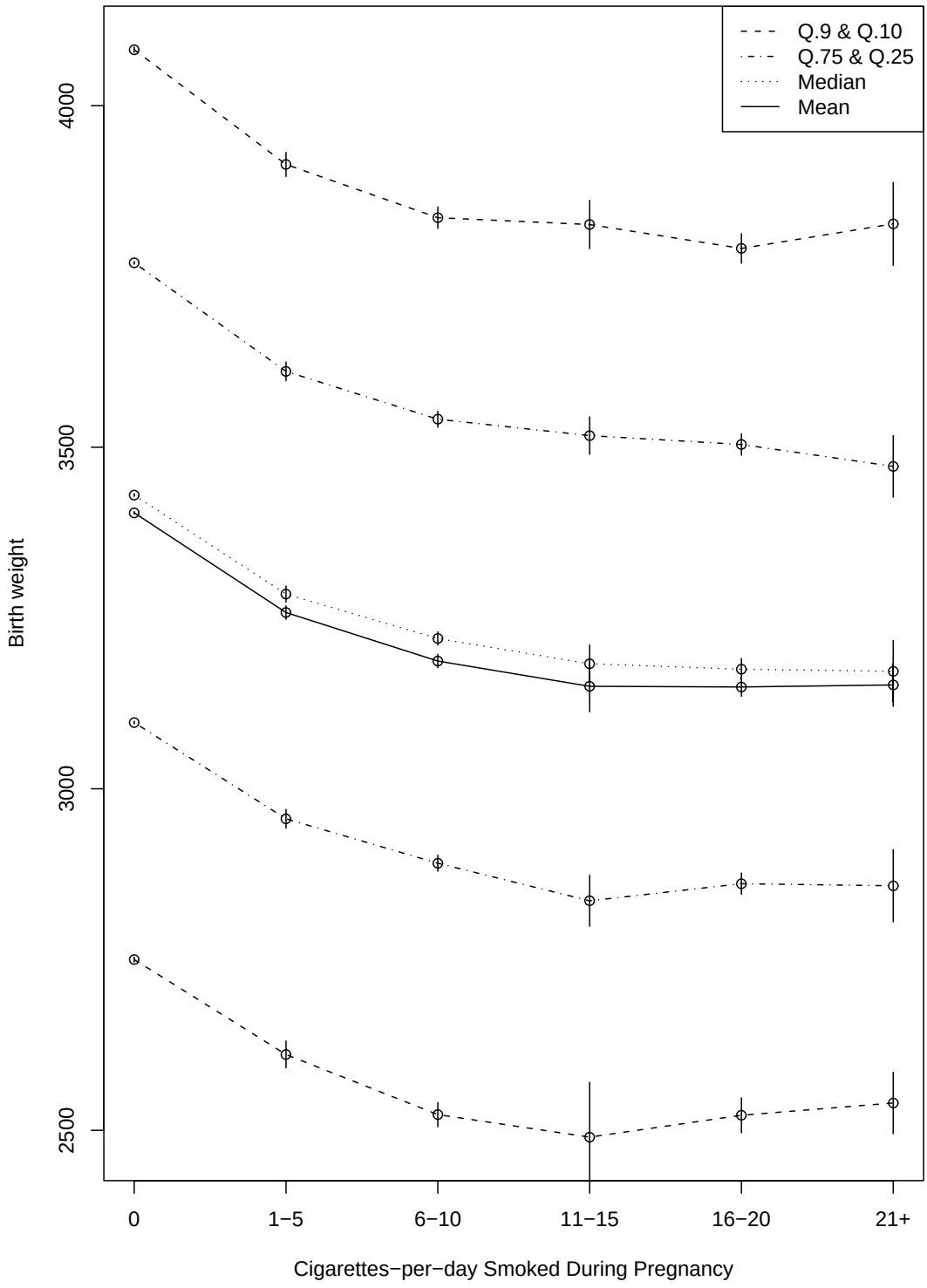

Figure 1: Effect of Maternal Smoking Intensity on Birth Weight (5-cigarette bins)

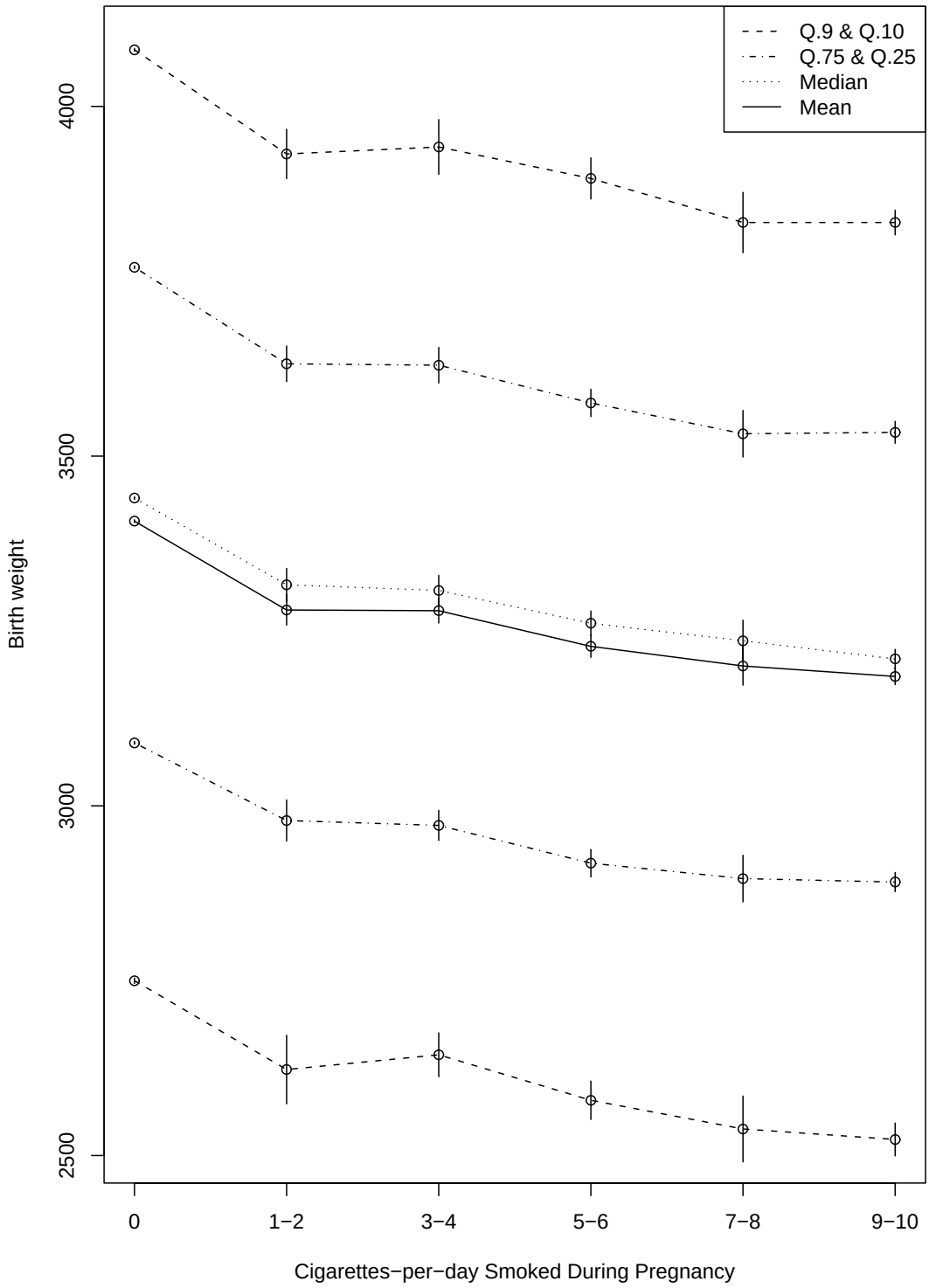

Figure 2: Effect of Maternal Smoking Intensity on Birth Weight (2-cigarette bins)

Table 1: Effect of Maternal Smoking Intensity on Birth Weight

	DRE							**IPWE**							**EIFE**						
	Q.9	Q.75	Q.5	Mean	Q.25	Q.1	SD	Q.9	Q.75	Q.5	Mean	Q.25	Q.1	SD	Q.9	Q.75	Q.5	Mean	Q.25	Q.1	SD
0	4082	3771	3430	3417	3119	2778	576	4082	3770	3430	3404	3097	2750	586	4081	3770	3431	3405	3091	2750	582
	(1)	(1)	(1)	(2)	(1)	(2)	n.a.	(2)	(1)	(1)	(1)	(1)	(2)	(1)	(1)	(1)	(1)	(1)	(1)	(2)	(1)
1-5	3872	3572	3232	3189	2892	2520	601	3914	3611	3285	3258	2956	2611	577	3914	3611	3285	3258	2954	2608	578
	(7)	(5)	(5)	(25)	(5)	(9)	n.a.	(9)	(7)	(6)	(5)	(7)	(10)	(5)	(9)	(7)	(6)	(5)	(7)	(10)	(5)
6-10	3814	3503	3175	3133	2835	2466	593	3836	3541	3220	3187	2891	2523	568	3836	3549	3231	3189	2891	2529	567
	(4)	(3)	(3)	(17)	(4)	(6)	n.a.	(8)	(6)	(5)	(5)	(6)	(9)	(4)	(8)	(6)	(5)	(5)	(6)	(9)	(5)
11-15	3800	3515	3175	3161	2863	2523	565	3826	3517	3183	3150	2836	2490	598	3826	3535	3201	3162	2865	2526	594
	(12)	(9)	(8)	(44)	(10)	(14)	n.a.	(18)	(14)	(14)	(19)	(19)	(41)	(26)	(18)	(13)	(13)	(19)	(19)	(35)	(26)
16-20	3780	3487	3153	3119	2807	2460	581	3791	3504	3175	3149	2861	2522	565	3805	3503	3183	3156	2869	2540	560
	(5)	(4)	(4)	(21)	(5)	(8)	n.a.	(11)	(8)	(8)	(7)	(8)	(13)	(8)	(11)	(8)	(8)	(7)	(8)	(12)	(8)
21+	3799	3459	3147	3105	2780	2438	585	3827	3472	3172	3152	2858	2540	560	3845	3487	3175	3165	2877	2549	546
	(12)	(9)	(9)	(46)	(10)	(15)	n.a.	(31)	(23)	(23)	(16)	(27)	(23)	(16)	(31)	(24)	(23)	(16)	(25)	(22)	(17)

Notes: (i) DRE = Dummy Regression Estimator, IPWE = Inverse Probability Weighting Estimator, EIFE = Eficient Influence Function Estimator.
(ii) Q.9, Q.75, Q.5, Q.25 and Q.1 are the 90%, 75%, 50%, 25% and 10% quantiles, respectively, and SD is the standard deviation.
(iii) standard errors in parentheses.

Table 2: Hypothesis Tests for Pairwise Differences and Difference-in-Differences Effects

	T1-T0	T2-T0	T3-T0	T4-T0	T5-T0	T2-T1	T3-T1	T4-T1	T5-T1	T3-T2	T4-T2	T5-T2	T4-T3	T5-T3	T5-T4
T1-T0	-146*					75*	38	37*	40*	109*	108*	111*	145*	148*	149*
T2-T0		-217*				146*	109*	108*	111*	180*	179*	182*	216*	219*	220*
T3-T0			-254*			183*	146*	145*	148*	217*	216*	219*	253*	256*	257*
T4-T0				-255*		184*	147*	146*	149*	218*	217*	220*	254*	257*	258*
T5-T0					-252*	181*	144*	143*	146*	215*	214*	217*	251*	254*	255*
T2-T1						-71*				34	33*	36	70*	73*	74*
T3-T1							-108*			71*	70*	73*	107*	110*	111*
T4-T1								-109*		72*	71*	74*	108*	111*	112*
T5-T1									-106*	69*	68*	71*	105*	108*	109*
T3-T2										-37			36	39	40
T4-T2											-38*		37	40	41
T5-T2												-35*	34	37	38*
T4-T3													-1		4
T5-T3														2	1
T5-T4															3

Notes: (i) treatments T0, T1, T2, T3, T4 and T5 are 0, 1-5, 6-10, 11-15, 16-20 and 21+ cigarettes-per-day smoked, respectively.
(ii) pairwise differences are reported on the diagonal, and difference-in-differences are reported outside the diagonal.
(iii) in all cases the null hypothesis is zero differential effect; (iv) * significant at 5%.

Table 3: Joint Hypotheses Tests (IPWE)

Joint Null Hypotheses	Number of Restrictions	Wald Test Statistic	p-value
Equal treatment effects (mean, quantiles, spread) for (11-15,16-20,21+)	14	16.60	0.2781
Equal treatment effects (mean, quantiles, spread) for (6-10,11-15,16-20,21+)	21	55.86	0.0001
Equal treatment effects (mean, quantiles, spread) for (1-5,6-10,11-15,16-20,21+)	28	246.88	0.0000
Equal mean and median for each treatment	6	1402.62	0.0000
Equal mean-median difference (MMD) across treatments	5	3.78	0.5809
Equal standard deviation across treatments	5	25.38	0.0001
Equal interquartile range (IQR) across treatments	5	25.32	0.0001
Equal Q.9-Q.1 range (Q.9-Q.1) across treatments	5	21.98	0.0005
Equal MMD, IQR and Q.9-Q.1 across treatments	15	38.59	0.0007

Note: all tests have been computed using the IPWE and its corresponding limiting distribution.