Journal of Econometrics 155 (2010) 138–154

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Efficient semiparametric estimation of multi-valued treatment effects under ignorability☆

Matias D. Cattaneo*

Department of Economics, University of Michigan, United States

ARTICLE INFO

Article history:
Received 22 January 2009
Received in revised form
12 July 2009
Accepted 22 September 2009
Available online 7 October 2009

JEL classification:
C14
C21
C31

Keywords:
Multi-valued treatment effects
Unconfoundedness
Semiparametric efficiency
Efficient estimation

ABSTRACT

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 the treatment assignment is assumed to be ignorable. Two estimators are introduced together with a set of sufficient conditions that ensure their $\sqrt{n}$-consistency, asymptotic normality and efficiency. Under mild assumptions, these conditions are satisfied for the *Marginal Mean Treatment Effect* and the *Marginal Quantile Treatment Effect*, estimands of particular importance for empirical applications. Previous results for average and quantile treatments effects are encompassed by the methods proposed here when the treatment is dichotomous. The results are illustrated by an empirical application studying the effect of maternal smoking intensity during pregnancy on birth weight, and a Monte Carlo experiment.

1. Introduction

A large fraction of the literature on program evaluation focuses on efficient, flexible 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 anti-poverty 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.

In addition, considering multi-valued treatment effects allows for potential efficiency gains in the estimation whenever additional information is available. Restrictions on the treatment effects between and across treatment levels are usually justified by the underlying economic theory (or other sources of knowledge about the data generating process) in specific applications, and are by construction ignored in the binary treatment effect literature. For example, in labor economics it is often assumed that the relationship between log-income and education is linear, leading to a simple restriction between different levels of educational attainment, or, in public finance, different levels of marginal tax rates may have a proportional effect on labor supply whenever the corresponding elasticity is assumed to be constant. In cases such as these, simple restrictions across treatment effects are available that may be exploited to improve efficiency in the estimation of the multi-valued treatment effects.

This paper is concerned with optimal joint inference for a general class of finite multi-valued treatment effects when the

☆ 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, Sebastian Galiani, Bryan Graham, Enrico Moretti, Salvador Navarro, Tom Rothenberg, Paul Ruud, Jeff Smith, Rocio Titiunik, seminar participants at Austin, Berkeley, Brown, BU, Duke, Harvard, Michigan, UPenn, and WashU, and conference participants at the 2009 AEA meeting for valuable comments and suggestions. I also thank three reviewers and an associate editor for their detailed comments and suggestions that improved this paper. Douglas Almond, Ken Chay and David Lee generously provided the data used in the empirical illustration of this paper.

* Corresponding address: Department of Economics, University of Michigan, 238 Lorch Hall, 611 Tappan Street, Ann Arbor, MI 48109-1220, United States. Tel.: +1 734 763 1306; fax: +1 734 764 2769.
E-mail address: cattaneo@umich.edu.

0304-4076/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2009.09.023

treatment assignment is assumed to be ignorable, that is, when treatment is assigned at random conditional on a set of observable characteristics and a common support condition holds. Results available in the literature for ignorable binary treatment effects may be applied to the context of multiple treatments, leading to efficient estimators of one treatment effect at the time. However, an important limitation of these results is that they do not allow for either joint inferences across and between multiple treatment levels, or efficiency gains in the estimation obtained from exploiting over-identification restrictions. The results presented in this paper overcome these limitations by allowing for the joint efficient estimation of multiple treatments.

Two estimation procedures for a population parameter implicitly defined by a possibly over-identified non-smooth collection of moment conditions are proposed, together with 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 the *Marginal Mean Treatment Effect* and the *Marginal Quantile Treatment Effect*, 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, measures of inequality, differential treatment effects, and heterogeneous treatment effects may be easily constructed by considering different functions of means and quantiles such as pairwise differences, interquantile ranges and incremental ratios. In addition, by allowing for over-identification, the main results of the paper may provide further efficiency gains and cover other situations of interest such as the explicit use of cross-equation restrictions.

A unified framework for the efficient estimation of a large class of multi-valued treatment effects is developed, which not only includes as particular cases important results from the program evaluation literature when the treatment is binary, but also allows for the efficient estimation of other estimands of interest. The theoretical results are developed in the context of a two-step semiparametric GMM model, where the treatment effects are implicitly defined through possibly non-smooth over-identified moment conditions and the first step estimation procedure is fully nonparametric. This general model has the advantage of being flexible and covering typical problems in econometrics, but it requires more technical machinery to handle both the potential lack of smoothness in the moment conditions as well as the preliminary nonparametric estimators. These technical issues are resolved by resorting to empirical process theory that, combined with semiparametric theory in the context of GMM estimation, allows for the semiparametric efficient estimation of a large class of population parameters of interest. Thus, the results are general in that they impose only typical restrictions in the class of moment functions and infinite-dimensional nuisance parameters.

The analysis begins by deriving the Efficient Influence Function (EIF) and Semiparametric Efficiency Bound (SPEB) for the general population parameter of interest using the methodology outlined in Bickel et al. (1993). Based on these results, two estimators of multi-valued treatment effects are introduced and motivated as the solution to a general GMM model, which circumvents the fundamental problem of causal inference by forming sample analogues of two moment conditions that depend only on observed data. For the first estimator, the observed moment condition is obtained by means of an inverse probability weighting scheme based on the Generalized Propensity Score (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 are derived in two basic stages. In the first stage, the consistency, asymptotic normality and efficiency of both estimators are established by means of imposing a set of mild sufficient conditions concerning the underlying moment identification functions, and well-known high-level conditions involving the nonparametric estimators. The former conditions are easily verified in applications, as shown in the examples discussed below, while the latter generally require additional work. For this reason, in the second stage a detailed discussion of the nonparametric estimation of the two nuisance parameters for the particular case of series estimation is provided. Since both nuisance parameters are conditional expectations, results from the nonparametric series estimation literature may be applied directly. However, since the GPS is a conditional probability, a nonparametric estimator is proposed which is based on series estimation and captures the specific features of this nuisance parameter. Using these nonparametric estimators, simple primitive conditions that guarantee the efficient estimation of general multi-valued treatment effects are provided.

Using these results, 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 et al. (2003), and Firpo (2007) may be seen as particular cases of the procedures developed here. Furthermore, the results also allow for the efficient estimation of restricted treatment effects by means of a simple minimum distance estimator based on efficiently estimated, unrestricted treatment effects.

The theoretical results are illustrated by means of an empirical application and a Monte Carlo experiment. The empirical application shows how joint inference using multi-valued treatment effects may be conducted. It is based on the analysis of Almond et al. (2005), who studied the effect of maternal smoking on birth weight, defining maternal smoking as a binary treatment. Exploiting the fact that their database includes the number of cigarettes-per-day smoked by the mother, their analysis is extended to a multi-valued treatment setup in order to study the effect of maternal smoking intensity on birth weight. The new findings suggest the presence of a non-linear negative effect where two thirds of the full impact of smoking on birth weight are due to the first five cigarettes, while the remaining third is explained by the next five cigarettes, with no important effects beyond the tenth cigarette-per-day smoked. Moreover, these effects appear to be additive, shifting the entire distribution of birth weight in parallel along the smoking intensity. To complement the empirical illustration, a Monte Carlo experiment is also reported that shows how the methods proposed here achieve efficiency gains in the estimation of multi-valued treatment effects when over-identification and cross-equation restrictions are available.

This paper contributes to the large portion of the program evaluation literature that focuses on the identification and semiparametric (efficient) estimation of different population parameters of interest using a conditional independence assumption. (Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) provide detailed recent reviews.) Although for concreteness the discussion in this paper uses terminology from the program evaluation literature, the results are also closely related (and contribute) to other literatures in econometrics and statistics that rely on a conditional independence assumption such as the missing data, measurement error and data combination literatures. (A review and discussion of the links between these literatures may be found in Tsiatis (2006),

Bang and Robins (2005), Chen et al. (2005), Chen et al. (2004, 2008), and Wooldridge (2007), among others.)

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 using either nonparametric regression methods (Hahn, 1998; Heckman et al., 1998; Imbens et al., 2006), matching techniques (Abadie and Imbens, 2006), or procedures based on the nonparametric estimation of the propensity score (Hirano et al., 2003). Recently, Firpo (2007) considered a different population parameter by studying the efficient estimation of Quantile Treatment Effects (QTEs) for dichotomous treatment assignments using a nonparametrically estimated propensity score. In the closely related context of missing data, Robins et al. (1994), Robins and Rotnitzky (1995) and Robins et al. (1995) develop a general (locally) efficient estimation strategy for models where the missingness indicator is binary that involves the parametric estimation of both a regression function and the propensity score, while Chen et al. (2005) and Chen et al. (2008) study efficient GMM estimation in the context of (non-classical) measurement error models when two samples are available (i.e., with a binary missingness indicator).

Considerably less work is available in these literatures for the case of multiple treatment assignments. In the context of program evaluation under ignorability, Imbens (2000) derives a generalization of the propensity score and shows that the results of Rosenbaum and Rubin (1983) continue to hold when the treatment is multi-valued. (See also Hirano and Imbens (2004) and Imai and van Dyk (2004) for extensions of this idea.) 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. Heckman and Vytlacil (2007) and Imbens and Wooldridge (2009) provide recent reviews of the results available in this literature, while Bang and Robins (2005) discuss similar results in the context of missing data. The results presented here include and extend those available in these literatures by considering the joint semiparametric efficient estimation of a large class of multi-valued treatment effects, which allows one to conduct joint inference within and across treatment effects and to obtain efficiency gains whenever over-identification restrictions are available.

The rest of the paper is organized as follows. Section 2 introduces the model and discusses identification. Section 3 includes the semiparametric efficiency calculations and presents the EIF and SPEB. Section 4 describes the two estimators proposed. Section 5 presents the large sample results. Section 6 discusses two leading examples and presents the empirical illustration and Monte Carlo study. Section 7 concludes. All proofs are collected in Appendices A and B.

2. Statistical model and identification

This section describes the multi-valued treatment effect model, discusses identification of the general population parameter of interest, and introduces the notation.

2.1. The model

A finite collection of multiple treatment status (categorical or ordinal) indexed by $t \in \mathcal{T}$ is assumed where, without loss of generality, $\mathcal{T} = \{0, 1, 2, \ldots, J\}$ with $J \in \mathbb{N}$ fixed. The random variables $Y(0), Y(1), \ldots, Y(J)$, with $Y(t) \in \mathcal{Y} \subset \mathbb{R}$ for all $t \in \mathcal{T}$, denote the collection of potential outcomes under each treatment level, 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 $Y = \sum_{t\in\mathcal{T}} D_t Y(t)$, where $D_t = \mathbf{1}\{T = t\}$ for all $t \in \mathcal{T}$ and $\mathbf{1}\{\cdot\}$ is the indicator function. There exists a random vector $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$, $d_x \in \mathbb{N}$, which is always observed. It is assumed that a random sample of size n from (Y, T, X) is available, denoted by (Y_i, T_i, X_i), $i = 1, 2, \ldots, n$, and with $D_{t,i} = \mathbf{1}\{T_i = t\}$ for $t \in \mathcal{T}$. This leads to a cross-sectional random sampling scheme where only the potential outcome corresponding to $T = t$ is observed, which implies that effectively the sample comes 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 of functionals of the latter distribution. In this model the fundamental problem of causal inference is exacerbated since only one of the $J+1$ potential outcomes is observed for each unit (Holland, 1986).

The population parameter of interest is the vector $\beta^* = [\beta_0^{*\prime}, \beta_1^{*\prime}, \ldots, \beta_J^{*\prime}]'$, where $\beta_t^* \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$ for $t \in \mathcal{T}$ and $d_\beta \in \mathbb{N}$. This parameter is assumed to uniquely solve a collection of $J+1$ (possibly over-identified, non-smooth) identifying moment conditions denoted by $m : \mathcal{Y} \times \mathcal{B} \to \mathbb{R}^{d_m}$ with $d_m \geq d_\beta$.[1]

Assumption 1. For all $t \in \mathcal{T}$, β^* satisfies $\mathbb{E}[m(Y(t); \beta_t)] = 0$ if and only if $\beta_t = \beta_t^*$.

Assumption 1 imposes a conventional high-level identification condition for GMM estimation. As discussed in more detail below, this model covers several examples of particular relevance for applied work. For instance, the mean response of some outcome of interest to each treatment level may be recovered by considering the collection of moment conditions $m(Y(t); \mu_t) = Y(t) - \mu_t$, $t \in \mathcal{T}$, which leads to $\mu_t^* = \mathbb{E}[Y(t)]$. This estimand, labeled the *Marginal Mean Treatment Effect* (MMTE), may be used to generalize the idea of average treatment effect and is sometimes called the Dose–Response Function in the statistical literature, and the Average Structural Function in the econometrics literature. Similarly, it is also possible to focus on the effect of multiple treatments for the τ-th quantile of the underlying potential outcome distributions by employing the collection of moment conditions $m(Y(t); q_t(\tau)) = \mathbf{1}\{Y(t) \leq q_t(\tau)\} - \tau$, $t \in \mathcal{T}$, 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 $Y(t)$. This alternative estimand, labeled the *Marginal Quantile Treatment Effect* (MQTE), may be used to capture and extend the idea of quantile treatment effects.

2.2. Identification

The population parameter of interest defined in Assumption 1 is not identifiable from the data on (Y, T, X). Following the program evaluation literature, this paper considers a "selection on observables" assumption to achieve identification:

Assumption 2. For all $t \in \mathcal{T}$: (a) $Y(t) \perp\!\!\!\perp D_t | X$; and (b) $0 < p_{\min} \leq \mathbb{P}[T = t|X]$.

[1] This model corresponds to a specialized case of a general GMM model with multi-level missing data. The results presented here apply without major changes to a more general model where the potential outcomes may be multi-dimensional, $Y(t)$ may include some components of X, and the moment conditions may depend on $t \in \mathcal{T}$. The discussion is based on the multi-valued treatment effect model only for simplicity.

In the context of multi-valued treatment effects, Assumption 2 is sometimes referred to as Ignorability and the conditional probabilities $p_t^*(X) \equiv \mathbb{P}[T = t|X]$, $t \in \mathcal{T}$, are known as the Generalized Propensity Score (Imbens, 2000).

Part (a) of Assumption 2 is usually called Unconfoundedness (or Missing at Random) and it ensures that the distribution of each potential outcome and the treatment level indicator are conditionally independent. 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. Part (b) of Assumption 2 is important for identification (in the absence of functional form restrictions), and is also a necessary condition for finiteness of the semiparametric efficiency bound for regular estimators of β^*.

Assumptions 1 and 2 provide identification of β^* because, for example, it is easily verified that

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

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

and

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

leading to three moment conditions based solely on observed random variables.

These assumptions lead to a collection of alternative, asymptotically equivalent efficient estimators. This paper studies two of these estimators for the case of multi-valued treatment effects. The first estimator is based on Eq. (2), while the second estimator is based on a different moment condition that may be constructed as a linear combination of Eqs. (1)–(3). The first estimator is motivated by its simplicity, while the second estimator is motivated from the semiparametric efficiency calculations, as discussed further below. In the special case of binary treatment effects, these estimators are asymptotically equivalent to those available in the literature.

2.3. Notation

Two important functions are the $J+1$ vector-valued function representing the GPS, denoted by $p^*(\cdot) = [p_0^*(\cdot), \ldots, p_J^*(\cdot)]'$, and the $(J+1)\,d_m$ vector-valued function of conditional expectations denoted by $\mathrm{e}^*(\cdot;\beta) = [e_0^*(\cdot;\beta_0)', \cdots, e_J^*(\cdot;\beta_J)']'$, where $e_t^*(X;\beta_t) = \mathbb{E}[m(Y(t);\beta_t)|X]$. It is assumed 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 two subspaces of (smooth) functions on $\mathcal{X}$, endowed with the supremum norm. These classes of functions will be further restricted later in the paper to enable the nonparametric estimation of these nuisance parameters. For simplicity, the arguments of the functions considered are dropped whenever they are clear from the context. In addition, $|\cdot|$ denotes the matrix norm given by $|A| = \sqrt{\mathrm{trace}(A'A)}$ for any matrix A, while $\|\cdot\|_\infty$ denotes the sup-norm in all arguments for functions. In particular, for all $t \in \mathcal{T}$, $\|p_t\|_\infty = \sup_{x\in\mathcal{X}} |p_t(x)|$ for any $p_t \in \mathcal{P}$, $\|e_t(\beta_t)\|_\infty = \sup_{x\in\mathcal{X}} |e_t(x;\beta_t)|$ and $\|e_t\|_\infty = \sup_{\beta_t\in\mathcal{B}, x\in\mathcal{X}} |e_t(x;\beta_t)|$ for any $e_t(\beta_t) \in \mathcal{E}$, and similarly for the vector-valued functions p and e. Finally, define

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

and

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

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

This section provides semiparametric efficiency calculations essential for the construction of efficient estimators of β^*. (See Bickel et al. (1993), Newey (1990) and van der Vaart (1998) for surveys.) The semiparametric efficiency 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, it provides a way of constructing efficient estimators using the EIF 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 model and then verifying that the proposed estimator admits an asymptotic linear representation based on this function.

Several semiparametric efficiency calculations are available when some form of Assumption 2 is imposed. In the context of program evaluation with binary treatments, efficient influence functions and efficiency bounds have been computed by Hahn (1998), Hirano et al. (2003) and Firpo (2007) for average and quantile treatment effects. In models of missing data, Robins et al. (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. Chen et al. (2004, 2008) provide semiparametric efficiency calculations for GMM models when the treatment or missingness indicator is binary. Section 5.5 discusses how the efficiency bounds for different population parameters, including those from the binary treatment program evaluation literature, may be recovered from the calculations presented here.

Assumption 3. For all $t \in \mathcal{T}$: (a) $\mathbb{E}[|m(Y(t);\beta_t)|^2] < \infty$ and $\mathbb{E}[m(Y(t);\beta_t)]$ is differentiable in $\beta_t \in \mathcal{B}$ at β_t^*; and (b) $\mathrm{rank}(\Gamma_*) = (J+1)\,d_\beta$, 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

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

The main role of Assumption 3 (together with part (b) of Assumption 2) is to ensure that the bound is finite, while the full column rank assumption on the gradient matrix Γ_* ensures a local identification condition. 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 2 and 3 hold. Then the EIF for any regular estimator of β^* is given by*

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

where $V_* = \mathbb{V}[\psi(Y,T,X;\beta^*,p^*,\mathrm{e}^*(\beta^*))]$ *and* $\psi(y,t,x;\beta^*,p^*,\mathrm{e}^*(\beta^*)) = m(y,t,x;\beta^*,p^*) - \alpha(t,x;\beta^*,p^*,\mathrm{e}^*(\beta^*))$. *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 of 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 (2).

To provide additional intuition on the structure of the SPEB, note that

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

Using this decomposition, it is seen that the results in Theorem 1 include the SPEB presented in Theorem 1 of Chen et al. (2004, 2008) in the context of measurement error with "verify-in-sample" auxiliary data. In addition, by proceeding as in Hahn (1998), Hirano et al. (2003) or Chen et al. (2004, 2008), it is possible to verify that (i) the GPS is ancillary for the estimation of β^*, and (ii) if the distribution of X is known or correctly specified the SPEB is reduced.

The calculations presented here explicitly allow for the components $\beta_0^*, \ldots, \beta_J^*$ of the population parameter β^* to be different. As illustrated by the Monte Carlo experiment, when further information about the components of β^* is available, this may be incorporated in the model to obtain restricted treatment effects and a reduction in the SPEB.

One important simplification in Theorem 1 is achieved in the 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* $\Psi(y,t,x;\beta^*,p^*,\mathrm{e}^*(\beta^*)) = \Gamma_*^{-1}\psi(y,t,x;\beta^*,p^*,\mathrm{e}^*(\beta^*))$. *Consequently, the SPEB for any regular estimator of* β^* *is given by* $V^* = \Gamma_*^{-1} V_* \Gamma_*^{\prime -1}$.

In the just-identified case $\Gamma_* = \mathrm{diag}\left(\Gamma_0^*, \ldots, \Gamma_J^*\right)$ and Corollary 1 implies that the EIF may be constructed by collecting the efficient influence functions corresponding to each $\beta_0^*, \ldots, \beta_J^*$. Thus, in this case it is possible to estimate β^* efficiently by estimating each $\beta_0^*, \ldots, \beta_J^*$ separately.

4. Estimation procedures

This section briefly describes the two estimators for the multivalued treatment effects considered. For simplicity, in the overidentified case, the construction of the two-step semiparametric GMM estimators employs a consistent estimator of the corresponding weighting matrix. In particular, it is assumed that A_n is a $(J+1)\,d_\beta \times (J+1)\,d_m$ (random) matrix such that $A_n = A + o_p(1)$ for some positive semidefinite matrix $W = A'A$. (A generalization to a continuously updated two-step semiparametric GMM model is straightforward provided the corresponding additional regularity conditions are imposed.)

4.1. Inverse probability weighting estimator (IPWE)

Eq. (2) leads to a moment condition based only on observed random variables, which involves both the finite-dimensional parameter of interest, β^*, and an infinite-dimensional nuisance parameter (GPS). This suggests that if a preliminary estimator for the GPS that converges to the true GPS sufficiently fast is available, it would still be possible to consistently estimate the finite-dimensional parameter of interest.

These ideas lead to 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. To save notation, define the moment condition $M^{IPW}(\beta,p) = \mathbb{E}[m(Y,T,X;\beta,p)]$, and its sample analogue

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

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

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

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 it needs to be estimated 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. Because it only involves the first part of the EIF, to ensure its semiparametric efficiency the nonparametric estimator $\hat{p}$ will have to play two roles simultaneously: it has to approximate p^* fast enough and it also has to approximate the absent term in the moment condition so that the limiting GMM problem becomes a GMM problem based on the EIF. This point was made by Hirano et al. (2003) in the binary treatment effects model; they showed that $\hat{p} = p^*$ will in general not lead to an efficient estimator. Newey (1994) provides a general discussion of this type of situations in semiparametric models.

4.2. Efficient influence function estimator (EIFE)

This estimator is based on the EIF derived in Theorem 1, which provides another collection of moment conditions that can be exploited to obtain a GMM 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

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

The EIFE may be described by the following steps. First, construct a nonparametric estimator of the GPS, denoted $\hat{p} = \left[\hat{p}_0, \ldots, \hat{p}_J\right]'$, and for each $\beta \in \mathcal{B}^{J+1}$ construct a nonparametric estimator of $e(\beta)$, denoted $\hat{e}(\beta) = [\hat{e}_0(\beta)', \ldots, \hat{e}_J(\beta)']'$. Second, the EIFE for β^* is given by

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

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 be required to approximate well only 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 discussed below (cf. Chen et al. (2004, 2008)). Furthermore, the additional term included in this estimation procedure may be interpreted as a "bias-correction" term that may lead to finite sample performance improvements. A theoretical comparison between these two estimators (and other first-order asymptotically equivalent estimators) is beyond the scope of this paper, and is a topic of future research. Section 6.3 reports a comparison between these two estimators using simulations.

5. Large sample properties

This section presents the main large sample results of the paper using the general theory of Pakes and Pollard (1989).[2] All references to the literature of empirical processes are based on van der Vaart and Wellner (1996). (For reviews on this literature see, e.g., Andrews (1994) and van der Vaart (1998).)

5.1. Consistency

Two mild conditions imposed on the underlying identifying function $m(\cdot;\beta)$ are sufficient to establish consistency of the IPWE.

Assumption 4. For all $t \in \mathcal{T}$: (a) the class of functions $\{\beta_t \mapsto m(\cdot;\beta_t) : \beta_t \in \mathcal{B}\}$ is Glivenko–Cantelli, and (b) $\mathbb{E}[\sup_{\beta_t\in\mathcal{B}} |m(Y(t);\beta_t)|] < \infty$.

Part (a) of Assumption 4 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 uniformly consistent in β for their population mean. Although consistency may be established by other means, requiring a uniform consistency property of the underlying sample moment conditions is standard in the GMM literature (Newey and McFadden, 1994). A simple set of sufficient conditions for Assumption 4(a) are $\mathcal{B}$ compact, $m(\cdot;\beta_t)$ continuous in β_t, and Assumption 4(b). Although this set of conditions is reasonably weak, it is still stronger than necessary. In fact, to cover interesting non-smooth cases 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.[3] Part (b) of Assumption 4 is a usual dominance condition.

Theorem 2 (*Consistency of IPWE*). *Let Assumptions 1, 2 and 4 hold. Assume that the following additional condition holds:*

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

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

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

The following additional assumption is required for consistency of the EIFE.

Assumption 5. For all $t \in \mathcal{T}$: the class of functions $\{\beta_t \mapsto e_t^*(\cdot;\beta_t) : \beta_t \in \mathcal{B}\}$ is Glivenko–Cantelli.

Assumption 5 captures the ideas implied by Assumption 4(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 that they are smooth in β_t.

Theorem 3 (*Consistency of EIFE*). *Let Assumptions 1, 2, 4 and 5 hold. Assume that the following additional condition holds:*

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

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

Since this estimator uses the full form of the EIF, Theorem 3 also requires the nonparametric estimator $\hat{e}$ to be uniformly consistent for e^* in both arguments. This condition is still weak and reasonable for most nonparametric estimators.

5.2. Asymptotic normality and efficiency

The following assumption provides sufficient conditions for asymptotic normality and efficiency of the IPWE.

Assumption 6. For all $t \in \mathcal{T}$ and some $\delta > 0$: (a) $\{\beta_t \mapsto m(\cdot;\beta_t) : |\beta_t - \beta_t^*| < \delta\}$ is a Donsker class; (b) there exist constant $C > 0$ and $r \in (0,1)$ such that $\mathbb{E}[\sup_{|\beta_t-\tilde{\beta}_t|<\delta} |m(Y(t);\beta_t) - m(Y(t);\tilde{\beta}_t)|^2] \le C\delta^{2r}$ for all $\tilde{\beta}_t \in \mathcal{B}$; and (c) $\mathbb{E}[\sup_{|\beta_t-\beta_t^*|<\delta} |m(Y(t);\beta_t)|^2] < \infty$.

Similar to the requirement for consistency, parts (a) and (b) of Assumption 6 restrict the class of identifying functions that may be considered. These restrictions are standard from the empirical process literature, ensuring that a uniform (in β_t) central limit theorem holds and a certain degree of smoothness (in β_t) is enjoyed by the moment conditions. In turn, these results guarantee that a stochastic equicontinuity condition applies, which allows one to obtain an asymptotic linear representation for the estimator. For most applications, Assumption 6(a) is already established or can be easily verified by some "permanence theorem", while Assumption 6(b) may be verified directly. Assumption 6(c) 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 2, 3 and 6 hold. Assume that the following additional conditions hold:*

(4.1) $\left\|\hat{p} - p^*\right\|_\infty = o_p\left(n^{-1/4}\right)$.
(4.2) $M_n^{IPW}\left(\beta^*, \hat{p}\right) = M_n^{EIF}(\beta^*, p^*, \mathrm{e}^*(\beta^*)) + o_p\left(n^{-1/2}\right)$.

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

Asymptotic normality of $\hat{\beta}^{IPW}$ follows directly from Theorem 4 while the 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, which imply certain restrictions in terms of smoothness for the class of functions $\mathcal{P}$ and $\mathcal{E}$, depending on the chosen nonparametric estimator and the dimension of $\mathcal{X}$.

Condition (4.1) is standard and imposes a lower bound in the uniform rate of convergence of $\hat{p}$, requiring this estimator to converge faster than $n^{-1/4}$. Condition (4.2) is crucial. This condition involves the sample moment condition (at $\beta = \beta^*$) and the nonparametric estimator, and requires that a particular linear expansion based on the efficient influence function holds (Newey, 1994). This assumption is 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 $M_n^{EIF}(\beta^*, p^*, \mathrm{e}^*(\beta^*))$, then the estimator cannot be efficient. For example, as mentioned above, if the GPS is known and $\hat{p}$ is replaced by p^* in $M_n^{IPW}\left(\beta^*, \hat{p}\right)$ when constructing the estimation procedure, the resulting estimator will not be efficient. In this sense, Condition (4.2) imposes an "upper bound" on the uniform rate of convergence of $\hat{p}$. Intuitively, the estimator $\hat{p}$ estimates p^* nonparametrically and simultaneously approximates the correction term $\alpha(\cdot; p, e(\beta))$ in

[2] It is also possible to apply the general large sample theory of Chen et al. (2003). However, since the criterion function is smooth in the infinite-dimensional nuisance parameter, the results from Pakes and Pollard (1989) turn out to be sufficient. The general theory of Ai and Chen (2003) does not apply directly to this problem since the moment conditions are non-smooth. In contrast, the recent results of Chen and Pouzo (2009) may be applied to this problem, leading to different (but asymptotically equivalent) estimators.

[3] 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.

the EIF nonparametrically. 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 efficient influence function when constructing the estimator, which leads to the EIFE. This estimator will be asymptotically normal if the following additional assumption holds.

Assumption 7. For all $t \in \mathcal{T}$, some $\delta > 0$, and for all $x \in \mathcal{X}$ and all β_t such that $\left|\beta_t - \beta_t^*\right| < \delta$: (a) $e_t^*(x;\beta_t)$ is continuously differentiable with derivative given by $\partial_{\beta_t} e_t^*(x;\beta_t) \equiv \frac{\partial}{\partial \beta_t} e_t^*(x;\beta_t)$ with $\mathbb{E}[\sup_{|\beta_t-\beta_t^*|<\delta} \left|\partial_{\beta_t} e_t^*(X;\beta_t)\right|] < \infty$; and (b) there exists $\epsilon > 0$ and a measurable function $b(x)$, with $\mathbb{E}[|b(X)|] < \infty$, such that $\left|\partial_{\beta_t} e_t(x;\beta_t) - \partial_{\beta_t} e_t^*(x;\beta_t)\right| \leq b(x) \left\|e_t - e_t^*\right\|_\infty^\epsilon$ for all functions $e_t(\beta_t) \in \mathcal{E}$ such that $\left\|e_t - e_t^*\right\|_\infty < \delta$.

Assumption 7 restricts the class of functions $\mathcal{G} = \{e_t : e_t(\beta) \in \mathcal{E}, \left\|e_t - e_t^*\right\|_\infty < \delta \text{ and } \left|\beta_t - \beta_t^*\right| < \delta\}$, where $e_t^* \in \mathcal{G}$ by construction. Part (a) of this assumption only imposes mild smoothness conditions on the conditional expectation $e_t(\beta_t)$ in β_t as well as a usual dominance condition. This will imply the smoothness requirement in Assumption 3 whenever integration and differentiation may be interchanged. Part (b) of Assumption 7 further restricts the possible class of functions by requiring that functions that are uniformly close also have their derivatives close.

Theorem 5 (*Asymptotic Linear Representation of EIFE*). *Let* $\beta^* \in \text{int}\left(\mathcal{B}^{J+1}\right)$, $\hat{\beta}^{EIF} = \beta^* + o_p(1)$ *and Assumptions 2, 3, 6 and 7 hold. Assume that the following additional conditions hold:*

(5.1) $\left\|\hat{p} - p^*\right\|_\infty = o_p\left(n^{-1/4}\right)$.
(5.2) $\sup_{|\beta-\beta^*|<\delta} \left\|\hat{e}(\beta) - \mathrm{e}^*(\beta)\right\|_\infty = o_p(1)$, *for some* $\delta > 0$.
(5.3) $M_n^{EIF}\left(\beta^*, \hat{p}, \hat{e}(\beta^*)\right) = M_n^{EIF}(\beta^*, p^*, \mathrm{e}^*(\beta^*)) + o_p(n^{-1/2})$.

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

Asymptotic normality of $\hat{\beta}^{EIF}$ also follows directly from Theorem 5, where 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.

The efficiency of the estimators follows 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 then the estimators are efficient without further work. Alternatively, if the problem is over-identified then a consistent estimator of the matrix V_*^{-1} is needed, inducing an intermediate step in the construction of the GMM problems for the IPWE and EIFE. A consistent estimator for V_*^{-1} is easy to construct without further assumptions, as shown in the next section.

5.3. Optimal weighting matrix and uncertainty estimation

This section considers the estimation of V_* and Γ_*, the variance of the efficient influence function and the "sandwich" matrix appearing in the SPEB, respectively.

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

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

for some consistent estimator $\hat{\beta}$ of β^*. Theorem 6 gives a set of simple sufficient conditions that ensure that $\hat{V}$ is consistent for V_*.

Theorem 6 (*Consistent Estimator of* V^*). *Let Assumptions 2, 3, 6 and 7(a) with* $\mathbb{E}[\sup_{|\beta_t-\beta_t^*|<\delta} \left|\partial_{\beta_t} e_t^*(X;\beta_t)\right|^2] < \infty$ *hold. If* $\hat{\beta} = \beta^* + o_p(1)$, $\left\|\hat{p} - p^*\right\|_\infty = o_p(1)$ *and* $\sup_{|\beta-\beta^*|<\delta} \left\|\hat{e}(\beta) - \mathrm{e}^*(\beta)\right\|_\infty = o_p(1)$, *for some* $\delta > 0$, *then* $\hat{V} = V_* + o_p(1)$.

The conditions imposed in Theorem 6 are the same as those assumed in Theorem 4 plus a mild smoothness and dominance condition on e^*.

For the estimation of Γ_* there are several alternatives. First, it is possible to consider a numerical derivative approach directly applied to the sample analogue (e.g., Pakes and Pollard (1989)). Second, in some cases, the estimator may be constructed by taking into consideration the explicit form of the matrix (e.g., $\Gamma_t\left(\beta_t^*\right) = f_{Y(t)}^*\left(q_t^*\right)$ for MQTE). As a third alternative, under the assumptions already imposed, it is also possible to construct a generic estimator provided that integration and differentiation can be interchanged. In this case, for all $t \in \mathcal{T}$,

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

which suggests the plug-in estimator given by

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

where in applications the derivative operator may be implemented by means of a finite difference operator (e.g., a numerical derivative approach). The following theorem provides the sufficient conditions needed and establishes the consistency of this plug-in estimator.

Theorem 7 (*Consistent Estimator of* Γ_*). *Let Assumptions 2, 3 and 7 hold. If* $\hat{\beta} = \beta^* + o_p(1)$ *and* $\sup_{|\beta-\beta^*|<\delta} \left\|\hat{e}(\beta) - \mathrm{e}^*(\beta)\right\|_\infty = o_p(1)$, *for some* $\delta > 0$, *then* $\hat{\Gamma}_t = \Gamma_t^* + o_p(1)$.

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

5.4. Nonparametric estimation of nuisance parameters

The results presented so far have been obtained by imposing high-level assumptions concerning the nonparametric estimators of the infinite-dimensional nuisance parameters. This section considers a particular form of such estimators and discusses explicitly the nonparametric estimation of p^* and e^*, and verifies the high-level conditions imposed in Theorems 4 and 5.

Since both p^* and e^* are (possibly high-dimensional) conditional expectations, this section considers a nonparametric series estimator (Newey, 1994; Chen, 2007). Let $g(X) = \mathbb{E}[Z|X]$ be the unknown regression function of interest for some random variable Z and random vector X, and let $\{r_k(x)\}_{k=1}^{\infty}$ be a sequence of known

approximating functions with the property that a linear combination of $R_K(x) = [r_1(x), \ldots, r_K(x)]'$ can approximate $g(x)$ for $K = 1, 2, \ldots$. 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), $i = 1, 2, \ldots, n$, is given by $\hat{g}(X) = g\left(X; \hat{\gamma}_K\right)$, with $\hat{\gamma}_K = \arg\min_{\gamma_K} \sum_{i=1}^{n} (Z_i - g(X_i; \gamma_K))^2$, where the closed-form solution is

$$\hat{\gamma}_K = \left(\sum_{i=1}^{n} R_K(X_i) R_K(X_i)' \right)^{-} \sum_{i=1}^{n} R_K(X_i) Z_i \tag{4}$$

with A^- denoting a generalized inverse of the matrix A.

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.

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 $\hat{\gamma}_{t,K}(\beta_t)$ be defined as in Eq. (4), but when only the data for $T = t$ are 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

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

A nonparametric series estimator for p^* may be constructed in a similar way. However, the GPS is not only a conditional expectation but also a conditional probability, which imposes additional restrictions that cannot be captured by this standard nonparametric estimator. Thus, in this case it is natural to consider a nonparametric estimator consistent with these additional requirements. In particular, this paper studies a generalization of the estimator introduced by Hirano et al. (2003) in the context of binary treatments, labeled the Multinomial Logistic Series Estimator (MLSE), which may be interpreted as a non-linear sieve estimation procedure.

Using the notation above, for all $t \in \mathcal{T}$, let $g\left(X; \gamma_{t,K}\right) = R_K(X)' \gamma_{t,K}$ be the approximating function, and for notational simplicity let $\gamma_K = (\gamma_{0,K}', \gamma_{1,K}', \ldots, \gamma_{J,K}')'$. When the coefficients $\gamma_{t,K}$, $t \in \mathcal{T}$, are chosen as in Eq. (4) with $Z = D_t$ the resulting estimator is the usual series estimator for the components of p^*. Alternatively, the MLSE chooses all the vectors in γ_K simultaneously by solving the maximum likelihood multinomial logistic problem

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

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

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

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, this section considers the special case of power series and splines.

Assumption 8. (a) For all $t \in \mathcal{T}$, $p_t^*(\cdot)$ and $e_t^*\left(\cdot, \beta_t^*\right)$ 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 $\delta > 0$, $\mathbb{V}[m(Y(t); \beta_t) | X = x]$ is uniformly bounded for all $x \in \mathcal{X}$ and all β_t such that $\left|\beta_t - \beta_t^*\right| < \delta$.

Part (a) of Assumption 8 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 8 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 some components of X to be discretely distributed, and to permit $\mathcal{X}$ to be unbounded by changing the norm used and restricting the tail behavior of the density of X (see Chen et al. (2005) for an example). Part (c) of Assumption 8 is standard.

Theorem 8 (*Nonparametric Estimation*). *Let Assumptions 2 and 8 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* $K = n^v$ *with* $4s/d_x - 6\eta > 1/v > 4\eta + 2$, *where* $\eta = 1$ *or* $\eta = 1/2$ *depending on whether power series or splines are used as basis functions, respectively.*

5.5. Other population parameters and hypothesis testing

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 provides a procedure to easily recover any collection of treatment effects that can be written as (or approximated by) a differentiable function of the marginal treatment effects.

As an application of this idea, and because the ATE and QTEs are continuous transformationsof the MMTE and MQTE, respectively, it is also possible to obtain the important results of Hahn (1998), Hirano et al. (2003) and Firpo (2007) from the binarytreatment effect literature as particular cases of the procedures discussed here. For instance, assuming that $\mathbb{E}[Y(t)^2] < \infty$ and noting that $\Gamma_t^* = 1$ for all $t \in \mathcal{T}$ in the case of MMTE, Theorem 1 implies that the SPEB for the MMTE is given by

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

where $\sigma_t^2(X) = \mathbb{V}[Y(t) | X]$, $\mu_t(X) = \mathbb{E}[Y(t) | X]$, for all $t \in \mathcal{T} = \{0, 1\}$. Since the ATE can be written as $\Delta^{ATE} \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = v'\mu^*$, where $v = (-1, 1)'$, it follows directly using Theorem 4 or Theorem 5 that $\sqrt{n}(\hat{\Delta}^{ATE} - \Delta^{ATE}) \xrightarrow{d} \mathcal{N}\left[0, v'V^*v\right]$, where $\hat{\Delta}^{ATE} = \hat{\mu}_1 - \hat{\mu}_2$,

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

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 et al. (2003) (see also Imbens et al. (2006) for another similar modification of this estimator). The same result can be verified for the case of quantiles since the QTE may also be written as $\Delta^{QTE} \equiv q_1^*(\tau) - q_0^*(\tau) = v'q^*(\tau)$. 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 QTE estimator.

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

$$\hat{\pi} = \arg\min_{\pi}[\hat{\beta} - \beta(\pi)]'(\hat{\Gamma}'\hat{V}^{-1}\hat{\Gamma})[\hat{\beta} - \beta(\pi)],$$

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

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

where $\partial\beta(\pi^*) = \frac{\partial}{\partial\pi}\beta(\pi)\,|_{\pi=\pi^*}$. A consistent estimator of the covariance matrix of $\hat{\pi}$ may also be constructed using a plug-in approach, as discussed previously.

To fix ideas, consider the case where the underlying distribution of the potential outcomes is assumed to be symmetric. This assumption may be incorporated to form an over-identified GMM problem for the estimation of the MMTE. Alternatively, it is possible to first jointly estimate $(\mu^*, q^*(.5))$ using either of the procedures discussed previously and then solve

$$\hat{\pi} = \arg\min_{\pi}\begin{bmatrix}\hat{\mu} - \pi \\ \hat{q}(0.5) - \pi\end{bmatrix}'(\hat{\Gamma}'\hat{V}^{-1}\hat{\Gamma})\begin{bmatrix}\hat{\mu} - \pi \\ \hat{q}(0.5) - \pi\end{bmatrix},$$

which leads to a more efficient estimator of the multi-valued treatment effects for location under symmetry. This idea may be used to incorporate other restrictions.

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 may be done within and across treatment levels for marginal treatment effects, for treatment effects obtained by means of some transformation of these parameters, and for restricted treatment effects by relying on classical testing strategies.

6. Examples and illustrations

This section analyzes two leading examples of particular importance for applied work, the Marginal Mean Treatment Effect and the Marginal Quantile Treatment Effect, and illustrates the main results of this paper by means of an empirical application and a simulation study.

6.1. Leading examples: MMTE and MQTE

To avoid the discussion of technical regularity conditions, this section presents the main results in two propositions that do not include detailed primitive assumptions. The discussion focuses on the implementation of the theoretical results previously discussed to these examples.

First consider the Marginal Mean Treatment Effect. This estimand is denoted by $\mu^* = \left[\mu_0^*, \mu_1^*, \ldots, \mu_J^*\right]'$, and it solves the moment condition in Assumption 1 with $m(Y(t); \mu_t) = Y(t) - \mu_t$, for all $t \in \mathcal{T}$, leading to $\mu_t^* = \mathbb{E}[Y(t)]$. In this case identification follows immediately after assuming finite first moments of the potential outcomes. The exact form of the SPEB was given in the previous section, after assuming that $\mathbb{E}[Y(t)^2] < \infty$, which is denoted here by V_μ^* with typical (i,j)-th element

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

where $\sigma_t^2(X) = \mathbb{V}[Y(t)|X]$, $\mu_t(X) = \mathbb{E}[Y(t)|X]$, for all $t \in \mathcal{T}$. The estimators may be expressed in closed form as

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

and similarly for $\hat{\mu}_t^{EIF}$, for $t \in \mathcal{T}$. Notice that in this case the IPWE corresponds to a properly re-weighted average for each treatment level. Under regularity conditions, these estimators satisfy:

Proposition 1. *$\sqrt{n}(\hat{\mu}^{IPW} - \mu^*) \xrightarrow{d} \mathcal{N}\left(0, V_\mu^*\right)$ and $\sqrt{n}(\hat{\mu}^{EIF} - \mu^*) \xrightarrow{d} \mathcal{N}\left(0, V_\mu^*\right)$.*

This proposition gives root-n consistency (and asymptotic equivalence) of both the IPWE and EIFE for the estimation of the MMTE. These results are obtained as a direct consequence of Theorems 4 and 5, respectively, together with Theorem 8 when the nonparametric procedures described in Section 5.4 are employed. For this example, a simple set of primitive conditions includes $\mathcal{B}$ compact, $\sup_{x\in\mathcal{X}}\mathbb{E}[|Y(t)|^2|X=x] < \infty$ for all $t \in \mathcal{T}$, and the other conditions listed in Assumption 8. In this example Assumptions 6 and 7 are easily satisfied.

Next consider the Marginal Quantile Treatment Effect. For some $\tau \in (0,1)$, this estimand is denoted by $q^*(\tau) = [q_0^*(\tau), q_1^*(\tau), \ldots, q_J^*(\tau)]'$, and it solves the moment condition in Assumption 1 with $m(Y(t); q_t(\tau)) = \mathbf{1}\{Y(t) \le q_t(\tau)\} - \tau$, for all $t \in \mathcal{T}$, which leads to the estimand $q_t^*(\tau) \in \inf\left\{q : F_{Y(t)}(q) \ge \tau\right\}$ with $F_{Y(t)}$ the c.d.f. of $Y(t)$. In this case, a sufficient condition for identification is that $Y(t)$ be a continuous random variable with density $f_{Y(t)}\left(q_t^*(\tau)\right) > 0$. To compute the SPEB note that using Leibniz's rule $\Gamma_t^* = f_{Y(t)}^*\left(q_t^*(\tau)\right) > 0$ for $t \in \mathcal{T}$. Thus, Assumption 3 is satisfied and Theorem 1 implies that the SPEB for the MQTE is given by $V_{q(\tau)}^*$ with typical (i,j)-th element

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

where $\sigma_i^2(X;\tau) = \mathbb{V}\left[\mathbf{1}\left\{Y(i) \le q_i^*(\tau)\right\}|X\right]$, $q_i(X;\tau) = \mathbb{E}[\mathbf{1}\{Y(i) \le q_i^*(\tau)\} - \tau|X]$, for all $i \in \mathcal{T}$. In this case it is not possible to obtain a closed-form solution to the minimization problem and therefore the IPWE is implicitly defined by

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

and similarly for $\hat{q}_t^{EIF}(\tau)$, with $e_t^*(X;\beta_t) = F_{Y(t)}^*(q_t(\tau)|X) - \tau$ and $\hat{F}_{Y(t)}(y|x)$ representing some nonparametric estimator of $F_{Y(t)}^*(y|x)$, the c.d.f. of $Y(t)|X$, for $t \in \mathcal{T}$. Under standard regularity conditions, these estimators satisfy:

Proposition 2. $\sqrt{n}(\hat{q}^{IPW}(\tau)-\mu^*) \xrightarrow{d} \mathcal{N}(0, V^*_{q(\tau)})$ *and* $\sqrt{n}(\hat{q}^{EIF}(\tau) - \mu^*) \xrightarrow{d} \mathcal{N}(0, V^*_{q(\tau)})$.

As in the case of the MMTE, this proposition gives root-n consistency (and asymptotic equivalence) of both the IPWE and EIFE for the estimation of the MQTE. These results are also obtained as a direct a consequence of Theorems 4 and 5, respectively, together with Theorem 8 when the nonparametric procedures described in Section 5.4 are employed. This example requires a different set of regularity conditions, including $F^*_{Y(t)}(y|x)$ continuous in y for every x, and the other conditions listed in Assumptions 7 and 8. Assumption 6 is easily satisfied, while Assumption 7 requires further restrictions on the conditional distributions of $Y(t)|X, t \in \mathcal{T}$.

6.2. Empirical application

In a recent paper, Almond et al. (2005) study the economic costs of low birth weight using different non-experimental techniques. Using a rich database of singletons in Pennsylvania, the authors find a strong negative effect of about 200–250 g of maternal smoking on birth weight using both subclassification on the propensity score and regression adjusted methods. Their results may be extended by considering the effect of maternal smoking intensity during pregnancy on birth weight, since the database records the number of cigarettes-per-day smoked by the mother during pregnancy. This additional information allows one to consider multivalued treatment effects and address several interesting questions, including whether the effect of smoking is constant across levels of smoking and whether there exist differential and/or heterogeneous treatment effects.

This empirical illustration uses the same database as in Almond et al. (2005, Section IV.C). The pre-intervention covariates include age, education and health indicators for the mother and father, among others. Approximately 80% of mothers in the sample did not smoke during pregnancy, while for the remaining 20% the data exhibit important mass points approximately every 5 cigarettes ranging from 1 to 25. This suggests collapsing the number of smoked cigarettes into six 5-cigarette-bin categories ($J = 5$) {0, 1–5, 6–10, 11–15, 16–20, 21+}. Five quantiles (.9,.75,.5,.25,.1), the mean and standard deviation for each potential outcome are jointly estimated, leading to 42 treatment effects. For the estimation of both nonparametric nuisance parameters, the analysis uses cubic B-splines and to reduce the computational burden an additive separability assumption on the approximating functions is used. The results appear to be robust to different choices of these specifications and tuning parameters.

Since this model is exactly identified, it is possible to estimate each treatment effect separately and then form the full EIF to estimate the SPEB. The point and uncertainty estimates for the 42 treatment effects were calculated using both the IPWE and the EIFE. Very similar results were obtained for both IPWE and EIFE, as well as for estimates of the gradient matrix Γ_* when using a plug-in estimator of its exact form and the general numerical derivative approach. Fig. 1 shows the point estimates and their 95% (marginal) confidence intervals for the case of the MMTE and MQTE using the IPWE and a plug-in estimator of the exact form of Γ_*. Interestingly, this figure shows a parallel shift in the entire distribution of birth weight along the smoking intensity. There is a large reduction of about 150 g when the mother starts to smoke (1–5 cigarettes-per-day), an additional reduction of approximately 70 g 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, a close-to-symmetric distribution with approximately constant dispersion (as measured by both interquartile ranges and standard deviation) is observed.

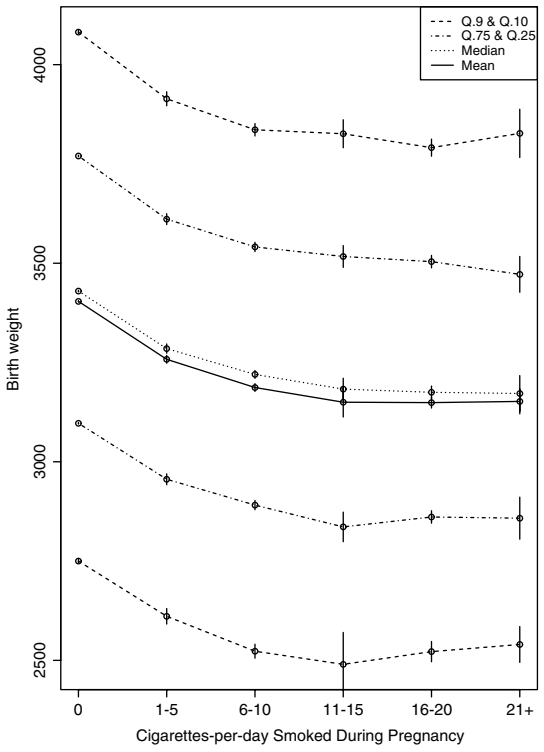

Fig. 1. Effect of maternal smoking intensity on birth weight (5-cigarette bins).

The qualitative results summarized in Fig. 1 may be formally tested. Table 1 presents a collection of hypothesis tests regarding pairwise differences and difference-in-differences of marginal mean treatment effects. On the diagonal, pairwise differences across treatment levels are reported. For example, the reduction in birth weight induced by increasing maternal smoking from 0 to 1–5 cigarettes is 146 g (statistically significant), while the corresponding reduction induced by increasing maternal smoking from 6–10 to 11–15 cigarettes-per-day is 37 g (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-per-day 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-per-day. This differential effect is statistically significant and provides formal evidence of non-linear treatment effects. Importantly, the non-linearities disappear beyond the tenth cigarette-per-day smoked during pregnancy. Similar results are obtained when analyzing the MQTE.

Table 2 illustrates additional joint hypothesis tests. In the first row, a joint test for the hypothesis of no treatment effect (as measured by mean, quantile and spread) for the highest three treatment levels is reported, while in the second and third rows analogous tests considering the highest four and highest five treatment levels are shown, respectively. As exhibited in this table, increasing the 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. Small but statistically significant differences on the interquantile ranges are found.

Table 1
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.
* Significant at 5%.

Table 2
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.

Table 3
Estimated models in the Monte Carlo experiment.

Model 1 (no restrictions)	Model 3 (between restrictions)
$\mu_1, \mu_2, \mu_3, q_1, q_2, q_3$ unrestricted	$\mu_2 = \mu_1 + \Delta_\mu, \mu_3 = \mu_2 + \Delta_\mu$, $q_2 = q_1 + \Delta_q, q_3 = q_2 + \Delta_q$ $\mu_1, \Delta_\mu, q_1, \Delta_q$ unrestricted
Model 2 (within restrictions)	Model 4 (within & between restrictions)
$\mu_1 = q_1, \mu_2 = q_2, \mu_3 = q_3$	$\mu_1 = q_1, \mu_2 = q_2, \mu_3 = q_3$, $\mu_2 = \mu_1 + \Delta_\mu, \mu_3 = \mu_2 + \Delta_\mu$
μ_1, μ_2, μ_3 unrestricted	μ_1, Δ_μ unrestricted

6.3. Monte Carlo evidence

To complement the evidence provided by the above empirical illustration, this section presents a small Monte Carlo study that shows how efficiency in the estimation of multi-valued treatment effects may be increased by incorporating over-identification restrictions. A multi-valued treatment model is considered where the potential outcomes distributions have certain restrictions. In particular, the data generating process (DGP) leads to distributions with equal mean and median, and constant incremental changes along the treatment levels. As discussed in Section 1, these restrictions may be naturally justified by an underlying economic model. For example, in some returns-to-schooling models, changes in log-income are assumed to be proportional and constant across different levels of educational attainment.

The simulations consider a DGP with three treatment levels, $\mathcal{T} = \{0, 1, 2\}$ ($J = 2$), where $\mu_t = q_t$ for all $t \in \mathcal{T}$ (letting $q_t(0.5) = q_t$ to save notation), and $\mu_2 = \mu_1 + \Delta$ and $\mu_3 = \mu_2 + \Delta$. In this case, Δ may be interpreted as the "treatment effect" for location. If these restrictions were ignored, results from the literature on binary treatment effects would still provide consistent and asymptotically normal estimators for Δ. However, such estimators would not enjoy the efficiency gains of the procedures proposed in this paper because they would not exploit the over-identification restrictions within and across treatment levels. The simulation study reported here confirms these results and shows that considering joint (over-identified) multi-valued treatment effects provides efficiency gains in the estimation.

The Monte Carlo experiment uses $S = 5000$ replications, $n = 1000$ observations, and an i.i.d. sample of mutually independent random variables $(X_{1,i}, X_{2,i}, \varepsilon_{0,i}, \varepsilon_{1,i}, \varepsilon_{2,i}, v_{0,i}, v_{1,i}, v_{2,i})$, $i = 1, 2, \ldots, n$. The observed characteristics X_i, with $d_x = 2$, are generated by $X_i = (X_{1i}, X_{2i})'$ with $X_{1,i}$ and $X_{2,i}$ independent $\texttt{uniform}(-1/2, 1/2)$ random variables. The treatment assignment is generated by $T_i = \arg\max_{t\in\mathcal{T}}\{T^*_{t,i}\}$, where $T^*_{t,i} = W_i'\gamma_{t,i} + \varepsilon_{t,i}$, $t \in \mathcal{T}$, $W_i = (1, X_{1,i}, X_{2,i}, X^2_{1,i}, X_{1,i}X_{2,i})'$, $\gamma_0 = (0, 0, 0, 0, 0)'$, $\gamma_1 = (1, 1, 1, 1, 1)'$, $\gamma_2 = (2, 2, 2, 2, 2)'$, and where $\varepsilon_{0,i}$, $\varepsilon_{1,i}$ and $\varepsilon_{2,i}$ are independent $\texttt{gumbel}(0, 1)$ random variables, leading to the multinomial logistic model. The potential outcomes are generated by $Y_i(t) = \mu_t + 0.5X_{1,i} + 0.5X_{2,i} + v_{t,i}$, $t \in \mathcal{T}$, with $\mu_0 = 0$, $\mu_1 = 1$, $\mu_2 = 2$, and $v_{0,i}$, $v_{1,i}$ and $v_{2,i}$ independent $\texttt{laplace}(0, 1)$ random variables. This setup leads to a DGP with $\mu_0 = q_0 = 0$, $\mu_1 = q_1 = 1$, $\mu_2 = q_2 = 2$, and $\Delta = \mu_2 - \mu_1 = \mu_3 - \mu_2 = 1$. (Different variants of this DGP were considered, and in all cases similar results were obtained.)

Four different models were estimated using data simulated from the DGP described above. These models are summarized in Table 3.

Model 1 corresponds to the unrestricted MMTE and MQTE, which leads to the estimation of Δ as it would be done in the context of binary treatment effects. Model 2 imposes over-identification restrictions within each treatment level, while Model 3 imposes over-identification restrictions across treatment

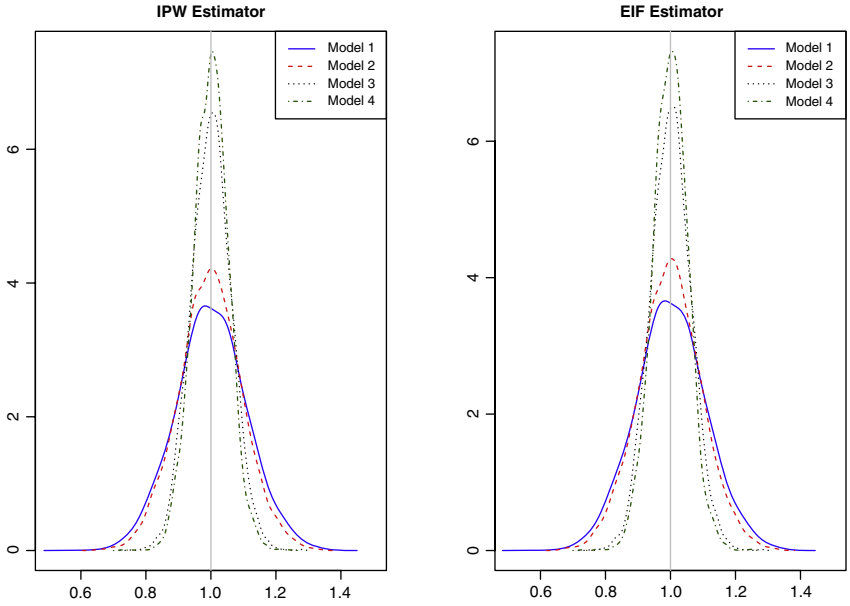

Fig. 2. Kernel density estimates of IPW and EIF estimators.

levels (i.e., cross-equation restrictions) for each estimand of location. Finally, Model 4 simultaneously imposes all over-identification restrictions considered in Table 3. This setup leads to a partial ordering in terms of efficiency, since all restrictions described in Table 3 are simultaneously true in the DGP. In particular, the estimators from Model 1 are dominated by the estimators from Model 2 and Model 3, while in turn these estimators are dominated by those obtained from Model 4. Of course, there is no general ranking between the estimators from Model 2 and Model 3.

The results are presented in Fig. 2, which reports both the IPW and the EIF estimators for $\mu_3 - \mu_2$ in Models 1 and 2, and Δ_μ in Models 3 and 4. These estimators are constructed using the nonparametric estimators described in Section 5.4 with a polynomial basis of approximation with a second-order tensor product based on X_i. (The results reported here were robust to different choices of K and to the use of splines instead of polynomials.) In Fig. 2, the solid line corresponds to the semiparametric average treatment effect estimator that would be constructed using classical results from the binary treatment effect literature, while the other three lines are obtained by the results presented in this paper. The efficiency gains from over-identification restrictions are substantial, being particularly important the cross-equation restrictions across treatment levels. Interestingly, the IPWE performed as well as the EIFE, despite the fact that the latter includes the correction term that may be interpreted as a bias-correction procedure. Of course, these results are design specific, and it remains an open question whether this correction term is important from a theoretical (and empirical) perspective. Further theoretical research comparing these estimators is underway.

7. Final remarks

This paper has studied 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. Two alternative estimators were proposed based on standard GMM arguments combined with the corresponding modifications needed to circumvent the fundamental problem of causal inference. The resulting estimators are of the two-stage semiparametric GMM variety, where the first step is fully nonparametric. Under regularity conditions, these estimators were shown to be root-n consistent, asymptotically normal and efficient for the general population parameter of interest. Using these estimators it was also shown how other estimands of interest may be efficiently estimated, allowing the researcher to recover a rich class of population parameters. Important results in the literature of program evaluation with binary treatment assignments may be seen as particular cases of the procedures discussed in this paper 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, 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 it is natural to expect multiple differential impacts within and across treatments, which highlights the relevance of considering multi-valued treatments, when possible, for making informed policy decisions.

Although this paper has focused on estimands based on the marginal distribution of the potential outcomes, it may be also be of interest to consider multi-valued weighted treatment effects (Hirano et al., 2003) leading to estimands based on the conditional distribution of the potential outcome given some treatment level. Efficient estimation procedures for these estimands may be derived directly by following and extending the work presented here.

Appendix A. Proofs of theorems

Let C denote a generic positive constant which may vary depending on the context. For any vector v, its t-th element is denoted by $v_{[t]}$, while $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalue of the matrix A, respectively. Standard qualifications such as "almost surely", "with probability approaching one", or "for n large enough" are omitted to conserve space.

Proof of Theorem 1. The proof given is based on the theoretical approach described in Bickel et al. (1993), and follows the results in Hahn (1998) and Chen et al. (2004, 2008). 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

$$\log f(y, t, x; \theta) = \sum_{j \in \mathcal{T}} \mathbf{1}\{t = j\} \left[\log f_j(y|x; \theta) + \log p_j(x; \theta)\right] + \log f_X(x; \theta),$$

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

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

where

$$S_y(y, t, x) = \sum_{j \in \mathcal{T}} \mathbf{1}\{t = j\} s_j(y, x),$$

$$s_j(y, x) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_j(y|x; \theta)|_{\theta_0},$$

$$S_p(t, x) = \sum_{j \in \mathcal{T}} \mathbf{1}\{t = j\} \frac{\dot{p}_j^*(x)}{p_j^*(x)}, \qquad \dot{p}_j^*(x) = \frac{\mathrm{d}}{\mathrm{d}\theta} p_j(x; \theta)|_{\theta_0},$$

$$S_x(x) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log f_X(x; \theta)|_{\theta_0}.$$

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

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

where $\mathbb{E}\left[S_p(T, X)|X\right] = \sum_{t \in \mathcal{T}} \dot{p}_t^*(X)$ and $\mathbb{E}[S_p(T, X)^2|X] = \sum_{t \in \mathcal{T}} \dot{p}_t^*(X)^2 / p_t^*(X)$, and hence it is required that $p_t^*(x)$ and $\dot{p}_t(x; \theta_0)$ be measurable functions such that $\sum_{t \in \mathcal{T}} \dot{p}_t^*(X) = 0$ and $\sum_{t \in \mathcal{T}} \dot{p}_t^*(X)^2 / p_t^*(X) < \infty$. 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 2(b) and the fact that T is finite.

Next, let $\underline{m}(\beta) = [m(Y(0); \beta_0)', \ldots, m\left(Y(J); \beta_J\right)']$ and note that, for any $(d_\beta(J+1) \times d_m(J+1))$ positive semidefinite matrix A, the population parameter satisfies $A\mathbb{E}[\underline{m}(\beta)] = 0$ if and only if $\beta = \beta^*$. Thus, the implicit function theorem implies that

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

where $\Gamma_* = \frac{\partial}{\partial \beta}\mathbb{E}\left[\underline{m}(\beta)\right]|_{\beta=\beta^*}$, $\Upsilon(\theta_0) = \frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\underline{m}(\beta^*)\right]|_{\theta=\theta_0} = \frac{\partial}{\partial \theta}\int \underline{m}(\beta^*)\,\mathrm{d}F(y, t, x; \theta)|_{\theta=\theta_0}$, and observe that $\Upsilon(\theta_0) = [\frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[m(Y(0); \beta_0)'\right]|_{\theta=\theta_0}, \ldots, \frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[m\left(Y(J); \beta_J\right)'\right]|_{\theta=\theta_0}]$ with typical element $j \in \mathcal{T}$,

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

The parameter is pathwise differentiable if there exits a $d_\beta(J+1)$-valued function $\Psi_\beta(y, t, x; A) \in \mathcal{H}$ such that for all regular parametric submodels

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

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

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

for a fixed choice of the matrix A.

Finally, the EIF is obtained when $A = \Gamma_*' V_*^{-1}$, leading to the SPEB given by $V^* = (\Gamma_* V_*^{-1} \Gamma_*')^{-1}$. ■

Proof of Theorem 2. The result follows from Corollary 3.2 in Pakes and Pollard (1989) after setting $\theta = \beta, \theta_0 = \beta^*, G_n(\beta) = A_n M_n^{IPW}\left(\beta, \hat{p}\right), G(\beta) = AM^{IPW}(\beta, p^*)$, and verifying their condition (iii). Because $A_n - A = o_p(1)$, to verify condition (iii) it is enough to show that $\sup_{\beta \in \mathcal{B}} |M_{[t],n}^{IPW}\left(\beta, \hat{p}_t\right) - M_{[t]}^{IPW}\left(\beta, p_t^*\right)| = o_p(1)$, for all $t \in \mathcal{T}$. The result follows because

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

by Assumption 4(b), and

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

because the class of functions $\mathcal{F}_t = \{\beta_t \mapsto \mathbf{1}\{\cdot = t\} m(\cdot; \beta_t) / p_t^*(\cdot) : \beta_t \in \mathcal{B}\}$ is Glivenko–Cantelli by Assumptions 2(b) and 4 (van der Vaart and Wellner, 2000). ■

Proof of Theorem 3. The result also follows from Corollary 3.2 in Pakes and Pollard (1989) after setting $\theta = \beta, \theta_0 = \beta^*, G_n(\theta) = A_n M_n^{EIF}\left(\beta, \hat{p}, \hat{e}\right), G(\theta) = AM^{EIF}(\beta, p^*, \mathrm{e}^*)$, and verifying their sufficient condition (iii). Using the proof and the conclusion of Theorem 2, the conclusion follows after noting that, for all $t \in \mathcal{T}$,

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

because $\mathbb{E}[\sup_{\beta \in \mathcal{B}} |e_t^*(X; \beta)|] < \infty$ (Assumption 4(b)), $\left\|\hat{e}_t - e_t^*\right\|_\infty = o_p(1)$ and the class of functions $\mathcal{F}_t = \{\beta_t \mapsto e_t^*(\cdot; \beta_t)\left(\mathbf{1}\{\cdot = t\} - p_t^*(\cdot)\right) / p_t^*(\cdot) : \beta_t \in \mathcal{B}\}$ is Glivenko–Cantelli by Assumptions 2(b) and 4 (van der Vaart and Wellner, 2000). ■

Proof of Theorem 4. The result follows from Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*, G_n(\beta) = A_n M_n^{IPW}\left(\beta, \hat{p}\right), G(\beta) = AM^{IPW}(\beta, p^*)$, and verifying stochastic equicontinuity (condition (iii)), since their other sufficient conditions hold by the construction of the estimator, Assumptions 3 and 6, and Condition (2.1) in this case. To establish the sufficient condition, it suffices to show that, for all positive sequences $\delta_n = o(1)$,

$$\begin{aligned} &\sup_{|\beta_t - \beta_t^*| \leq \delta_n} \frac{\sqrt{n}\left|M_{[t],n}^{IPW}\left(\beta, \hat{p}\right) - M_{[t]}^{IPW}(\beta, p^*) - M_{[t],n}^{IPW}\left(\beta^*, \hat{p}\right)\right|}{1 + C\sqrt{n}|\beta_t - \beta_t^*|} \\ &\quad = o_p(1), \end{aligned} \tag{A.1}$$

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

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

and note that the left-hand side of (A.1) is bounded by $R_{1n} + R_{2n} + R_{3n} + R_{4n}$, where

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

Box I.

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

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

$$R_{3n} = \sup_{|\beta_t-\beta_t^*|\leq\delta_n} \times \frac{\sqrt{n}\left|M_{[t],n}^{IPW}\left(\beta^*,\hat{p}\right)+M_{[t],n}^{IPW}(\beta^*,p^*)-\Delta_{[t],n}\left(\beta^*,\hat{p}-p^*\right)\right|}{1+C\sqrt{n}\left|\beta_t-\beta_t^*\right|},$$

$$R_{4n} = \sup_{|\beta_t-\beta_t^*|\leq\delta_n} \times \frac{\sqrt{n}\left|M_{[t],n}^{IPW}\left(\beta^*,\hat{p}\right)+M_{[t],n}^{IPW}(\beta^*,p^*)-\Delta_{[t],n}\left(\beta^*,\hat{p}-p^*\right)\right|}{1+C\sqrt{n}\left|\beta_t-\beta_t^*\right|}.$$

Now, $R_{1n} = o_p(1)$ because the class of functions $\mathcal{F}_t = \{\beta_t \mapsto \mathbf{1}\{\cdot = t\} m(\cdot;\beta_t)/p_t^*(\cdot) : |\beta_t-\beta_t^*| \leq \delta\}$ is Donsker with finite integrable envelope by Assumption 6 (Theorem 2.10.6 of van der Vaart and Wellner (1996)) and L_2 continuous by Assumptions 3 and 6, while $R_{2n} = o_p(1)$ and $R_{3n} = o_p(1)$ using elementary inequalities and Condition (2.1) and Assumption 2. Finally, $R_{4n} = o_p(1)$ by the triangular inequality and because the class of functions $\mathcal{F}_t = \{\beta_t \mapsto \mathbf{1}\{\cdot = t\}\left|m(\cdot;\beta_t)-m\left(\cdot;\beta_t^*\right)\right|/p_t^*(\cdot) : \left|\beta_t-\beta_t^*\right| \leq \delta\}$ is Donsker with finite integrable envelop by Assumption 6 (Theorem 2.10.6 of van der Vaart and Wellner (1996)) and L_2 continuous by Assumptions 3 and 6. This establishes condition (iii) of Theorem 3.3 in Pakes and Pollard (1989). ■

Proof of Theorem 5. The proof follows the same logic of the proof of Theorem 4. Apply Theorem 3.3 and Lemma 3.5 in Pakes and Pollard (1989) after setting $\theta = \beta$, $\theta_0 = \beta^*$, $G_n(\theta) = A_n M_n^{EIF}\left(\beta,\hat{p},\hat{e}\right)$, $G(\theta) = AM^{EIF}(\beta,p^*,\mathrm{e}^*)$, and verifying their condition (iii). This condition is verified if for all sequences $\delta_n = o(1)$,

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

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

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

Now, the left-hand side of (A.2) is bounded by $R_{1n}+R_{2n}$, where the expression shown in Box I holds for some convex linear combination $\tilde{\beta}_t$ (between β_t and β_t^*) and

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

Finally,

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

and $R_{1n} = o_p(1)$ because the first term is $o_p(1)$ by Assumption 7(b), the second term is $o_p(1)$ because the class of functions $\mathcal{F}_t = \{\beta_t \mapsto \left(\partial_{\beta_t}e_t^*(\cdot;\beta_t)-\partial_{\beta_t}e_t^*\left(\cdot;\beta_t^*\right)\right)\left(\mathbf{1}\{\cdot = t\}-p_t^*(\cdot)\right)/p_t^*(\cdot) : \left|\beta_t-\beta_t^*\right| \leq \delta\}$ is Glivenko–Cantelli for some $\delta > 0$ by Assumption 7(a) (van der Vaart and Wellner, 2000), and the third term is $o_p(1)$ by Assumption 7(a). Similarly, $R_{2n} = o_p(1)$ by Assumption 7(a). This establishes condition (iii) of Theorem 3.3 in Pakes and Pollard (1989). ■

Proof of Theorem 6. Let $V_n = n^{-1}\sum_{i=1}^{n}\psi(Y_i,T_i,\beta^*,p^*,\mathrm{e}^*(\beta^*))\psi(Y_i,T_i,\beta^*,p^*,\mathrm{e}^*(\beta^*))'$. Using Holder's Inequality it follows that $|\hat{V}_n-V_*| \leq |\hat{V}_n-V_n|+|V_n-V_*| = o_p(1)$, provided that, for all sequences $\delta_n = o(1)$ and for all $t \in \mathcal{T}$,

$$\frac{1}{n}\sum_{i=1}^{n}\left|m(Y_i,T_i,X_i;\hat{\beta},\hat{p})-m\left(Y_i,T_i,X_i;\beta^*,p^*\right)\right|^2 = o_p(1) \quad (A.3)$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left|\alpha(T_i,X_i;\hat{p},\hat{e}(\hat{\beta}))-\alpha\left(T_i,X_i;p^*,\mathrm{e}^*\left(\beta^*\right)\right)\right|^2 = o_p(1). \quad (A.4)$$

The first (A.3) follows 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), while the second (A.4) is verified using the assumptions of the theorem. ■

Proof of Theorem 7. Follows directly by the same arguments given in the proof of Theorem 5. ■

Proof of Theorem 8. Using the notation and results of Appendix B, for the case of power series and splines $\zeta(K) = K^\eta$ with $\eta = 1$ and

$\eta = 1/2$, respectively, while Assumption 8 implies Assumption B-1 with $\alpha = s/d_x$ (Newey, 1994). Theorem B-1 gives

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

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

Next, to verify Condition (4.2) in Theorem 4 it is enough to consider the typical t-th component, $\sqrt{n}|M^{IPW}_{[t],n}(\beta^*_t, \hat{p}_t) - M^{EIF}_{[t],n}(\beta^*_t, p^*_t, e^*_t(\beta^*_t))| \leq R_{1n} + R_{2n} + R_{3n}$, where

$$R_{1n} = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{\hat{p}_t(X_i)} - \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)} + \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \left(\hat{p}_t(X_i) - p^*_t(X_i)\right) \right\} \right|,$$

$$R_{2n} = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \left(\hat{p}_t(X_i) - p^*_t(X_i)\right) + \frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} \left(\hat{p}_t(X_i) - p^*_t(X_i)\right) \right\} \right|,$$

$$R_{3n} = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ -\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} \left(\hat{p}_t(X_i) - p^*_t(X_i)\right) + \frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} \left(D_{t,i} - p^*_t(X_i)\right) \right\} \right|.$$

For the first term,

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

For the second term,

$$R_{2n} \leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \right) \left(\hat{p}_t(X_i) - p^0_{K,t}(X_i)\right) \right| \tag{A.5}$$

$$+ \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \right) \left(p^0_{K,t}(X_i) - p^*_t(X_i)\right) \right|, \tag{A.6}$$

using the notation introduced in Appendix B. To obtain a bound on the term ((A.5)), first notice that by a second-order Taylor expansion and using the results in Appendix B it follows that, for some $\tilde{\gamma}_K$ such that $\left|\tilde{\gamma}_K - \gamma^0_K\right| \leq \left|\hat{\gamma}_K - \gamma^0_K\right|$,

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \right) \left(\hat{p}_t(X_i) - p^0_{K,t}(X_i)\right) \right|$$

$$\leq \left|\hat{\gamma}_K - \gamma^0_K\right| \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \right) \times \left[\dot{\mathbf{L}}_t\left(g_{-0}\left(X_i, \gamma^0_K\right)\right) \otimes R_K(X_i)'\right] \right|$$

$$+ \sqrt{n}\left|\hat{\gamma}_K - \gamma^0_K\right|^2 \frac{1}{n} \sum_{i=1}^{n} \left| \frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \right| \times \left|\mathbf{I}_J \otimes R_K(X_i) R_K(X_i)'\right|$$

$$= O_p\left(K^{1/2} n^{-1/2} + K^{1/2} K^{-s/d_x}\right) O_p\left(K^{1/2}\right) + O_p\left(\sqrt{n}(K^{1/2} n^{-1/2} + K^{1/2} K^{-\alpha})^2\right) O_p(K),$$

while for the term (A.6),

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - \frac{D_{t,i} m\left(Y_i; \beta^*_t\right)}{p^*_t(X_i)^2} \right) \left(p^0_{K,t}(X_i) - p^*_t(X_i)\right) \right| = O_p\left(K^{-s/d_x}\right) = o_p(1).$$

For the last term, using the first-order condition for MLSE, which implies that $\sum_{i=1}^{n}(D_{t,i} - \hat{p}_t(X_i)) R_K(X_i) = \mathbf{0}$, and by choosing $\theta \in \mathbb{R}^K$ appropriately,

$$R_{3n} \leq \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - R_K(X_i)'\theta \right) \left(D_{t,i} - p^*_t(X_i)\right) \right| + \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{e^*_t\left(X_i; \beta^*_t\right)}{p^*_t(X_i)} - R_K(X_i)'\theta \right) \left(p^*_t(X_i) - \hat{p}_t(X_i)\right) \right|$$

$$\leq O_p\left(K^{-s/d_x}\right) + n^{1/2} O\left(K^{-s/d_x}\right) O_p \times \left(K^{\eta} K^{1/2} n^{-1/2} + K^{\eta} K^{1/2} K^{-s/d_x}\right).$$

Condition (4.2) in Theorem 4 follows directly under the assumptions of this theorem.

Next, consider Theorem 5. Conditions (5.1) and (5.2) follow directly from previous calculations and the first part of Proposition A1 in Chen et al. (2005), respectively. It remains to show Condition (5.3) in Theorem 5. From Newey (1994),

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

To establish the final condition is enough to show the result for the typical t-th component. From the previous calculations and using the identity $\hat{a}/\hat{b} = a/b + (\hat{a} - a)/b - a(\hat{b} - b)/b^2 + a(\hat{b} - b)^2/(b^2\hat{b}) - (\hat{a} - a)(\hat{b} - b)/(b\hat{b})$ it follows that $\sqrt{n}|M^{EIF}_{[t],n}(\beta^*_t, \hat{p}_t, \hat{e}_t(\beta^*_t)) - M^{EIF}_{[t],n}(\beta^*_t, p^*_t, e^*_t(\beta^*_t))| \leq R_{4n} + R_{5n} + R_{6n} + o_p(1)$, where

$$R_{4n} = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_{t,i}\left(m\left(Y_i; \beta^*_t\right) - e^*_t\left(X_i; \beta^*_t\right)\right)}{p^*_t(X_i)^2} \left(\hat{p}_t(X_i) - p^*_t(X_i)\right) \right|,$$

$$R_{5n} = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{D_{t,i} - p^*_t(X_i)}{p^*_t(X_i)} \left(\hat{e}_t\left(X_i; \beta^*_t\right) - e^*_t\left(X_i; \beta^*_t\right)\right) \right|,$$

$$R_{6n} = \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\hat{e}_t\left(X_i; \beta^*_t\right) - e^*_t\left(X_i; \beta^*_t\right)\right) \right|.$$

Finally, by the same arguments used for term R_{2n} above it is verified that $R_{4n} = o_p(1)$, while by similar arguments but for the case of series it is also verified $R_{6n} = o_p(1)$ under the assumptions of this theorem. Therefore Condition (5.3) in Theorem 5 holds. ■

Appendix B. Multinomial Logistic Series Estimator

This appendix derives uniform rates of convergence for the non-linear sieve estimator proposed for the estimation of the Generalized Propensity Score. These results encompass those in Hirano et al. (2003), and further allow for an arbitrary number

of outcomes, arbitrary choice of approximating basis, and less stringent smoothness requirements.

Under the conditions imposed below and by choosing an appropriate non-singular linear transformation, assume without loss of generality that $\mathbb{E}\left[R_K(X)R_K(X)'\right] = \mathbf{I}_K$, where $\mathbf{I}_K$ is the $K \times K$ identity matrix (see Newey (1994) for details). Let $\zeta(K) = \sup_{x\in\mathcal{X}}|R_K(x)|$ and define $p_{-0}(X) = \left[p_1(X), \ldots, p_J(X)\right]' \in \mathbb{R}^J$, $\gamma_{-0,K} = [\gamma_{K,1}', \ldots, \gamma_{K,J}']' \in \mathbb{R}^{JK}$, and $g_{-0}(X, \gamma_K) = [R_K(X)'\gamma_{K,1}, \ldots, R_K(X)'\gamma_{K,J}]' \in \mathbb{R}^J$. Recall that $p_0^*(X) = 1 - \sum_{j=1}^J p_j^*(X)$ by construction. In addition, define for a vector $z \in \mathbb{R}^J$, $z = \left[z_1, \ldots, z_J\right]'$, the functions $L_t : \mathbb{R}^J \to \mathbb{R}$ and $L_t^{-1} : \mathbb{R}^J \to \mathbb{R}$, for all $t = 1, 2, \ldots, J$, $L_t(z) = \exp(z_t)/(1 + \sum_{j=1}^J \exp(z_j))$, $L_t^{-1}(z) = \log(z_t/(1 - \sum_{j=1}^J \exp(z_j)))$ 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 denoted by $\dot{L}_t(z)$ and satisfies $\sup_z|\dot{L}_t(z)| < C$ since $\left|L_t(z)L_j(z)\right| < 1$ and $L_t(z)(1 - L_t(z)) < 1/4$. Define the vector-valued function $\mathbf{L}(z) = [L_1(z), \ldots, L_J(z)]'$ and $\mathbf{L}^{-1}(z) = [L_1^{-1}(z), \ldots, L_J^{-1}(z)]'$ and observe that the function $\mathbf{L}(\cdot)$ is differentiable with gradient (matrix) $\dot{\mathbf{L}}(z) = [\dot{L}_1(z), \ldots, \dot{L}_J(z)] \in \mathbb{R}^{J\times J}$ and $\sup_z|\dot{\mathbf{L}}(z)| < C$, for some constant C that only depends on J. With this notation, $p_t\left(X; \gamma_{t,K}\right) = L_t\left(g_{-0}(X, \gamma_K)\right)$ for $t \in \mathcal{T}$ (recall $\gamma_{K,0} = \mathbf{0}_K$ for identification purposes).

The multinomial logistic log-likelihood is $\ell_n(\gamma_K) = \sum_{i=1}^n \sum_{t=0}^J D_{t,i}\log\left(L_t\left(g_{-0}(X_i, \gamma_K)\right)\right)$, with solution $\hat{\gamma}_K = \arg\max_{\gamma_K}\ell_n(\gamma_K)$ and estimated probabilities given by $\hat{p}_t(X) = L_t(g_{-0}(X_i, \hat{\gamma}_K))$, for all $t \in \mathcal{T}$. Letting $\mathbf{D}_i = [D_{1,i}, D_{2,i}\cdots, D_{J,i}]'$, it follows that

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

$$\frac{\partial^2}{\partial\gamma_K\partial\gamma_K'}\ell_n(\gamma_K) = -\sum_{i=1}^n \mathbf{H}(X_i, \gamma_K)\otimes R_K(X_i)R_K(X_i)',$$

where $\mathbf{H}(X_i, \gamma_K) = \operatorname{diag}\left(\mathbf{L}\left(g_{-0}(X_i, \gamma_K)\right)\right) - \mathbf{L}\left(g_{-0}(X_i, \gamma_K)\right)\mathbf{L}(g_{-0}(X_i, \gamma_K))'$.

The followings conditions are sufficient to derive the uniform rates of convergence.

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 $\sup_{x\in\mathcal{X}}|R_K(x)| \le \zeta(K)$, 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 assumptions, 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, such a sequence is denoted by $p_{t,K}^0(X) = L_t\left(g_{-0}\left(X, \gamma_K^0\right)\right)$, for all $t \in \mathcal{T}$, so that $p_K^0 = [p_{0,K}^0, \ldots, p_{J,K}^0]'$.

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

Theorem B-1 (*Uniform Rate of Convergence of MLSE*). *Let Assumptions 2 (b) and B-1. 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 hence $\left\|\hat{p} - p^*\right\|_\infty = O_p(\zeta(K)K^{1/2}n^{-1/2} + \zeta(K)K^{1/2}K^{-\alpha})$.

Proof of Theorem B-1. Since the map $\mathbf{L}(\cdot)$ is differentiable with $\sup_z|\dot{\mathbf{L}}(z)| < C$, the mean value theorem and Assumption B-1(c) give

$$\begin{aligned}&\sup_{x\in\mathcal{X}}\left|p_{-0}^*(x) - \mathbf{L}\left(g_{-0}\left(x, \gamma_K^0\right)\right)\right|\\ &\quad\le C\sup_{x\in\mathcal{X}}\left|\mathbf{L}^{-1}\left(p_{-0}^*(x)\right) - g_{-0}\left(x, \gamma_K^0\right)\right|\\ &\quad= O\left(K^{-\alpha}\right),\end{aligned}$$

giving part (i).

For part (ii), recall that $L_t\left(g_{-0}(x, \gamma)\right) > 0$, for all $t \in \mathcal{T}$, and $\sum_{t=1}^J L_t\left(g_{-0}(x, \gamma)\right) < 1$. The form of the matrix $\mathbf{H}(x, \gamma)$ and Theorem 1 in Tanabe and Sagae (1992) show that $\mathbf{H}(x, \gamma)$ is symmetric positive definite with $0 < \lambda_{\min}(\mathbf{H}(x, \gamma)) \le \lambda_{\max}(\mathbf{H}(x, \gamma)) < 1$, which implies that $\mathbf{H}(x, \gamma) \ge \lambda_{\min}(\mathbf{H}(x, \gamma))\mathbf{I}_J$ and $\lambda_{\min}(\mathbf{H}(x, \gamma)) \ge \det(\mathbf{H}(x, \gamma))$. These results and the exact Cholesky decomposition of $\mathbf{H}(x, \gamma)$ give $\inf_{x\in\mathcal{X}}\mathbf{H}(x, \gamma) \ge \inf_{x\in\mathcal{X}}\prod_{t=0}^J L_t\left(g_{-0}(x, \gamma)\right)\mathbf{I}_J$, in a positive semidefinite sense.

For $\hat{\Omega}_K = n^{-1}\sum_{i=1}^n R_K(X_i)R_K(X_i)'$, Newey (1994) showed that $|\hat{\Omega}_K - \mathbf{I}_K| = O_p(\zeta(K)K^{1/2}n^{-1/2})$. Define the event $\mathcal{A}_n = \{\lambda_{\min}(\hat{\Omega}_K) > 1/2\}$, and note that by Assumption B-1(b) $\mathbb{P}[\mathcal{A}_n] \to 1$. Let $\partial\ell_n(\gamma)/\partial\gamma = \dot{\ell}_n(\gamma)$ and note that

$$\begin{aligned}\mathbb{E}\left[\left|\frac{1}{n}\dot{\ell}_n\left(\gamma_K^0\right)\right|\right] &\le C\left(\frac{1}{n}\mathbb{E}\left[\left|\left(\mathbf{D}_i - p_{-0}^*(X_i)\right)\otimes R_K(X_i)\right|^2\right]\right)^{1/2}\\ &\quad + C\sup_{x\in\mathcal{X}}\left|p_{-0}^*(x) - \mathbf{L}\left(g_{-0}\left(x, \gamma_K^0\right)\right)\right|\mathbb{E}\left[|R_K(X)|\right]\\ &= O\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right);\end{aligned}$$

then by Markov's Inequality it follows that $|\frac{1}{n}\dot{\ell}_n\left(\gamma_K^0\right)| = O_p\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right)$. This implies that for any fixed constant $\varsigma > 0$ the probability of the event $\mathcal{B}_n(\varsigma) = \{|\frac{1}{n}\dot{\ell}_n\left(\gamma_K^0\right)| < \varsigma(K^{1/2}n^{-1/2} + K^{-\alpha+1/2})\}$ approaches one, i.e., $\mathbb{P}[\mathcal{B}_n(\varsigma)] \to 1$.

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 Assumption 2(b). Define the sets $\Gamma_K^\delta = \{\gamma \in \mathbb{R}^{JK} : \inf_{x\in\mathcal{X}}\prod_{t=0}^J L_t\left(g_{-0}(x, \gamma)\right) > \delta/2\}$, and $\Gamma_K^0(\varrho) = \left\{\gamma \in \mathbb{R}^{JK} : \left|\gamma - \gamma_K^0\right| \le \varrho\left(K^{1/2}n^{-1/2} + K^{1/2}K^{-\alpha}\right)\right\}$ for any $\varrho > 0$. Because (for some intermediate point $\tilde{\gamma}_K$)

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

by Assumption B-1(b) and (c), it follows that for n for large enough $\Gamma_K^\delta \subset \Gamma_K^0(\varrho)$.

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

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

which implies that with probability at least $(1-\varepsilon)$, $\lambda_{\min}(-\partial\ell_n(\gamma_K)/\partial\gamma\partial\gamma') \geq \delta/4$. Moreover, under the same conditions (i.e., also with probability at least $(1-\varepsilon)$) and for any $\gamma_K \in \Gamma_K^0 \setminus \{\gamma_K^0\}$ it is verified that

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

for some $\tilde{\gamma}_K$ such that $\left|\tilde{\gamma}_K - \gamma_K^0\right| \leq \left|\gamma_K - \gamma_K^0\right|$. Since $\ell_n(\gamma_K)$ is continuous and concave, it follows that $\hat{\gamma}_K$ maximizes $\ell_n(\gamma_K)$ and $\hat{\gamma}_K$ satisfies the first-order condition with probability approaching one. Now the result follows directly. ■

References

Abadie, A., 2005. Semiparametric difference-in-differences estimators. Review of Economic Studies 72, 1–19.

Abadie, A., Imbens, G.W., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74, 235–267.

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795–1843.

Almond, D., Chay, K.Y., Lee, D.S., 2005. The costs of low birth weight. Quarterly Journal of Economics 120, 1031–1083.

Andrews, D.W.K., 1994. Empirical process methods in econometrics. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier Science B.V, pp. 2247–2294.

Bang, H., Robins, J.M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–972.

Bickel, P.J., Klaasen, C.A.J., Ritov, Y., Wellner, J.A., 1993. Efficient and Adaptive Estimation for Semiparametric Models. Springer, New York.

Chen, X., 2007. Large Sample Sieve Estimation of Semi-Nonparametric Models. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. VI. Elsevier Science B.V, pp. 5549–5632.

Chen, X., Hong, H., Tamer, E., 2005. Measurement error models with auxiliary data. Review of Economic Studies 72, 343–366.

Chen, X., Hong, H., Tarozzi, A., 2004. Semiparametric efficiency in GMM models of nonclassical measurement errors, missing data and treatment effects. Cowles Foundation Discussion Paper No. 1644.

Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Annals of Statistics 36, 808–843.

Chen, X., Linton, O., van Keilegom,, 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71, 1591–1608.

Chen, X., Pouzo, D., 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. Journal of Econometrics 152, 46–60.

Firpo, S., 2007. Efficient semiparametric estimation of quantile treatment effects. Econometrica 75, 259–276.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315–331.

Heckman, J., Ichimura, H., Todd, P., 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261–294.

Heckman, J.J., Vytlacil, E.J., 2007. Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. VI. Elsevier Science B.V, pp. 4780–4874.

Hirano, K., Imbens, G.W., 2004. The propensity score with continuous treatments. In: Gelman, A., Meng, X.-L. (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. Wiley, New York, pp. 73–84.

Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–960.

Imai, K., van Dyk, D.A., 2004. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association 99, 854–866.

Imbens, G.W., 2000. The role of the propensity score in estimating dose–response functions. Biometrika 87, 706–710.

Imbens, G.W., Newey, W.K., Ridder, G., 2006. Mean-squared-error calculations for average treatment effects. Working Paper.

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

Lechner, M., 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Physica/Springer, Heidelberg, pp. 43–58.

Newey, W.K., 1990. Semiparametric efficiency bounds. Journal of Applied Econometrics 5, 99–135.

Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.

Newey, W.K., 1994. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics 79, 147–168.

Newey, W.K., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. IV. Elsevier Science B.V, pp. 2112–2245.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1057.

Robins, J.M., Rotnitzky, A., 1995. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association 90, 122–129.

Robins, J.M., Rotnitzky, A., Zhao, L., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846–866.

Robins, J.M., Rotnitzky, A., Zhao, L., 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90, 846–866.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Tanabe, K., Sagae, M., 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, 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., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer, New York.

van der Vaart, A.W., Wellner, J.A., 2000. Preservation theorems for Glivenko–Cantelli and uniform Glivenko–Cantelli classes. In: Giné, E., Mason, D., Wellner, J.A. (Eds.), High Dimensional Probability II. Bikhäuser, Boston, pp. 115–134.

Wooldridge, J.M., 2007. Inverse probability weighted estimation for general missing data problems. Journal of Econometrics 141, 1281–1301.