Article:

Bravo, Francesco orcid.org/0000-0002-8034-334X and Jacho-Chavez, David T. (2016) Semiparametric quasi-likelihood estimation with missing data. Communications in Statistics, Theory and Methods. pp. 1345-1369. ISSN 0361-0926

https://doi.org/10.1080/03610926.2013.863928

Semiparametric Quasi-likelihood Estimation with Missing Data

Francesco Bravo[*] David T. Jacho-Chávez[†]
University of York Emory University

Abstract

This paper develops quasi-likelihood estimation for generalized varying coefficient partially linear models when the response is not always observable. The paper considers two estimation methods and shows that under the assumption of selection on the observables the resulting estimators are asymptotically normal. As an application of these results the paper proposes a new estimator for the average treatment effect parameter. A simulation study illustrates the finite sample properties of the proposed estimators.

Keywords: Backfitting; Double Robustness; Inverse Probability Weighting; Profiling; Unconfoundness.

JEL classification: C13; C14; C21

1 Introduction

Quasi-likelihood estimation is routinely used in econometrics and statistics to estimate known index structure models for binary, counts and fractional responses, see for example McCullagh & Nelder (1989), Gourieroux, Monfort & Trognon (1984) and especially Wooldridge (2010) for a comprehensive review of models and applications). Quasi-likelihood estimation can also be used in the context of semiparametric regression models and in particular for generalized varying coefficients partially linear models. These models are semiparametric extensions of the classical generalized linear models and include many important semiparametric regression models such as the kernel generalized linear model of Fan, Heckman & Wand (1995), the generalized partially linear model of Carroll, Fan, Gijbels & Wand (1997), and the varying-coefficient model of Hastie & Tibshirani (1993) and of Cai, Fan & Li (2000). Compared to the popular partially linear model considered by Engle, Granger, Rice & Weiss (1986) and Robinson (1988) generalized varying partially linear models offer additional flexibility and allow interaction effects between covariates and the nonparametric components while avoiding the curse of dimensionality typically associated with partially linear models. Furthermore as with classical (i.e. parametric) generalized linear models using a canonical link function ensures that the final estimates have always the correct range (e.g. Logit link leads to a probability), however as opposed to classical generalized linear models the choice of the link function is less important, making them therefore more robust to potential misspecification of the conditional mean.

In this paper we consider quasi-likelihood estimation for generalized varying coefficients partially linear models when the responses are partially observable. Under the assumption of selection on the observables we propose a new estimator for the unknown parameters based on *inverse probability weighting method* (Horvitz & Thompson 1952). This method has been used for regression models with missing data, see for example Robins, Rotnitzky &

[*]Department of Economics, University of York, Heslington, York YO10 5DD, UK. E-mail: francesco.bravo@york.ac.uk. Web Page: https://sites.google.com/a/york.ac.uk/francescobravo/.

[†]Department of Economics, Emory University, Rich Building 306, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: djachocha@emory.edu. Web Page: http://userwww.service.emory.edu/~djachoc/.

Zhao (1994) and Robins & Rotnitzky (1995), in the treatment effect literature, see for example Hirano, Imbens & Ridder (2003), in nonclassical measurement error models, see for example Robins, Hsieh & Newey (1995) and Chen, Hong & Tamer (2005), attrition in panel data, see for example Wooldridge (2002), and by Wooldridge (1999) and Wooldridge (2007) for M-estimation with missing data. The probabilities of the weighting method are typically unknown and therefore have to be estimated either with parametric or with nonparametric methods. In this paper we consider the parametric approach because as opposed to the nonparametric one it does not suffer from the curse of dimensionality and it is less negatively affected by a high proportion of missing data in the sample, making it perhaps more useful from an empirical point of view. Furthermore, as noted by Wooldridge (2007), as long as the conditional mean is correctly specified and the assumption of selection on the observables holds misspecification of the parametric estimator for probabilities does not cause inconsistency of the weighted estimator for the parameters of the generalized varying coefficient partially linear estimator.

The results of this paper are rather general and can be seen as a semiparametric extension of some of the results obtained by Wooldridge (2007). The results are based on backfitting and profiling, which are the two main approaches to estimate parameters for general semiparametric models and differ in the way they deal with the infinite dimensional parameter. To be specific, backfitting involves iterating between the estimation of the infinite dimensional parameter and that of the finite dimensional one until convergence, see for example Hastie & Tibshirani (1990), Mammen, Linton & Nielsen (1999) and Opsomer (2000). Profiling involves reparameterizing the infinite dimensional parameter as a certain function of the finite dimensional parameter and then estimate simultaneously the resulting reparameterized infinite dimensional parameter as well as the finite dimensional one, see for example Severini & Staniswalis (1994), Murphy & Van der Vaart (2000) and Lam & Fan (2008). A similar procedure, albeit without reparameterization is considered by Ai & Chen (2003) for semiparametric moment conditions models. Opsomer & Ruppert (1999) and more recently Van Keilegom & Carroll (2007) compare backfitting and profiling and note that in certain situations they result in asymptotically equivalent estimators as long as different level of smoothing is applied.

The new results of the paper are the following: First we show that the proposed estimators defined as the solutions to a set of local quasi-scores are consistent. This result is based on a generalization to infinite dimensional parameters of the same approach used by Foutz (1977), and complements the standard approach based on the global concavity of the quasi-likelihood function. Second, we show that both backfitting and profiling lead to estimators that are asymptotically normal but they are not asymptotically equivalent even if we consider different level of smoothing. Third, as an application of these results we propose a new semiparametric estimator for the average treatment effect parameter. This new estimator is motivated by some recent literature in health economics (see e.g. Basu, Polsky & Manning (2008) and references therein) advocating the use of parametric generalized linear models to capture potential nonlinear effects and interactions between outcomes and covariates as well as specific structures of the outcomes. Our estimator is flexible enough to capture these important features while preserving some of the advantages of using parametric methods. Furthermore for Normal, Bernoulli and Poisson quasi-likelihoods the new estimator enjoys the so-called doubly-robust property as noted by Wooldridge (2007). Finally we use simulations to investigate the finite sample properties of the estimators based on backfitting and profiling and for the new average treatment effect estimator. The latter are compared with those based on commonly used alternatives.

The results of this paper generalize and/or complement a number of results including those obtained by Cai et al. (2000), Wooldridge (2002), Chen, Fan, Li & Zhou (2006), Lam & Fan (2008), and Wooldridge (2007) among others. The results can be used to show consistency and asymptotic normality for estimators defined as the solutions to a set of semiparametric smooth estimating equations, which could be, for example, the result of some economic theory restriction. The results can also be used to characterize the asymptotic behavior of the solutions to a set of local first order conditions that are often easier to find than those corresponding to global maximum in models with an infinite dimensional parameters.

The rest of the paper is structured as follows: Section 2 introduces the basic model and discusses the two general estimation approaches. Section 3 contains the main theoretical results. Section 4 considers average treatment effect estimation and proposes a novel estimator based on the results of the previous sections. Section 5 illustrates the results with three examples and related simulations. Finally Section 6 contains some concluding remarks. An Appendix contains all the proofs.

2 The Model and the Estimators

The model we consider is a generalized varying coefficient partially linear model (GVCPL henceforth)

$$E\left(Y|X\right) = g^{-1}[X_1^\top \beta_0 + X_2^\top \alpha_0\left(X_3\right)], \tag{1}$$

where $g^{-1}\left(\cdot\right)$ is the inverse function of the known link function $g\left(\cdot\right)$, X_1 and X_2 are respectively a k_1 and k_2-dimensional vectors, β_0 is a vector of unknown parameters, $\alpha\left(\cdot\right)$ is a vector of unknown smooth functions, and X_3 is a scalar covariate. GVCPL includes a number of important semiparametric regression models including the kernel generalized linear model of Fan et al. (1995) (specification (1) without X_1, X_2, and β), the generalized partially linear model of Carroll et al. (1997) (specification (1) with $X_2 = 1$), the varying-coefficient model of Hastie & Tibshirani (1993) and of Cai et al. (2000) (specification (1) without X_1 and β).

Let $W_i^\top = \left(Y_i, X_i^\top\right)$ $(i = 1, ..., n)$ denote an i.i.d. sample from $W^\top = \left(Y, X^\top\right)$; when the response Y_i is always observable the unknown parameters in (1) can be estimated by the same quasi-likelihood approach used by Severini & Staniswalis (1994), Fan et al. (1995), Carroll et al. (1997) and many others. To be specific, let $Q\left(g^{-1}\left(\cdot\right), Y\right)$ denote a quasi-likelihood that is defined by

$$\frac{\partial Q\left(\mu, Y\right)}{\partial \mu} = \frac{Y - \mu}{V\left(\mu\right)},$$

where the variance function $V\left(\cdot\right)$ is known and may depend on an unknown scale parameter σ^2 (see e.g. McCullagh & Nelder (1989) for examples), and let

$$\alpha_{0j}\left(v\right) = a_j + b_j\left(v - u\right) \quad j = 1, ..., k_2$$

for v in a neighbourhood of u and $a_j = \alpha_j\left(u\right)$, $b_j = \alpha_j'\left(u\right)$ denote a linear[1] approximation for $\alpha_j\left(v\right)$. Then for a fixed x_3

$$Q_n\left(\beta, \alpha, x_3\right) := \sum_{i=1}^{n} Q\left[g^{-1}\left(X_{1i}^\top \beta + X_{2i}^\top\left(a + b\left(X_{3i} - x_3\right)\right)\right), Y_i\right] K_{h_1}\left(X_{3i} - x_3\right), \tag{2}$$

where defines a local quasi-likelihood function that can be used to estimate $\alpha_0\left(\cdot\right)$ and β_0 using either the backfitting or profiling method. If however, some of the responses are missing and this fact is not taken into account into the estimation process, both approaches might result in inconsistent estimators.

We characterize missing data with a binary indicator $T = \{0, 1\}$ so that we have an i.i.d. sample $\left(W_i^\top, T_i\right)$ from $\left(W^\top, T\right)$ and the Y_i are not observed if T_i is zero. The key of our results is that the covariates are good predictors of the selection as the following assumption specifies:

S1 The vector W is always observed when $T = 1$;

[1] The results of the paper can be easily extended to the case of a polynomial approximation. The only change would be in Lemma A.1 in Appendix A in which the order of approximation would change to $h^{(p+1)q}$ where p is the degree of the polynomial approximation. As a result the order of the bias in Theorems 1 and 3 would also change to h^{p+1}.

S2 (i) $Y \perp T|X$, (ii) $0 < \Pr(T=1|X) \leq 1$.

Assumption S2(i) corresponds to the *missing at random* in the statistical literature, and it is related to the so-called *unconfoundness* in the programme evaluation literature. A fundamental implication of S2 is that if the selection probabilities $\pi(X_i)$ were known, then the generalized varying coefficient partially linear model specification (1) for the missing Y's can be recovered by weighting the selected observations by the inverse of the probability of selection. This suggests the following inverse probability weighting (IPW henceforth) modification of (2)

$$Q_n(\beta, \alpha, \widehat{\pi}, x_3) := \sum_{i=1}^{n} \frac{T_i}{\widehat{\pi}(X_i)} Q\left[g^{-1}\left(X_{1i}^\top \beta + X_{2i}^\top (a + b(X_{3i} - x_3))\right), Y_i\right] K_{h_1}(X_{3i} - x_3), \tag{3}$$

where $K_{h_1}(\cdot) = K(\cdot/h_1)$, $K(\cdot)$ is a kernel function, $h_1 =: h_1(n)$ is the bandwidth and the $\widehat{\pi}(X_i)$'s are consistent estimates of the typically unknown selection probabilities $\pi(X_i)$. Also let

$$Q_n(\beta, \alpha, \widehat{\pi}) := \sum_{i=1}^{n} \frac{T_i}{\widehat{\pi}(X_i)} Q\left[g^{-1}\left(X_{1i}^\top \beta + X_{2i}^\top \alpha(X_{3i})\right), Y_i\right] \tag{4}$$

denote the inverse probability weighting quasi-likelihood.

The estimation of the unknown $\alpha_0(\cdot)$ and β_0 is based on both (3) and (4), and can be carried out using either the backfitting or profiling algorithm. The estimators can be defined either as maximizers of (3) and (4) or as the solution $\widehat{\beta}$ and $\widehat{\alpha}$ to the quasi-score equations defined by the first order conditions from (3) and (4), that is

$$\begin{aligned} \partial Q_n(\beta, \alpha, \widehat{\pi}, x_3) / \partial(\beta^\top, a^\top, b^\top)^\top &= 0, \\ \partial Q_n(\beta, \widehat{\alpha}, \widehat{\pi}) / \partial\beta &= 0. \end{aligned} \tag{5}$$

The results of the paper are valid for both cases and with simple modifications in the proofs also for estimators $\widehat{\beta}$ and $\widehat{\alpha}$ defined as the solution of

$$\sum_{i=1}^{n} \frac{T_i}{\widehat{\pi}(X_i)} \varphi\left(Y_i; X_{1i}^\top \beta + X_{2i}^\top (a + b(X_{3i} - x_3))\right) \left[X_{1i}^\top, X_{2i}^\top \otimes [1, (X_{3i} - x_3)]^\top K_{h_1}(X_{3i} - x_3)\right]^\top = 0,$$

$$\sum_{i=1}^{n} \frac{T_i}{\widehat{\pi}(X_i)} \varphi\left(Y_i; X_{1i}^\top \beta + X_{2i}^\top \widehat{\alpha}\right) X_{1i} = 0,$$

where φ is a known scalar function. In what follows we consider the case of estimators defined as solution to quasi-score equations (5).

2.1 Backfitting Estimation

The idea of backfitting, often called two-step procedure, is to use first use a set of local first order conditions (5) based on (3) to obtain local estimates of all the unknown parameters, and then to use the global set of first order condition (5) based on (4) to improve the estimation of the finite dimensional parameter. To be specific, the procedure consists of the following steps:

B1 Either find $\widehat{\beta}$, $\widehat{a}$ and $\widehat{b}$ that solve the $(k_1 + 2k_2) \times 1$ vector of local first-order conditions $\partial Q_n(\beta, \alpha, \widehat{\pi}, x_3)/\partial(\beta^\top, a^\top, b^\top)^\top = 0$, or for a fixed $\overline{\beta}$ find $\widehat{a}$ and $\widehat{b}$ that solve the $2k_2 \times 1$ vector of local first-order conditions $\partial Q_n(\beta, \alpha, \widehat{\pi}, x_3) / \partial(a^\top, b^\top)^\top = 0$;

B2 Let $\widehat{\alpha} := \widehat{a}$ found at B1; find $\widehat{\beta}$ that solves the $k_1 \times 1$ vector of first-order conditions $\partial Q_n(\beta, \widehat{\alpha}, \widehat{\pi}) / \partial\beta = 0$.

The above two steps can then be iterated until convergence if needed. Note that the final estimate $\widehat{\alpha}$ obtained at the end of B2 can be improved by considering a third-step which involves solving the $k_2 \times 1$ vector of local first-order conditions $\partial Q_n(\widehat{\beta}, \alpha, \widehat{\pi}, x_3)/\partial a = 0$. Unless the functions α are of particular interest, this last step may be omitted.

Backfitting delivers $n^{1/2}$-consistent estimators for β_0; however, in order to achieve the $n^{1/2}$-rate, they require undersmoothing (see Theorem (3.2) below for details). To avoid undersmoothing, we propose an alternative method that is computationally more involved.

2.2 Profiling Estimation

The method of profiling, or one-step estimation, is based on the notion of least favourable curve that is defined to be the parameterization $\alpha_\beta(\cdot)$ of $\alpha(\cdot)$ which has the smallest possible (Fisher) information for β and such that at β_0, $\alpha_{\beta_0}(\cdot) = \alpha(\cdot)$. As long as this curve can be estimated, it can be used to compute the least favorable quasi-score for β, which coincides with the efficient one. The procedure consists of the following steps:

P1 For a given β let $\widehat{\alpha}_\beta := \widehat{a}$ that solve the $2k_2 \times 1$ vector of local first-order conditions $\partial Q_n(\beta, \alpha_\beta, \widehat{\pi}, x_3)/\partial(a^\top, b^\top)^\top = 0$;

P2 Find $\widehat{\beta}$ that solves the $k_1 \times 1$ vector of first-order conditions $\partial Q_n(\beta, \widehat{\alpha}_\beta, \widehat{\pi})/\partial\beta = 0$.

It is important to note that the IPW profile quasi-score for β is

$$\frac{\partial Q_n(\beta, \widehat{\alpha}_\beta, \widehat{\pi})}{\partial\beta} = \sum_{i=1}^n \frac{T_i}{\widehat{\pi}_i(X_i)} q_1\left(g^{-1}\left(X_{1i}^\top\beta + X_{2i}^\top\widehat{\alpha}_\beta(X_{3i})\right), Y_i\right)\left(X_{1i} + \left(\frac{\partial\widehat{\alpha}_\beta(X_{3i})}{\partial\beta^\top}\right)^\top X_{2i}\right),$$

where $q_1(x,y) = \partial Q\left[g^{-1}(x), y\right]/\partial x$. This involves the difficult computation of the $k_2 \times k_1$ matrix $\partial\widehat{\alpha}_\beta(X_{3i})/\partial\beta^\top$ (the so-called least favorable direction) using, for example, numerical derivatives. To overcome this difficulty we can use as in Severini & Staniswalis (1994) and Lam & Fan (2008) a simple estimator that is based on a local version of its explicit expression (given in $(A-22)$ of the Appendix) that is

$$\begin{aligned}\partial\widehat{\alpha}_\beta(x_3)/\partial\partial\beta^\top = &\left(\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\widehat{\pi}_i(X_i)} q_2\left(g^{-1}(X_{1i}^\top\beta + X_{2i}^\top\widehat{\alpha}_\beta(X_{3i})), Y_i\right) X_{2i}X_{2i}^\top K_h(X_{3i} - x_3)\right)^{-1} \times \\ &\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\widehat{\pi}_i(X_i)} q_2\left(g^{-1}(X_{1i}^\top\beta + X_{2i}^\top\widehat{\alpha}_\beta(X_{3i})), Y_i\right) X_{2i}X_{1i}^\top K_h(X_{3i} - x_3),\end{aligned}$$

where $q_2(x,y) = \partial^2 Q\left[g^{-1}(x), y\right]/\partial x^2$; see Section 4 for further details on the computation of this estimator.

3 Main Results

We begin this section by introducing some auxiliary notation and the following convention: A quantity with a superscript π indicates that the relevant expectation is weighted by the inverse of the propensity score, so for example $\Delta(x) = E[g(x)]$ and $\Delta_\pi(x) = E[g(x)/\pi(x)]$. For $j = 0, 1, \ldots$ let $q_j(x,y) = \partial^j Q\left[g^{-1}(x), y\right]/\partial x^j$, $\rho_j(x) = \left(\partial g^{-1}(x)/\partial x\right)^j/\mathrm{var}(y|x)$, $\kappa_j = \int t^j K(t)\,dt$, $\upsilon_j = \int t^j K^2(t)\,dt$ and $\eta = X_1^\top\beta + X_2^\top\alpha(X_3)$. Let $B(\beta_0)$ denote an open neighbourhood of β_0; and assume that:

A1 The random variable X_3 has compact support $\mathcal{X}_3$, and its density $f(x_3)$ is twice continuously differentiable and is uniformly bounded away from 0 on $\mathcal{X}_3$;

A2 The functions $\alpha_j''(\cdot)$ $(j = 1, \ldots, k_2)$ are continuous in $\mathcal{X}_3$; the functions $V(\cdot)$ and $g(\cdot)$ are, respectively, twice and three times continuously differentiable in $B(\beta_0)$;

A3 The matrices $E\left[q_1^2(\eta, Y) X_j X_k^\top | X_3 = x_3\right]$ $(j, k = 1, 2)$ are twice continuously differentiable in $x_3 \in \mathcal{X}_3$; the least favourable curve $\alpha_\beta(\cdot)$ is three times continuously differentiable in $x_3 \in \mathcal{X}_3$ and $B(\beta_0)$;

A4 The matrices $E\left\{\rho_2(\eta_0) X_j X_j^\top | X_3 = x_3\right\}$ $(j = 1, 2)$ are nonsingular, $E\{\rho_2(\eta_0) X_j X_j^\top | X_3 = x_3\}$ are negative definite for each $x_3 \in \mathcal{X}_3$, $E\left[\rho_2(\eta_0) X_j X_j^\top\right]$ are negative definite, and for some $\gamma > 0$, $E[\| Tq_1(\eta_0, Y)\left[X_1^\top, X_2^\top\right]^\top / \pi(X) \|^{2+\gamma}] < \infty$, $E[\| \rho_2(\eta_0) X_j X_j^\top \|^{2+\gamma}] < \infty$, $E[\| \rho_2(\eta_0) X_j X_j^\top \|^{2+\gamma} | X_3 = x_3] < \infty$, $E[\sup_{x_3 \in \mathcal{X}_3, \beta \in B(\beta_0)} \| q_3(\eta) X_j X_j^\top X_{jl} \|] < \infty$, $(j = 1, 2, l = 1, \ldots, k = k_1 + k_2)$;

A5 The kernel K is a bounded symmetric density function with bounded support.

Assumptions A1-A5 are standard moment and smoothness conditions in the literature on nonparametric/semiparametric estimation with quasi-likelihood functions, see e.g. Severini & Staniswalis (1994), Carroll et al. (1997) and Cai et al. (2000). Note that we do not require the quasi-likelihood to be globally concave and thus we allow for possible misspecification of the variance. These conditions ensure the consistency and asymptotic normality of a unique solution to the quasi-score equations (5).

The computation of $\widehat{\pi}(X_i)$ can be done using binary maximum likelihood under the following additional standard regularity conditions. Let $\pi(X, \gamma)$ denote a parametric model for $\pi(X)$ where $\gamma \in \Gamma \subset \mathbb{R}^{d_\gamma}$, and assume that

A6 (i) $\pi(X, \gamma) > 0$ for all X and all $\gamma \in \Gamma$, (ii) $\pi(X, \gamma_0) = \pi(X)$, (iii) $\widehat{\gamma}$ has the following stochastic expansion:

$$n^{1/2}(\widehat{\gamma} - \gamma_0) = I^{-1}(\gamma_0) \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{\partial \pi_i(\gamma_0)}{\partial \gamma} \frac{(T_i - \pi_i(\gamma_0))}{\pi_i(\gamma_0)(1 - \pi_i(\gamma_0))} + o_p(1).$$

Let

$$\Sigma(\alpha, \beta, x_3) = E\left\{\rho_2(X_1^\top \beta + X_2^\top \alpha(X_3))[X_1^\top, X_2^\top]^\top \left[X_1^\top, X_2^\top\right] | X_3 = x_3\right\},$$
$$\Gamma(\alpha, \beta, x_3) = E\left\{\rho_2(X_1^\top \beta + X_2^\top \alpha(X_3))[X_1 X_2^\top, X_2 X_2^\top]^\top \alpha''(X_3) | X_3 = x_3\right\}.$$

The following theorem establishes the asymptotic distribution of the local estimators used in the backfitting procedure described in step B1.

Theorem 3.1 *Under S1, S2 and A1-A6. Then*

$$(nh_1)^{1/2}\left[\begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\alpha}(x_3) - \alpha_0(x_3) \end{pmatrix} - \frac{h_1^2 b_1(\alpha_0, \beta_0, x_3)}{2}\right] \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{v_0 A(\beta_0, \alpha_0, \pi, x_3)}{f(x_3)}\right),$$

where

$$b_1(\alpha, \beta, x_3) = \kappa_2 \Sigma(\alpha, \beta, x_3)^{-1} \Gamma(\alpha, \beta, x_3),$$
$$A(\beta, \alpha, \pi, x_3) = \Sigma(\alpha, \beta, x_3)^{-1} \Sigma_\pi(\alpha, \beta, x_3) \Sigma(\alpha, \beta, x_3)^{-1}.$$

Theorem 3.1 is a direct generalization of results of Chen et al. (2006). It can be used to characterize the distribution of estimators for semiparametric quasi-likelihood models, and semiparametric estimating equations models with data missing at random.

For $j,k=1,2$ let

$$\begin{aligned}
B_{jk}(\alpha,\beta,x_3) &= E\left[\rho_2\left(X_1^\top\beta + X_2^\top\alpha(X_3)\right)X_jX_k^\top | X_3 = x_3\right],\\
B_{jk}(\alpha,\beta) &= E\left[\rho_2\left(X_1^\top\beta + X_2^\top\alpha(X_3)\right)X_jX_k^\top\right],\\
D(\alpha,\beta,x_3) &= B_{11\pi}(\alpha,\beta,x_3)B_{11}(\alpha,\beta,x_3)^{-1}B_{12}(\alpha,\beta,x_3)\Delta(\alpha,\beta,x_3)B_{21}(\alpha,\beta,x_3) - \\
&\quad B_{12}(\alpha,\beta,x_3)\Delta(\alpha,\beta,x_3),\\
\Delta(\alpha,\beta,x_3) &= B_{22}(\alpha,\beta,x_3) - B_{21}(\alpha,\beta,x_3)B_{11}(\alpha,\beta,x_3)^{-1}B_{12}(\alpha,\beta,x_3).
\end{aligned}$$

The following theorem establishes the $n^{1/2}$-consistency of $\widehat{\beta}$ obtained in step B2.

Theorem 3.2 *Under S1, S2 and A1-A6. If $nh_1^4 \to 0$, then*

$$n^{1/2}(\widehat{\beta}-\beta_0) \xrightarrow{d} N\left(0, B^b(\alpha_0,\beta_0,\pi)\right),$$

where

$$\begin{aligned}
B^b(\alpha_0,\beta_0,\pi) &= B_{11}(\alpha_0,\beta_0)^{-1}\Omega^b(\alpha_0,\beta_0,\pi)B_{11}(\alpha_0,\beta_0)^{-1},\\
\Omega^b(\alpha_0,\beta_0,\pi) &= B_{11\pi}(\alpha_0,\beta_0) + E\left[D(\alpha,\beta,X_3)\right] + E\left[D(\alpha,\beta,X_3)\right]^\top + \\
&\quad E\Big[B_{12}(\alpha_0,\beta_0,X_3)S_\alpha\Sigma_\kappa(\alpha_0,\beta_0,X_3)^{-1}\times\\
&\quad \Sigma_\pi(\alpha_0,\beta_0,X_3)\Sigma_\kappa(\alpha_0,\beta_0,X_3)^{-1}S_\alpha^\top B_{12}(\alpha_0,\beta_0,X_3)^\top\Big],
\end{aligned}$$

*where $S_\alpha = [0, I, 0]$ and $\Sigma_\kappa(\alpha_0,\beta_0,x_3) =$*diag$[\Sigma(\alpha,\beta,x_3), \kappa_2 B_{22}(\alpha,\beta,x_3)]$.

Theorem 3.2 shows that to achieve $n^{1/2}$-consistency using the backfitting method we need to undersmooth. This is typical for a number of semiparametric models as noted for example by Van Keilegom & Carroll (2007). Note that if one is interested in α_0, then because of the undersmoothing it might be desirable to consider a third estimation which uses $\widehat{\beta}$ found in step B2, and is defined by the local quasi-score equations $\partial Q_n(\widehat{\beta},\alpha,\widehat{\pi},x_3))/\partial(a^\top,b^\top)^\top = 0$. Note also that since $\widehat{\beta}$ is $n^{1/2}$-consistent this estimation can be carried out as if β was known. This result is summarized in the following theorem. Let

$$\begin{aligned}
\Phi(\alpha,\beta,x_3) &= E\left[\rho_2(X_1^\top\beta + X_2^\top\alpha(X_3))[X_2X_2^\top, 0^\top]^\top\alpha''(X_3)|X_3 = x_3\right],\\
\Psi_\kappa(\alpha,\beta,x_3) &= \mathrm{diag}\left[B_{22}(\alpha,\beta,x_3), \kappa_2 B_{22}(\alpha,\beta,x_3)\right],\\
\Psi_\upsilon(\alpha,\beta,x_3) &= \mathrm{diag}\left[\upsilon_0 B_{22}(\alpha,\beta,x_3), \upsilon_2 B_{22}(\alpha,\beta,x_3)\right].
\end{aligned}$$

Theorem 3.3 *Under S1-S2 and A1-A6. Then*

$$(nh_2)^{1/2}\left[\begin{pmatrix}\widehat{\alpha}(x_3)-\alpha_0(x_3)\\ h_2\left(\widehat{\alpha}'(x_3)-\alpha_0'(x_3)\right)\end{pmatrix} - \frac{h_2^2}{2}b_2(\alpha,\beta,x_3)\right] \xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix}, \frac{C(\alpha_0,\beta_0,x_3)}{f(x_3)}\right),$$

where

$$\begin{aligned}
b_2(\alpha,\beta,x_3) &= \kappa_2\Psi_\kappa(\alpha,\beta,x_3)^{-1}\Phi(\alpha,\beta,x_3),\\
C(\alpha,\beta,x_3) &= \Psi_\kappa(\alpha,\beta,x_3)^{-1}\Psi_{\pi\upsilon}(\alpha,\beta,x_3)\Psi_\kappa(\alpha,\beta,x_3)^{-1}.
\end{aligned}$$

We now establish the $n^{1/2}$-consistency of the estimator $\widehat{\beta}$ based on P2 step of the profile algorithm. Note that unlike the backfitting approach, there is no need to undersmooth here to achieve $n^{1/2}$-consistency.

Theorem 3.4 *Under S1-S2 and A1-A6. Then*

$$n^{1/2}(\widehat{\beta} - \beta_0) \xrightarrow{d} N\left(0, B^p\left(\alpha_0, \beta_0, \pi\right)\right)$$

where

$$\begin{aligned}
&B^p\left(\alpha_0, \beta_0, \pi\right) = \Xi\left(\alpha_0, \beta_0\right)^{-1} \Omega^p\left(\alpha_0, \beta_0, \pi\right) \Xi\left(\alpha_0, \beta_0\right)^{-1}, \\
&\Xi\left(\alpha_0, \beta_0\right) = B_{11}\left(\alpha_0, \beta_0\right) - E\left[B_{12}\left(\alpha_0, \beta_0, X_3\right) B_{22}\left(\alpha_0, \beta_0, X_3\right)^{-1} B_{12}\left(\alpha_0, \beta_0, X_3\right)^{\top}\right], \\
&\Omega^p\left(\alpha_0, \beta_0, \pi\right) = B_{11\pi}\left(\alpha_0, \beta_0\right) - E\left[B_{12\pi}\left(\alpha_0, \beta_0, X_3\right) B_{22}\left(\alpha_0, \beta_0, X_3\right)^{-1} B_{12}\left(\alpha_0, \beta_0, X_3\right)^{\top}\right] - \\
&E\left[B_{12}\left(\alpha_0, \beta_0, X_3\right) B_{22}\left(\alpha_0, \beta_0, X_3\right)^{-1} B_{12\pi}\left(\alpha_0, \beta_0, X_3\right)^{\top}\right] + \\
&E\left[B_{12}\left(\alpha_0, \beta_0, X_3\right) B_{22}\left(\alpha_0, \beta_0, X_3\right)^{-1} B_{22\pi}\left(\alpha_0, \beta_0, X_3\right) B_{22}\left(\alpha_0, \beta_0, X_3\right)^{-1} B_{12}\left(\alpha_0, \beta_0, X_3\right)^{\top}\right].
\end{aligned}$$

4 Average Treatment Effect Estimation

As an application of the results of the previous section we consider the problem of estimating the average treatment effect parameter, see e.g. Imbens (2004) for a recent review. We propose a novel semiparametric estimator that is a middle ground between the parametric specifications recently used in some health economics literature (see e.g. Basu et al. (2008)) and the fully nonparametric approach of Hahn (1998) and Hirano et al. (2003). The estimator combines the regression adjustment approach with the GVCPL specification of the conditional mean, and enjoy a somewhat stronger version of the same double robustness property noted by Wooldridge (2007), because of the semiparametric specification of the conditional mean of the outcomes as opposed to the fully parametric one proposed by Wooldridge (2007).

We follow the standard potential-outcome notation and use $Y(1)$ and $Y(0)$ to denote the potential outcome for an experimental unit with and without the treatment, which is indicated by the dummy variable $T \in \{0, 1\}$. We are interested in the *average treatment effect* parameter[2]

$$\tau_0 = E\left[Y(1) - Y(0)\right], \tag{6}$$

As in Section 2 let $\{W_i^{\top}, T_i\}_{i=1}^{n}$ be an i.i.d. sample and let

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0),$$

denote the realized outcome. Assume that

S2* (i) $Y(1), Y(0) \perp T|X$, (ii) $0 < \Pr(T = 1|X) < 1$;

S3 $E\left[Y(\delta)|X\right] = g^{-1}\left(X_1^{\top}\beta_0^{\delta} + X_2^{\top}\alpha_0^{\delta}(X_3)\right)$ for $\delta = 0, 1$.

Assumptions S2*(i) and S3 imply that τ_0 can be estimated by the sample analogue of the mean regressions difference

$$\tau_0 = E\left[g^{-1}\left(X_1^{\top}\beta_0^{1} + X_2^{\top}\alpha_0^{1}(X_3)\right) - g^{-1}\left(X_1^{\top}\beta_0^{0} + X_2^{\top}\alpha_0^{0}(X_3)\right)\right], \tag{7}$$

that is

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[g^{-1}\left(X_{1i}^{\top}\widehat{\beta}^{1} + X_{2i}^{\top}\widehat{\alpha}^{1}(X_{3i})\right) - g^{-1}\left(X_{1i}^{\top}\widehat{\beta}^{0} + X_{2i}^{\top}\widehat{\alpha}^{0}(X_{3i})\right)\right], \tag{8}$$

[2] Although similar results can also be obtained for the so-called *average treatment effect on the treated* parameter

$$\tau_{0,t} = E\left[Y(1) - Y(0)|T = 1\right].$$

where $\widehat{\beta}^{\delta}$ and $\widehat{\alpha}^{\delta}(\cdot)$ are the solutions to (5) with $T_i/\widehat{\pi}_i$ and $(1-T_i)/(1-\widehat{\pi}_i)$ respectively for $\delta=1$ and $\delta=0$ computed using both backfitting and profiling methods. Let $S_\alpha=[0,I,0]$, $\pi_\delta=\delta\pi+(1-\delta)(1-\pi)$ and

$$\begin{aligned}
G_1(\alpha^{\delta},\beta^{\delta}) &= E\left[\frac{\partial g^{-1}\left(X_1^\top\beta^{\delta}+X_2^\top\alpha^1(X_3)\right)X_1}{\partial\left(\beta^{\delta\top},\alpha^{\delta\top}\right)^\top}\right],\\
G_2(\alpha^{\delta},\beta^{\delta},X_3) &= E\left[\frac{\partial g^{-1}\left(X_1^\top\beta^{\delta}+X_2^\top\alpha^1(X_3)\right)X_2}{\partial\left(\beta^{\delta\top},\alpha^{\delta\top}\right)^\top}|X_3\right],\\
F(\alpha^{\delta},\beta^{\delta},X_3) &= S_\alpha\Sigma_\kappa^{-1}(\alpha^{\delta},\beta^{\delta},X_3)\Sigma_{\pi_\delta}(\alpha^{\delta},\beta^{\delta},X_3)S_\alpha\Sigma_\kappa^{-1}(\alpha^{\delta},\beta^{\delta},X_3).
\end{aligned}$$

Theorem 4.1 *(I) Under S1, S2*, S3, A1-A6 and if $nh_1^4\to0$ then for the backfitting method*

$$n^{1/2}(\widehat{\tau}^b-\tau_0)\xrightarrow{d}N\left(0,V^b(\alpha_0,\beta_0)\right),$$

where

$$\begin{aligned}
&V^b(\alpha_0,\beta_0)=\operatorname{var}\left[g^{-1}\left(X_1^\top\beta_0^1+X_2^\top\alpha_0^1(X_3)\right)-g^{-1}\left(X_1^\top\beta_0^0+X_2^\top\alpha_0^0(X_3)\right)\right]+\\
&\sum_{\delta=1,0}\left[\Lambda_{1\delta}^b(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{2\delta}^b(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{3\delta}^b(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{3\delta}^{b\top}(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{4\delta}^b(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{4\delta}^{b\top}(\alpha_0^\delta,\beta_0^\delta)\right],
\end{aligned}$$

and

$$\begin{aligned}
\Lambda_{1\delta}^b(\alpha_0^\delta,\beta_0^\delta) &= G_1^\top(\alpha_0^\delta,\beta_0^\delta)B^b(\alpha_0^\delta,\beta_0^\delta,\pi_\delta)G_1(\alpha_0^\delta,\beta_0^\delta),\\
\Lambda_{2\delta}^b(\alpha_0^\delta,\beta_0^\delta) &= E\left[G_2^\top(\alpha_0^\delta,\beta_0^\delta,X_3)F(\alpha_0^\delta,\beta_0^\delta,X_3)G_2(\alpha_0^\delta,\beta_0^\delta,X_3)\right],\\
\Lambda_{3\delta}^b(\alpha_0^\delta,\beta_0^\delta) &= G_1(\alpha_0^\delta,\beta_0^\delta)B_{11}^{-1}(\alpha_0,\beta_0)E\left[-B_{11\pi_\delta}(\alpha_0^1,\beta_0^1,X_3)B_{11}^{-1}(\alpha_0^1,\beta_0^1,X_3)B_{12}(\alpha_0^1,\beta_0^1,X_3)\times\right.\\
&\qquad\left.\Delta(\alpha_0^1,\beta_0^1,X_3)G_2(\alpha_0^1,\beta_0^1,X_3)+B_{12}(\alpha_0^1,\beta_0^1,X_3)\Delta(\alpha_0^1,\beta_0^1,X_3)G_2(\alpha_0^1,\beta_0^1,X_3)\right]\\
\Lambda_{4\delta}^b(\alpha_0^\delta,\beta_0^\delta) &= -G_1(\alpha_0^\delta,\beta_0^\delta)B_{11}^{-1}(\alpha_0,\beta_0)E\left[B_{12}(\alpha_0,\beta_0,X_3)F(\alpha_0^\delta,\beta_0^\delta,X_3)G_2(\alpha_0^\delta,\beta_0^\delta,X_3)\right].
\end{aligned}$$

(II) Under S1, S2, S3 and A1-A6, then for the profiling method*

$$n^{1/2}(\widehat{\tau}^p-\tau_0)\xrightarrow{d}N\left(0,V^p(\alpha_0,\beta_0)\right),$$

where

$$\begin{aligned}
&V^p(\alpha_0,\beta_0)=\operatorname{var}\left[g^{-1}\left(X_1^\top\beta_0^1+X_2^\top\alpha_0^1(X_3)\right)-g^{-1}\left(X_1^\top\beta_0^0+X_2^\top\alpha_0^0(X_3)\right)\right]+\\
&\sum_{\delta=1,0}\left[\Lambda_{1\delta}^p(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{2\delta}^p(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{3\delta}^p(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{3\delta}^{p\top}(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{4\delta}^p(\alpha_0^\delta,\beta_0^\delta)+\Lambda_{4\delta}^{p\top}(\alpha_0^\delta,\beta_0^\delta)\right],
\end{aligned}$$

and

$$\begin{aligned}
\Lambda_{1\delta}^p(\alpha_0^\delta,\beta_0^\delta) &= G_1^\top(\alpha_0^\delta,\beta_0^\delta)B^p(\alpha_0^\delta,\beta_0^\delta,\pi_\delta)G_1(\alpha_0^\delta,\beta_0^\delta),\quad \Lambda_{2\delta}^p(\alpha_0^\delta,\beta_0^\delta)=\Lambda_{2\delta}^b(\alpha_0^\delta,\beta_0^\delta),\\
\Lambda_{3\delta}^p(\alpha_0^\delta,\beta_0^\delta) &= G_1^\top(\alpha_0^\delta,\beta_0^\delta)\Xi^{-1}(\alpha_0^\delta,\beta_0^\delta)E\left[B_{11\pi_\delta}(\alpha_0^\delta,\beta_0^\delta,X_3)-B_{12}(\alpha_0^\delta,\beta_0^\delta,X_3)B_{22}^{-1}(\alpha_0^\delta,\beta_0^\delta,X_3)\times\right.\\
&\qquad\left.B_{21\pi_\delta}(\alpha_0^\delta,\beta_0^\delta,X_3)\right]B_{11}^{-1}(\alpha_0^1,\beta_0^1,X_3)\Delta(\alpha_0^1,\beta_0^1,X_3),\\
\Lambda_{4\delta}^p(\alpha_0^\delta,\beta_0^\delta) &= G_1^\top(\alpha_0^\delta,\beta_0^\delta)\Xi^{-1}(\alpha_0^\delta,\beta_0^\delta)E\left[B_{12\pi}(\alpha_0^1,\beta_0^1,X_3)-B_{12}(\alpha_0^1,\beta_0^1,X_3)B_{22}^{-1}(\alpha_0^1,\beta_0^1,X_3)\times\right.\\
&\qquad\left.B_{22\pi}(\alpha_0^1,\beta_0^1,X_3)\right]\Delta(\alpha_0^1,\beta_0^1,X_3).
\end{aligned}$$

5 Numerical Experiments

In this section we illustrate the results of the previous section with some examples and simulations. We consider three models commonly used in quasi-likelihood estimation of (1), namely the Normal, the Poisson and the Logit, for which the link functions are given, respectively, by

$$\begin{array}{ll} \text{Normal:} & g(u) = u, \\ \text{Poisson:} & g(u) = \ln(u), \\ \text{Logit:} & g(u) = \ln\left(\frac{u}{1-u}\right). \end{array}$$

We first consider two separate cases corresponding to $\delta = 0$ and 1, and then use the same two cases to consider average treatment effect estimation.

For the Normal design, we set $X_2 \sim U[-2,2]$, $X_3 \sim U[-2,2]$ and $X_1 = [X_{11}, X_{12}]^\top$ with $X_{11} \sim U[-1,0]$ and $X_{12} \sim U[0,1]$, where we have used the notation $V \sim U[a,b]$ to denote that V follows an uniform distribution between a and b. We set $\beta_0^1 = [\beta_{10}^1, \beta_{20}^1]^\top = [1,3]^\top$, $\beta_0^0 = [\beta_{10}^0, \beta_{20}^0]^\top = [1,1]^\top$, $\alpha_0^1(u) = 3\cos(2u)$, and $\alpha_0^0(u) = 3\sin(2u)$. We also set $T = I\{X^\top\theta_0 - u > 0\}$, where $I\{\cdot\}$ is the standard indicator function that equals one if its argument is true and zero otherwise, $X = [X_1^\top, X_2, X_3]^\top$, $\theta_0 = [1/4, 1/4, 1/4, 1/4]^\top$ and u follows a standard normal distribution. For this specification the proportion of missing responses is 0.50.

In the Poisson and Logit designs, we set $\beta_{10}^1 = \beta_{10}^0 = 0$, $\beta_{20}^1 = \beta_{20}^0 = -1$, $\alpha_0^1(u) = \alpha_0^0(u) = \sin(\pi u)$. The binary indicator is set as $T = I\{X^\top\theta_0 - u > 0\}$, where u is a standard normal as in the previous case but with $\theta_0 = [0, 1/3, 1/3, 1/3]^\top$. For both designs we set $X_2 \sim Beta[2,4]$, $X_3 \sim U[-1,1]$ and $X_{12} \sim 2 \times Beta[4,2]$, where $Beta[a,b]$ denotes a Beta distribution with shape parameters a and b. For this specification the proportion of missing responses is approximately 0.30. Note also that the average treatment effect parameter τ_0 is 0 by construction.

In each of 500 replications we generated n pseudo-random numbers from these three designs for $n \in \{100, 200, 400\}$. For $\delta = 1$ and $\delta = 0$, we implement the estimators discussed in Section 2.1 and Section 2.2, using a second order Gaussian kernel with bandwidth chosen by Silverman's rule-of-thumb and a correctly specified Probit model for π_i in each replication.

Tables 1, 3 and 5 report the median bias (Bias) and the interquartile range (IQR) for the backfitting and profile estimators of β_{20}^δ - 'Backfitting' and 'Profile' respectively in the tables. The tables also report the root average mean square error (RAMSE) of the backfitting and profile estimators of the nonparametric component α_0^δ. We first note that all finite sample biases and interquartile range decreases as the sample size increases uniformly across all specifications and designs. More interestingly we observe that the profile estimator of β_{20}^δ outperforms that based on backfitting across all designs and δs both in terms of absolute median bias and spread in all of the three specifications. The improvement is particularly evident in the case of the Poisson specification, where the finite sample bias of the profile estimator is roughly half that of the backfitting one for $\delta = 0$, and up to 10 times less for $\delta = 1$ and $n = 200$. The finite sample interquartile range is also considerably smaller especially for $n = 100$, where it is roughly a quarter for $\delta = 0$ and $\delta = 1$. The profile estimator of α_0^δ also outperforms its backfitting counterpart in terms of RAMSE across all designs and for both δs.

Tables 2, 4 and 6 present the results of the implied backfitting and profile estimators of τ_0 discussed in Section 4. For comparison purposes, the efficient inverse probability weighted (Eff. IPW) estimator of Hirano et al. (2003) is also calculated. The tables clearly show that the implied profile estimator of τ_0 does better than its backfitting counterpart in terms of finite sample bias and spread across all designs and sample sizes, especially in the case of the interquartile range for the Poisson specification. This is perhaps not surprising given the results of Tables 1, 3 and 5. Comparing now both implied estimators with the efficient inverse probability weighted one of Hirano et al. (2003) we note that in the case of the Gaussian specification both estimators are characterized by smaller finite sample bias and interquartile range. In the Poisson case the profile estimator has smaller bias and interquartile range whereas the Backfitting one is less precise and have bigger spread. Finally for the Logit specification the efficient inverse probability weighted estimator has the smallest finite sample bias for $n = 100$ and $n = 200$ and is characterized by a smaller spread than the backfitting implied estimator, but it is dominated as in the previous other two cases by the profile estimator in terms of interquartile range. Taken together the results of Tables 1-6 seem to suggest that the proposed estimators perform well in finite samples and can be effectively used in situations where there are missing observations and selection on observables can be assumed.

6 Conclusions

This paper proposes a new estimator for the parameters of generalized varying coefficients partially linear models when the responses are not perfectly observable but selection on the observables can be assumed. The estimator is based on an inverse probability weighting quasi-likelihood method with probability weights calculated using a parametric specification. The resulting estimator enjoys the double robustness property for three important link functions and can be used with many covariates, which makes it very useful from an applied point of view. The paper considers two general estimating techniques, namely backfitting and profiling, which yield estimators that are not asymptotically equivalent. Simulations seems to suggest that the estimators are characterized by good finite sample properties and that the one based on profiling dominates that based on backfitting both in terms of bias and spread.

References

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

Basu, A., Polsky, D. & Manning, W. (2008). Use of propensity scores in nonlinear response models: The case for health care expenditures. NBER Working paper 14086.

Cai, Z., Fan, J. & Li, R. (2000). Efficient estimation and inference for varying-coefficient models, *Journal of the American Statistical Association* **95**: 888–902.

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

Chen, J., Fan, J., Li, K. & Zhou, H. (2006). Local quasi-likelihood estimation with data missing at random, *Statistica Sinica* **16**: 1071–1100.

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

Engle, R., Granger, C., Rice, J. & Weiss, A. (1986). Nonparametric estimation of the relation between weather and electricity sales, *Journal of the American Statistical Association* **81**: 310–320.

Fan, J., Heckman, N. & Wand, M. (1995). Local polynomial kernel regression for generalized linear models and quasilikelihood functions, *Journal of the American Statistical Association* **90**: 141–150.

Foutz, R. (1977). On the unique consistent solution to the likelihood equations, *Journal of the American Statistical Association* **72**: 147–148.

Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory, *Econometrica* **52**: 681–700.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica* **66**(2): 315–331.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman and Hall.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models, *Journal of the Royal Statistical Society B* **55**: 757–796.

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

Horvitz, D. & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**: 663–685.

Imbens, G. (2004). Nonparametric estimation of average effects under exogeneity: A review, *Review of Economics and Statistics* **86**: 4–29.

Lam, C. & Fan, J. (2008). Profile-kernel inference with diverging number of parameters, *Annals of Statistics* **36**: 2232–2260.

Mammen, E., Linton, O. & Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions, *Annals of Statistics* **27**: 1443–1490.

Masry, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates, *Journal of Time Series Analysis* **17**: 571–599.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Murphy, S. & Van der Vaart, A. (2000). On profile likelihood, *Journal of the American Statistical Association* **95**: 449–485.

Newey, W. (1994). Kernel estimation of partial means and a general variance estimator, *Econometric Theory* **10**: 233–253.

Opsomer, J. (2000). Asymptotic properties of backfitting estimators, *Journal of Multivariate Analysis* **73**: 166–179.

Opsomer, J. & Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling, *Journal of Computational and Graphical Statistics* **8**: 715–732.

Robins, J., Hsieh, F. & Newey, W. (1995). Semiparametric efficient estimation of a conditional density function with missing or mismeasured covariates, *Journal of the Royal Statistical Society B* **57**: 409–424.

Robins, J. & Rotnitzky, A. (1995). Analysis of semiparametric models for repeated outcomes and missing data, *Journal of the American Statistical Association* **90**: 106–121.

Robins, J., Rotnitzky, A. & Zhao, L. (1994). Estimation of regression coefficients when ssome regressors are not always observed, *Journal of the American Statistical Association* **89**: 846–866.

Robinson, P. (1988). Root-n consistent semiparametric regression, *Econometrica* **56**: 931–954.

Severini, T. & Staniswalis, J. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of The American Statistical Association* **89**: 501–511.

Van Keilegom, I. & Carroll, R. (2007). Backfitting versus profiling in general criterion functions, *Statistica Sinica* **17**: 797–816.

Wooldridge, J. (1999). Asymptotic properties of weighted M-estimators for variable probability samples, *Econometrica* **67**: 1385–1406.

Wooldridge, J. (2002). Inverse probability weighted M-estimators for sample selection, attrition and stratification, *Portuguese Economic Journal* **1**: 117–139.

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

Wooldridge, J. (2010). *Econometric Analysis of Cross Sections and Panel Data*, Mit Press.

Appendix A

Let $c_n = (nh_1)^{-1/2}$ and "CLT", "CMT", "LLN" stand, respectively, for "central limit theorem", "continuous mapping theorem" and "law of large numbers". Let

$$X_i^* = \left[X_{1i}^\top, X_{2i}^\top, X_{2i}^\top (X_{3i} - x_3)/h_1\right]^\top,$$
$$\eta_i^h = X_{1i}^\top \beta + X_{2i}^\top \alpha(x_3) + X_{2i}^\top \alpha'(x_3)(X_{3i} - x_3)/h_1,$$

and note that the (scaled) local quasi-score $\partial Q_n(\beta, \alpha, \widehat{\pi}, x_3)/\partial\left(\beta^\top, a^\top, b^\top\right)^\top = 0$ as given in (5) is

$$S_n(\alpha, \beta, \widehat{\pi}, x_3) = \left(\frac{h_1}{n}\right)^{1/2} \sum_{i=1}^n \frac{T_i}{\widehat{\pi}_i} q_1(\eta_i^h, Y_i) X_i^* K_h(X_{3i} - x_3),$$

where for notational simplicity $\widehat{\pi}(X_i) := \widehat{\pi}_i$; also let $\partial S_n(\beta, \alpha, \widehat{\pi}, x_3)/\partial\left(\beta^\top, a^\top, b^\top\right)^\top = H_n(\alpha, \beta, \pi, x_3)$ and

$$H_n(\alpha, \beta, \pi, , x_3) = \frac{1}{n} \sum_{i=1}^n \frac{T_i}{\pi_i} q_2(\eta_i^h, Y_i) X_i^* X_i^{*\top} K_h(X_{3i} - x_3).$$

A1 Auxiliary lemmas

Lemma A.1 *Let $Z_i = \left(Y_i, X_i^\top\right)$ be i.i.d. $\mathbb{R}^p$ and $\mathbb{R}^q$-valued random vectors such that $E\|Y\|^s < \infty$, $E(\|Y\|^s|X) < \infty$ for some $s > 2$ and $E(Y|X = x)$ is continuously differentiable in C_x, a compact set such that $f(x) > 0$. Let K be a bounded positive function with bounded support satisfying a Lipschitz condition, and let $K_h(\cdot) = K(\cdot/h)$, where $h := h(n)$ is the bandwidth. Then for $n^{1-(2/s)}h^q/\log(n) \to \infty$ and*

$$\sup_{x \in C_x}\left|\frac{1}{n}\sum_{i=1}^n K_h\left(\frac{X_i - x}{h}\right) Y_i - E\left[K_h\left(\frac{X - x}{h}\right) Y\right]\right| = O_p\left(\left(\frac{\log(n)}{nh^q}\right)^{1/2}\right), \tag{A-9}$$

$$\sup_{x \in C_x}\left|\frac{1}{n}\sum_{i=1}^n K_h\left(\frac{X_i - x}{h}\right) Y_i - E[Y|X = x] f(x)\right| = O_p\left(h^{2q} + \left(\frac{\log(n)}{nh^q}\right)^{1/2}\right). \tag{A-10}$$

Proof. For (A-9) see Lemma B1 of Newey (1994) or Theorem 1 of Masry (1996). (A-10) follows by (A-9), the standard bias calculation for kernels and the triangle inequality. ∎

Lemma A.2 *Let C_x be a compact set, $B(\theta_0, \delta)$ be a closed ball of radius δ centered at θ_0, and let $\widehat{\theta}(x)$ denote the solution of $f_n(x, \widehat{\theta}(x)) = 0$ for each $x \in C_x$. Assume that (i) $f(x, \theta)$ and $\partial f(x, \theta)/\partial\theta^\top$ are continuous functions in x and θ, (ii) $f(x, \theta_0) = 0$ for each $x \in C_x$, (iii) $\partial f(x, \theta_0)/\partial\theta^\top$ is negative definite for each $x \in C_x$, (iv) $\sup_{\theta \in B(\theta_0, \delta), x \in C_x}\left\|\partial f_n(x, \theta)/\partial\theta^\top - \partial f(x, \theta)/\partial\theta^\top\right\| = o_p(1)$. Then there exists a unique $\widehat{\theta}(x)$ in $B(\theta_0, \delta)$ such that*

$$\sup_{x \in C_x}\left\|\widehat{\theta}(x) - \theta_0(x)\right\| = o_p(1).$$

Proof. The proof relies on the inverse function theorem as in Foutz (1977). Firstly, let $\lambda(x) = 1/\left(4\left\|\partial f(x,\theta_0(x))/\partial\theta^\top\right\|\right)$ and choose δ small enough so that

$$\left\|\partial f(x,\theta(x))/\partial\theta^\top - \partial f(x,\theta_0(x))/\partial\theta^\top\right\| < \lambda(x)$$

uniformly in $x \in C_x$, whenever $\theta \in B(\theta_0,\delta)$. Let $\lambda_n(x) = 1/\left(4\left\|\partial f_n(x,\theta(x))/\partial\theta^\top\right\|\right)$ and note that by (iv)

$$\sup_{x\in C_x} |\lambda_n(x) - \lambda(x)| = o_p(1). \tag{A-11}$$

Then by triangle inequality

$$\|\partial f_n(x,\theta(x))/\partial\theta' - \partial f_n(x,\theta_0(x))/\partial\theta'\| \leq \lambda(x) < 2\lambda_n(x)$$

uniformly in $x \in C_x$ with probability tending to 1. By (i) and (iii) the inverse function theorem implies that $f_n(x,\theta(x))$ is a one-to-one function from $B(\theta_0,\delta)$ to $f_n(x,B(\theta_0,\delta))$ for each $x \in C_x$ with probability tending to 1 and the image set contains an open ball of radius $\lambda_n(x)\delta$ around $f_n(x,\theta_0(x))$. By (A-11) $f_n(x,B(\theta_0,\delta))$ also contains a ball of radius $\lambda(x)\delta/2$ around $f_n(x,\theta_0(x))$ for each $x \in C_x$ with probability tending to 1. By (ii) $0 \in f_n(x,B(\theta_0,\delta))$ with probability tending to 1. Let $f_n^{-1} : f_n(x,B(\theta_0,\delta)) \to B(\theta_0,\delta)$, which exists with probability tending to 1 for each $x \in C_x$. Since $0 \in f_n(x,B(\theta_0,\delta))$ and C_x is compact it follows that $\widehat{\theta}(x) = f_n(x,0)$ exists in $B(\theta_0(x),\delta)$ with probability tending to 1 uniformly in $x \in C_x$. Moreover since δ is arbitrary small the conclusion follows. To show the uniqueness note that by the one-to-one property any other sequence $\widetilde{\theta}(x)$ of $f_n(x,\widetilde{\theta}(x))$ necessarily lies outside $B(\theta_0,\delta)$ with probability tending to 1 and by the compactness of C_x this result holds uniformly in C_x. ■

Lemma A.3 *Let*

$$Z_n(\widehat{\pi},x_3) = S_n(\alpha_0,\beta_0,\widehat{\pi},x_3) - \frac{h_1^2}{2}\Gamma(x_3),$$

and $\Sigma_{v,\pi}(\alpha_0,\beta_0,x_3) = diag[\Sigma_\pi(\alpha_0,\beta_0,x_3), v_2 B_{22\pi}(\alpha_0,\beta_0,x_3)]$. *Under A1-A6*

$$Z_n(\widehat{\pi},x_3) \xrightarrow{d} N(0, f(x_3)\Sigma_{v,\pi}(\alpha_0,\beta_0,x_3)).$$

Proof. Let $S_n(\alpha_0,\beta_0,\widehat{\pi},x_3) := S_n(\widehat{\pi},x_3)$, and note that $S_n(\widehat{\pi},x_3) = S_n(\pi,x_3) + S_{1n}(\widehat{\pi},x_3)$ where

$$S_{1n}(\widehat{\pi},x_3) = \left(\frac{h_1}{n}\right)^{1/2}\sum_{i=1}^n \frac{T_i(\widehat{\pi}_i - \pi_i)}{\widehat{\pi}_i\pi_i} q_1(\eta_i^h, Y_i) X_i^* K_h(X_{3i} - x_3).$$

Let $\eta_0 := X_1^\top\beta_0 + X_2^\top\alpha_0(X_3)$; by iterated expectation and Taylor expansion it can be shown that

$$\begin{aligned} E[S_n(\pi,x_3)] &= \frac{c_n}{2}h_1^2 f(x_3)\, E\left\{\rho_2(\eta_0)\left[X_1X_2^\top, X_2X_2^\top, 0^\top\right]^\top \alpha''(X_3)\,|X_3 = x_3\right\} + o(c_n h) \\ &:= \frac{c_n h_1^2 f(x_3)}{2}\Gamma(x_3) + o(c_n h), \end{aligned} \tag{A-12}$$

and that

$$\begin{aligned} &\mathrm{var}[S_n(\pi,x_3)] = h_1 E\left[\left(\frac{T}{\pi}\right)^2 q_1(\eta_0,Y)^2 X^* X^{*\top} K_h(X_3 - x_3)^2\right] + O\left(h_1^4\right) = \\ &f(x_3)\, E\left\{E\left[\left(\frac{T}{\pi}\right)^2 q_1(\eta_0,Y)^2 \begin{bmatrix} X_1X_1^\top v_0 & X_1X_2^\top v_0 & 0 \\ X_2X_1^\top v_0 & X_2X_2^\top v_0 & 0 \\ 0 & 0 & X_2X_2^\top v_2 \end{bmatrix} |X\right] |X_3 = x_3\right\} + o(1) \\ &= f(x_3)\, E\left\{\frac{\rho_2(\alpha_0,\beta_0)}{\pi}\begin{bmatrix} X_1X_1^\top v_0 & X_1X_2^\top v_0 & 0 \\ X_2X_1^\top v_0 & X_2X_2^\top v_0 & 0 \\ 0 & 0 & X_2X_2^\top v_2 \end{bmatrix} |X_3 = x_3\right\} + o(1) \\ &= f(x_3)\,\Sigma_{v,\pi}(\alpha_0,\beta_0,x_3) + o(1). \end{aligned}$$

Furthermore noting that $E[\|Tq_1(\eta_0, Y) X^* K_h(X_3 - x_3)/\pi\|^{2+\gamma}] = O\left(h^{-(1+\gamma)}\right)$ it follows that

$$E\left[\left(d^\top Z_i(\pi, x_3)\right)^2 I\left(\left|d^\top Z_i(\pi, x_3)\right| \geq \varepsilon d^\top f(x_3) \Sigma_{v,\pi}(\alpha_0, \beta_0, x_3) d n^{1/2}\right)\right] \leq$$
$$d^\top f(x_3) \Sigma_{v,\pi}(\alpha_0, \beta_0, x_3) d O((nh)^{-1-\gamma/2}) \to 0,$$

for any unit vector $d \in \mathbb{R}^k$ hence $Z_n(\pi, x_3) \xrightarrow{d} N(0, f(x_3) \Sigma_{v,\pi}(\alpha_0, \beta_0, x_3))$ by Lindeberg-Feller CLT and the Cramér-Wold device. By Assumption A6 and Taylor expansion

$$\frac{T_i(\widehat{\pi}_i - \pi_i)}{\widehat{\pi}_i \pi_i} = \frac{T_i}{\pi_i^2} \frac{\partial \pi_i}{\partial \gamma^\top} (\widehat{\gamma} - \gamma_0) + o_p(1), \tag{A-13}$$

hence by the same argument of (A-12)

$$\|S_{1n}(\widehat{\pi}, x_3)\| = O_p(nh_1 c_n \|\widehat{\gamma} - \gamma_0\|) = o_p(1).$$

■

Lemma A.4 *Let* $\Sigma_\kappa(\alpha, \beta, x_3) =$*diag*$[\Sigma(\alpha, \beta, x_3), \kappa_2 B_{22}(\alpha, \beta, x_3)]$; *under A1-A6*

$$\|H_n(\widehat{\pi}, x_3) - f(x_3) \Sigma_\kappa(\alpha_0, \beta_0, x_3)\| = o_p(1).$$

Proof. By the same decomposition used in Lemma A.3 $H_n(\widehat{\pi}, x_3) = H_n(\pi, x_3) + H_{1n}(\widehat{\pi}, x_3)$ where

$$H_{1n}(\widehat{\pi}, x_3) = \frac{1}{n} \sum_{i=1}^n \frac{T_i(\widehat{\pi}_i - \pi_i)}{\widehat{\pi}_i \pi_i} q_2(\eta_i, Y_i) X_i^* X_i^{*\top} K_h(X_{3i} - x_3).$$

By iterated expectations and Taylor expansion

$$\begin{aligned} & E\left\{E\left[\frac{T}{\pi} q_2(\eta_0, Y) X^* X^{*\top} K_h(X_3 - x_3) | X\right]\right\} = \\ & -E\left\{E\left[\rho_2\left(X_1^\top \beta_0 + X_2^\top \alpha(x_3)\right) X^* X^{*\top} K_h(X_3 - x_3)\right] | X_3\right\} + \\ & O(\|a - \alpha\|) + O\left(h_1^2\right) + o(1) = \\ & -f(x_3) E\left\{\rho_2\left(X_1^\top \beta_0 + X_2^\top \alpha(x_3) + O(h_1)\right) \begin{bmatrix} X_1 X_1^\top & X_1 X_2^\top & 0 \\ X_2 X_1^\top & X_2 X_2^\top & 0 \\ 0 & 0 & X_2 X_2^\top \kappa_2 \end{bmatrix} | X_3 = x_3\right\} = \\ & f(x_3) \Sigma_\kappa(\alpha_0, \beta_0, x_3) + O(h_1). \end{aligned} \tag{A-14}$$

Similarly it is possible to show that var$[H_n(\pi, x_3)] = O((nh)^{-1} + O(h)) \to 0$ hence by LLN $\|H_n(\pi, x_3) - \Sigma_\kappa(x_3)\| = o_p(1)$. By (A-13) and the same arguments as those used in (A-14) it follows that

$$\|H_{1n}(\widehat{\pi}, x_3)\| \leq \|\widehat{\gamma} - \gamma_0\| \|\Sigma_\kappa(\partial\pi/\partial\gamma_j, x_3)\| + o_p(1) = o_p(1),$$

where $\Sigma_\kappa(\partial\pi/\partial\gamma_l, x_3) = O(1)$ $(l = 1, 2, ...p)$ are $k \times k$ matrices whose structure is as that of $\Sigma_\kappa(\alpha_0, \beta_0, x_3)$ with generic (j_1, j_2) term given by $X_{j_1} X_{j_2} \partial\pi/\partial\gamma_l$. The conclusion follows by the triangle inequality. ■

Lemma A.5 *Let* $g_{ij}(Z, W) := g_1(Z_i) g_2(W_i) K_h(Z_j - Z_i)/f(Z_i)$, $h_i(Z, W) := h(Z_i, W_i)$ *such that* $E[h_i(Z, W)] = 0$, $f(Z_i)$ *denote the marginal density of* Z, *and let* $G(Z_j) = E[g_1(Z_j) g_2(W_i) | Z_j]$. *Then*

$$\left\|\frac{1}{n^{3/2}} \sum_{i \neq, j}^n h_j(Z, W) g_{ij}(Z, W) - \frac{1}{n^{1/2}} \sum_{j=1}^n h_j(Z, W) G(Z_j)\right\| = o_p(1).$$

Proof. Without loss of generality we assume the scalar case. Note that

$$E\left[g_{ij}\left(Z,W\right)|Z_j\right]=E\left[g_1\left(Z_i\right)g_2\left(W_i\right)K_h\left(Z_j-Z_i\right)|Z_i,Z_j\right]= \tag{A-15}$$
$$\int\int g_1\left(Z_j+uh\right)g_2\left(W_i\right)K\left(u\right)f\left(W_i|Z_j+uh\right)dw_i du=E\left[g_1\left(Z_j\right)g_2\left(W_i\right)|Z_j\right]+O_p\left(h^2\right)$$

by a standard Taylor expansion. Next let $h_j\left(Z,W\right)g_{ij}\left(Z,W\right)=h_jg_{ij}$, $G\left(Z_j\right)=G_j$ and note that by independence

$$E\left[\frac{1}{n^{3/2}}\sum_{i=1}^{n}\sum_{j=1}^{n}h_jg_{ij}-\frac{1}{n^{1/2}}\sum_{j=1}^{n}h_jG_j\right]^2=\frac{1}{n^3}\sum_{\substack{i,j,k,l=1\\ i\neq j,k\neq l}}^{n}E\left[\left(h_jg_{ij}-h_jG_j\right)\left(h_lg_{lk}-h_lG_l\right)\right].$$

Clearly when all indices are different all the terms in the summation are 0 because $E\left(h_jG_j\right)=0$ by iterated expectations. It remains to consider the case when at most two indices are equal. In this case there are two types of relevant combinations: (1) $i=k$ and (2) $i\neq k$. For (1) a standard kernel calculation shows that $E\left[\left(h_jg_{ij}-h_jG_j\right)\left(h_lg_{lk}-h_kG_k\right)|Z_j,Z_l\right]=O\left(h\right)$; for (2) by iterated expectations it follows similarly to (A-15) that each term in the summation is of order $O\left(h^2\right)$. Thus in both cases the summation is at most of order $n^2\left(n-1\right)O\left(h\right)/n^3$ hence the result. ∎

Lemma A.6 *(A) Let* $f_n\left(x,\theta\right):=\sum_{i=1}^{n}g\left(X_i,\theta\right)K_h\left(X_i,x\right)/n$ *and* θ_0 *is such that* $f\left(x,\theta_0\right)=0$ *for each* $x\in C_x$. *Correspondingly let* $\widehat{\theta}\left(x\right)$ *denote the solution to* $0=f_n(x,\widehat{\theta}\left(x\right))$. *Assume that (i)* C_x *and* C_θ *are a compact sets, (ii)* $\partial^k f_n\left(x,\theta\right)/\partial\theta^\top\partial\theta_j$ $(k=0,1,2)$, $(j=1,...,q)$ *are continuous functions in* x *and* θ, *(iii)* $F\left(x\right):=\partial f\left(x,\theta_0\right)/\partial\theta^\top$ *is negative definite and invertible for each* $x\in C_x$, *(iv) for some* $s>2$ $E\left\|\partial^2 g\left(X,\theta_0\right)/\partial\theta^\top\partial\theta_j\right\|^s<\infty$, $E\left(\left\|\partial^2 g\left(x,\theta_0\right)/\partial\theta^\top\partial\theta_j\right\|^s X=x\right)<\infty$ *(v)* $\sup_{\theta\in C_{\theta_0},x\in C_x}\left\|\partial f_n\left(x,\theta\right)/\partial\theta^\top\partial\theta_j-\partial f\left(x,\theta\right)/\partial\theta^\top\partial\theta_j\right\|=o_p\left(1\right)$. *Then*

$$\sup_{x\in C_x}\left\|\widehat{\theta}\left(x\right)-\theta_0\left(x\right)-F^{-1}\left(x\right)f_n\left(x,\theta_0\left(x\right)\right)\right\|=O_p\left(h^{2q}+\left(\frac{\log\left(n\right)}{nh^q}\right)^{1/2}\right). \tag{A-16}$$

(B) Consider a curve $\beta\rightarrow\theta_\beta\left(\cdot\right)$ *such that at* β_0 $\theta_{\beta_0}\left(\cdot\right)=\theta_0\left(\cdot\right)$ *and* β *is finite dimensional. Let* $f_n\left(x,\theta_\beta\right):=\sum_{i=1}^{n}g\left(X_i,\theta_\beta\right)K_h\left(X_i,x\right)/n$ *and assume that (i)-(v) assumptions used in (A) with* θ *replaced by* θ_β *hold, and that (v)* $\partial^k\theta_\beta\left(x\right)/\partial\beta_{j_1}...\partial\beta_{j_k}$ *are continuous functions in* x. *Then*

$$\sup_{x\in C_x}\left\|\frac{\partial^k\widehat{\theta}_\beta\left(x\right)}{\partial\beta_{j_1}...\partial\beta_{j_k}}-\frac{\partial^k\theta_{\beta_0}\left(x\right)}{\partial\beta_{j_1}...\partial\beta_{j_k}}\right\|=O_p\left(h^{2q}+\left(\frac{\log\left(n\right)}{nh^q}\right)^{1/2}\right). \tag{A-17}$$

Proof. (A) Assumptions (i), (ii) and (v) imply that $\widehat{\theta}\left(x\right)$ satisfies the conditions of Lemma A.2 hence $\widehat{\theta}\left(x\right)$ is unique and $\sup_{x\in C_x}\|\,\widehat{\theta}\left(x\right)-\theta_0\left(x\right)\|=o_p\left(1\right)$. Taylor expanding $0=f_n(x,\widehat{\theta}\left(x\right))$ we have

$$0=f_n\left(x,\theta_0\left(x\right)\right)+\frac{\partial f_n\left(x,\theta_0\right)}{\partial\theta^\top}\left[\widehat{\theta}\left(x\right)-\theta_0\left(x\right)\right]+\sum_{j=1}^{q}\frac{\partial^2 f_n\left(x,\theta^*\right)}{\partial\theta^\top\partial\theta_j}\left[\widehat{\theta}\left(x\right)-\theta_0\left(x\right)\right]\times \tag{A-18}$$
$$\left[\widehat{\theta}\left(x\right)_j-\theta_0\left(x\right)_j\right],$$

where θ^* is the mean value. Then, by Lemma A.1 and LLN we have that

$$0=f_n\left(x,\theta_0\left(x\right)\right)+\left(\frac{\partial f_n\left(x,\theta_0\right)}{\partial\theta^\top}-F\left(x\right)\right)\left[\widehat{\theta}\left(x\right)-\theta_0\left(x\right)\right]+F\left(x\right)\left[\widehat{\theta}\left(x\right)-\theta_0\left(x\right)\right]+$$
$$o_p(\|\,\widehat{\theta}\left(x\right)-\theta_0\left(x\right)\|),$$
$$=f_n\left(x,\theta_0\left(x\right)\right)+F\left(x\right)\left[\widehat{\theta}\left(x\right)-\theta_0\left(x\right)\right]\left(1+O_p\left(h^{2q}+\left(\frac{\log\left(n\right)}{nh^q}\right)^{1/2}\right)\right)+o_p\left(1\right)$$

uniformly in C_x hence the first conclusion. (B) For $k=0$ the result follows by the arguments used in (A). For $k=1$ by differentiating (A-18) with respect to β_l $(l=1,...,k)$

$$0=\frac{\partial f_n(x,\theta_0)}{\partial\theta_\beta^\top}\frac{\partial\theta_\beta}{\partial\beta_l}+\sum_{j=1}^{q}\frac{\partial^2 f_n(x,\theta_0)}{\partial\theta_\beta^\top\partial\theta_{\beta j}}\frac{\partial\theta_{\beta j}}{\partial\beta_l}\left(\widehat{\theta}_\beta(x)_j-\theta_0(x)_j\right)+$$

$$\frac{\partial f_n(x,\theta_0)}{\partial\theta^\top}\left(\frac{\partial\widehat{\theta}_\beta(x)}{\partial\beta_l}-\frac{\partial\theta_{\beta_0}(x)}{\partial\beta_l}\right)+o_p(1),$$

$$=\frac{\partial f_n(x,\theta_0)}{\partial\theta_\beta^\top}\frac{\partial\theta_\beta}{\partial\beta_l}+o_p\left(h^{2q}+(\log(n)/nh^q)^{1/2}\right)+$$

$$F(x)\left(\frac{\partial\widehat{\theta}_\beta(x)}{\partial\beta_l}-\frac{\partial\theta_{\beta_0}(x)}{\partial\beta_l}\right)\left(1+O_p\left(h^{2q}+\left(\frac{\log(n)}{nh^q}\right)^{1/2}\right)\right)+o_p(1),$$

uniformly in C_x hence noting that by Lemma A.1

$$\left\|(\partial f_n(x,\theta_0)/\partial\theta_\beta^\top)(\partial\theta_\beta/\partial\beta_l)\right\|=O_p(h^{2q}+(\log(n)/nh^q)^{1/2})$$

the result follows. For $k\geq 2$ the result follows by repeated differentiation with respect to β using recursively the fact that

$$\left\|\frac{\partial^{k-1}\widehat{\theta}_\beta(x)}{\partial\beta_{l_1}....\partial\beta_{l_{k-1}}}-\frac{\partial^{k-1}\theta_{\beta_0}(x)}{\partial\beta_{l_1}....\partial\beta_{l_{k-1}}}\right\|=O_p\left(h^{2q}+\left(\frac{\log(n)}{nh^q}\right)^{1/2}\right).$$

■

A2 Proof of the Main Results

Proof of Theorem 3.1. Let $\theta(x_3)=\left[(\beta-\beta_0)^\top,(a(x_3)-\alpha_0(x_3))^\top,h(b(x_3)-\alpha_0'(x_3))^\top\right]^\top$ and $\eta_{0i}^h=X_{1i}^\top\beta_0+X_{2i}^\top[\alpha_0(x_3)+\alpha_0'(x_3)(X_{3i}-x_3)/h]$; by Assumptions A2 and A3 the solution $\widehat{\theta}(x_3)$ satisfies Lemma A.2 hence $\widehat{\theta}(x_3)=o_p(1)$ uniformly in $B(\beta_0)$ and $\mathcal{X}_3$. Let $\widehat{\theta}_n(x_3):=\widehat{\theta}(x_3)c_n$; by a Taylor expansion of the local version of (5) about 0 we have

$$0=\frac{h_1^{1/2}}{n^{1/2}}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}q_1\left(\eta_{0i}^h+X_i^{*\top}\widehat{\theta}_n(x_3),Y_i\right)X_{1i}^*=S_n(\alpha_0,\beta_0,\widehat{\pi},x_3)+H_n(\alpha_0,\beta_0,\widehat{\pi},x_3)\widehat{\theta}(x_3)+$$

$$\frac{c_n^2}{2}\left(\frac{h_1}{n}\right)^{1/2}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}q_3\left(\eta_{0i}^h+X_i^{*\top}\theta^*(x_3),Y_i\right)X_{1i}^*\left(X_i^{*\top}\widehat{\theta}_n(x_3)\right)^2K_h(X_{3i}-x_3),$$

where $\theta^*(x_3)$ is the mean value. By Assumptions A2, A4 and the same arguments as those used in Lemma A.4 the last term in the above expansion is $O_p(c_n)\to 0$, hence by Lemmas A.4 and A.6 we have that

$$\sup_{x_3\in\mathcal{X}_3,\beta\in B(\beta_0)}\left\|\widehat{\theta}_n(x_3)-\Sigma_\kappa(\alpha_0,\beta_0,x_3)^{-1}S_n(\widehat{\pi},x_3)\right\|=O_p\left(h^2+\left(\frac{\log(n)}{nh}\right)^{1/2}\right).\qquad\text{(A-19)}$$

Thus the result follows by Lemma A.3, CMT and simple algebra. ■

Proof of Theorem 3.2. The consistency of the solution $\widehat{\beta}$ on $B(\beta_0)$ follows by Assumption (A3) which combined with the uniform consistency of $\widehat{\alpha}(\cdot)$ as given in the proof of Theorem 3.1 implies a global version of Lemma A.2. Let $\widehat{\eta}_i=X_{1i}^\top\beta_0+X_{2i}^\top\widehat{\alpha}(X_{3i})$, $b_n=n^{1/2}(\beta-\beta_0)$; as in the proof of Theorem 3.1 a Taylor expansion of $\beta-\beta_0$ about 0 gives

$$0=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}q_1\left(\widehat{\eta}_i+X_{1i}^\top b_n/n^{1/2},Y_i\right)X_{1i}=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}q_1(\widehat{\eta}_i,Y_i)X_{1i}+$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}q_2(\widehat{\eta}_i,Y_i)X_{1i}X_{1i}^\top\widehat{b}_n+\frac{1}{2n^{3/2}}\sum_{i=1}^{n}\frac{T_i}{\widehat{\pi}_i}q_3(\widehat{\eta}_i+\xi_i,Y_i)X_{1i}(X_{1i}^\top\widehat{b}_n)^2,$$

where ξ_i is the mean value. By the consistency of $\widehat{\beta}$, $\widehat{\alpha}(\cdot)$ and $\widehat{\pi}_i$, and A3-A4 it follows by dominated convergence that $\|\sum_{i=1}^n T_i q_3(\widehat{\eta}_i+\xi_i,Y_i)X_{1i}X_{1i}X_{1ij}/n\widehat{\pi}_i\| = O_p(1)$ uniformly in $\mathcal{X}_3$ and $B(\beta_0)$, hence the last term is $o_p(1)$. Similarly

$$\left\|\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i} q_2(\widehat{\eta}_i,Y_i)X_{1i}X_{1i}^\top - B_{11}(\alpha_0,\beta_0)\right\| = o_p(1).$$

By Taylor expansion and A6

$$\begin{aligned}
&\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\widehat{\pi}_i} q_1(\widehat{\eta}_i,Y_i)X_{1i} = \frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i} q_1(\eta_{0i},Y_i)X_{1i} + \\
&\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i} q_2(\eta_{0i},Y_i)X_{1i}(\widehat{\eta}_i-\eta_{0i}) + O_p\left(n^{1/2}\|\widehat{\eta}-\eta_0\|^2\right) + \\
&\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^2} q_1(\eta_{0i},Y_i)X_{1i}\frac{\partial \pi_i}{\partial\gamma^\top} n^{1/2}(\widehat{\gamma}-\gamma_0) + \\
&\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^2} q_2(\eta_{0i},Y_i)X_{1i}(\widehat{\eta}_i-\eta_{0i})\frac{\partial \pi_i}{\partial\gamma^\top} n^{1/2}(\widehat{\gamma}-\gamma_0) + o_p(1) = \sum_{j=1}^4 I_{1jn} + o_p(1),
\end{aligned}$$

uniformly in $\mathcal{X}_3$ and Γ. Lemma A.6 and the fact that $\|\widehat{\eta}_i-\eta_i\| = O(\|X_j-X_i\|) = O_p(h^2)$ imply

$$\begin{aligned}
I_{12n} = &\frac{1}{n^{3/2}}\sum_{i=1}^n \frac{T_i}{\pi_i f(X_{3i})} q_2(\eta_{0i},Y_i)X_{1i}X_{2i}^\top \sum_{j=1}^n \frac{T_j}{\pi_j} q_1(\eta_{0j},Y_j)S_\alpha\Sigma_\kappa^{-1}(\alpha_0,\beta_0,x_3)X_j^*\times \\
&K_{h_1}(X_{3j}-X_{3i}) + O_p\left(n^{1/2}h_1^2\right) + O_p\left(h^2+\left(\frac{\log(n)}{nh}\right)^{1/2}\right),
\end{aligned}$$

where $S_\alpha = [0, I, 0]$. Conditional on X_{3j}, the law of iterated expectations and Taylor expansion yields

$$\begin{aligned}
&E\left[\frac{T_i}{\pi_i f(X_{3i})} q_2(\eta_{0i},Y_i)X_{1i}X_{2i}^\top K_{h_1}(X_{3j}-X_{3i})\,|X_{3j}\right] = \\
&\quad -E\left[\frac{1}{f(X_{3i})}\rho_2(\eta_{0i})X_{1i}X_{2i}^\top K_{h_1}(X_{3j}-X_{3i})\,|X_{3j}\right] = -B_{12}(\alpha_0,\beta_0,X_{3j}),
\end{aligned}$$

hence by Lemma A.5

$$I_{12n} = -\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i} B_{12}(\alpha_0,\beta_0,X_{3i})q_1(\eta_{0i},Y_i)S_\alpha\Sigma_\kappa(x_3)^{-1}\left[X_{1i}^\top,X_{2i}^\top,0^\top\right]^\top + O_p\left(n^{1/2}h_1^2\right).$$

By iterated expectations $E\left[T_i q_1(\eta_{0i},Y_i)X_{1i}\left(\partial\pi_i/\partial\gamma^\top\right)/\pi_i^2\right] = 0$ hence $\|I_{13n}\| = o_p(1)$ by LLN. The same arguments as those used for I_{12n} can be used to show that $\|I_{14n}\| = o_p(1)$. Thus we have that

$$\begin{aligned}
0 = &\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i} q_1(\eta_{0i},Y_i)X_{1i} - \\
&\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i}\left\{B_{12}(\alpha_0,\beta_0,X_{3i})q_1(\eta_{0i},Y_i)S_\alpha\Sigma_\kappa(\alpha_0,\beta_0,x_3)^{-1}\left[X_{1i}^\top,X_{2i}^\top,0^\top\right]\right\}^\top - \\
&B_{11}(\alpha_0,\beta_0)\widehat{b}_n + o_p(1),
\end{aligned}$$

so that

$$\begin{aligned}
\widehat{b}_n = &B_{11}(\alpha_0,\beta_0)^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i}\left[q_1(\eta_{0i},Y_i)X_{1i} - \right. \\
&\left. B_{12}(\alpha_0,\beta_0,X_{3i})q_1(\eta_{0i},Y_i)S_\alpha\Sigma_\kappa(\alpha_0,\beta_0,x_3)^{-1}\left[X_{1i}^\top,X_{2i}^\top,0^\top\right]^\top\right] + o_p(1).
\end{aligned} \tag{A-20}$$

The conclusion follows by CLT noting that by conditional expectations and some algebra

$$E\left\{\frac{T_i^2}{\pi_i^2}q_1\left(\eta_0,Y\right)^2 X_1\left[X_1^\top,X_2^\top,0^\top\right]\Sigma_\kappa\left(\alpha_0,\beta_0,x_3\right)^{-1}S_\alpha^\top B_{12}\left(\alpha_0,\beta_0,X_{3i}\right)^\top\right\}=$$
$$E\left\{E\left[\frac{T_i^2}{\pi_i^2}q_1\left(\eta_0,Y\right)^2 X_{1i}\left[X_{1i}^\top,X_{2i}^\top,0^\top\right]|X_{3i}\right]\Sigma_\kappa\left(\alpha_0,\beta_0,x_3\right)^{-1}S_\alpha^\top B_{12}\left(\alpha_0,\beta_0,X_{3i}\right)^\top\right\}=$$
$$E\left\{E\left(\frac{\rho_2\left(\alpha_0,\beta_0\right)\left[X_1X_1^\top,X_1X_2^\top,0^\top\right]}{\pi}|X_3\right)\times\right.$$
$$\left[-B_{11}\left(\alpha_0,\beta_0,X_{3i}\right)^{-1}B_{12}\left(\alpha_0,\beta_0,X_3\right)\Delta\left(\alpha_0,\beta_0,X_3\right)^{-1},\Delta\left(\alpha_0,\beta_0,X_3\right)^{-1},0\right]^\top\times$$
$$\left.B_{12}\left(\alpha_0,\beta_0,X_{3i}\right)^\top\right\},$$

where

$$\Delta\left(\alpha_0,\beta_0,X_3\right)=B_{22}\left(\alpha_0,\beta_0,X_3\right)-B_{21}\left(\alpha_0,\beta_0,X_3\right)B_{11}\left(\alpha_0,\beta_0,X_3\right)^{-1}B_{12}\left(\alpha_0,\beta_0,X_3\right).$$

■

Proof of Theorem 3.3. Let $\widehat{\eta}_i=X_{1i}^\top\widehat{\beta}+X_{2i}^\top\left[a\left(x_3\right)+b\left(x_3\right)\left(X_{3i}-x_3\right)\right]$, $\theta_{2n}\left(x_3\right)=c_n^{-1}[(a\left(x_3\right)-\alpha_0\left(x_3\right))^\top$, $h_2\left(b\left(x_3\right)-\alpha_0'\left(x_3\right)\right)^\top]^\top$, $X_{2i}^*=\left[X_{2i}^\top,X_{2i}^\top\left(X_{3i}-x_3\right)/h_2\right]^\top$ and let $\widehat{\theta}_2\left(x_3\right)$ denotes the solution to the local first order conditions $\partial Q_n(\widehat{\beta},\alpha,\widehat{\pi},x_3)/\partial(\beta^\top,a^\top,b^\top)^\top=0$. Consistency of $\widehat{\theta}_2\left(x_3\right)$ follows by the same arguments as those used in the proof of Theorem 3.1. Then by Taylor expansion we have

$$0=S_{2n}\left(\alpha_0,\beta_0,\pi,x_3\right)+H_{2n}\left(\alpha_0,\beta_0,\pi,x_3\right)\theta_{2n}\left(x_3\right)+$$
$$O_p(nh_2c_n[\|\widehat{\beta}-\beta_0\|+\|\widehat{\gamma}-\gamma_0\|])+O_p\left(c_n\right),$$

where

$$S_{2n}\left(\alpha_0,\beta_0,\pi,x_3\right)=\left(\frac{h_2}{n}\right)^{1/2}\sum_{i=1}^n\frac{T_i}{\pi_i}q_1\left(\eta_{i0},Y_i\right)X_{2i}^*K_{h_2}\left(X_{3i}-x_3\right)$$

and

$$H_{2n}\left(\alpha_0,\beta_0,\pi,x_3\right)=\frac{1}{n}\sum_{i=1}^n\frac{T_i}{\pi_i}q_2\left(\eta_{i0},Y\right)X_2^*X_2^{*\top}K_{h_2}\left(X_3-x_3\right).$$

The conclusion follows as in the proof of Theorem 3.1 using Lemmas A.3, A.4 and some algebra. ■

Proof of Theorem 3.4. Let $\eta_\beta=X_1^\top\beta+\alpha_\beta\left(X_3\right)^\top X_2$; by definition the least favourable curve $\alpha_\beta\left(\cdot\right)$ satisfies

$$\frac{\partial}{\partial\zeta}E\left\{Q\left[g^{-1}\left(X_1^\top\beta+X_2^\top\zeta\right),Y\right]|X_3=x_3\right\}=0 \tag{A-21}$$

Differentiating (A-21) with respect to β and evaluating at β_0

$$0=E\left\{\left[Y-g^{-1}\left(\eta_\beta\right)\right]\rho_1'\left(\eta_\beta\right)\times\left[X_1^\top+X_2^\top\partial\alpha_\beta\left(X_3\right)/\partial\beta^\top\right]-\right.$$
$$\left.\rho_2\left(\eta_\beta\right)X_2^\top\left[X_1^\top+X_2^\top\partial\alpha_\beta\left(X_3\right)/\partial\beta^\top\right]|X_3=x_3\right\}|_{\beta=\beta_0},$$

which implies that the so-called least favourable direction is

$$\frac{\partial\alpha_\beta\left(x_3\right)}{\partial\beta^\top}=-\left\{E\left[\rho_2\left(\eta_{\beta_0}\right)X_2X_2^\top|X_3=x_3\right]\right\}^{-1}\times \tag{A-22}$$
$$E\left[\rho_2\left(\eta_{\beta_0}\right)X_2X_1^\top|X_3=x_3\right]=-\left[B_{22}\left(\alpha_0,\beta_0,x_3\right)\right]^{-1}B_{21}\left(\alpha_0,\beta_0,x_3\right),$$

where $\eta_{\beta_0}=X_1^\top\beta_0+\alpha_{\beta_0}^\top X_2$ and by definition $\alpha_{\beta_0}\left(x_3\right)=\alpha_0\left(x_3\right)$. As in the proof of Theorem 3.2, Assumption A3-A4 and Lemma A.2 imply the consistency and uniqueness of the solution $\widehat{\beta}$ to $0=\partial Q_n\left(\alpha_\beta,\beta,\widehat{\pi}\right)/\partial\beta$. By

Taylor expansion of $0 = \partial \widehat{Q}_n(\alpha_{\widehat{\beta}}, \widehat{\beta}, \widehat{\pi})/\partial\beta$ we have

$$0 = S_n\left(\pi, \beta_0, \alpha_{\beta_0}\right) + S_n\left(\widehat{\pi}, \beta_0, \alpha_{\beta_0}\right) + \left[\widehat{H}_n\left(\alpha_0, \beta_0, \pi\right) + \widehat{H}_n\left(\alpha_0, \beta_0, \widehat{\pi}\right)\right] n^{1/2}\left(\widehat{\beta} - \beta_0\right) + O_p\left(n^{1/2}\left\|\widehat{\beta} - \beta_0\right\|^2\right), \tag{A-23}$$

where

$$S_n\left(\pi, \beta_0, \alpha_{\beta_0}\right) = \frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i} q_1\left[g^{-1}\left(\eta_{i\beta_0}\right), Y_i\right]\left[X_{1i} + \left(\frac{X_{2i}^\top \partial\alpha_{\beta_0}\left(X_{3i}\right)}{\partial\beta^\top}\right)^\top\right] +$$
$$\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i} q_2\left[g^{-1}\left(\eta_{i\beta_0}\right), Y_i\right]\left[X_{1i} + \left(\frac{X_{2i}^\top \partial\alpha_{\beta_0}\left(X_{3i}\right)}{\partial\beta^\top}\right)^\top\right] X_{2i}^\top\left(\widehat{\alpha}_{\beta_0}\left(X_{3i}\right) - \alpha_{\beta_0}\left(X_{3i}\right)\right) +$$
$$\frac{1}{n^{1/2}}\sum_{i=1}^n \frac{T_i}{\pi_i} q_1\left[g^{-1}\left(\eta_{i\beta_0}\right), Y_i\right]\left[X_{2i}^\top\left(\frac{\partial\widehat{\alpha}_{\beta_0}\left(X_{3i}\right)}{\partial\beta^\top} - \frac{\partial\alpha_{\beta_0}\left(X_{3i}\right)}{\partial\beta^\top}\right)\right]^\top := \sum_{j=1}^3 I_{2jn},$$
$$S_n\left(\widehat{\pi}, \beta_0, \alpha_{\beta_0}\right) = \sum_{j=1}^3 \widehat{I}_{2jn} + o_p(1),$$

and each of the $\widehat{I}_{2jn}$ is as that of I_{2jn} with T_i/π_i replaced by (A-13). By (A-22) and CLT we have that $I_{21n} \xrightarrow{d} N\left(0, \Omega^p\left(\alpha_0, \beta_0, \pi\right)\right)$. By the least favourable property

$$E\left\{q_2\left[g^{-1}\left(\eta_\beta\right), Y\right]\left[X_1 + \left(\frac{X_2^\top \partial\alpha_\beta\left(X_3\right)}{\partial\beta^\top}\right)^\top\right] X_2^\top | X_3 = x_3\right\} = 0$$

and hence

$$\left\|I_{22n}\right\| \leq O_p(1)\left\|\left(\widehat{\alpha}_\beta\left(X_3\right) - \alpha_0\left(X_3\right)\right)\right\| = O_p\left(h^2 + \left(\frac{\log(n)}{nh}\right)^{1/2}\right),$$

uniformly in $\mathcal{X}_3$ by Lemma A.6(B) and similarly for I_{23n}. By the same arguments as those used in Theorem 3.2 we have $\left\|\widehat{I}_{2jn}\right\| = o_p(1)$ for $j = 1$ and 3. For $\widehat{I}_{22n}$ note that by Lemma A.6

$$\left\|\widehat{I}_{22n}\right\| \leq n^{1/2}\left\|\widehat{\gamma} - \gamma_0\right\|\left\|\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \frac{T_i}{\pi_i^2} q_2\left[g^{-1}\left(\eta_{i\beta_0}\right), Y_i\right] \times\right.$$
$$\left.\left[X_{1i} + \left(\frac{X_{2i}^\top \partial\alpha_{\beta_0}\left(X_{3i}\right)}{\partial\beta^\top}\right)^\top\right] X_{2i}^\top \frac{T_j}{\pi_j} q_1\left(\eta_j, Y_j\right) S_\alpha \Sigma_\kappa\left(x_3\right)^{-1} X_j^* K_{h_1}\left(X_{3j} - X_{3i}\right)\right\| +$$
$$O_p\left(h^2 + \left(\frac{\log(n)}{nh}\right)^{1/2}\right) =$$
$$n^{1/2}\left\|\widehat{\gamma} - \gamma_0\right\|\left\|I_{24n}\right\| + O_p\left(h^2 + \left(\frac{\log(n)}{nh}\right)^{1/2}\right).$$

By Lemma A.5 it follows $\left\|I_{24n} - I_{25n}\right\| = o_p(1)$ where

$$I_{25n} = -\frac{1}{n}\sum_{i=1}^n \frac{T_i}{\pi_i^2} B_{3\pi}\left(\alpha_0, \beta_0, X_{3i}\right) q_1\left(\eta_{0i}, Y_i\right) S_\alpha \Sigma_\kappa\left(X_{3i}\right)^{-1}\left[X_{1i}^\top, X_{2i}^\top, 0^\top\right]^\top,$$

and

$$B_{3\pi}\left(\alpha_0, \beta_0, X_3\right) = E\left[\frac{1}{\pi}\rho_2\left(\alpha_0, \beta_0\right)\left[X_1 + \left(\frac{X_2^\top \partial\alpha_{\beta_0}\left(X_3\right)}{\partial\beta^\top}\right)^\top\right] X_2^\top | X_3\right].$$

Note that $\|I_{25n}\| = o_p(1)$ by LLN, hence $\left\|\widehat{I}_{22n}\right\| \leq n^{1/2}\|\widehat{\gamma}-\gamma_0\|\|I_{14n}\| = o_p(1)$. We now consider the third term in (A-23). By Taylor expansion, LLN, Lemma A.6 and triangle inequality

$$
\begin{aligned}
&\left\|\widehat{H}_n(\alpha_0,\beta_0,\pi) - H_n(\alpha_0,\beta_0,\pi)\right\| \leq \left\|\sum_{j=1}^{k_2}\sum_{i=1}^{n}\frac{1}{n}q_3\left(\left[g^{-1}(\eta_{i\beta_0}),Y_i\right]\right)\times\right.\\
&\left.\left[X_{1i}+\left(\frac{X_{2i}^{\top}\partial\alpha_{\beta_0}(X_{3i})}{\partial\beta^{\top}}\right)^{\top}\right]\left[X_{1i}+\left(\frac{X_{2i}^{\top}\partial\alpha_{\beta_0}(X_{3i})}{\partial\beta^{\top}}\right)^{\top}\right]^{\top}X_{2ij}\right\|\times\\
&\|\widehat{\alpha}_\beta(X_{3i})-\alpha_0(X_{3i})\| + 2\left\|\sum_{j=1}^{k_2}H_n(\alpha_0,\beta_0,\pi)X_{2ij}\right\|\\
&\left\|\frac{\partial\widehat{\alpha}_\beta(X_{3i})}{\partial\beta}-\frac{\partial\alpha_{\beta_0}(X_{3i})}{\partial\beta}\right\| + \left\|\sum_{i=1}^{n}\frac{1}{n}q_1\left(\left[g^{-1}(\eta_{i\beta_0}),Y_i\right]\right)X_{1i}\right\|\times\\
&\left\|\sum_{j=1}^{k_1}\frac{\partial^2\widehat{\alpha}_\beta(X_{3i})}{\partial\beta\partial\beta_j}-\frac{\partial\alpha_{\beta_0}(X_{3i})}{\partial\beta\partial\beta_j}\right\| +\\
&\left\|\sum_{j=1}^{k_1}\sum_{i=1}^{n}\frac{1}{n}q_1\left(\left[g^{-1}(\eta_{i\beta_0}),Y_i\right]\right)X_{1i}\frac{\partial\alpha_{\beta_0}(X_3)}{\partial\beta\partial\beta_j}\right\|\|\widehat{\alpha}_\beta(X_{3i})-\alpha_\beta(X_{3i})\| =\\
&O_p(1)\,O_p\left(h^2+\left(\frac{\log(n)}{nh}\right)^{1/2}\right) = o_p(1)
\end{aligned}
$$

uniformly in $\mathcal{X}_3$. Since

$$
\begin{aligned}
H_n(\alpha_0,\beta_0,\pi) = \frac{1}{n}\sum\frac{T_i}{\pi_i}q_2\left(\left[g^{-1}(\eta_{i\beta_0}),Y_i\right]\right)&\left[X_{1i}+\left(\frac{X_{2i}^{\top}\partial\alpha_{\beta_0}(X_{3i})}{\partial\beta^{\top}}\right)^{\top}\right]\times\\
&\left[X_{1i}+\left(\frac{X_{2i}^{\top}\partial\alpha_{\beta_0}(X_{3i})}{\partial\beta^{\top}}\right)^{\top}\right]^{\top} + o_p(1),
\end{aligned}
$$

it follows by LLN that

$$
\|H_n(\alpha_0,\beta_0,\pi) - \Xi(\alpha_0,\beta_0)\| = o_p(1). \tag{A-24}
$$

Next by (A-13) and (A-24) it follows that

$$
\|H_n(\alpha_0,\beta_0,\widehat{\pi})\| \leq \|\widehat{\gamma}-\gamma\|\,O_p(1) = o_p(1)
$$

hence the result follows by CMT. ■

Proof of Theorem 4.1. Let $\widehat{\tau}^m$ denote the estimator based on either backfitting ($m=b$) or profiling ($m=p$);

by Taylor expansion

$$
\begin{aligned}
&n^{1/2}\left(\widehat{\tau}^{m}-\tau\right)=\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[g^{-1}(X_{1i}^{\top}\widehat{\beta}^{1}+X_{2i}^{\top}\widehat{a}^{1}\left(X_{3i}\right))-g^{-1}(X_{1i}^{\top}\widehat{\beta}^{0}+X_{1i}^{\top}\widehat{a}^{0}\left(X_{3i}\right))-\tau\right]= \\
&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left[g^{-1}\left(X_{1i}^{\top}\beta^{1}+X_{2i}^{\top}\alpha^{1}\left(X_{3i}\right)\right)-g^{-1}\left(X_{1i}^{\top}\beta^{0}+X_{2i}^{\top}\alpha^{0}\left(X_{3i}\right)\right)-\tau\right]+ \\
&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{\partial g^{-1}\left(X_{1i}^{\top}\beta^{1}+X_{2i}^{\top}\alpha^{1}\left(X_{3i}\right)\right)}{\partial\left(\beta^{1\top},\alpha^{1\top}\right)^{\top}}\left[X_{1i}^{\top}(\widehat{\beta}^{1}-\beta_{0}^{1}),X_{2i}^{\top}(\widehat{\alpha}^{1}\left(X_{3i}\right)-\alpha^{1}\left(X_{3i}\right))\right]- \\
&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{\partial g^{-1}\left(X_{1i}^{\top}\beta^{0}+X_{2i}^{\top}\alpha^{0}\left(X_{3i}\right)\right)}{\partial\left(\beta^{0\top},\alpha^{0\top}\right)^{\top}}\left[X_{1i}^{\top}(\widehat{\beta}^{1}-\beta_{0}^{1}),X_{2i}^{\top}(\widehat{\alpha}^{1}\left(X_{3i}\right)-\alpha^{1}\left(X_{3i}\right))\right]+o_{p}\left(1\right) \\
&:=\sum_{j=1}^{3}I_{3j1}^{m}.
\end{aligned}
$$

For the backfitting estimator $\widehat{\tau}^{b}$ using (A-20), (A-19), Lemma A.5 and LLN we have

$$
\begin{aligned}
&I_{32n}^{b}=G_{1}\left(\alpha_{0}^{1},\beta_{0}^{1}\right)^{\top}B_{11}\left(\alpha_{0},\beta_{0}\right)^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{T_{i}}{\pi_{i}}q_{1}\left(\eta_{0i},Y_{i}\right)\left[X_{1i}-B_{12}\left(\alpha_{0},\beta_{0},X_{3i}\right)\times\right. \\
&\left.S_{\alpha}\Sigma_{\kappa}\left(\alpha_{0},\beta_{0},X_{3}\right)^{-1}\left[X_{1i}^{\top},X_{2i}^{\top},0^{\top}\right]^{\top}\right]+ \\
&\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{T_{i}}{\pi_{i}}G_{2}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3i}\right)^{\top}q_{1}\left(\eta_{0i},Y_{i}\right)S_{\alpha}\Sigma_{\kappa}\left(\alpha_{0},\beta_{0},X_{3}\right)^{-1}\times \\
&\left[X_{1i}^{\top},X_{2i}^{\top},0^{\top}\right]^{\top}+o_{p}\left(1\right),
\end{aligned}
$$

and likewise for I_{33n}^{b} with α_{0}^{1}, β_{0}^{1} and π replaced by α_{0}^{0}, β_{0}^{0} and $1-\pi$. Note that

$$
\begin{aligned}
&\operatorname{var}\left(I_{31n}^{b}\right)=\operatorname{var}\left[g^{-1}\left(X_{1i}^{\top}\beta_{0}^{1}+X_{2i}^{\top}\alpha_{0}^{1}\left(X_{3i}\right)\right)-g^{-1}\left(X_{1i}^{\top}\beta_{0}^{0}+X_{1i}^{\top}\alpha_{0}^{0}\left(X_{3i}\right)\right)\right], \\
&\operatorname{var}\left(I_{32n}^{b}\right)=G_{1}^{\top}\left(\alpha_{0}^{1},\beta_{0}^{1}\right)B^{b}\left(\alpha_{0},\beta_{0},\pi\right)G_{1}\left(\alpha_{0}^{1},\beta_{0}^{1}\right)+E\left[G_{2}^{\top}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)S_{\alpha}\Sigma_{\kappa}^{-1}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)\times\right. \\
&\left.\Sigma_{\pi}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)\Sigma_{\kappa}^{-1}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)S_{\alpha}G_{2}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)\right]+2G^{\top}\left(\alpha_{0}^{1},\beta_{0}^{1}\right)B_{11}^{-1}\left(\alpha_{0},\beta_{0}\right)\times \\
&E\left[-B_{11\pi}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)B_{11}^{-1}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)B_{12}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)\Delta\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)G_{2}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)+\right. \\
&\left.B_{12}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)\Delta\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)G_{2}^{\top}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)\right]-2G_{1}\left(\alpha_{0}^{1},\beta_{0}^{1}\right)B_{11}\left(\alpha_{0},\beta_{0}\right)^{-1}\times \\
&E\left[B_{12}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)S_{\alpha}\Sigma_{\kappa}^{-1}\left(\alpha_{0}^{1},\beta_{0}^{1},x_{3}\right)\Sigma_{\pi}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)\Sigma_{\kappa}^{-1}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)S_{\alpha}G_{2}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3}\right)^{\top}\right],
\end{aligned}
$$

$\operatorname{var}\big(I_{33n}^{b}\big)$ is as $\operatorname{var}\big(I_{32n}^{b}\big)$ with α_{0}^{1}, β_{0}^{1} and π replaced by α_{0}^{0}, β_{0}^{0} and $1-\pi$ and $\operatorname{cov}(I_{3jn}^{b},I_{3kn}^{b})=0$ for $j\neq k=1,2,3$ and the conclusion follows by CLT and CMT. Similarly for the profile estimator $\widehat{\tau}^{p}$ using (A-20), (A-19), Lemma A.5 and LLN we have

$$
\begin{aligned}
&I_{32n}^{p}=G_{1}\left(\alpha_{0}^{1},\beta_{0}^{1}\right)^{\top}\Xi\left(\alpha_{0},\beta_{0}\right)^{-1}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{T_{i}}{\pi_{i}}q_{1}\left[g^{-1}\left(X_{1i}^{\top}\beta_{0}+X_{2i}^{\top}\alpha_{\beta_{0}}\left(X_{3i}\right)\right),Y_{i}\right]\times \\
&\left[X_{1i}+\left(\frac{X_{2i}^{\top}\partial\alpha_{\beta_{0}}\left(X_{3i}\right)}{\partial\beta^{\top}}\right)^{\top}\right]+\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{T_{i}}{\pi_{i}}G_{2}\left(\alpha_{0}^{1},\beta_{0}^{1},X_{3i}\right)^{\top}q_{1}\left(X_{1i}^{\top}\beta_{0}+X_{2i}^{\top}\alpha_{\beta_{0}}\left(X_{3i}\right),Y_{i}\right)\times \\
&S_{\alpha}^{-1}\Sigma_{\kappa}\left(\alpha_{0},\beta_{0},X_{3}\right)^{-1}\left[X_{1i}^{\top},X_{2i}^{\top},0^{\top}\right]^{\top}+o_{p}\left(1\right),
\end{aligned}
$$

hence

$$\begin{aligned}\operatorname{var}\left(I_{32n}^{p}\right) = {} & G_1\left(\alpha_0^1,\beta_0^1\right)^{\top} B^p\left(\alpha_0,\beta_0,\pi\right) G_1\left(\alpha_0^1,\beta_0^1\right) + E\Big[G_2\left(\alpha_0^1,\beta_0^1,X_3\right)^{\top} S_\alpha \Sigma_\kappa\left(\alpha_0^1,\beta_0^1,X_3\right)^{-1} \\ & \Sigma_\pi\left(\alpha_0^1,\beta_0^1,X_3\right)\Sigma_\kappa\left(\alpha_0^1,\beta_0^1,X_3\right)^{-1} S_\alpha G_2\left(\alpha_0^1,\beta_0^1,X_3\right)\Big] + \\ & 2D_1\left(\alpha_0^1,\beta_0^1\right)^{\top}\Xi\left(\alpha_0,\beta_0\right)^{-1} E\left[\Delta_{11}\left(\alpha_0^1,\beta_0^1,X_3\right) + \right. \\ & \left.\Delta_{12}\left(\alpha_0^1,\beta_0^1,X_3\right) G_2\left(\alpha_0^1,\beta_0^1,X_3\right)\right],\end{aligned}$$

where

$$\begin{aligned}\Delta_{11}\left(\alpha_0^1,\beta_0^1,X_3\right) = {} & \Big[B_{11\pi}\left(\alpha_0^1,\beta_0^1,X_3\right) - B_{12}\left(\alpha_0^1,\beta_0^1,X_3\right) B_{22}\left(\alpha_0^1,\beta_0^1,X_3\right)^{-1} B_{21\pi}\left(\alpha_0^1,\beta_0^1,X_3\right)\Big] \times \\ & B_{11}\left(\alpha_0^1,\beta_0^1,X_3\right)^{-1}\Delta\left(\alpha_0^1,\beta_0^1,X_3\right), \\ \Delta_{12}\left(\alpha_0^1,\beta_0^1,X_3\right) = {} & \Big[B_{12\pi}\left(\alpha_0^1,\beta_0^1,X_3\right) - B_{12}\left(\alpha_0^1,\beta_0^1,X_3\right) B_{22}\left(\alpha_0^1,\beta_0^1,X_3\right)^{-1} B_{22\pi}\left(\alpha_0^1,\beta_0^1,X_3\right)\Big] \\ & \Delta\left(\alpha_0^1,\beta_0^1,X_3\right),\end{aligned}$$

and $\operatorname{var}(I_{33n}^p)$ is as $\operatorname{var}(I_{32n}^p)$ with α_0^1, β_0^1 and π replaced by α_0^0, β_0^0 and $1-\pi$ and $\operatorname{var}(I_{3jn}^p, I_{3kn}^p) = 0$ for $j \neq k = 1,2,3$. Thus the conclusion follows by CLT and CMT. ■

Table 1: Monte Carlo - Gaussian Design

		β_2^δ		α^δ
$\delta = 0$	n	Bias	IQR	RAMSE
Backfitting	100	-0.028	0.785	0.654
	200	-0.005	0.513	0.497
	400	0.003	0.353	0.390
Profile	100	-0.003	0.668	0.623
	200	0.009	0.457	0.487
	400	0.007	0.324	0.383
$\delta = 1$				
Backfitting	100	0.019	0.693	0.588
	200	0.027	0.457	0.473
	400	-0.007	0.281	0.400
Profile	100	-0.023	0.582	0.564
	200	0.030	0.424	0.467
	400	-0.006	0.284	0.394

Note: Gaussian design with $\beta_{20}^0 = 1$ and $\beta_{20}^1 = 3$. Similarly, $\alpha_0^0(u) = 3\sin(2u)$ and $\alpha_0^1(u) = 3\cos(2u)$. IQR stands for Inter Quartile range and RAMSE stands for Root Average Mean Square Error.

Table 2: Average Treatment Effect - Gaussian Design

	Backfitting		Profile		Eff. IPW	
n	Bias	IQR	Bias	IQR	Bias	IQR
100	0.0169	0.5580	-0.0167	0.5179	-0.0737	0.7503
200	-0.0158	0.3978	-0.0153	0.3720	-0.0466	0.5488
400	-0.0146	0.2870	-0.0143	0.2554	-0.0238	0.3939

Note: Gaussian design with $\tau_0 = 0$. Eff. IPW stands for the efficient semiparametric estimator of Hirano et al. (2003).

Table 3: Monte Carlo - Poisson Design

		β_2^δ		α^δ
$\delta = 0$	n	Bias	IQR	RAMSE
Backfitting	100	-0.916	4.409	3.025
	200	-0.235	0.939	1.742
	400	-0.080	0.458	1.189
Profile	100	-0.428	1.110	2.071
	200	-0.135	0.736	1.209
	400	-0.044	0.324	0.772
$\delta = 1$				
Backfitting	100	-0.400	1.959	1.462
	200	-0.138	0.757	1.050
	400	-0.071	0.396	0.783
Profile	100	-0.096	0.536	1.010
	200	-0.012	0.318	0.676
	400	-0.007	0.207	0.494

Note: Poisson design with $\beta_{20}^0 = \beta_{20}^1 = -1$ and $\alpha_0^0(u) = \alpha_0^1(u) = \sin(\pi u)$. IQR stands for Inter Quartile range and RAMSE stands for Root Average Mean Square Error.

Table 4: Average Treatment Effect - Poisson Design

	Backfitting		Profile		Eff. IPW	
n	Bias	IQR	Bias	IQR	Bias	IQR
100	-0.0182	0.5838	-0.0106	0.2493	0.0121	0.1765
200	-0.0085	0.3177	0.0038	0.1384	0.0102	0.1244
400	0.0041	0.1806	0.0011	0.0820	0.0033	0.0848

Note: Poisson design with $\tau_0 = 0$. Eff. IPW stands for the efficient semiparametric estimator of Hirano et al. (2003).

Table 5: Monte Carlo - Logit Design

$\delta = 0$	n	β_2^δ Bias	IQR	α^δ RAMSE
Backfitting	100	0.121	1.470	5.369
	200	-0.056	0.911	2.787
	400	-0.073	0.565	1.763
Profile	100	0.110	1.416	2.810
	200	-0.134	0.831	1.703
	400	-0.080	0.536	1.072
$\delta = 1$				
Backfitting	100	-0.099	0.763	2.213
	200	-0.068	0.464	1.393
	400	-0.038	0.307	0.990
Profile	100	-0.132	0.753	1.536
	200	-0.070	0.438	0.957
	400	-0.044	0.285	0.651

Note: Logit design with $\beta_{20}^0 = \beta_{20}^1 = -1$ and $\alpha_0^0(u) = \alpha_0^1(u) = \sin(\pi u)$. IQR stands for Inter Quartile range and RAMSE stands for Root Average Mean Square Error.

Table 6: Average Treatment Effect - Logit Design

	Backfitting		Profile		Eff. IPW	
n	Bias	IQR	Bias	IQR	Bias	IQR
100	-0.0132	0.1997	0.0072	0.1399	-0.0035	0.1466
200	-0.0044	0.0994	0.0039	0.0940	0.0033	0.0981
400	0.0019	0.0668	0.0012	0.0636	0.0013	0.0642

Note: Logit design with $\tau_0 = 0$. Eff. IPW stands for the efficient semiparametric estimator of Hirano et al. (2003).