Klein, Roger; Shen, Chan; Vella, Francis

Working Paper

Semiparametric selection models with binary outcomes

Working Paper, No. 2014-03

Provided in Cooperation with:
Department of Economics, Rutgers University

Suggested Citation: Klein, Roger; Shen, Chan; Vella, Francis (2014) : Semiparametric selection models with binary outcomes, Working Paper, No. 2014-03, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:
https://hdl.handle.net/10419/123813

Semiparametric Selection Models with Binary Outcomes

Roger Klein
Rutgers University

Chan Shen
The University of Texas
MD Anderson Cancer Center

Francis Vella
Georgetown University

Abstract

This paper addresses the estimation of a semiparametric sample selection index model where both the selection rule and the outcome variable are binary. Since the marginal effects are often of primary interest and are difficult to recover in a semiparametric setting, we develop estimators for both the marginal effects and the underlying model parameters. The marginal effect estimator uses only observations where the selection probability is above a certain threshold. A key innovation is that this high probability set is adaptive to the data. The model parameter estimator is a quasi-likelihood estimator based on regular kernels with bias corrections. We establish their large sample properties and provide simulation evidence confirming that these estimators perform well in finite samples.

1 Introduction

There is an extensive literature extending the sample selection model of Heckman (1974, 1979) to relax parametric restrictions.[1] For example, for models with continuous outcomes Ahn and Powell (1993) allowed for a nonparametric selection control function whereas more recently Das et al. (2003) allowed for a fully nonparametric treatment. There are also several papers which investigate the semiparametric treatment of the related model which features a binary response with a continuous endogenous explanatory variable (see, for example, Blundell and Powell, 2004 and Rothe, 2009). Despite this, there are relatively few papers that focus explicitly on the sample selection model with both binary outcomes and a binary selection rule. Moreover, there is no treatment of the semiparametric estimation of the marginal effects in such a model. This represents a significant void as important empirical examples exist in many areas of microeconomics.

In the fully parametric setting, both the model parameters and the marginal effects, which are generally the objects of primary interest, are easily obtainable. However, those are not easily attainable in the case of the semiparametric index model. The marginal effect is especially difficult to estimate because it cannot be directly derived from parameter estimates as the error distribution is unknown. The main issue arises from partial observability in that the outcome is only observed in the selected sample. Consequently, it is difficult to study the conditional probability of the outcome given that the individual is not selected as there are no observations available to estimate it. And this conditional probability is essential to estimating the outcome probability given exogenous variables by Bayes rule. A more complete discussion is provided in Section 3.1. Some important developments in this area include the work of Chesher (2005), Vytlacil and Yildiz (2007), and Shaikh and Vytlacil (2011), which discuss identification of the marginal impact of a discrete endogenous variable. However, detailed estimation of marginal effects has not been addressed for the case of sample selection.

In this paper we develop semiparametric estimators for both the marginal effects and the index parameters underlying them. We make no distributional assumptions and allow for a model structure that is more general than threshold-crossing. Our primary focus is upon the marginal effects as they have not been addressed in this setting. To deal with the partial observability issue mentioned above, we propose to estimate the relevant probabilities by focusing on those observations in an estimated high probability set where the selection probability tends to one. The framework of this approach is developed in pioneering

[1] For a survey see Vella (1998).

papers of Heckman (1990) and Andrews and Schafgans (1998) for a known high probability set. This set depends on the tail behavior of the index and error distributions. Therefore, in practice it is important to study the empirical tail behavior so as to find the appropriate high probability set. In this paper we characterize the high probability set as one where the probability exceeds a cutoff that approaches one as the sample size increases. We propose and establish the theoretical properties for an estimator of this cutoff that depends on the empirical tail behavior. Based on the estimated high probability set, we formulate a marginal effect estimator and provide the theory for it which takes the estimation of this set into account. This data-dependent feature of the high probability set underlying the marginal effect estimator poses a number of theoretical challenges, but is essential in empirical studies.

Estimation of the marginal effects requires estimates of the index parameters. To estimate them, we propose a likelihood-based procedure employing a double index formulation. Identification issues are explicitly treated in Newey (2007), although that paper does not address estimation. Our index parameter estimator employs bias adjustment mechanisms similar to those developed by Klein and Shen (2010) for single index regression models. We develop an estimator based on regular kernels and show that it has both desirable theoretical properties and good finite sample performance.[2] It is possible to develop index parameter estimators within various frameworks (see, e.g. Gallant and Nychka ,1987, Klein and Spady, 1993, Ichimura and Lee, 1991, Lee, 1995, and Klein and Vella, 2009). The estimator for index parameters provided here was initially developed in an earlier unpublished working paper.[3] Most recently Escanciano, Jacho-Chavez and Lewbel (2012) have proposed semiparametric estimators for the index parameters. Their estimator differs from ours in two main respects. First, to obtain asymptotic normality, we exploit a property of semiparametric derivatives (due to Whitney Newey) to control for the bias under regular kernels. Escanciano et. al. control for the bias using higher order kernels. Second, identification of index parameters here is based on exclusion restrictions, while Escanciano et. al. provide an alternative identification strategy. As mentioned above, there are also alternative frameworks for developing estimators for index parameters. However, the estimation of marginal effects remains unaddressed.

We describe the model in Section 2, and motivate the estimators for the marginal effects and the index parameters in Section 3. Assumptions and definitions are in Section 4. Section 5 provides a brief proof strategy, an illustrative example and asymptotic results for the marginal effect estimator. Similarly, the

[2] There are other alternative methods that control for the bias under regular kernels. For example, Honore and Powell (2005) employed a jackknife approach where the final estimator is a linear combination of estimators using different windows.

[3] See http://www.iza.org/conference_files/SPEAC2010/vella_f1653.pdf

proof strategy and asymptotic results for our index parameter estimator are presented in Section 6. We provide simulation evidence in Section 7 and offer concluding comments in Section 8. The Appendix contains all proofs.

2 Model

The model we address in this paper is a semiparametric variant on the Heckman (1974, 1979) selection model where the outcome of interest is binary. More explicitly:

$$Y_1 = I\{g(X\beta_0, \epsilon) > 0\} \quad (1)$$

$$Y_2 = I\{h(Z\pi_0, u) > 0\}, \quad (2)$$

where Y_1 is only observed for the subsample for which $Y_2 = 1$. Here $I\{.\}$ is an indicator function; X and Z are vectors of exogenous variables where Z includes at least one absolutely continuous element excluded from X; ϵ and u are error terms with a non-zero correlation; $g(.)$ and $h(.)$ are unknown functions. While the estimator for the index parameters is developed for a model of the above generality, large sample theory for the marginal effect requires us to characterize individuals with high selection probabilities. To this end, we assume $h(Z\pi_0, u) = Z\pi_0 - u$.[4] As in most semiparametric models the parameters are identified up to location and scale. Writing

$$X\beta_0 = b_1(X_1 + X_2\theta_{10}) + c_1 \equiv b_1 V_{10} + c_1$$

$$Z\pi_0 = b_2(Z_1 + Z_2\theta_{20}) + c_2 \equiv b_2 V_{20} + c_2,$$

the $\theta_0's$ are identified, while the $b's$ and $c's$ are not identified. We refer to V_{10} and V_{20} as indices and assume that the model satisfies the following index restrictions:

$$\Pr(Y_1 = d_1, Y_2 = d_2|X, Z) = \Pr(Y_1 = d_1, Y_2 = d_2|V_{10}, V_{20}) \quad (3)$$

$$\Pr(Y_2 = d_2|X, Z) = \Pr(Y_2 = d_2|V_{20}) \quad (4)$$

$$\Pr(Y_1 = d_1|X, Z) = \Pr(Y_1 = d_1|V_{10}). \quad (5)$$

[4] Since the error distribution is unknown, this threshold-crossing model is unchanged if we replace $Z\pi_0$ with any monotonic function of it.

We note that the above conditions hold if the errors are independent of X and Z. We impose this index structure, as opposed to a nonparametric one, to improve the performance of the estimators.

3 Motivation

3.1 Marginal Effects

The marginal effect of interest is the change in $\Pr(Y_1 = 1|X) = \Pr(Y_1 = 1|V_{10})$ due to a change in one of the explanatory X-variables. To motivate this marginal effect, let Y_2 denote whether or not an individual decides to have a diagnostic test for a particular genetic disease and let Y_1 denote whether or not an individual has that disease. We would like to know how a change in one of the X-variables affects the probability of having the disease for the entire population and not just the subgroup that received the diagnostic test. In the fully parametric case (e.g., bivariate probit with selection) the probability of having the disease $\Pr(Y_1 = 1|V_{10})$ is a known function, and the corresponding marginal effect of interest can be directly calculated once the parameters of the model are estimated.

Now consider the semiparametric case where the functional form of this probability function is not known. Under index restrictions, the probability of interest can be written as

$$\begin{aligned}\Pr(Y_1 = 1|V_{10}) &= \Pr(Y_1 = 1|V_{10}, V_{20}) \\ &= \Pr(Y_1 = 1|Y_2 = 1, V_{10}, V_{20})\, P_2 \\ &\quad + \Pr(Y_1 = 1|Y_2 = 0, V_{10}, V_{20})\,(1 - P_2),\end{aligned}$$

where $P_2 = \Pr(Y_2 = 1|V_{10}, V_{20}) = \Pr(Y_2 = 1|V_{20})$. We can recover the first argument on the right-hand side semiparametrically. The question then becomes how to recover the second part: $\Pr(Y_1 = 1|Y_2 = 0, V_{10}, V_{20})\,(1 - P_2)$. In general, this is not estimable because we do not observe Y_1 (genetic disease) when $Y_2 = 0$ (no testing). However, if $P_2 = 1$ this second term disappears and we can estimate the marginal effect of interest based on only the first term. In an approach related to that of Heckman (1990) and Andrews and Schafgans (1998, hereafter referred to as A&S), we estimate the marginal effect by using only those observations for which the selection probability P_2 is high. With N as the full sample size, F_U as the distribution function for the

selection error u, $Y_2 = I\{V_{20} > u\}$, and $a > 0$, the high probability set is defined as

$$\left\{v_{20} : F_U(v_{20}) > 1 - N^{-a}\right\}.$$

The probability of being in this high probability set is given by $P_h = \Pr(V_{20} > F_U^{-1}(1 - N^{-a})) = 1 - G_V\left(F_U^{-1}(1 - N^{-a})\right)$, where G_V is the distribution function for the selection index V_{20}. For example, when the index has a standard Weibull distribution $G_V = 1 - exp(-v_{20})$, and the error follows a Weibull distribution $F_U = 1 - exp(-u^c), c > 1$, $P_h = \exp(-[-\ln(N^{-a})]^{1/c})$. As the error tails become thinner relative to the index tail (c increases), P_h increases. This example demonstrates that the appropriate value of a depends on the thickness of the index tail relative to that of the error. As these tails are unknown, we propose a data dependent value for a and establish its asymptotic properties.

Assume that we are interested in the marginal impact of a particular exogenous variable X_m ceteris paribus. If we write the vector of exogenous variables as: $X = \left[X_{[m]}, X_m\right]$, we are basically studying the impact of moving X_m from a fixed baseline level x_{mb} to a fixed evaluation level x_{me}, keeping all other exogenous variables fixed at $x_{[m]}$. Define the corresponding baseline and evaluation levels for the outcome equation index as:

$$vb \equiv v\left(x_{[m]}, x_{mb}\right) \text{ and } ve \equiv v\left(x_{[m]}, x_{me}\right) \tag{6}$$

where $v(x)$ is the index value at $x = \left(x_{[m]}, x_m\right)$.

Employing the definitions above, let $\zeta_0(\bar{v}) \equiv \Pr(Y_1 = 1 | V_{10} = \bar{v})$ and define the true marginal effect as:

$$ME = \zeta_0(ve) - \zeta_0(vb). \tag{7}$$

When X_m is discrete, the above definition is natural. When X_m is continuous, it is also possible to define a marginal impact as a derivative. However, in applications we are more often than not more interested in measuring the impact of a discrete change than that of an infinitesimal change. For example, suppose that X_m denotes income and that income is continuous. Then we might want to know the impact of a 10% increase in income beyond a base level (e.g. median income). We also note that the marginal effect estimator for discrete changes converges to the truth faster than derivatives. Therefore we use the above definition of a marginal effect for examining perturbations in both discrete and continuous variables.

To motivate the marginal effect estimator, notice that a traditional semiparametric estimator for the

probability of interest would have the following form without sample selection

$$\widehat{\Pr}\left(Y=1|V=\bar{v}\right)=\frac{\sum_j\left\{\frac{1}{Nh}Y_{1j}K\left[\left(\bar{v}-V_{1j}\right)/h\right]\right\}}{\sum_j\left\{\frac{1}{Nh}K\left[\left(\bar{v}-V_{1j}\right)/h\right]\right\}}$$

where the sums are nonparametric kernel estimates with K a regular symmetric kernel and h a window parameter.

The estimator $\hat{\zeta}\left(\bar{v}\right)$, as in (D1) of Section 4, differs from this estimated probability in two respects. First, as we only observe Y_1 when $Y_2=1$, we need to have Y_{2j} in both the numerator and the denominator so as to select observations for which $Y_2=1$. This introduces sample selection bias that we eliminate (asymptotically) by adding the smooth high probability indicator $\hat{S}_j$. The definition of the S-function is in (D3) of Section 4 with a discussion of it at the end of that section. Further, we provide an illustrative example in Section 5.2.

3.2 Index Parameters

While the proposed marginal effect estimator is the primary focus, it depends on estimated index parameters. The index parameter estimators are obtained by maximizing a quasi or estimated likelihood. The true likelihood has the following form

$$L\left(\theta\right)\equiv\sum_{i=1}^{N}\sum_{d_1\leq d_2}Y_i(d_1,d_2)\ln\left(P_i\left(d_1,d_2;\theta\right)\right),$$

where

$$\begin{aligned}Y_i(d_1,d_2) &= \begin{cases} I\{Y_{1i}=d_1,Y_{2i}=d_2\} & \text{for } d_2=1\\ I\{Y_{2i}=d_2\} & \text{for } d_2=0\end{cases}\\ P_i\left(d_1,d_2;\theta\right) &\equiv \Pr\left(Y_i(d_1,d_2)=1|V_i\left(\theta\right)=v_i\left(\theta\right)\right).\end{aligned}$$

Here $V_i\left(\theta\right)=\left(V_{1i}\left(\theta\right),V_{2i}\left(\theta\right)\right)'$. In practice, we do not have the true P_i and need to estimate it ($\hat{P}_i$). The properties of our index parameter estimator depend on how these likelihood probabilities are estimated. We employ regular kernels and several bias-reducing mechanisms to ensure that the estimator has desirable large sample properties and also performs well in finite samples.

To motivate these mechanisms we show below that the gradient to the quasi-likelihood is a product of terms, one of which is the derivative of the probability function, $\nabla_\theta\hat{P}_i(d_1,d_2;\theta_0)$, where θ_0 denotes the true

parameter value. Subject to some issues that we address below, the key to our bias reduction mechanisms is the result due to Whitney Newey (see Klein and Shen, 2010, Theorem 0) that:

$$E\left(\nabla_\theta P_i(d_1, d_2; \theta_0) \mid V_i\left(\theta_0\right)\right) = 0. \tag{8}$$

It can be shown that we need to trim on the basis of estimated indices to take advantage of Newey's result. Accordingly, we develop a two stage estimation strategy. In the first stage, we estimate the index with trimming based on the continuous exogenous variables. We then trim observations on the basis of this estimated index in the second stage. This type of argument poses two problems which we solve. First, the consistency argument requires that estimated probability functions converge uniformly in the parameters to the corresponding true functions. When trimming is based on an estimated indices, the rate at which the density denominators vanish is only controlled in a neighborhood of the true parameter values. We solve this problem by using adjusted probabilities so as to control the rate at which density denominators vanish away from the truth. We set the adjustment so that it vanishes slowly when parameters are evaluated away from the truth but vanishes rapidly at the truth. The second problem is that we must be able to replace estimated probability derivatives with the corresponding true ones if we want to use Newey's result as a bias reducing mechanism. Here, we employ an adjustment to ensure that it is asymptotically valid to treat estimated probability derivatives as known.

4 Assumptions and Definitions

Here we provide the assumptions and definitions that we employ to establish the asymptotic properties for the estimators.

A1. The observations are i.i.d. from the model in (1)-(2), where the matrices X and Z have full rank with probability 1. For the marginal effects estimator, we restrict the function h to have the additively separable form given following (2).

A2. The vector of the true parameter values $(\theta_{10}, \theta_{20})$ lies in the interior of a compact parameter space, Θ.

A3. The indices V_1 and V_2 each contains at least one absolutely continuous exogenous variable. Further,

V_2 contains at least one absolutely continuous variable that does not enter V_1 in any form. The model satisfies index restrictions as in (3-5).

A4. Let $g_{V|Y}(v_1, v_2|y_1, y_2)$ be the conditional density for the indices. Letting $\nabla^p g_{V|Y}$ be any of the partials or cross partials of $g_{V|Y}$ up to order p, with $\nabla^0 g_{V|Y} = g_{V|Y}$, assume $g_{V|Y} > 0$ on all fixed compact subsets of the support for the indices, and $\nabla^p g_{V|Y}$, $\frac{\partial}{\partial\theta}\left(\nabla^p g_{V|Y}\right)$, and $\frac{\partial^2}{\partial\theta\partial\theta}\left(\nabla^p g_{V|Y}\right)$ are bounded for $p = 0, 1, 2, 3, 4$.

A5. Let F_U be the marginal distribution for the selection error, G_{V_2} the marginal distribution function for the selection index, and $G_{V_2|V_1}(v_2|\bar{v})$ the conditional distribution of $v_2|V_1 = \bar{v}$. With a_{0N} defined in (D4) below, characterize the high probability set as $\{v_2 : P_2 \equiv F_U(v_2) \geqslant 1 - N^{-a_{0N}}\}$. Assume: (a) For all $t_2 > t_1 > T$ sufficiently large,

$$\begin{aligned} G_{V_2}(t_2) - G_{V_2}(t_1) &> F_U(t_2) - F_U(t_1) \\ G_{V_2|V_1}(t_2|\bar{v}) - G_{V_2|V_1}(t_2|\bar{v}) &> F_U(t_2) - F_U(t_1). \end{aligned}$$

(b) The marginal density for the selection index $g_{V_2}(v_2)$, is decreasing in the tail. With $g_{V_2}(v_u) \equiv O(N^{-\varepsilon})$ where ε is a small positive number, and $H(v_u) \equiv \frac{g_{V_2}(v_u)}{1-G_{V_2}(v_u)}$ as the hazard for V_2: $\frac{1-F_U(v_u)}{1-G_{V_2}(v_u)} < H(v_u)N^{-a_{0N}}$.

A6. Let $g_{V_2|V_1}(v_2|\bar{v})$ be the density for V_2 conditioned on $V_1 = \bar{v}$. For all $t > T$ sufficiently large, assume that $O\left(g_{V_2}(t)\right) \geq O\left(g_{V_2|V_1}(t|\bar{v})\right)$.

A7. Assume $\Pr(Y_1 = d_1|Y_2 = d_2, V_1 = v_1, V_2 = v_2)$ has up to four bounded derivatives with respect to v_1 at $\bar{v}$.

A8. Assume (a) $\Pr(Y_1 = d_1|Y_2 = 1) > 0$, $\Pr(Y_2 = d_2) > 0$ and with g_X as the joint density for X, $\sup_\theta \left|\ln\left[P_i\left(d_1, d_2; \theta\right)\right]\right| g_X$ is integrable. Further, (b) $\sup_\theta E\left[\ln^2\left[P_i\left(d_1, d_2; \theta\right)\right]\right]$ is finite.

The first three assumptions are standard in index models. Assumption (A4) provides required smoothness conditions for determining the order of the bias for density estimators. As is well known in the literature (see e.g. Khan and Tamer (2010)), tail conditions are needed to develop the large sample distribution of these types of estimators. These conditions are provided in (A5a). Notice that the error and index supports can be finite provided these tail conditions hold. For example, when the error has a bounded support that

is a subset of that for the index, this assumption holds. However, when the index support is a subset of that for the error, this assumption will not hold. In Section 5.2, we illustrate the problem that results when these conditions do not hold.

Assumption (A5b) is required for the trimming arguments. Let v_l be a value of the selection index such that the selection probability $P_2(v_2) \equiv F_U(v_2) \geq 1 - N^{-a}$ for $v_2 \geq v_l$. Let v_u be a value of the selection index such that $g_{V_2}(v_2) \geq N^{-\iota}$ for $v_2 \leq v_u$. To avoid a conflict in these conditions, we need to guarantee that $v_l < v_u$. This inequality will hold provided that

$$1 - F_U(v_u) < 1 - F_U(v_l) \equiv N^{-a.}.$$

Dividing both sides by $1 - G_{V_2}(v_u)$ and noting that $g_{V_2}(v_u) \equiv N^{-\iota} = [1 - G_{V_2}(v_u)]\, H(v_u)$, where H is the hazard function for V_2, it suffices that

$$\frac{1 - F_U(v_u)}{1 - G_{V_2}(v_u)} < \frac{N^{-a}}{1 - G_{V_2}(v_u)} = H(v_u) N^{-(a-\iota)}.$$

Assumptions (A6-7) are used to derive the order of the bias in estimating the marginal effect components. For purposes of establishing consistency irrespective of whether the X's are bounded, Assumption (A8a) implies that the following expected log-likelihood is bounded and continuous in θ:

$$L(\theta) \equiv \sum_{i=1}^{N} I\{V_i \epsilon \Psi_v\} \sum_{d_1 \leq d_2} Y_i(d_1, d_2) \ln\left(P_i\left(d_1, d_2; \theta\right)\right).$$

while (A8b) is useful for obtaining consistency for the case in which several of the explanatory variables are unbounded**.** In addition to the above assumptions, we also need a number of definitions for densities, probability functions and estimators. These are given below.

D1. The estimator for marginal effects. Define:

$$\hat{\zeta}(\bar{v}, \hat{a}) \equiv \frac{\sum_j \frac{1}{N h_s} Y_{1j} K\left[\left(\bar{v} - V_{1j}\right)/h_s\right] Y_{2j} \hat{S}_j}{\sum_j \frac{1}{N h_s} K\left[\left(\bar{v} - V_{1j}\right)/h_s\right] Y_{2j} \hat{S}_j},$$

where K is a regular symmetric kernel, window $h_s = O(N^{-.2-3\varepsilon})$,[5] ε is a small positive number,

[5] Throughout the window is set to be $\sigma(V) N^{-r}$, where $\sigma(V)$ is the standard devation of the index in the kernel. The order of the window needs to be N^{-r} were $r < .2$. To simplify the form of other expressions that depend on this window, we set $r = .2 - 3\varepsilon$.

which is set to be 0.01 in the Monte Carlo simulations; and $\hat{S}_j$ is a smoothed trimming function on an estimated high probability set that depends on $\hat{a}$ (see (D3-4)).

D2. Expectation and Density Components. For $k = 1, 2$, and $d_k = 0, 1$, define:

$$\hat{f}_k\left(t_k; d_k, h\right) \equiv \sum_{j=1}^{N} \frac{Y_{kj}^{d_k}(1 - Y_{kj})^{1-d_k}}{Nh} K\left[\frac{t_k - V_{kj}}{h}\right]$$

D3. The S-function. With $b > 0$, define $S(\tau, x) \equiv \tau\left(\omega, g_{V_2}\left(v_2\right)\right) T(x)$,

$$\begin{aligned} T(x) &= \begin{cases} 0, & x \epsilon R1 \equiv \{x : x \leq 0\} \\ 1 - \exp \frac{-x^k}{b^k - x^k}, & x \epsilon R2 \equiv \{x : 0 < x < b\} \\ 1, & x \epsilon R3 \equiv \{x : x \geq b\} \end{cases} \\ \tau\left(\omega, g_{V_2}\left(v_2\right)\right) &\equiv \frac{1}{1 + \exp\left[N^{\alpha}\left(\omega - g_{V_2}\left(v_2\right)\right)\right]}, \omega = E(g_{V_2}) \frac{N^{-\varepsilon\prime}}{\ln N}, 0 < \varepsilon\prime < \varepsilon < \alpha \leq .2 \end{aligned}$$

The integer k is set to ensure that S is as many times differentiable as is needed at $x = 0$. The $T(x)$ is adapted from A&S.

D4. True and estimated high probability parameters a_{0N} and $\hat{a}$. Let $S_j \equiv S(\tau\left(\omega, g_{V_2}\left(v_{2j}\right)\right), x(a, P_{2j}))$where

$$x\left(a, P_2\right) \equiv \left[\ln\left(\frac{1}{1 - P_2}\right) - \ln\left(N^a\right)\right]$$

To define the estimator corresponding to S_j, let

$$\hat{P}_{aj} \equiv \frac{\sum_i \frac{1}{Nh_T} Y_{2i} K_T\left[\left(V_{2j} - V_{2i}\right)/h_T\right]}{\sum_i \frac{1}{Nh_T} K_T\left[\left(V_{2j} - V_{2i}\right)/h_T\right]},$$

where K_T a normal twicing kernel (Newey, Hsieh and Robins, 2004) and $h_T = O(N^{-.1})$. Next, in the notation of (D2), define the density estimator:

$$\hat{g}_{V_2}(t_2) = \hat{f}_2\left(t_2; d_2 = 1, h_2\right) + \hat{f}_2\left(t_2; d_2 = 0, h_2\right) = \sum_{j=1}^{N} \frac{1}{Nh_2} K(\frac{t_2 - V_{2j}}{h_2}),$$

with window $h_2 = O(N^{-.2})$. When the value t_k is replaced by the observation V_{ik}, the above averages are taken over the $(N - 1)$ observations for which $j \neq i$**.** Referring to (D3), define the estimated

S-function as:

$$\hat{S}_j \equiv S(\tau_j(\bar{w}, \hat{g}_{V_2}), x(a, \hat{P}_{aj})), \bar{w} \equiv \frac{N^{-\varepsilon'}}{\ln N}, \sum \hat{g}_{V_2}(\hat{V}_{2i})/N.$$

Further, for $\kappa = 1, 2$, let:

$$\hat{E}_2\left(\hat{S}\right) \equiv \frac{1}{N}\sum_j \hat{S}_j;\ \hat{E}_2\left(\hat{S}^\kappa|\bar{v}\right) \equiv \frac{\sum_j \frac{1}{Nh_T}\hat{S}_j^\kappa K_T\left[\left(\bar{v} - V_{1j}\right)/h_T\right]}{\sum_j \frac{1}{Nh_T}K_T\left[\left(\bar{v} - V_{1j}\right)/h_T\right]}.$$

Suppressing j subscripts, assume $\frac{[E(S)]^2}{E(S^2|\bar{v})} \leq 1$ and is an increasing function of N^{-a} for a sufficiently large N. Further, for p below a positive finite bound $\bar{p}$, $\frac{[E(S)]^2}{E(S^2|\bar{v})} = O_p(N^{-2ap})$. Then for $\varepsilon' > 0$, $\mathcal{A} = \{a : 0 < \varepsilon' < a \leq .4 - \varepsilon\}$ and we define:[6]

$$\begin{aligned} a_{0N} &= \arg\min_{a \epsilon \mathcal{A}} \left[h_s N^{1-2a+\varepsilon} \frac{[E(S)]^2}{E(S^2|\bar{v})} - 1\right]^2 \\ \hat{a} &= \arg\min_{a \epsilon \mathcal{A}} \left[h_s N^{1-2a+\varepsilon} \frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} - 1\right]^2 \end{aligned}$$

where $h_s = O(N^{-.2-3\varepsilon})$ from (D1).

D5. Unadjusted Probabilities and Densities. Let σ_k be the standard deviation for V_k, $k = 1, 2$. For the Y_2-model, and employing (D2), let

$$\widehat{\Pr}\left(Y_{2i} = d_2|V_{2i} = t_2\right) \equiv \hat{f}_2\left(t_2; d_2, h_m\right) / \sum_{d_2=0}^{1} \hat{f}_2\left(t_2; d_2, h_m\right),$$

where $h_m \equiv O(N^{-r_m})$, $r_m = \frac{1}{6+\varepsilon}$.

[6] In the definition of a_{0N}, substitute N^{-2ap} for $\frac{[E(S)]^2}{E(S^2|\bar{v})}$. Then, for N large:

$$a = \frac{.4 - \varepsilon}{1 + p}$$

Replacing p with its upper bound yields the lower bound for a with $\varepsilon' = \frac{.4-\varepsilon}{1+\bar{p}}$. The upper bound follows as p becomes small.

For the Y_1-model, conditioned on $Y_2 = 1$, let:

$$\begin{aligned}\widehat{\Pr}\left(Y_{1i} = d_1 | Y_{2i} = 1, V_i = t\right) &\equiv \hat{f}\left(t; d_1, h_{1c}, h_{2c}\right) / \sum_{d_1=0}^{1} \hat{f}\left(t; d_1, h_{1c}, h_{2c}\right) \\ \hat{f}\left(t; d_1, h_{1c}, h_{2c}\right) &\equiv \sum_{j=1}^{N} \frac{Y_{1j}^{d_1}(1 - Y_{1j})^{1-d_1} Y_{2j}}{N h_{1c} h_{2c}} K(\frac{t_1 - V_{1j}}{h_{1c}}) K(\frac{t_2 - V_{2j}}{h_{2c}}),\end{aligned}$$

where $h_{1c} = O(N^{-r_c}), h_{2c} = O(N^{-r_c}), r_c = \frac{1}{8+\varepsilon}$. The joint probability estimator is then given as:

$$\hat{P}_i(d_1, d_2; \theta) \equiv \widehat{\Pr}\left(Y_{1i} = d_1 | Y_{2i} = 1, V_i = t\right) \widehat{\Pr}\left(Y_{2i} = d_2 | V_{2i} = t_2\right).$$

Let $g_{V_1}(t_1)$ be the marginal density for V_1 at t_1 and the corresponding estimate:

$$\hat{g}_{V_1}(t_1) = \hat{f}_1\left(t_1; d_1 = 1, h_1\right) + \hat{f}_1\left(t_1; d_1 = 0, h_1\right) = \sum_{j=1}^{N} \frac{1}{N h_1} K(\frac{t_1 - V_{1j}}{h_1}),$$

where $h_1 = O(N^{-.2})$.When the value t_k is replaced by the observation V_{ik}, the above averages are taken over the $(N-1)$ observations for which $j \neq i$.[7]

D6. Interior Index Trimming. Let $\hat{V}_k^U$ and $\hat{V}_k^L$ be upper and lower sample quantiles for the indices: $V_k \equiv V_k(\theta)$, $k = 1, 2$. Define smooth interior trimming functions as

$$\begin{aligned}\hat{\tau}_I(t_k) &\equiv \left[1 + \exp\left(\ln(N)\left[\hat{V}_k^L - t_k\right]\right)\right]^{-1} \\ &\quad * \left[1 + \exp\left(\ln(N)\left[t_k - \hat{V}_k^U\right]\right)\right]^{-1}.\end{aligned}$$

D7. Adjusted Semiparametric Probability Functions. Referring to (D5), let $\hat{q}_{2L}(d_2)$ be a lower sample quantile for $\hat{f}_2\left(V_2; d_2\right)$, and $\hat{q}_L(d_1)$ a lower sample quantile for $\hat{f}\left(V; d_1\right)$. Then, define the adjusted estimates as

$$\begin{aligned}\hat{f}_2^*\left(t_2; d_2, h_m\right) &= \hat{f}_2\left(t_2; d_2, h_m\right) + \Delta_2(\hat{\tau}_I, \hat{q}_{2L}(d_2)) \\ \text{with } \Delta_2(\hat{\tau}_I, \hat{q}_{2L}(d_2)) &\equiv N^{-r_m/2}\left[1 - \hat{\tau}_I\left(t_2\right)\right] \hat{q}_{2L}(d_2)\end{aligned}$$

[7] It can easily be shown that all estimators with windows depending on population standard deviations are asymptotically the same as those based on sample standard deviations. For notational simplicity, we employ population standard deviations throughout.

$$\begin{aligned}
\hat{f}^{*}\left(t;d_1,h_{1c},h_{2c}\right) &= \hat{f}\left(t;d_1,h_{1c},h_{2c}\right)+\Delta(\hat{\tau}_I,\hat{q}_L(d_1)) \\
\text{with } \Delta(\hat{\tau}_I,\hat{q}_L(d_1)) &\equiv N^{-r_c/2}\left[1-\hat{\tau}_I\left(t_1\right)\hat{\tau}_I\left(t_2\right)\right]\hat{q}_L(d_1),
\end{aligned}$$

where h_m, h_{1c} and h_{2c} are defined as in (D5). Adjusted probabilities are now given as:

$$\begin{aligned}
\widehat{\Pr}^{*}\left(Y_{2i}=d_2|V_{2i}=t_2\right) &\equiv \hat{f}_2^{*}\left(t_2;d_2,h_m\right)/\sum_{d_2=0}^{1}\hat{f}_2^{*}\left(t_2;d_2,h_m\right) \\
\widehat{\Pr}^{*}\left(Y_{1i}=d_1|Y_{2i}=1,V_i=t\right) &\equiv \hat{f}^{*}\left(t;d_1,h_{1c},h_{2c}\right)/\sum_{d_1=0}^{1}\hat{f}^{*}\left(t;d_1,h_{1c},h_{2c}\right) \\
\hat{P}_i^{*}\left(d_1,d_2;\theta\right) &\equiv \widehat{\Pr}^{*}\left(Y_i(d_1,d_2)=1|V_i\left(\theta\right)=v_i\left(\theta\right)\right) \\
&\equiv \widehat{\Pr}^{*}\left(Y_{1i}=d_1|Y_{2i}=d_2,V_i=t\right)*\widehat{\Pr}^{*}\left(Y_{2i}=d_2|V_{2i}=t_2\right).
\end{aligned}$$

D8. Optimal Semiparametric Probability Functions. Referring to (D5), let

$$\hat{f}_2^{o}\left(t_2;d_2\right)\equiv\hat{f}_2\left(t_2;d_2,h_o\right);\ \hat{f}^{o}\left(t;d_1\right)\equiv\hat{f}\left(t;d_1,h_{1o},h_{2o}\right)$$

where $h_o=O(N^{-1/5})$ and $h_{1o}=O(N^{-1/6}), h_{2o}=O(N^{-1/6})$. Then, define

$$\begin{aligned}
\widehat{\Pr}^{o}\left(Y_{2i}=d_2|V_{2i}=t_2\right) &\equiv \hat{f}_2^{o}\left(t_2;d_2\right)/\sum_{d_2=0}^{1}\hat{f}_2^{o}\left(t_2;d_2\right) \\
\widehat{\Pr}^{o}\left(Y_{1i}=d_1|Y_{2i}=1,V_i=t\right) &\equiv \hat{f}^{o}\left(t;d_1\right)/\sum_{d_1=0}^{1}\hat{f}^{o}\left(t;d_1\right) \\
\hat{P}_i^{o}\left(d_1,d_2;\theta\right) &\equiv \widehat{\Pr}^{o}\left(Y_i(d_1,d_2)=1|V_i\left(\theta\right)=v_i\left(\theta\right)\right) \\
&\equiv \widehat{\Pr}^{o}\left(Y_{1i}=d_1|Y_{2i}=d_2,V_i=t\right)*\widehat{\Pr}^{o}\left(Y_{2i}=d_2|V_{2i}=t_2\right).
\end{aligned}$$

D9. The Initial Estimator for Index Parameters. Define the initial or first stage estimator as

$$\begin{aligned}
\hat{\theta} &\equiv \arg\max_{\theta}\hat{L}\left(\theta\right), \\
\hat{L}\left(\theta\right) &\equiv \sum_{i=1}^{N}I\{X_{\mathrm{Ci}}\epsilon\hat{\Psi}_x\}\sum_{d_1\le d_2}Y_i(d_1,d_2)\ln\left(\hat{P}_i\left(d_1,d_2;\theta\right)\right).
\end{aligned}$$

Let

$$\hat{\delta}_i\left(d_1, d_2; \theta\right) \equiv \nabla_\theta \hat{P}_i(d_1, d_2; \theta) / \hat{P}_i(d_1, d_2; \theta)$$

With X_{Ci} as a vector of the continuous variable in X_i, $\hat{q}_{Lx}$ as a vector of lower sample quantiles for X_{Ci}, and $\hat{q}_{Ux}$ as a vector of upper sample quantiles for X, define $\hat{\Psi}_x$ as:

$$\hat{\Psi}_x \equiv \{x : \hat{q}_{Lx} < x < \hat{q}_{Ux}\}.$$

Define a gradient correction as:

$$\hat{C}_X\left(\hat{\theta}_I\right) \equiv \sum_{i=1}^{N} I\{X_{\text{Ci}} \epsilon \hat{\Psi}_x\} \sum_{d_1 \leq d_2} \left[\hat{P}_i\left(d_1, d_2; \hat{\theta}^*\right) - \hat{P}_i^o\left(d_1, d_2; \hat{\theta}^*\right)\right] \hat{\delta}_i\left(d_1, d_2; \hat{\theta}\right).$$

With $\hat{H}\left(\hat{\theta}_I\right)$ as the estimated Hessian matrices from $\hat{L}$, the adjusted initial estimator estimator is defined as

$$\hat{\theta}_I^o \equiv \hat{\theta}_I - \hat{H}\left(\hat{\theta}_I\right)^{-1} \hat{C}_X\left(\hat{\theta}_I\right).$$

D10. The Final Estimator for Index Parameters. With $\hat{V}_i \equiv X_{1i} + X_{2i}\hat{\theta}_I^o$, $\hat{q}_{Lv}$ as a vector of lower sample quantiles for the $\hat{V}_i's$, and $\hat{q}_{Uv}$ as the corresponding vector of upper sample quantiles, define $\hat{\Psi}_v$ as:

$$\hat{\Psi}_v \equiv \{v : \hat{q}_{Lv} < v < \hat{q}_{Uv}\}.$$

Define the second stage estimator as follows:

$$\begin{aligned} \hat{\theta}^* &\equiv \arg\max_\theta \hat{L}^*(\theta). \\ \hat{L}^*(\theta) &\equiv \sum_{i=1}^{N} I\{\hat{V}_i \epsilon \hat{\Psi}_v\} \sum_{d_1 \leq d_2} Y_i(d_1, d_2) \ln\left(\hat{P}_i^*\left(d_1, d_2; \theta\right)\right). \end{aligned}$$

Let

$$\hat{\delta}_i^*\left(d_1, d_2; \theta\right) \equiv \nabla_\theta \hat{P}_i^*(d_1, d_2; \theta) / \hat{P}_i^*(d_1, d_2; \theta),$$

and define $\hat{P}^o(d_1, d_2; \theta)$ as in (D8). Then, define a gradient correction as:

$$\hat{C}_V\left(\hat{\theta}^*\right) \equiv \sum_{i=1}^{N} I\{\hat{V}_i \epsilon \hat{\Psi}_v\} \sum_{d_1 \leq d_2} \left[\hat{P}_i^*\left(d_1, d_2; \hat{\theta}^*\right) - \hat{P}_i^o\left(d_1, d_2; \hat{\theta}^*\right)\right] \hat{\delta}_i^*\left(d_1, d_2; \hat{\theta}^*\right).$$

With $\hat{H}^*\left(\hat{\theta}^*\right)$ as the estimated Hessian matrix from $\hat{L}^*$ above, the adjusted final estimator is defined as

$$\hat{\theta}^o \equiv \hat{\theta}^* - \hat{H}^*\left(\hat{\theta}^*\right)^{-1} \hat{C}_V\left(\hat{\theta}^*\right).$$

As mentioned in Section 3.1, the S-function in (D3) plays an important role in the marginal effect estimation by smoothly restricting observations to a high probability set and at the same time ensuring that density denominators are not too small. Assumption (A5) ensures that there is no conflict in these restrictions, which is further illustrated in Section 5.2. By restricting observations in this manner, we affect both the bias and the variance of the estimator. For example, if the high probability set parameter is set too high, the bias will be small but the variance will be large due to the small "effective" sample size. Similarly, if the high probability parameter is too small, then there will be a substantial bias in the estimator. The moment conditions in (D4) reflect the bias-variance trade-off in estimating the marginal effect. From Lemmas 3-4,

$$O\left(\frac{bias^2}{var}\right) \leq \frac{O(h_s N^{1-2a})\left[E(S)\right]^2}{E(S^2|\bar{v})}. \tag{9}$$

To maximize the rate at which the mean-squared error tends to zero, the order of the bias squared and the variance are usually set to be the same. However, to ensure normality we require that the bias squared vanishes slightly faster than the variance. This requirement generates the moment condition in (D4).

5 Marginal Effect Estimators

5.1 Proof Strategy

Before providing theorems on consistency and normality, we discuss the proof strategy for the marginal effect estimator. To facilitate the discussion, we begin with the following simplified (and infeasible) estimator:

$$\gamma(\bar{v}, a) \equiv \frac{\sum_j \left\{\frac{1}{Nh_s} Y_{1j} K\left[\left(\bar{v} - V_{1j}\right)/h_s\right]\right\} \quad (Y_{2j} S_j)}{E\left(S|\bar{v}\right) g_{V_1}(\bar{v})}, \tag{10}$$

which differs from the original estimator (D1) in that it depends on a known high probability parameter, an S-function that depends on a known semiparametric probability function, and a denominator which is the population expectation corresponding to the denominator of the original estimator. Our strategy is to first analyze the infeasible estimator and then establish its relation to the original estimator. In the Appendix, we establish the orders of the squared bias (Lemma 3) and the variance (Lemma 4) for the simplified estimator:

$$Bias^2\left[\gamma\left(\bar{v},a\right)\right]=O\left[b_N^2\left(\bar{v},a\right)\right];\quad Var\left[\gamma\left(\bar{v},a\right)\right]=O\left[v_N(\bar{v},a)\right].$$

To balance the bias/variance trade-off, we could set a_{0N} such that

$$b_N^2\left(\bar{v},a_{0N}\right)=v_N(\bar{v},a_{0N}).$$

Then, with $\hat{b}_N$ and $\hat{v}_N$ as estimators for the bias and variance functions, we could set $\hat{a}$ to satisfy:

$$\hat{b}_N^2\left(\bar{v},\hat{a}\right)=\hat{v}_N(\bar{v},\hat{a}).$$

However, to establish asymptotic normality, we actually set a_{0N} such that the square bias vanishes at a rate close to but slightly faster than the variance and similarly for $\hat{a}$.

With estimated and true high probability parameters set as above, we show that it suffices to analyze the simplified, infeasible estimator by showing that:

$$\sqrt{1/v_N(\bar{v},a_{0N})}\left[\gamma\left(\bar{v},a_{0N}\right)-\hat{\zeta}\left(\bar{v},\hat{a}\right)\right]=o_p(1).$$

It is then possible to establish asymptotic results for the marginal effect estimator. The proofs are in the Appendix.

We note that unlike the index case, the proof of asymptotic normality for the marginal effect estimator is not based on $\sqrt{N}$-asymptotics. Employing the characterization above, we show that

$$\frac{\widehat{ME}-ME}{\sqrt{\hat{V}_N}}\xrightarrow{d}Z\tilde{}N(0,1),$$

where with V_N interpreted as a deterministic variance sequence, $\hat{V}_N-V_N\xrightarrow{p}0$. The convergence rate for

the marginal effect estimator is given by the order of $\sqrt{V_N}$ and is not known a priori. It depends in part on tail properties of the index and error distributions in the selection equation. Consequently, proofs are quite different than in the case of $\sqrt{N}$-asymptotics. Nevertheless, as in A & S, the limiting standard normal result makes it possible to conduct inference as usual. We provide an illustrative example in which we determines the convergence rate below.

5.2 Illustrative Example

In this subsection, we provide an illustrative example to show the importance of the tail assumptions and their implications. When these assumptions hold, we will determine the high probability set and the convergence rate for the marginal effect estimator. For this example, let F_U and G_{V_2} be the following Weibull distribution functions for the selection error and index, respectively:

$$F_U(u) = 1 - \exp(-u^c);\ G_{V_2}(v_2) = 1 - \exp(-v_2). \tag{11}$$

The theory depends on being able to control for density denominators while still having a sufficiently large number of observations in the high probability set. To ensure that estimated selection probabilities converge to the truth, we trim observations to ensure that:

$$g_{V_2}(v_2) = \exp(-v_2) > N^{-\varepsilon} \Leftrightarrow v_2 < \varepsilon \ln(N). \tag{12}$$

On the other hand, to be in the high probability set, we select those observations for which

$$P_2 = 1 - \exp(-v_2^c) > 1 - N^{-a} \Leftrightarrow v_2 > [a \ln(N)]^{1/c} \tag{13}$$

If error tails are fatter than index tails ($c < 1$), (12) and (13) cannot hold simultaneously. In this case, for a large N the set of v_2 values satisfying (12) and (13) will be empty.

If $c > 1$, then index tails will be fatter than error tails as we have assumed. In this case for a large N, there will be a set with positive probability (calculated below) on which both conditions hold. For this case, we proceed to find the high probability set and the rate of convergence for the estimator. Assuming for simplicity that selection and outcome indices are independent, from (D4), the high probability set parameter,

a_{0N}, satisfies:

$$\frac{\ln\left(h_s N^{1-2a_{0N}}\right)}{\ln N} + \frac{1}{\ln N}\ln\left(\frac{[E(S)]^2}{E(S^2)}\right) = -\varepsilon. \tag{14}$$

Notice that the second left-hand term converges to zero from the proposition below. Then, with $h_s = N^{-.2-3\varepsilon}$, a_{0N} satisfies:

$$.8 - 2\varepsilon - 2a_{0N} \to 0.$$

Therefore, in large samples, $a_{0N} = .4 - \varepsilon$. Further, it immediately follows from the proposition below that the convergence rate of the marginal effect estimator will be

$$O\left(\sqrt{\frac{Nh_s\left[E(S)\right]^2}{E(S^2)}}\right) = O\left(\sqrt{Nh_s \exp[(-a_{0N}\ln N)^{1/c}}\right).$$

Proposition 1 $.\frac{[E(S)]^2}{E(S^2)} = O(\exp[-(a\ln N)^{1/c}])$.

Proof. *We establish the order* $\frac{[E(S)]^2}{E(S^2)}$ *by bounding the S-function for large N. We begin by bounding the* $\tau-$*component of the S-function that controls density denominators. Since* τ *is an increasing function of the index density,* $g_{V_2}(v_2)$*, suppressing the* ω *argument of the* τ *function, with any* $0 < \delta < 1$*, we have:*

$$\begin{aligned}
\tau(g_{V2}) &= \tau(g_{V2}) * 1\{g_{V2} > (1-\delta)\omega\} + \tau(g_{V2}) * 1\{g_{V2} \leq (1-\delta)\omega\} \\
&\leq 1\{g_{V2} > (1-\delta)\omega\} + \tau\left(g_{V2} = (1-\delta)\omega\right),
\end{aligned}$$

where $\tau\left(g_{V2} = (1-\delta)\omega\right)$ *vanishes exponentially fast. For the lower bound on* τ *:*

$$\begin{aligned}
\tau(g_{V2}) &= \tau(g_{V2}) * 1\{g_{V2} > (1+\delta)\omega\} + \tau(g_{V2}) * 1\{g_{V2} \leq (1+\delta)\omega\} \\
&\geq \tau\left(g_{V2} = (1+\delta)\omega\right) 1\{g_{V2} > (1+\delta)\omega\},
\end{aligned}$$

where $\tau\left(g_{V2} = (1+\delta)\omega\right)$ *converges exponentially to 1. It now follows from the definition of S in (D3-4) and*

(12-13) that:

$$\begin{aligned}
S^- &\leq S \equiv \tau(g_{V2})T(X) \leq S^+, \\
S^- &\equiv \tau\left(g_{V2} = (1+\delta)\omega\right) 1\{g_{V2} > (1+\delta)\omega\} 1\left\{X > b\right\} \\
&= \tau\left(g_{V2} = (1+\delta)\omega\right) 1\left\{[a\ln N + b]^{1/c} \leq V_2 \leq -\ln((1+\delta)\omega)\right\} \\
S^+ &\equiv 1\{g_{V2} > (1-\delta)\omega\} 1\left\{X > 0\right\} + \tau\left(g_{V2} = (1-\delta)\omega\right) \\
&= 1\left\{[a\ln N]^{1/c} \leq V_2 \leq -\ln((1-\delta)\omega\right\} + \tau\left(g_{V2} = (1-\delta)\omega\right).
\end{aligned}$$

Recall (11),when $c > 1$,

$$\begin{aligned}
E\left(S^-\right) &= \left[e^{-[a\ln N+b]^{1/c}} - (1+\delta)\omega\right] \tau\left(g_{V2} = (1+\delta)\omega\right) \\
E\left[\left(S^-\right)^2\right] &= \left[e^{-[a\ln N+b]^{1/c}} - (1+\delta)\omega\right] \tau\left(g_{V2} = (1+\delta)\omega\right)^2 \\
E\left(S^+\right) &= e^{-[a\ln N]^{1/c}} - (1-\delta)\omega + \tau\left(g_{V2} = (1-\delta)\omega\right) \\
E\left[\left(S^+\right)^2\right] &= \left[e^{-[a\ln N]^{1/c}} - (1-\delta)\omega\right] \left[1 + 2\tau\left(g_{V2} = (1-\delta)\omega\right)\right] + \tau^2\left(g_{V2} = (1-\delta)\omega\right).
\end{aligned}$$

Since $\tau\left(g_{V2} = (1+\delta)\omega\right)$ *and* $\tau\left(g_{V2} = (1-\delta)\omega\right)$ *converge exponentially fast to 1 and 0 respectively and since* $(1-\delta)\omega$ *converges to 0 faster than does* $e^{-[a\ln N]^{1/c}}$*, each of the above expectations has order* $O(\exp[-(a\ln N)^{1/c}])$*. Therefore, as* $\frac{\left[E(S^-)\right]^2}{E[(S^+)^2]} \leq \frac{[E(S)]^2}{E(S^2)} \leq \frac{\left[E(S^+)\right]^2}{E[(S^-)^2]}$*, it follows that* $\frac{[E(S)]^2}{E(S^2)} = O(\exp[-(a\ln N)^{1/c}])$*.* ■

5.3 Asymptotic Results

Theorem 1 (Consistency). Refer to (6), select the high probability set as in (D4), and assume that

$$NhE(S|V_1 = \bar{v}) \to \infty$$

as N increases. Then, under A1-4,A5(b),A6-7,

$$\hat{\zeta}\left(\bar{v}, \hat{a}\right) \xrightarrow{p} \zeta_0\left(\bar{v}\right).$$

Theorem 2 (Normality). Refer to (6) and (7), let

$$\begin{aligned}\hat{V}_N &= \widehat{Var}(\gamma(ve, a_{0N})) + \widehat{Var}(\gamma(vb, a_{0N})) \\ \text{where } \widehat{Var}(\gamma(\bar{v}, a_{0N})) &= \frac{\hat{\zeta}(\bar{v}, \hat{a})\left[1 - \hat{\zeta}(\bar{v}, \hat{a})\right] \sum_j \frac{1}{Nh_s} K^2\left[(\bar{v} - V_{1j})/h_s\right] \hat{S}_j^2}{Nh_s \hat{E}_2^2\left(\hat{S}|\bar{v}\right) \hat{g}_{V_1}^2(\bar{v})}.\end{aligned}$$

Then under A1-7,

$$\frac{\widehat{ME} - ME}{\sqrt{\hat{V}_N}} \xrightarrow{d} Z \tilde{} N(0,1).$$

6 Index Parameter Estimators

6.1 Proof Strategy

To provide an overview of the theoretical arguments, we note that the consistency argument is rather standard except that we need to accommodate the bias controls used in the normality arguments. The main challenge is in the proof for normality. As discussed earlier and as shown below, we employ a two-stage construction in order to exploit Newey's result on probability derivatives for controlling the bias. In the first stage, we trim on continuous X's and obtain parameter estimates by maximizing the quasi-likelihood in (D9) and then make a bias correction. In the second stage, with trimming now based on the estimated index and with probabilities "adjusted" to ensure consistency (by controlling density denominators away from the truth), we maximize the quasi-likelihood in (D10). From standard Taylor series arguments, with $\hat{H}\left(\theta^+\right)$ as the Hessian to the objective function at an intermediate point, the estimator $\hat{\theta}^*$ satisfies:

$$\sqrt{N}\left(\hat{\theta}^* - \theta_0\right) = -\hat{H}^*\left(\theta^+\right)^{-1} \sqrt{N}\left[\hat{A}^* + \hat{B}^*\right],$$

where

$$\begin{aligned}\hat{A}^* &\equiv \frac{1}{N}\sum_{i=1}^N I\{\hat{V}_i \epsilon \hat{\Psi}_v\} \sum_{d_1 \le d_2} \left[Y_i(d_1, d_2) - P_i(d_1, d_2; \theta_0)\right] \hat{\delta}_i^*(d_1, d_2; \theta_0) \\ \hat{B}^* &\equiv \frac{1}{N}\sum_{i=1}^N I\{\hat{V}_i \epsilon \hat{\Psi}_v\} \sum_{d_1 \le d_2} \left[P_i(d_1, d_2; \theta_0) - \hat{P}_i^*(d_1, d_2; \theta_0)\right] \hat{\delta}_i^*(d_1, d_2; \theta_0) \\ \hat{\delta}_i^*(d_1, d_2; \theta) &\equiv \nabla_\theta \hat{P}_i^*(d_1, d_2; \theta) / \hat{P}_i^*(d_1, d_2; \theta).\end{aligned}$$

While it is relatively straight forward to analyze the Hessian and $\hat{A}^*$ components, the $\hat{B}^*$ component involves bias issues. We deal with this problem by employing a bias correction that effectively enables us to replace this component with one with desirable bias properties. Employing the correction factor in (D10), our final estimator is

$$\hat{\theta}^o \equiv \hat{\theta}^* - \hat{H}^* \left(\hat{\theta}^*\right)^{-1} \hat{C}_V \left(\hat{\theta}^*\right).$$

We show that it has the following convenient form:

$$\sqrt{N}\left(\hat{\theta}^o - \theta_0\right) = -\hat{H}^* \left(\theta^+\right)^{-1} \sqrt{N}\left[\hat{A}^* + \hat{B}^o\right] + o_p(1)$$

where with $\hat{P}_i^o$ as an estimated probability function based on an optimal window,

$$\hat{B}^o \equiv \frac{1}{N}\sum_{i=1}^{N} I\{\hat{V}_i \epsilon \hat{\Psi}_v\} \sum_{d_1 \leq d_2} \left[P_i\left(d_1, d_2; \theta_0\right) - \hat{P}_i^o\left(d_1, d_2; \theta_0\right)\right] \hat{\delta}_i^*\left(d_1, d_2; \theta_0\right).$$

To analyze $\sqrt{N}\left(\hat{\theta}^o - \theta_0\right)$, from standard uniform convergence arguments, the Hessian component converges to a fixed matrix. To analyze the $\hat{A}^*$ component above, we employ Lemma 2.17 in Pakes and Pollard (1989) to deal with the indicator on an estimated set and a mean-square convergence argument to deal with $\hat{\delta}_i^*$. We show that with $\delta_i \equiv \nabla_\theta P_i(d_1, d_2; \theta_0)/P_i(d_1, d_2; \theta_0)$,

$$\sqrt{N}\hat{A}^* = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} I\{V_i \epsilon \Psi_v\} \sum_{d_1 \leq d_2} \left[Y_i(d_1, d_2) - P_i\left(d_1, d_2; \theta_0\right)\right] \delta_i + o_p(1).$$

For the $\hat{B}^o$ component, employing convergence rates for estimated indicators, probabilities and probability derivatives, we show that:

$$\sqrt{N}\hat{B}^o = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} I\{V_i \epsilon \Psi_v\} \sum_{d_1 \leq d_2} \left[P_i\left(d_1, d_2; \theta_0\right) - \hat{P}_i^o\left(d_1, d_2; \theta_0\right)\right] \delta_i + o_p(1).$$

We show that the expression on the right-hand-side is asymptotically equivalent to a U-statistic with expectation 0. This follows from Newey's result which ensures that $E\left[\delta_i | V\right] = 0$. From standard projection arguments we are able to show that this U-statistic vanishes in probability. Asymptotic normality then follows. Consistency is relatively easy to establish. Detailed proofs are in the Appendix.

6.2 Asymptotic Results

In the theorems below, we note that the consistency holds under much weaker window conditions than does normality.

Theorem 3 (Consistency). Assume that each index satisfies the identifying assumptions required for single index models.[8] Then, under (A1-4,8),

$$\hat{\theta} \xrightarrow{p} \theta_0, \hat{\theta}^* \xrightarrow{p} \theta_0, \hat{\theta}^o \xrightarrow{p} \theta_0.$$

Theorem 4 (Normality). With $L(\theta)$ as the limiting likelihood of $\hat{L}^*(\theta)$ defined in (D10) and with H as its Hessian matrix, define $H_0 \equiv EH(\theta_0)$. Then under (A1-4,8):

$$\sqrt{N}\left[\hat{\theta}^o - \theta_0\right] \xrightarrow{d} Z \tilde{} N(0, -H_0^{-1}).$$

7 Simulation Evidence

In this section, we consider the finite sample performance of the estimator in four different models. These differ according to: i) whether or not the model is threshold-crossing; and ii) whether the continuous variables and errors follow a Normal or Weibull distribution. The first two models we consider have threshold-crossing structures. The first model (TNorm design) has normal errors and is given as

$$\begin{aligned} Y_1 &= I\left\{\sqrt{2}(X_1 + X_3) > \epsilon\right\} \\ Y_2 &= I\left\{\sqrt{2}(X_2 - X_3) > u\right\}, \end{aligned}$$

where Y_1 is observed when $Y_2 = 1$. The errors and the continuous $X\imath s$ (X_1, X_2) are generated as

$$\begin{aligned} u, X_2 &\sim N(0,1) \\ \epsilon &= 2u + z, \;\; z \sim N(0,1) \\ X_1 &= X_2 + 2z_1, \;\; z_1 \sim N(0,1), \end{aligned}$$

[8] See, for example, Ichimura (1993) or Klein and Spady (1993).

and rescaled to each have variance one, while X_3 is a binary variable with a probability of .5 on each of its support points, {-1,1}. Notice that the indices have a standard deviation of 2. For the second index, this ensures that the index has fatter tails than the error, which is theoretically needed in estimating the marginal effect.

In a second model (TWeibull design), the selection error is non-normal while the model structure stays the same. The error u follows a Weibull (1,1.5) distribution giving a right tail probability of $exp(-u^{1.5})$. We set X_2 to follow Weibull (1,1) distribution so that the tail comparison condition is satisfied. As stated above, all the variables and errors are rescaled to have zero mean and variance one.

In the third (NTNorm design) and fourth (NTWeibull design) models, the Y_1 equation has the following non-threshold-crossing structure:

$$Y_1 = I\left\{X_1 + X_3 > s\left[1 + (X_1 + X_3)^2/4\right]\epsilon\right\},$$

where the variables are generated as in the previous models. Note that s is chosen to ensure the right-hand-side of the inequality is rescaled to have variance one as above. Similar to the first two models above, here the third and fourth models differ according to whether Normal or Weibull distributions are employed.

For all models we set $N = 2000$ and conduct 1000 replications. We compare the finite sample performance of the semiparametric marginal effect estimator with the bivariate probit counterpart. We also compare the parameter estimates upon which these marginal effects are based. Finally, we provide results for the estimation of the high probability set. Notice that there is an infinite number of marginal effects because there is an infinite number of base levels and evaluation levels. Here we report the marginal effect of moving X_1 from its median level to one unit above while keeping the binary variable X_3 at zero. Finally, we set $b = 0.01$, with b as a parameter in the definition of the high probability trimming function T in (D3).

Table 1. Estimation of Marginal Effects

	Truth		Bivariate Probit	Semiparametric
TNorm	0.33	mean	0.34	0.32
		std	0.03	0.05
		RMSE	0.03	0.05
TWeibull	0.29	mean	0.37	0.31
		std	0.06	0.06
		RMSE	0.09	0.07
NTNorm	0.47	mean	0.36	0.48
		std	0.08	0.05
		RMSE	0.13	0.05
NTWeibull	0.43	mean	0.47	0.48
		std	0.17	0.06
		RMSE	0.18	0.08

With results for the marginal effects shown in Table 1, overall the semiparametric estimator performs well in all designs with a small bias and standard deviation. In contrast, the bivariate probit counterpart does not perform well outside of the TNorm design where bivariate probit is correct. In the TNorm case, where bivariate probit is the correct specification, it does indeed have a small bias and standard deviation and performs better than the semiparametric method. In the TWeibull case, the semiparametric method shows significant advantage in terms of the bias. The bias of the semiparametric marginal effect estimator is .02, while the bivariate probit counterpart has a bias of .08, which is almost 30% of the truth (.29). When we move onto the non-threshold-crossing designs, we continue to see the semiparametric estimator performing significantly better. In the NTNorm case, the semiparametric estimator has both smaller bias (.01 vs .11) and smaller standard deviation (.05 vs .08). In the NTWeibull case, the semiparametric estimator still performs much better than the bivariate probit in terms of both bias and variance, resulting in a much smaller RMSE (.08 vs .18). Most of the advantage comes from the standard deviation (.06 vs .17). We have also explored the sensitivity of the results to the point at which the marginal effect was calculated and found that a larger sample size is needed for the marginal effects estimator to perform well when we evaluate it further away from the center of the distribution.

Table 2. Estimation of Index Parameters

	Bivariate Probit						Semiparametric	
	Outcome			Selection			Outcome	Selection
	Coef(X1)	Coef(X3)	Ratio_{31}	Coef(X2)	Coef(X3)	Ratio_{32}	Ratio_{31}	Ratio_{32}
TNorm								
mean	1.01	1.02	1.01	1.00	-1.00	-1.00	0.98	-1.02
std	0.07	0.13	0.11	0.05	0.04	0.03	0.07	0.04
RMSE	0.07	0.13	0.11	0.05	0.04	0.03	0.07	0.04
TWeibull								
mean	1.20	1.20	1.00	1.02	-1.06	-1.05	0.94	-1.04
std	0.09	0.23	0.16	0.06	0.04	0.04	0.06	0.04
RMSE	0.22	0.30	0.16	0.06	0.08	0.06	0.09	0.06
NTNorm								
mean							0.96	-1.02
median	1.07	1.13	1.04	1.00	-1.00	-1.00	0.97	-1.02
std							0.05	0.04
MAD	0.11	0.14	0.08	0.03	0.03	0.02	0.04	0.03
RMSE							0.06	0.04
NTWeibull								
mean							0.93	-1.04
median	1.51	1.46	0.95	1.03	-1.04	-1.05	0.93	-1.04
std							0.05	0.04
MAD	0.52	0.47	0.08	0.05	0.06	0.05	0.07	0.04
RMSE							0.08	0.06

We provide the index parameter estimation results in Table 2. For semiparametric estimation, the parameters are identified up to location and scale, hence we report $\text{Ratio}_{31}=\frac{coef(X_3)}{coef(X_1)}$ in the outcome equation and $\text{Ratio}_{32}=\frac{coef(X_3)}{coef(X_2)}$ in the selection equation. Notice that for the non-threshold-crossing designs, we report the median and median absolute deviation (MAD) for the bivariate probit estimators because there were a number of replications where bivariate probit performed extremely poorly. The semiparametric estimator,

however, does not have this issue, hence we report not only the median and MAD but also the mean, standard deviation, and RMSE. For the parametric case, we also report ratios to facilitate direct comparisons with the semiparametric method. For the selection equation, over all designs, both parametric and semiparametric estimators perform quite well. Regarding the outcome equation, the semiparametric estimator performed better in all designs and usually by a substantial amount in terms of MAD or RMSE.

We also investigated using higher order kernels for estimating index parameters as an alternative to the bias controls implemented here.[9] Due to convergence problems, we found it necessary to examine this estimator on a two-dimensional grid, which was quite time-consuming. Accordingly, we only examined 100 replications for each design (at which point the estimator seemed quite stable). For the selection equation, the RMSEs were similar with the exception of the TWeibull design. For this design and for all designs pertaining to the outcome equation, the RMSEs under higher order kernels were more than twice as large as that under regular kernels.

Lastly, we provide the estimation results for the high probability set parameters. The means of $\hat{a}$ with standard deviations in parentheses are as follows: .32(.004), .29(.005), .29(.004), .29(.005) for TNorm, TWeibull, NTNorm, NTWeibull, respectively. While the variances for all of the estimates are quite small, it is difficult to evaluate the performance of the estimator without knowing a_{0N}. Accordingly, we examined the performance of the estimator for the following example where the error and index densities have the following Weibull form:

$$1 - F_U(u) = \exp(-u^{c_u});\ 1 - G_{V_2}(v) = \exp(-v^{c_v}),$$

with $c_u = 1.5$ and $c_v = 1$ so that the index tails are fatter than those of the selection error. It can be shown that by setting $b = 0$ the moment condition is approximately equivalent to

$$\left[2 + (a_{0N} \ln N)^{\frac{1}{c_u}-1}\right] a_{0N} = .8 - 2\varepsilon.$$

Since a_{0N} depends on the sample size, we examined three different sample sizes: $N = 500, 1000$,and 2000. At each of these sample sizes, we solved the above equation for a_{0N} and conducted a Monte Carlo experiment with 100 replications to evaluate the performance of $\hat{a}$ at the base level of the index.

[9] In our Monte Carlo studies, the higher order kernel we use is the twicing kernel for both index parameter estimation and estimation of the high probability set parameter.

Table 3. Estimation of High Probability Parameter

Sample Size	a_{0N}	$\|Bias\|$	Standard Deviation	RMSE
500	0.279	0.037	0.027	0.046
1000	0.280	0.025	0.025	0.035
2000	0.283	0.019	0.009	0.020

Table 3 shows the results. The bias, standard deviation and RMSE are standardized by the truth a_{0N}. It shows that the estimator $\hat{a}$ performs very well in terms of absolute bias, standard deviation and RMSE. It also confirms that the absolute bias, standard deviation and RMSE all decline as the sample size increases. As expected, a_{0N} increases slowly as the sample size increases.

8 Conclusions

This paper studies the binary outcome model with sample selection in a semiparametric framework. As marginal effects are often of primary interest in this type of model, we propose a semiparametric marginal effect estimator. This marginal effect estimator is based on observations in a high probability set where the selection probabilities are above a cutoff. We propose an estimator for this cutoff and establish its large sample properties. Based on that, we establish the large sample properties for our marginal effect estimator, which takes into account that the cutoff and the selection probability are estimated. In a Monte Carlo study we find that our marginal effect estimator based on the estimated high probability set performs quite well in finite samples.

This marginal effect estimator is developed under an index framework so as to achieve good performance in finite samples. Accordingly, it depends on an estimator for index parameters. In this paper, we propose an index parameter estimator based on regular kernels with bias control mechanisms and show that the estimator is consistent and asymptotically distributed as normal. While retaining these desirable large sample properties, the Monte Carlo results show that this estimator performs very well in finite samples.

References

[1] Ahn, H. and J.L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics,* 58, 3-29.

[2] Andrews, D. and M.Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model" *Review of Economic Studies*, 65, 497-517.

[3] Bhattacharya , P.K. (1967): "Estimation of a Probability Density Function and its Derivatives," *The Indian Journal of Statistics,* Series 4, v. 29, 373-382.

[4] Blundell, R. W. and James L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies, v. 71, No. 3, 655-679.*

[5] Chesher, A. (2005): "Nonparametric Identification under Discrete Variation", *Econometrica,* 73, 1525-1550.

[6] Das, M., W.Newey and F.Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies,* 70, 33-58.

[7] Escanciano, J. C., D. T. Jacho-Chavez and A. Lewbel (2012): "Uniform Convergence of Weighted Sums of Non- and Semi-parametric Residuals for Estimation and Testing," working paper.

[8] Gallant, A. and D. Nychka (1987): "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 15, 363-390.

[9] Heckman, J. (1974): "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42(4), 679-94.

[10] Heckman, J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica,* 47(1), 153-61.

[11] Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313-18.

[12] Honore, B. E. and J. L. Powell (2005): "Pairwise Difference Estimation of Nonlinear Models." *D. W. K. Andrews and J. H. Stock, eds., Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg* (Cambridge: Cambridge University Press), 520–53.

[13] Ichimura, H. (1993):"Semiparametric least squares (SLS) and weighted SLS estimation of single index models," *Journal of Econometrics*, 58, 71-120.

[14] Ichimura, H., and L. F. Lee (1991): "Semiparametric least squares (SLS) and weighted SLS estimation of multiple index models: Single equation estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics,* ed. W.Barnett, J.Powell and G.Tauchen, Cambridge University Press.

[15] Khan, S. and E.Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation, *Econometrica*, 6, 2021-2042.

[16] Klein, R. and C. Shen (2010): "Bias Corrections in Testing and Estimating Semiparametric, Single Index Models," *Econometric Theory,* 1683-1718.

[17] Klein, R. and R. Spady (1993): "An Efficient Semiparametric Estimator for the Binary Response Model," *Econometrica*, 61, 387-421.

[18] Klein, R. and F.Vella (2009): "A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity," *Journal of Applied Econometrics,* 24, 735-762.

[19] Klein, R. (1993), Specication tests for binary choice models based on index quantiles, Journal of Econometrics 59, 343-375.

[20] Lee, L.F (1995): "Semi-Parametric Estimation of Polychotomous and Sequential Choice Models", *Journal of Econometrics,* 65, 381-428.

[21] Newey, W, (2007): "Nonparametric continuous/discrete choice models", *International Economic Review*, 48: 1429–1439.

[22] Newey, W., F. Hsieh, and J. Robins (2004): "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947-962.

[23] Pakes, A., and D.Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1058

[24] Rothe, C. (2009): "Semiparametric Estimation of Binary Response Models with Endogenous Regressors, *Journal of Econometrics,* 253, 51-64.

[25] Shaikh, A. M. and E. J. Vytlacil (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica* 79(3), 949–955.

[26] Vella, F. (1998): "Estimating Models with Sample Selection Bias: A Survey." *The Journal of Human Resources*, 33:1, 127-169.

[27] Vytlacil, E. and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757-779.

9 Appendix

9.1 Main Results

9.1.1 Marginal Effect

Proof of Theorem 1. By Lemma 10,

$$C_N(\bar{v})\left[\hat{\zeta}\left(\bar{v},\hat{a}\right)-\zeta_0\left(\bar{v}\right)\right]=C_N(\bar{v})\gamma(\bar{v},a_{0N})+o_p(1).$$

Lemma 3 characterizes the order of the bias of the estimator. Recalling the definition of the high probability parameter in (D4), the bias in the estimator vanishes. From Lemma 4, the reciprocal of the estimator variance has the following order:

$$Nh_sE\left(S|\bar{v}\right)^2/E\left(S^2|\bar{v}\right)>Nh_sE\left(S|\bar{v}\right),$$

which completes the proof as $Nh_sE\left(S|\bar{v}\right)$ tends to ∞ as N increases.

Proof of Theorem 2. By definition, $\widehat{ME}-ME=\left[\hat{\zeta}(ve,\hat{a})-\zeta_0(ve)\right]-\left[\hat{\zeta}(vb,\hat{a})-\zeta_0(vb)\right]$. We begin by showing that the covariance between these two components vanishes. Notice that $\hat{\zeta}(\bar{v},\hat{a})-\zeta_0(\bar{v})$ is close to $\gamma(\bar{v},a_{0N})$ from Lemma 10, which we can write as a sample average $\sum_j \frac{1}{N}t_j\left(\bar{v}\right)$. The covariance is then of the form $E\left[t_j\left(ve\right)t_k\left(vb\right)\right]$. For $j\neq k$, from independence and the vanishing bias of the expectation of each term, this expectation vanishes. For $j=k$, the kernel function ensures that this expectation also vanishes faster than $N^{-1/2}$ as V_{1j} cannot be close to both ve and vb. Therefore, the variance is the sum of the variances of $\gamma(ve,a_{0N})$ and $\gamma(vb,a_{0N})$.

To provide the normality argument without assuming that the marginal effect components have variances of the same order, we consider several cases. First, if $O(Var(\gamma(ve,a_{0N})))>O(Var(\gamma(vb,a_{0N})))$, then with $V\equiv Var(\gamma(ve,a_{0N}))+Var(\gamma(vb,a_{0N}))$,

$$\begin{aligned}\frac{1}{\sqrt{V_N}} &= O\left(\frac{1}{\sqrt{Var(\gamma(ve,a_{0N}))}}\right)\\ &= O\left(C_N(ve)\right).\end{aligned}$$

Therefore, the characterization results in Lemma 10 apply to yield

$$\frac{\widehat{ME} - ME}{\sqrt{V_N}} = O\left(C_N(ve)\right)\left[\hat{\zeta}\left(ve, a_{0N}\right) - \zeta_0\left(ve, a_{0N}\right)\right] + o_p(1).$$

Asymptotic normality now follows from Lemma 9(b). A symmetric argument holds for the case where $O(Var(\gamma_N(ve, a_{0N}))) < O(Var(\gamma_N(vb, a_{0N})))$. For the case where $O(Var(\gamma_N(ve, a_{0N}))) = O(Var(\gamma_N(ve, a_{0N}))$, an argument similar to that in the proof of Lemma 9(b) shows that the relevant Lindeberg condition holds. Therefore, $\frac{\widehat{ME}-ME}{\sqrt{V_N}} \xrightarrow{d} Z \tilde{} N(0,1)$. Employing similar arguments as in Lemma 9(a), it can be shown that $\frac{V_N - \hat{V}_N}{V_N} \xrightarrow{p} 0$. Hence the theorem follows.

9.1.2 Index Parameters

Proof of Theorem 3. We provide the proof for $\hat{\theta}^*$, with the arguments for the other estimators being very similar. Lemma 11 proves that we can replace the $\hat{P}_i^*$ with P_i^* in the objective function $\hat{L}^*(\theta)$, and obtain $L^*(\theta)$ satisfying

$$\sup_{\theta} \left| \hat{L}^*(\theta) - L^*(\theta) \right| \xrightarrow{p} 0.$$

From Lemma 12,

$$\sup_{\theta} \left| L^*(\theta)] - E\left[L(\theta)\right] \right| \xrightarrow{p} 0.$$

To complete the argument, we must show that $E\left[L(\theta)\right]$ is uniquely maximized at θ_0. From standard arguments, θ_0 is a maximum, and the only issue is one of uniqueness. With θ^* as any potential maximizer, it can be shown that any candidate for a maximum must give correct probabilities for all three cells: $(Y_1 = 1, Y_2 = 1), (Y_1 = 0, Y_2 = 1)$, and $Y_2 = 0$. It then follows that for the $Y_2 = 0$ cell:

$$\Pr(Y_2 = 0 | V_2\left(\theta_2^*\right)) = \Pr(Y_2 = 0 | X) = \Pr(Y_2 = 0 | V_2\left(\theta_{20}\right)).$$

Under identifying conditions for single index models, $\theta_2^* = \theta_{20}$. For the $(Y_1 = 1, Y_2 = 1)$ cell:

$$\begin{aligned} \Pr(Y_1 &= 1 | Y_2 = 1, V_1\left(\theta_1^*\right), V_2\left(\theta_2^*\right)) \Pr(Y_2 = 1 | V_2\left(\theta_2^*\right)) = \\ \Pr(Y_1 &= 1 | Y_2 = 1, V_1\left(\theta_{10}\right), V_2\left(\theta_{20}\right)) \Pr(Y_2 = 1 | V_2\left(\theta_{20}\right)). \end{aligned}$$

Since $\theta_2^* = \theta_{20}$:

$$\Pr(Y_1 = 1|Y_2 = 1, V_1(\theta_{10}), V_2(\theta_{20})) = \Pr(Y_1 = 1|Y_2 = 1, V_1(\theta_1^*), V_2(\theta_{20})).$$

Solving the first probability function for $V_1(\theta_{10})$, for some function Υ we have:

$$V_1(\theta_{10}) = \Upsilon(V_1(\theta_1^*), V_2(\theta_{20})).$$

holding for X in some set. Since V_2 contains a continuous variable that does not affect V_1, differentiating both sides with respect to this variable yields $0 = \nabla_{v_2}\Upsilon$. Therefore, Υ must only be a function of the first index and is equal to $V_1(\theta_{10})$. Identification now follows from conditions that identify single index models.

Proof of Theorem 4. From a Taylor expansion, the second stage estimator has the following form:

$$\left(\hat{\theta}^* - \theta_0\right) = -\hat{H}^*\left(\theta^+\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\sum_{d_1 \leq d_2}\left[Y_i(d_1,d_2) - \hat{P}_i^*(d_1,d_2;\theta_0)\right]\hat{\delta}_i^*(d_1,d_2;\theta_0)\, I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\},$$

where θ^+ is an intermediate point. Using the notation in Lemma 15, the above can be written as

$$\left(\hat{\theta}^* - \theta_0\right) = -\hat{H}^*\left(\theta^+\right)^{-1}(A^* - B^*).$$

Referring to (D10), we begin by simplifying the adjustment factor by showing that:

$$\hat{H}^*\left(\hat{\theta}^*\right)^{-1}\hat{C}_V\left(\hat{\theta}^*\right) - \hat{H}^*\left(\theta^+\right)^{-1}\hat{C}_V(\theta_0) = o_p(N^{-1/2}).$$

Adopting the same strategy as in Lemma 14, the following two terms are $o_p(N^{-1/2})$:

$$\hat{H}^*\left(\theta^+\right)^{-1}\hat{H}^*\left(\hat{\theta}^*\right)^{-1}\left[\hat{H}^*\left(\theta^+\right) - \hat{H}^*\left(\hat{\theta}^*\right)\right]\hat{C}_V\left(\hat{\theta}^*\right)$$
$$\hat{H}^*\left(\theta^+\right)^{-1}\left[\hat{C}_V\left(\hat{\theta}^*\right) - \hat{C}_V(\theta_0)\right].$$

From the definition of $\hat{\theta}_I^o$ in (D10) and employing the result above:

$$\begin{aligned}\sqrt{N}\left(\hat{\theta}^o-\theta_0\right) &= \sqrt{N}\left(\hat{\theta}^o-\theta_0+\hat{H}^*\left(\theta^+\right)^{-1}\hat{C}_V\left(\theta_0\right)\right)+o_p(1)\\ &= -\hat{H}^*\left(\theta^+\right)^{-1}\sqrt{N}\left(A^*-B^o\right)+o_p(1),\end{aligned}$$

where

$$B^o\equiv\frac{1}{N}\sum_{i=1}^{N}\sum_{d_1\leq d_2}\left[\hat{P}_i^o\left(d_1,d_2;\theta_0\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\hat{\delta}_i^*\left(d_1,d_2;\theta_0\right)I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\}.$$

From Lemma 15a, $\sqrt{N}\left[A^*-A_0^*\right]=o_p(1)$. From Lemma 16, $\sqrt{N}B^o=o_p(1)$. The theorem now follows.

9.2 Intermediate Lemmas

This section provides three types of lemmas: 1) basic lemmas required by all estimators, 2) lemmas required to analyze the marginal effects estimator, and 3) lemmas relevant for the index estimator.

9.2.1 Basic Lemmas

Referring to (D5) and with an arbitrary small positive number ε and $k=1,2$, define

$$\mathcal{V}_{kN}(h)=\{v_2:a_k(d_k)+h^{1-\varepsilon}<v_k<b_k(d_k)-h^{1-\varepsilon}\} \tag{15}$$

$$\mathcal{V}_N(h)=\{v:a_{ck}\left(d_k\right)+h^{1-\varepsilon}<v_k<b_{ck}\left(d_k\right)-h^{1-\varepsilon}\}. \tag{16}$$

If the conditional density $g_{V_k|Y_k}\left(v_k|Y_k=d_k\right)$ has compact support, interpret $[a_k(d_2),\ b_k(d_2)]$ as that support. Similarly, if the conditional density $g_{V|Y}\left(v|Y_1=d_1,Y_2=d_2\right)$ is compact, interpret $[a_{ck}(d_k),\ b_{ck}(d_k)]$ as the corresponding support. When these supports are unbounded, we let $\mathcal{V}_{kN}$ and $\mathcal{V}_N$ approach all of R^1 and R^2 respectively, as N increases.

We begin with two basic lemmas on uniform and pointwise convergence rates. As the proofs of these lemmas are standard in the literature (see, for example Bhattacharya (1967) for a discussion of density estimators and derivatives and an extension in Klein (1993)), they are not provided here but are available upon request.

Lemma 1. Uniform Convergence. For ψ any p^{th} differentiable function of θ, let $\nabla_\theta^p(\psi)$ be the p^{th} partial derivative of ψ with respect to θ, $\nabla_\theta^0(\psi) \equiv \psi$. Let $\hat{f}_2$, $\hat{f}$, and $\hat{g}_{V_1}$,be the estimators in (D5) with respective probability limits f_2 , f and g_{V_2}; let $\hat{g}_{V_2}$ be the estimator in (D4) with probability limit g_{V_2}. Denote m as the kernel order, where $m = 1$ for regular kernels and $m = 2$ for twicing kernels. Then, under A1-4, for θ in a compact set), the following rates hold for $p = 0, 1, 2$:

$$a) \quad : \quad \sup_{t_k,\theta} \left| \nabla_\theta^p \left(\hat{f}_k(t_k; d_k, h) \right) - \nabla_\theta^p \left(f_k(t_k; d_k, h) \right) \right| = O_p \left(\min \left[h^{2m}, \frac{1}{\sqrt{N} h^{p+1}} \right] \right) \text{ with } t_k \epsilon \mathcal{V}_{kN}(h)$$

$$b) \quad : \quad \sup_{t,\theta} \left| \nabla_\theta^p \left(\hat{f}(t; d_1, d_2, h) \right) - \nabla_\theta^p \left(f(t; d_1, d_2, h) \right) \right| = O_p \left(\min \left[h^{2m}, \frac{1}{\sqrt{N} h^{p+2}} \right] \right) \text{ with } t \epsilon \mathcal{V}_N(h).$$

If $\sup_{t,\theta} \left| \hat{A}_{ij}(\theta) - A_{ij}(\theta) \right| = O_p(N^{-t})$ for $j = 1, 2$ and $\inf_{t,\theta} A_{2j}(\theta) > N^{-s}$ for $s < t$, then

$$c) : \sup_{t,\theta} \left| \frac{\hat{A}_{1j}(\theta)}{\hat{A}_{2j}(\theta)} - \frac{A_{1j}(\theta)}{A_{2j}(\theta)} \right| = O_p(N^{-(t-s)}).$$

Lemma 2. Pointwise Convergence. Using the same notation as above in Lemma 1, under A1-4

$$a) \quad : \quad \left| \nabla_\theta^p \left(\hat{f}_k(t_k; d_k, h) \right) - \nabla_\theta^p \left(f_k(t_k; d_k, h) \right) \right| = O_p \left(\min \left[h^{2m}, \frac{1}{\sqrt{N h^{2p+1}}} \right] \right)$$

$$b) \quad : \quad \left| \nabla_\theta^p \left(\hat{f}(t; d_1, d_2, h) \right) - \nabla_\theta^p \left(f(t; d_1, d_2, h) \right) \right| = O_p \left(\min \left[h^{2m}, \frac{1}{\sqrt{N} h^{p+1}} \right] \right).$$

If $\left| \hat{A}_{ij}(\theta) - A_{ij}(\theta) \right| = O_p(N^{-t})$ for $j = 1, 2$ and $A_{2j}(\theta) > N^{-s}$ for $s < t$, then

$$c) : \left| \frac{\hat{A}_{1j}(\theta)}{\hat{A}_{2j}(\theta)} - \frac{A_{1j}(\theta)}{A_{2j}(\theta)} \right| = O_p(N^{-(t-s)}).$$

9.2.2 Marginal Effects Lemmas

Lemma 3. Under A1-4,A5(b),A6-7, with $\zeta_0(\bar{v}) \equiv \Pr(Y_1 = 1 | V_1 = \bar{v})$ and $\gamma \equiv \gamma(\bar{v}, a)$ as in (10),

$$|E(\gamma_N)| \leq B_N = O(N^{-a} E(S)/E(S|\bar{v})).$$

Proof. With $P_2 = \Pr(Y_2 = 1|V_2)$ and $\mu_d(V_1, V_2) \equiv E[Y_1 - \zeta_0(\bar{v}) | Y_2 = d, V_1, V_2]$, and γ_A as the

numerator of γ :

$$\begin{aligned} E(\gamma_A) &= E(\frac{1}{h_s}\mu_1(V_1,V_2)K\left[(\bar{v}-V_1)/h_s\right]S)P_2 \\ &= \iint \mu_1(\bar{v}+h_s z,v_2)K(z)SP_2 g(\bar{v}+h_s z,v_2)dzdv_2. \end{aligned}$$

Using a Taylor series expansion,

$$|E(\gamma_A)| \leq \left| \int \mu_1(\bar{v},v_2)P_2 S g(\bar{v},v_2)dv_2 \right| + |RES|.$$

Note that $\mu_1(\bar{v},v_2)P_2 + \mu_0(\bar{v},v_2)(1-P_2) = E\left[Y_1 - \zeta_0(\bar{v})\,|V_1=\bar{v},V_2\right] = 0$, hence for the first term on the right-hand side,

$$\begin{aligned} \left| \int \mu_1(\bar{v},v_2)P_2 S g(\bar{v},v_2)dv_2 \right| &= \left| \int \mu_0(\bar{v},v_2)(1-P_2)\,S g(\bar{v},v_2)dv_2 \right| \\ &\leq O\left(N^{-a} \int S g(\bar{v},v_2)dv_2 \right) \\ &= O\left(N^{-a} g_{V_1}(\bar{v}) \int S g_{V_2|V_1}(v_2|\bar{v})dv_2 \right) \\ &= O\left(N^{-a} E(S|\bar{v}) \right). \end{aligned}$$

The second term on the right-hand side ($|RES|$) is a residual term from the Taylor series expansion. Under A4,A7, it is $O\left(h_s^2 E(S)\right)$. Therefore, combining those two terms, the slowest rate would be $|E(\gamma_A)| = O(N^{-a}E(S))$ since $O(h_s^2) < O\left(N^{-a}\right)$ and $O(E(S|\bar{v})) \leq O(E(S))$ from (A6).

Lemma 4. Under (A1-4,A5b,A6-7) for γ defined in Lemma 3 and a_{0N} defined in (D4),

$$\frac{1}{\sqrt{Var(\gamma)}} = O\left(C_N(\bar{v})\right).$$

where $C_N(\bar{v}) \equiv \frac{\sqrt{Nh_s}E(S|\bar{v})}{\sqrt{E(S^2|\bar{v})}}$.

Proof. For a_{0N} set as in (D4), $\frac{(E(\gamma_N))^2}{Var(\gamma_N)} \to 0$; hence

$$\begin{aligned} Var(\gamma) &= O\left(\frac{E([Y_1 - \zeta_0(\bar{v})]^2 K^2\left[(\bar{v}-V_1)/h_s\right]Y_2 S^2)}{Nh_s^2 \left(E\left(S|V_1=\bar{v}\right)\right)^2 g_{V_1}^2(\bar{v})} \right) \\ &= O\left(\frac{E(K^2\left[(\bar{v}-V_1)/h_s\right]S^2)}{Nh_s^2\left(E\left(S|V_1=\bar{v}\right)\right)^2} \right). \end{aligned}$$

Letting $z = (V_1 - \bar{v})/h_s$, the result follows from a Taylor series expansion about $h_s = 0$.

Lemma 5. Recall (D3), let

$$\hat{\omega}\left(\hat{\theta}\right) = \hat{E}(\hat{g}_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right))\frac{N^{-\varepsilon'}}{\ln N}$$

$$\begin{aligned}
\hat{\tau}_i &\equiv \tau\left(\hat{\omega}\left(\hat{\theta}\right), \hat{g}_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right)\right) \\
\tau_{i0} &\equiv \tau\left(\omega\left(\theta_0\right), g_{V_2}\left(V_{2i}\left(\theta_0\right)\right)\right) \\
\Delta_{1i} &\equiv \left[\hat{\omega}\left(\hat{\theta}\right) - \omega\left(\theta_0\right)\right] \\
\Delta_{2i} &\equiv \left[\hat{g}_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right) - g_{V_2}\left(V_{2i}\left(\theta_0\right)\right)\right]
\end{aligned}$$

Then, under (A1-4) there exists $\delta > 0$ and a large number K such that:

$$\hat{\tau}_i = \tau_{i0} + \tau_{i0} * o_p\left(N^{-\delta}\right) + o_p\left[N^{-K\delta}\right],$$

where each o_p term is uniform in i.

Proof. From a Taylor series expansion in $\hat{\omega}\left(\hat{\theta}\right)$ and $\hat{g}_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right)$ about $\omega\left(\theta_0\right)$ and $g_{V_2}\left(V_{2i}\left(\theta_0\right)\right)$:

$$\hat{\tau}_i = \tau_{i0} - \tau_{i0}(1 - \tau_{i0})N^{\alpha}\left[\Delta_{1i} + \Delta_{2i}\right] + ... \tag{17}$$

To complete the proof, we will show that there exists $\delta > 0$ such that $N^{\alpha}\Delta_{2i}$ is $o_p(N^{-\delta})$ uniformly in i. Since Δ_{1i} depends on the difference between a sample average of density estimators and the expectation of the true density, it can be shown that this term converges in probability, uniformly in i, to zero faster than Δ_{2i}. Suppose we stop the Taylor series expansion at K, the remainder term is $o_p\left[\left(N^{-\delta}\right)^K\right]$ uniformly in i from Lemma 1 under (A1-4). Therefore, $\hat{\tau}_i = \tau_{i0} + \tau_{i0} * o_p\left(N^{-\delta}\right) + o_p\left[N^{-K\delta}\right]$.

To establish the uniform convergence of $N^{\alpha}\Delta_{2i}$ upon which the above argument depends, note that

$$N^{\alpha}\Delta_{2i} = N^{\alpha}\left[\hat{g}_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right) - g_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right)\right] + N^{\alpha}\left[g_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right) - g_{V_2}\left(V_{2i}\left(\theta_0\right)\right)\right].$$

With $\hat{g}_{V_2}$ depending on a window $h_2 = O(N^{-.2})$, from Lemma 1, $\hat{g}_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right) - g_{V_2}\left(V_{2i}\left(\hat{\theta}\right)\right) = O_p(N^{-.3})$ uniformly. Hence as as $\alpha \leq .2$ in (D3), we can find $\delta > 0$ such that the first term is of order $N^{-\delta}$. Since g_{V_2} has bounded first derivatives under (A4), the second term vanishes faster than the first under a Taylor

series argument and an index parameter estimator $\hat{\theta}$ that satisfies $\left[\hat{\theta} - \theta_0\right] = O_p(N^{-.5})$.

Lemma 6 below provides a result needed to obtain the convergence rate of $\hat{a} - a_{0N}$.

Lemma 6. Let

$$\begin{aligned} c_N(a) &\equiv N^{-2(a-.4+\varepsilon)} \\ M_1(a) &\equiv E(S)^2 \\ M_2(a) &\equiv E(S^2|\bar{v}). \end{aligned}$$

and recall (D4). Then under A1-5, there exists $\delta > 0$ such that

$$c_N(\hat{a}) \left(\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} - \frac{M_1(\hat{a})}{M_2(\hat{a})} \right) = O_p\left(N^{-\delta}\right).$$

Proof. To prove the result, we show there exists $\delta > 0$ such that

$$c_N(\hat{a}) \left(\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} - \frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{M_2(\hat{a})} \right) = O_p\left(N^{-\delta}\right) \tag{18}$$

$$\text{and } c_N(\hat{a}) \left(\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{M_2(\hat{a})} - \frac{M_1(\hat{a})}{M_2(\hat{a})} \right) = O_p\left(N^{-\delta}\right). \tag{19}$$

Here, we provide the proof for (18); the proof of (19) is very similar. Recall the definition of $\hat{a}$ in (D4),

$$\begin{aligned} c_N(\hat{a}) \frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} &= N^{-2(a-.4+\varepsilon)} * h_s^{-1} N^{-(1-2a+\varepsilon)} \\ &= N^{-2a+.8-2\varepsilon} * O(N^{.2+3\varepsilon-(1-2a+\varepsilon)}) \\ &= O(1) \end{aligned} \tag{20}$$

because $h_s = O(N^{-.2-3\varepsilon})$ from (D1). The term in (18) can be written as

$$c_N(\hat{a}) \frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} \left(\frac{M_2(\hat{a}) - \hat{E}_2\left(\hat{S}^2|\bar{v}\right)}{M_2(\hat{a})} \right) = O(1) \left(\frac{M_2(\hat{a}) - \hat{E}_2\left(\hat{S}^2|\bar{v}\right)}{M_2(\hat{a})} \right) \tag{21}$$

Hence it suffices to determine the order of $\frac{M_2(\hat{a})-\hat{E}_2\left(\hat{S}^2|\bar{v}\right)}{M_2(\hat{a})}$. Recalling (D3), it is equal to

$$\frac{\left[\hat{E}_2\left(S^2\left[\tau, x(\hat{a}, P_2)\right]|\bar{v}\right) - \hat{E}_2\left(S^2\left[\hat{\tau}, x(\hat{a}, \hat{P}_a)\right]|\bar{v}\right)\right]}{M_2(\hat{a})} + \frac{\left[M_2(\hat{a}) - \hat{E}_2\left(S^2\left[\tau, x(\hat{a}, P_2)\right]|\bar{v}\right)\right]}{M_2(\hat{a})}. \tag{22}$$

To show the first term in (22) is $O_p(N^{-\delta})$ with $\delta > 0$, notice that

$$\begin{aligned} S^2\left[\tau, x(\hat{a}, P_2)\right] - S^2\left[\hat{\tau}, x(\hat{a}, \hat{P}_a)\right] &= \tau^2 T^2\left(x(\hat{a}, P_2)\right) - \hat{\tau}^2 T^2\left(x(\hat{a}, \hat{P}_a)\right) \\ &= \tau^2\left[T^2\left(x(\hat{a}, P_2)\right) - T^2\left(x(\hat{a}, \hat{P}_a)\right)\right] && (T_A) \\ &\quad + \left(\tau^2 - \hat{\tau}^2\right) T^2\left(x(\hat{a}, P_2)\right) && (T_B) \\ &\quad + \left(\hat{\tau}^2 - \tau^2\right)\left[T^2\left(x(\hat{a}, P_2)\right) - T^2\left(x(\hat{a}, \hat{P}_a)\right)\right]. && (T_C) \end{aligned}$$

We start by showing $\hat{E}_2\left[T_A|\bar{v}\right]/M_2(\hat{a}) = O_p\left(N^{-\delta}\right)$. From a Taylor series expansion in $\hat{P}_a$:

$$\begin{aligned} T^2\left(x(\hat{a}, \hat{P}_a)\right) - T^2(x(\hat{a}, P_2)) &= \sum_{k=1}^{m-1}\left[T^2\right]'^{(k)}\left(x(\hat{a}, P_2)\right)\left[\frac{\hat{P}_a - P_2}{1 - P_2}\right]^k /k! \\ &\quad + \left[T^2\right]'^{(m)}\left(x(\hat{a}, P_2^+)\right)\left[\frac{\hat{P}_a - P_2}{1 - P_2^+}\right]^m /m!\,, \end{aligned}$$

where $\left[T^2\right]'^{(m)}$ is the mth derivative of the function T^2 w.r.t. x, and P_2^+ is an intermediate point. Then, noting that trimming keeps the density for V_2 at $\bar{v}$ bounded away from 0 (D3), $\hat{E}_2\left[T_A|\bar{v}\right]/M_2(\hat{a}) = O_p\left(N^{-\delta}\right)$ if

$$\begin{aligned} A_k &\equiv \left|\sum_{i=1}^N \frac{1}{N}\tau_i^2\left[T^2\right]'^{(k)}\left(x(\hat{a}, P_{2i})\right)\left[\frac{\hat{P}_{ai} - P_{2i}}{1 - P_{2i}}\right]^k K_2\left(\frac{\bar{v} - v_i}{h_T}\right)/M_2(\hat{a})h_T\right| = O_p\left(N^{-\delta}\right) \\ A_m &\equiv \left|\sum_{i=1}^N \frac{1}{N}\tau_i^2\left[T^2\right]'^{(m)}\left(x(\hat{a}, P_{2i}^+)\right)\left[\frac{\hat{P}_{ai} - P_{2i}}{1 - P_{2i}^+}\right]^m K_2\left(\frac{\bar{v} - v_i}{h_T}\right)/M_2(\hat{a})h_T\right| = O_p\left(N^{-\delta}\right). \end{aligned}$$

For A_k, setting $k = 1$ for expositional purposes,[10] the term will be bounded above by

$$\sup\left|\left(\frac{\hat{P}_{ai} - P_{2i}}{1 - P_{2i}}\right)\tau_i T'^{(1)}(x(\hat{a}, P_{2i}))\right| * \sum_{i=1}^N \frac{1}{N}\tau_i * 2T(x(\hat{a}, P_{2i}))K_2\left(\frac{\bar{v} - v_i}{h_T}\right)/M_2(\hat{a})h_T, \tag{23}$$

[10] Higher order derivatives of the T-function are bounded and keep x in the $R2$ region same as does the original T-function in (D3). Moreover, higher order terms converge faster to 0 than lower order terms.

where $T'^{(m)}$ is the m^{th} derivative of T w.r.t. x. The sup component has order:

$$\frac{\sup_{i,\theta}\left|\tau_i\left(\hat{P}_{ai}-P_{2i}\right)\right|}{\inf_{\hat{a}} N^{-\hat{a}}}.$$

The numerator is $O_p\left(N^{-(4-\varepsilon')}\right)$ from Lemma 1 because $\hat{P}_{ai}$ is based on a twicing kernel with window $h_T = N^{-.1}$ with τ_i constraining the density denominator to be larger than $O(N^{-\varepsilon'})$. The denominator is $O\left(N^{-(.4-\varepsilon)}\right)$from (D4). As $\varepsilon - \varepsilon' > 0$ there exists δ such that the sup component will have order of $N^{-\delta}$. Turning our attention to the second component of (23), it is bounded above by

$$\sup_a \frac{\left|\sum_{i=1}^N \frac{1}{N}\tau_i T(x(a,P_{2i})K_2\left(\frac{\bar{v}-v_i}{h_T}\right)/h_T - E(\tau T(x(a,P_2))|\bar{v})\right|}{M_2(a)} + \sup_a \frac{E(\tau T(x(a,P_2))|\bar{v})}{M_2(a)}.$$

For the first term, denote a^* as the value for a such that this term takes on its supremum. Under a twicing kernel, from Lemma 1 it follows that with $h_T = O(N^{-.1})$, the numerator is $O(N^{-.4})$. For the denominator, referring to (D3), notice that the S^2 function is bounded below by an indicator function set to be zero when x is in either $R1$ or $R2$, and one when x is in $R3$. Hence $M_2(a)$ is bounded below by the conditional expectation of that indicator function, which is a probability that is of order N^{-a} as the tail of $v_2|\bar{v}$ is fatter than the error tail. Therefore, $M_2(a) \geqslant O(N^{-a}) \geqslant O(N^{-(.4-\varepsilon)})$ (see D4). Therefore, there exists $0 < \delta < \varepsilon$ such that

$$\sup_a \frac{\left|\sum_{i=1}^N \frac{1}{N}\tau_i T(x(a,P_{2i})K_2\left(\frac{\bar{v}-v_i}{h_T}\right)/h_T - E(\tau T(x(a,P_2))|\bar{v})\right|}{M_2(a)} = O_p(N^{-\delta}). \tag{24}$$

Hence it suffices to show that $\sup_a \frac{E(\tau T(x(a,P_2))|\bar{v})}{M_2(a)} = O(1)$. Referring to (D3), notice that

$$\frac{E(\tau T(x(a,P_2))|\bar{v})}{M_2(a)} \leq \frac{c_1 \Pr(R2|\bar{v}) + \Pr(R3|\bar{v})}{c_2 \Pr(R2|\bar{v}) + \Pr(R3|\bar{v})}$$

where

$$\begin{aligned} c_1 &\equiv E\left[1 - \exp\frac{-x^k}{b^k - x^k}|R2,\bar{v}\right] \\ c_2 &\equiv E\left[\left(1 - \exp\frac{-x^k}{b^k - x^k}\right)^2|R2,\bar{v}\right]. \end{aligned}$$

The above ratio converges to some constant irrespective of which of the regional probabilities converges faster to zero. For the remainder term A_m, it vanishes faster than $N^{-\delta}$ for a sufficiently large value of m.

Turning to $\hat{E}_2\left[T_B|\bar{v}\right]/M_2(\hat{a})$, from the expansion in Lemma 5 whose order K is set sufficiently large,

$$\hat{E}_2\left[T_B|\bar{v}\right]/M_2(\hat{a}) = o_p\left(N^{-2\delta}\right)\sum_{i=1}^{N}\frac{1}{N}\tau_i^2 T^2(x(\hat{a},P_{2i}))K_2\left(\frac{\bar{v}-v_i}{h_T}\right)/M_2(\hat{a})h_T$$

As $T^2(x(\hat{a},P_{2i})) \leq T(x(\hat{a},P_{2i}))$, it follows from (24) and the subsequent discussion that the second component is $O_p(1)$, which establishes the order for this component. Employing a very similar argument, there exists $\delta > 0$ such that $\hat{E}_2\left[T_C|\bar{v}\right]/M_2(\hat{a}) = O_p(N^{-\delta})$.

To complete the argument, we need to provide the order for the second term in (22). For this term:

$$\left|\frac{M_2(\hat{a}) - \hat{E}_2\left(S^2\left[\tau, x(\hat{a},P_2)\right]|\bar{v}\right)}{M_2(\hat{a})}\right| \leq \frac{\sup_a\left|M_2(a) - \hat{E}_2\left(S^2\left[\tau, x(a,P_2)\right]|\bar{v}\right)\right|}{M_2(\hat{a})}$$

Notice that the $\hat{E}_2$ is a twicing kernel and has an estimated density at $\bar{v}$ that converges to a positive value. Then, from Lemma 1, with window $h_T = O(N^{-.1})$, the expression above is $O(\frac{N^{-.4}}{M_2(\hat{a})})$. From above, $M_2(\hat{a}) = O(N^{-(.4-\varepsilon)})$. Then, there exists $0 < \delta < \varepsilon$ for which the result follows.

Lemma 7. Referring to Lemma 6, define

$$\begin{aligned} z(a) &\equiv N^{-a} \\ R\left(z\left(a\right)\right) &\equiv \tfrac{M_1(a)}{M_2(a)} \\ z_0 \equiv z(a_{0N}), \hat{z} &\equiv z\left(\hat{a}\right), z^+ \equiv z\left(a^+\right). \end{aligned}$$

Then, under (A1-5) there exists $\delta > 0$ such that $|\hat{a} - a_{0N}| = o_p\left(N^{-\delta}\right)$.

Proof. From Lemma 6 and a Taylor series expansion, there exists $\delta > 0$ such that

$$\begin{aligned} c_N\left(\hat{a}\right)\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} &= c_N\left(\hat{a}\right)R\left(z\left(\hat{a}\right)\right) + O_p\left(N^{-\delta}\right) \\ &= c_N\left(a_{0N}\right)R\left(z\left(a_{0N}\right)\right) + O_p\left(N^{-\delta}\right) - \\ &\quad \ln\left(N\right)c_N\left(z^+\right)\left[2R\left(z^+\right) + z^+R'\left(z^+\right)\right]\left[\hat{a} - a_{0N}\right]. \end{aligned}$$

Recall (20), notice that $c_N(\hat{a})\frac{\left[\hat{E}_2(\hat{S})\right]^2}{\hat{E}_2(\hat{S}^2|\bar{v})} = c_N(a_{0N})R(z(a_{0N}))$ and is bounded away from zero by definition (D4). Therefore,

$$[\hat{a} - a_{0N}] = \frac{O_p\left(N^{-\delta}\right)}{\ln(N)c_N(z^+)R(z^+) + z^+R'(z^+)}.$$

Since $R = \frac{[E(S)]^2}{E(S^2|\bar{v})}$ in (D4) is increasing, $z^+R'(z^+) > 0$; hence

$$[\hat{a} - a_{0N}] = O_p\left(N^{-\delta}/\ln(N)c_N\left(z^+\right)R\left(z^+\right)\right).$$

Suppose $\hat{a} < a_{0N}$ (the argument when $\hat{a} \geqslant a_{0N}$ is the same), then $\hat{a} < a^+ < a_{0N}$, and $z_0 < z^+ < \hat{z}$. Since R is increasing, $R(z_0) < R(z^+) < R(\hat{z})$. Therefore, $c_N(z_0)R(z_0) < c_N(z^+)R(z^+) < c_N(\hat{z})R(\hat{z})$. Both the upper and lower bounds are bounded away from zero. The proof now follows.

Lemma 8. Expectations of Kernel Products. Let $\{\varepsilon_{1j},\varepsilon_{2j},\varepsilon_{3j}\}$ be i.i.d. over j with each depending on an index, V_{2j}. Throughout this lemma and its proof, all expectations are conditioned on V_{2j} and hold uniformly in j. For expositional purposes, we suppress the conditioning notation. Assume the following properties:

$$\begin{aligned} a) \quad &: \quad E(\varepsilon_{\gamma j}) = O(h_\gamma^{2p_\gamma}) \\ b) \quad &: \quad E\left[\varepsilon_{\gamma j}^{\rho_\gamma}\right] = O(\frac{1}{h_\gamma^{\rho_\gamma - 1}}), \rho_\gamma > 1. \end{aligned}$$

where $\gamma = 1,2,3$. Set $h_\gamma^{4p_\gamma} = O(\frac{1}{Mh_\gamma})$ and denote $\bar{\varepsilon}_\gamma = \frac{1}{M}\sum_{j=1}^M \varepsilon_{\gamma j}$, then

$$E\left\{[\bar{\varepsilon}_1]^r[\bar{\varepsilon}_2]^s[\bar{\varepsilon}_3]^t\right\} = O\left(h_1^{2p_1 r}\right)O\left(h_2^{2p_2 s}\right)O\left(h_3^{2p_3 t}\right).$$

Proof. From the Cauchy–Schwarz inequality,

$$E\left\{[\bar{\varepsilon}_1]^r[\bar{\varepsilon}_2]^s[\bar{\varepsilon}_3]^t\right\} < \left\{E[\bar{\varepsilon}_1]^{4r}\right\}^{1/4}\left\{E[\bar{\varepsilon}_2]^{4s}\right\}^{1/4}\left\{E[\bar{\varepsilon}_3]^{2t}\right\}^{1/2}.$$

It suffices to order one of the three terms, hence we can study a general term: $E[\bar{\varepsilon}]^q$ with general h and

p. This general term has q types of terms, with kth ($k = 1, ..., q$) type:

$$\frac{1}{M^{q-k}} \underbrace{\sum \cdots \sum}_{k} \frac{1}{M^k} \varepsilon_{j_1}^{i_1} \cdots \varepsilon_{j_k}^{i_k}$$

$$\text{where } i_1 + ... + i_k = q \text{ and } j_1 \neq \cdots \neq j_k.$$

From the i.i.d. property of the $\varepsilon's$, the expectation of this term is given as

$$\frac{1}{M^{q-k}} E\left[\varepsilon_{j_1}^{i_1}\right] \cdots E\left[\varepsilon_{j_k}^{i_k}\right].$$

Suppose we study a term $E\left[\varepsilon_{j_t}^{i_t}\right]$, where $1 \leqslant t \leqslant k$. There are two types of expectations: single power of ε and multiple powers of ε. For the single power case, from property (a):

$$E\left[\varepsilon_{j_t}^{i_t}\right] = O(h^{2p}) \text{ for } i_t = 1.$$

For the multiple power case, from property (b):

$$E\left[\varepsilon_{j_t}^{i_t}\right] = O\left(\left[\frac{1}{h}\right]^{(i_t-1)}\right) \text{ for } i_t > 1.$$

There are different combinations of $i_1, \cdots, i_k$ for a given q and k. To order $\frac{1}{M^{q-k}} E\left[\varepsilon_{j_1}^{i_1}\right] \cdots E\left[\varepsilon_{j_k}^{i_k}\right]$, we next need to find the combination that yields the slowest convergence rate. One observation we make here is that the slowest term is the one with the least number of single power $\varepsilon's$.

When $k \leqslant \frac{q}{2}$, the slowest term would have no single power of ε in it (see below for an example). Therefore, from property (b) the convergence rate will be

$$O\left((\frac{1}{Mh})^{q-k}\right).$$

When $k > \frac{q}{2}$, the slowest term would include at least one single power ε in it, hence from property (a) and (b) the rate will be:

$$O\left((h^{2p})^{2k-q}\right) O\left((\frac{1}{Mh})^{q-k}\right).$$

To illustrate our proof strategy, suppose q is an even number (the case for an odd number is very similar),

for example $q = 6$. If we denote the type by the powers of the elements, for example 1122 would mean the $\varepsilon_{j_1}^1 \varepsilon_{j_2}^1 \varepsilon_{j_3}^2 \varepsilon_{j_4}^2$ term, then we have the following table:

$$\begin{bmatrix} k & Slowest\ Type & Rate \\ 1 & 6 & O\left[(\frac{1}{Mh})^5\right] \\ 2 & 33 & O\left[(\frac{1}{Mh})^4\right] \\ 3 & 222 & O\left[(\frac{1}{Mh})^3\right] \\ 4 & 1122 & O\left[(h^{2p})^2(\frac{1}{Mh})^2\right] \\ 5 & 11112 & O\left[(h^{2p})^4(\frac{1}{Mh})\right] \\ 6 & 111111 & O\left[(h^{2p})^6\right] \end{bmatrix}$$

For $k = \frac{q}{2}$, the slowest term would be the one with k squared terms,

$$\frac{1}{M^{q-k}} E\left[\varepsilon_{j_1}^2\right] \cdots E\left[\varepsilon_{j_k}^2\right].$$

Hence the rate would be

$$O\left(\frac{1}{M^{q-k}}(\frac{1}{h})^k\right) = O\left((\frac{1}{Mh})^{q-k}\right).$$

It can be shown that this same expression holds for all smaller k. For $k = \frac{q}{2} + 1$, the slowest term would have two single power $\varepsilon's$ and the rest would be squared terms, e.g.,

$$\frac{1}{M^{q-k}} E\left[\varepsilon_{j_1}\right] E\left[\varepsilon_{j_2}\right] E\left[\varepsilon_{j_3}^2\right] \cdots E\left[\varepsilon_{j_k}^2\right];$$

hence the rate would be

$$\frac{1}{M^{q-k}} O\left((h^{2p})^2(\frac{1}{h})^{k-2}\right) = O\left((h^{2p})^{2k-q}(\frac{1}{Mh})^{q-k}\right).$$

This same expression holds for all larger k.

We now need to find the k^{th} term with the slowest convergence rate. Set h optimally, i.e. $h^{4p} = O(\frac{1}{Mh})$

and substitute it in each term above, we have

$$O\left(\ (h^{4p})^{q-k}\right) = O\left(\ (h^{2p})^{2q-2k}\right) \quad \text{when } k \leqslant \tfrac{q}{2},$$
$$O\left(\ (h^{2p})^{2k-q}(h^{4p})^{q-k}\right) = O(h^{2pq}) \quad \text{when } k > \tfrac{q}{2}.$$

Hence the slowest convergence rate is $O(h^{2pq})$. The lemma follows.

Lemma 9. Define

$$\hat{E}\left(Y_2 S\left(\tau, x(a, P_2)\right)|\bar{v}\right) \equiv \sum_j \frac{1}{Nh} K\left[\left(\bar{v} - V_{1j}\right)/h\right] Y_{2j} S(\tau_j, x(a, P_{2j})).$$

Then under (A1-7) with $\gamma_0 \equiv \gamma(\bar{v}, a_{0N})$,

$$a) \quad : \quad \frac{E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v}) - \hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)}{\hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)} \xrightarrow{p} 0.$$
$$b) \quad : \quad \frac{\gamma_0}{\sqrt{Var(\gamma_0)}} \xrightarrow{d} Z \tilde{} N(0,1)$$

Proof. Since for any random variable X_N, $X_N \xrightarrow{p} 1$ iff $(1/X_N) \xrightarrow{p} 1$, the result in (a) is equivalent to:

$$\frac{E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v})}{\hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)} \xrightarrow{p} 1 \text{ iff } \frac{\hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)}{E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v})} \xrightarrow{p} 1$$
$$\text{iff } \frac{E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v}) - \hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)}{E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v})} \xrightarrow{p} 0.$$

Notice that the last result above holds iff:

$$\frac{\left[E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v}) - \hat{E}\left(Y_2 S\left(\tau, x(a_{0N}, P_2)\right)|\bar{v}\right)\right]}{E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v})} \xrightarrow{p} 0$$
$$\text{and} \quad \frac{\left[\hat{E}\left(Y_2 S\left(\tau, x(a_{0N}, P_2)\right)|\bar{v}\right) - \hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)\right]}{E\left(S|V_1 = \bar{v}\right) g_{V_1}(\bar{v})} \xrightarrow{p} 0.$$

The above terms are very similar to (22) in Lemma 6 and the argument for their convergence is very similar to that in the proof of Lemma 6. The first term converges to zero by a pointwise convergence argument using the properties of the S-function and the fact that the conditional expectation of Y_2 approaches

1 on the high probability set .Employing a Taylor series argument and utilizing the result from Lemma 7, it can be shown that under (A4-5) the second term goes to zero in probability.

To prove (b), let $E\left[S|\bar{v}\right] \equiv E\left[S|V_1 = \bar{v}\right]$, and define:

$$\gamma_{0j} \equiv \frac{\left[Y_{1j} - \zeta_0\left(\bar{v}\right)\right] K\left[\left(\bar{v} - V_{1j}\right)/h_s\right] Y_{2j} S_j}{h_s E\left(S|\bar{v}\right) g_{V_1}(\bar{v})}$$

Since $\frac{[Y_{1j}-\zeta_0(\bar{v})]Y_{2j}}{g_{V_1}(\bar{v})}$ is uniformly bounded in j, from Lemma 4:

$$Var\left(\sum_{i=1}^{N} \gamma_{0j}\right) = O\left[\frac{N}{h_s} \frac{E\left[S^2|\bar{v}\right]}{\left(E\left[S|\bar{v}\right]\right)^2}\right] \equiv s_N^2$$

The Lindberg Condition for normality requires that for every $\varepsilon > 0$:

$$\frac{1}{s_N^2} \sum_{i=1}^{N} E\left[\gamma_{0j}^2 1\left\{\gamma_{0j}^2 > \varepsilon^2 s_N^2\right\}\right] \to 0.$$

Substituting for s_N^2, it suffices to show that:

$$\frac{1}{E\left[S^2|\bar{v}\right]} E\left[\begin{array}{c} \frac{1}{h_s} K^2\left(\frac{\bar{v}-V_1}{h_s}\right) S^2 * \\ 1\left\{\frac{1}{h_s} K^2\left(\frac{\bar{v}-V_1}{h_s}\right) S^2 > O\left[NE\left[S^2|\bar{v}\right]\right]\right\} \end{array}\right] \to 0$$

To establish the above limit, it will be useful to bound $E\left[S^2|\bar{v}\right]$ from below. Recall from the proof of the proposition in Section 5.2 that:

$$S \geq \tau\left(g_{V2} = (1+\delta)\omega\right) 1\{g_{V2} > (1+\delta)\omega\} 1\left\{X > b\right\}$$

For the second indicator, from the definition of X in (D4):

$$1\left\{X > b\right\} = 1\{F_U(V_2) > 1 - N^{-a_{0N}} e^{-b}\} = 1\left\{V_2 > F_U^{-1}\left(1 - N^{-a_{0N}} e^{-b}\right)\right\}$$

For the first indicator, define v_U such that $g_{V2}\left(v_U\right) = (1+\delta)\omega\}$. Then, from the monotonicity assumption on g_{V2} in its tail,(A5b) it follows that

$$1\{g_{V2} > (1+\delta)\omega\} = 1\{V_2 < v_U\}$$

To bound S from below, it will suffice to characterize a lower bound on v_U. From (A5):

$$\begin{aligned}\frac{1-F_U\left(v_U\right)}{1-G_{V_2}\left(v_U\right)} &< H\left(v_U\right)N^{-a_{0N}} \Longleftrightarrow 1-F_U\left(v_U\right) < g_{V_2}\left(v_U\right)N^{-a_{0N}} \\ &\Longleftrightarrow F_U\left(v_U\right) > 1-(1+\delta)\omega N^{-a_{0N}} \\ &\Longleftrightarrow v_U > F_U^{-1}\left(1-(1+\delta)\omega N^{-a_{0N}}\right)\end{aligned}$$

Therefore

$$S \geq \tau\left(g_{V2}=(1+\delta)\omega\right)\left\{F_U^{-1}\left(1-N^{-a_{0N}}e^{-b}\right) < V_2 < F_U^{-1}\left(1-(1+\delta)\omega N^{-a_{0N}}\right)\right\}$$

and hence

$$E(S^2|\bar{v}) \geq \tau^2\left(g_{V2}=(1+\delta)\omega\right)\Pr\left[F_U^{-1}\left(1-N^{-a_{0N}}e^{-b}\right) < V_2 < F_U^{-1}\left(1-(1+\delta)\omega N^{-a_{0N}}\right)|\bar{v}\right]$$

From the tail assumption in (A5a) on the conditional distribution of V_2 given $V_1=\bar{v}$,

$$\begin{aligned}&\Pr\left[F_U^{-1}\left(1-N^{-a_{0N}}e^{-b}\right) < V_2 < F_U^{-1}\left(1-(1+\delta)\omega N^{-a_{0N}}\right)|\bar{v}\right] \\ = \;& G_{V_2}(F_U^{-1}\left(1-(1+\delta)\omega N^{-a_{0N}}\right)|\bar{v}) - G_{V_2}(F_U^{-1}\left(1-N^{-a_{0N}}e^{-b}\right)|\bar{v}) \\ > \;& \left(1-(1+\delta)\omega N^{-a_{0N}}\right) - \left(1-N^{-a_{0N}}e^{-b}\right) = N^{-a_{0N}}\left[e^{-b}-(1+\delta)\omega\right] = O\left(N^{-a_{0N}}\right)\end{aligned}$$

Since $\tau\left(g_{V2}=(1+\delta)\omega\right)$ approaches 1 as N increases, for N sufficiently large, we have established a lower bound of $O\left(N^{-a_{0N}}\right)$ for $E(S^2|\bar{v})$. Now the Lindeberg condition becomes:

$$L \equiv O\left(N^a\right)E\left[\begin{array}{c}\frac{1}{h_s}K^2\left(\frac{\bar{v}-V_1}{h_s}\right)S^2* \\ 1\left\{K^2\left(\frac{\bar{v}-V_1}{h_s}\right)S^2 > h_s O\left[N^{1-a_{0N}}\right]\right\}\end{array}\right] \to 0.$$

Since $K^2\left(\frac{\bar{v}-V_{1j}}{h_s}\right)S^2$ is uniformly bounded and since $h_s N^{1-a_{0N}} \to \infty$, the indicator above will be zero for N sufficiently large and the proof will follow.

Lemma 10. For notational simplicity, we denote $C_N = C_N\left(\bar{v}\right)$ as in Lemma 4. Then, under (A1-7)

$$C_N\left[\hat{\zeta}\left(\bar{v},\hat{a}\right)-\zeta_0\left(\bar{v}\right)\right] = C_N\gamma_0 + o_p(1).$$

Proof. Letting $\vartheta_j \equiv (Y_{1j} - \zeta_0(\bar{v}))\, Y_{2j} K\left[(\bar{v} - V_{1j})/h_s\right]/h_s$,

$$\begin{aligned}
& C_N\left[\hat{\zeta}(\bar{v}, \hat{a}) - \zeta_0(\bar{v})\right] - C_N\gamma_0 \\
= \; & C_N\left[\frac{\sum_j \frac{1}{N}\vartheta_j S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right)}{\hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)} - \frac{\sum_j \frac{1}{N}\vartheta_j S\left(\tau_j, x(a_{0N}, P_{2j})\right)}{E(S|V_1 = \bar{v})\, g_{V_1}(\bar{v})}\right] \\
= \; & C_N\left[\begin{array}{c} \frac{\sum_j \frac{1}{N}\vartheta_j\left[S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right) - S\left(\tau_j, x(a_{0N}, P_{2j})\right)\right]}{\hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)} - \\ \frac{\sum_j \frac{1}{N}\vartheta_j S\left(\tau_j, x(a_{0N}, P_{2j})\right)}{E(S|V_1 = \bar{v})\, g_{V_1}(\bar{v})} * \frac{\hat{E}\left(Y_2 S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right)|\bar{v}\right) - E(S|V_1 = \bar{v})\, g_{V_1}(\bar{v})}{\hat{E}\left(Y_2 S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right)|\bar{v}\right)} \end{array}\right].
\end{aligned}$$

where $P_{2j} = \Pr(Y_{2j} = 1|V_{2j})$. From Lemma 9,

$$\begin{aligned}
C_N \frac{\sum_j \frac{1}{N}\vartheta_j S\left(\tau_j, x(a_{0N}, P_{2j})\right)}{M_3(a_{0N})} &= O_p(1) \\
\frac{\hat{E}\left(Y_2 S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right)|\bar{v}\right) - E(S|V_1 = \bar{v})\, g_{V_1}(\bar{v})}{\hat{E}\left(Y_2 S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right)|\bar{v}\right)} &= o_p(1).
\end{aligned}$$

Therefore, we may ignore the product of these two terms and study

$$C_N\left[\frac{\sum_j \frac{1}{N}\vartheta_j\left[S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right) - S\left(\tau_j, x(a_{0N}, P_{2j})\right)\right]}{\hat{E}\left(Y_2 S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right)|\bar{v}\right)}\right].$$

From Lemma 9(a), we can multiply it with $\frac{\hat{E}\left(Y_2 S\left(\hat{\tau}, \hat{a}, \hat{P}_a\right)|\bar{v}\right)}{E(S|V_1 = \bar{v})\, g_{V_1}(\bar{v})}$ and study

$$C_N \frac{\sum_j \frac{1}{N}\vartheta_j\left[S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right) - S\left(\tau_j, x(a_{0N}, P_{2j})\right)\right]}{E(S|V_1 = \bar{v})\, g_{V_1}(\bar{v})}.$$

It remains to be shown that

$$C_N \frac{\sum_j \frac{1}{N}\vartheta_j\left[S\left(\hat{\tau}_j, x(\hat{a}, \hat{P}_{aj})\right) - S\left(\tau_j, x(a_{0N}, P_{2j})\right)\right]}{E(S|V_1 = \bar{v})\, g_{V_1}(\bar{v})} \xrightarrow{p} 0. \tag{25}$$

To study $S\left(\hat{\tau}, x(\hat{a}, \hat{P}_a)\right) - S\left(\tau, x(a_{0N}, P_2)\right) = \hat{\tau} T\left(x(\hat{a}, \hat{P}_a)\right) - \tau T\left(x(a_{0N}, P_2)\right)$, we use Taylor expansion on both $\hat{\tau}$ and $T\left(\hat{a}, \hat{P}_a\right)$. In doing these expansions, we are evaluating all indices at θ_0 to simplify the exposition. It is relatively straight forward to account for the estimation of the index parameters using a

further Taylor series expansion in θ and the fact that $\left(\hat{\theta}-\theta_0\right)$ is $O_p\left(N^{-1/2}\right)$.

$$\begin{aligned}\hat{\tau}(\hat{\omega},\hat{g}_{V_2}) &= \tau(\hat{\omega},g_{V_2})+\frac{\partial\tau}{\partial\omega}(\hat{\omega}-\omega)+\frac{\partial\tau}{\partial g_{V_2}}(\hat{g}_{V_2}-g_{V_2})+...\\ T\left(x(\hat{a},\hat{P}_a)\right) &= T\left(x(a_{0N},P_2)\right)+\frac{\partial T}{\partial x}\frac{\partial x}{\partial a}(\hat{a}-a_{0N})+\frac{\partial T}{\partial x}\frac{\partial x}{\partial P_2}(\hat{P}_a-P_2)+...\end{aligned}$$

Recalling Lemma 5, with the expansion going far enough the remainder terms can be ignored. Substituting the above series into (25), denote $\bar{\varepsilon}_{2j}\equiv\hat{g}_{V_2j}-g_{V_2j}$ it suffices to study:

$$C_N\frac{\sum_j\frac{1}{N}\vartheta_j\left[\frac{\partial\tau_j}{\partial g_{V_2j}}\bar{\varepsilon}_{2j}\right]\left[\frac{\partial T_j}{\partial x_j}\frac{\partial x_j}{\partial P_{2j}}(\hat{P}_{aj}-P_{2j})\right]}{E\left(S|V_1=\bar{v}\right)g_{V_1}(\bar{v})}$$

This term is representative in the sense that an analysis of it will make it clear that the same argument holds for the other terms.

$$C_N\frac{\sum_j\frac{1}{N}\vartheta_j\left[-\tau_j(1-\tau_j)N^{\alpha}\bar{\varepsilon}_{2j}\right]\left[\frac{\partial T_j}{\partial x_j}*\frac{\hat{P}_{aj}-P_{2j}}{1-P_{2j}}\right]}{E\left(S|V_1=\bar{v}\right)g_{V_1}(\bar{v})}$$

To show that the above term converges in probability to zero, we show that the expectation of its square converges to zero. There are two types of elements in it:

$$\frac{C_N^2}{N^2E^2\left(S|V_1=\bar{v}\right)g_{V_1}^2(\bar{v})}\sum_j\left\{\vartheta_j\tau_j(1-\tau_j)N^{\alpha}\bar{\varepsilon}_{2j}\frac{\partial T_j}{\partial x_j}\left[\frac{\hat{P}_{aj}-P_{2j}}{1-P_{2j}}\right]\right\}^2 \qquad \text{(Squared Term)}$$

and

$$\frac{C_N^2}{N^2E^2\left(S|V_1=\bar{v}\right)g_{V_1}^2(\bar{v})}\sum_j\sum_{i\neq j}\left[\begin{array}{c}\left\{\vartheta_i\tau_i(1-\tau_i)N^{\alpha}\bar{\varepsilon}_{2i}\frac{\partial T_i}{\partial x_i}\left[\frac{\hat{P}_{ai}-P_{2i}}{1-P_{2i}}\right]\right\}\\ *\left\{\vartheta_j\tau_j(1-\tau_j)N^{\alpha}\bar{\varepsilon}_{2j}\frac{\partial T_j}{\partial x_j}\left[\frac{\hat{P}_{aj}-P_{2j}}{1-P_{2j}}\right]\right\}\end{array}\right]. \qquad \text{(Cross-product Term)}$$

It can be shown that the expectation of the squared terms converge to zero much faster than the expectation of the cross-product terms and require a much simpler argument. Accordingly, in what follows we analyze the cross-product terms.

Recall (D5) and denote $\tilde{g}_{V_2j}\equiv\sum_{d_2=0}^{1}\hat{f}_2\left(t_2;d_2,h_m\right)$, write

$$\frac{\hat{P}_{aj}-P_{2j}}{1-P_{2j}}=\left[\frac{\bar{\varepsilon}_{1j}}{1-P_{2j}}\right]\left[\frac{g_{V_2j}}{\tilde{g}_{V_2j}}\right]\text{ where }\bar{\varepsilon}_{1j}\equiv\left(\hat{P}_{aj}-P_{2j}\right)\frac{\tilde{g}_{V_2j}}{g_{V_2j}}.$$

From a Taylor series expansion in $\frac{1}{\tilde{g}_{V_2 j}}$, we have

$$\frac{g_{V_2 j}}{\tilde{g}_{V_2 j}} = 1 - \frac{\tilde{g}_{V_2 j} - g_{V_2 j}}{g_{V_2 j}} + ...$$

As $\frac{\tilde{g}_{V_2 j} - g_{V_2 j}}{g_{V_2 j}}$ and the following terms converges to zero, in the following proof we consider the leading term, which is equal to 1. Define $\hat{P}_{aj}[i]$ by removing from $\hat{P}_{aj}$ its dependence on ith observation. Similarly, define $\hat{P}_{ai}[j], \bar{\varepsilon}_{1i}[j], \bar{\varepsilon}_{1j}[i], \bar{\varepsilon}_{2i}[j]$ and $\bar{\varepsilon}_{2j}[i]$. As the terms involving the removed observations converge to zero very fast, it suffices to study

$$\begin{aligned} & \frac{N^{2\alpha} C_N^2}{E^2(S|V_1 = \bar{v}) g_{V_1}^2(\bar{v})} E \left[\begin{array}{c} \left\{ \vartheta_i \tau_i (1-\tau_i) \bar{\varepsilon}_{2i}[j] \frac{\partial T_i}{\partial x_i} \left[\frac{\bar{\varepsilon}_{1i}[j]}{1-P_{2i}} \right] \right\} \\ * \left\{ \vartheta_j \tau_j (1-\tau_j) \bar{\varepsilon}_{2j}[i] \frac{\partial T_j}{\partial x_j} \left[\frac{\bar{\varepsilon}_{1j}[i]}{1-P_{2j}} \right] \right\} \end{array} \right] \\ = & \frac{N^{2\alpha} C_N^2}{E^2(S|V_1 = \bar{v}) g_{V_1}^2(\bar{v})} E \left[\begin{array}{c} E\{\vartheta_i \vartheta_j | X\} E\{\tau_i \tau_j \bar{\varepsilon}_{2i}[j] \bar{\varepsilon}_{2j}[i] \bar{\varepsilon}_{1i}[j] \bar{\varepsilon}_{1j}[i] | X\} \\ (\frac{1-\tau_i}{1-P_{2i}})(\frac{1-\tau_j}{1-P_{2j}}) * \frac{\partial T_j}{\partial x_j} \frac{\partial T_i}{\partial x_i} \end{array} \right]. \end{aligned}$$

Notice that $E\{\vartheta_i \vartheta_j | X\} = E\{\vartheta_i \vartheta_j | V\}$. Employing a similar argument as in Lemma 3, we have $E\{\vartheta_i \vartheta_j | V\}$ to be uniformly of the order $O(N^{-2a})$. From Lemma 8, $E\{\tau_i \tau_j \bar{\varepsilon}_{2i}[j] \bar{\varepsilon}_{2j}[i] \bar{\varepsilon}_{1i}[j] \bar{\varepsilon}_{1j}[i] | V\}$ is uniformly of the order $O(h_T^4 h_2^4) = O(N^{-1.2})$. As $C_N(\bar{v}) \equiv \frac{\sqrt{N h_s} E(S|\bar{v})}{\sqrt{E(S^2|\bar{v})}}$ and $O(E(S|\bar{v})) \leq O(E(S))$, we have $C_N^2 = O(\frac{N h_s [E(S)]^2}{E(S^2|\bar{v})}) = O(N^{2a-\varepsilon})$ by (D4). Further, $(\frac{1-\tau_i}{1-P_{2i}})(\frac{1-\tau_j}{1-P_{2j}}) * \frac{\partial T_j}{\partial x_j} \frac{\partial T_i}{\partial x_i} = O(N^{2a}) \frac{\partial T_j}{\partial x_j} \frac{\partial T_i}{\partial x_i}$, hence we can study

$$\begin{aligned} & N^{2\alpha} O(N^{2a-\varepsilon}) O(N^{-2a-1.2}) O(N^{2a}) \frac{E\left\{ \frac{\partial T_i}{\partial x_i} \frac{\partial T_j}{\partial x_j} \right\}}{E^2(S|V_1 = \bar{v}) g_{V_1}^2(\bar{v})} \\ = & O(N^{2\alpha + 2a - 1.2 - \varepsilon}) \frac{E\left\{ \frac{\partial T_i}{\partial x_i} \frac{\partial T_j}{\partial x_j} \right\}}{E^2(S|V_1 = \bar{v}) g_{V_1}^2(\bar{v})}. \end{aligned}$$

Because $\alpha \leq 0.2, a \leq 0.4 - \varepsilon$, it remains to be shown that $\frac{E\left\{ \frac{\partial T_i}{\partial x_i} \frac{\partial T_j}{\partial x_j} \right\}}{E^2(S|V_1 = \bar{v}) g_{V_1}^2(\bar{v})} = O(1)$ to complete the proof. Notice that $\frac{\partial T}{\partial x}$ is zero except in region $R2$ in (D3), while T is zero except in region $R2$ and $R3$, we have

$$\frac{E\left\{ \frac{\partial T_i}{\partial x_i} \frac{\partial T_j}{\partial x_j} \right\}}{E^2(S|V_1 = \bar{v}) g_{V_1}^2(\bar{v})} = O\left(\frac{c_1 \Pr(R2)}{c_2 \Pr(R2) + c_3 \Pr(R3)} \right)^2$$

where

$$\begin{aligned} c_1 &\equiv \left\{ E\left[\frac{\partial T_i}{\partial x_i} \frac{\partial T_j}{\partial x_j} | R2 \right] \right\} \\ c_2 &\equiv E\left[\left(1 - \exp \frac{-x^k}{b^k - x^k} \right) \left(\frac{1}{h} Y_2 K\left[\frac{\bar{v} - V_1}{h} \right] \right) | R2 \right] \\ c_3 &= E\left[\left(\frac{1}{h} Y_2 K\left[\frac{\bar{v} - V_1}{h} \right] \right) | R3 \right]. \end{aligned}$$

Therefore, $\frac{E\left\{ \frac{\partial T_i}{\partial x_i} \frac{\partial T_j}{\partial x_j} \right\}}{E^2(S|V_1=\bar{v}) g_{V_1}^2(\bar{v})} = O(1)$. The proof now follows.

9.2.3 Index Lemmas

The next lemma proves that the estimated second-stage objective function $\hat{L}^*(\theta)$ is uniformly close to a more tractable objective function $L^*(\theta)$ that does not depend on estimated functions. To define $L^*(\theta)$, we require additional notation. Let $q_L^*(d_1)$ be a lower sample quantile for $f(V(\theta); d_1)$ where:

$$f(V(\theta); d_1) \equiv \Pr(Y_1 = d_1 | Y_2 = 1)\, g_{V|Y_1=d_1, Y_2=1}(V(\theta))$$

Refer to (D7) and define conditional components of adjusted probabilities as:

$$\begin{aligned} f^*(t; d_1) &\equiv f(t; d_1) + \Delta(\tau_I(t), q_L^*(\theta, d_1)) \\ P_c^*(t; d_1) &\equiv f^*(t; d_1) / \sum_{d_1=0}^{1} f^*(t; d_1) \end{aligned}$$

For the marginal components of adjusted probabilities, let $q_{2L}^*(d_2)$ be a lower sample quantile for $f_2(V_2(\theta); d_2)$ where:

$$f_2(V_2(\theta); d_2) \equiv \Pr(Y_2 = d_2)\, g_{V_2}(V_2(\theta)).$$

Then, define:

$$\begin{aligned} f_2^*(t_2; d_2) &\equiv f_2(t_2, d_2) + \Delta_2(\tau_I(t_2), q_{2L}^*(\theta, d_2)) \\ P_2^*(t_2; d_2) &\equiv f_2^*(t_2; d_2) / \sum_{d_2=0}^{1} f_2^*(t_2; d_2). \end{aligned}$$

The adjusted probability function is then given as:

$$P_i^*(d_1, d_2; \theta) \equiv P_c^*(V_i(\theta); d_1) P_2^*(V_{2i}(\theta); d_2).$$

In the lemmas below, it will also be useful to provide a separate analysis depending on the set in which the index vector lies. To this end, denote $g_{V|Y_1=d_1,,Y_2=1}$ as the density for V conditioned on $Y_1 = d_1$ and $Y_2 = 1$. and $g_{V_2|,Y_2=d_2}$ as the density for V_2 conditioned on $Y_2 = d_2$, Referring to (D?), assume that there exists a set of index values, $\mathfrak{B}$. with complement $\mathfrak{B}^*$ such that: (a) For $v \equiv (v_1, v_2) \epsilon\, \mathfrak{B}^*$ all of these densities exceed $N^{-\delta^*}$, $\delta^* < r_c/2$ and (b) $Pr(V \epsilon\, \mathfrak{B}) = O(N^{-\delta}), \delta > 0$. The class of densities for which these conditions hold is very wide.

Lemma 11. Let $\hat{L}^*(\theta) \equiv \sum_i \hat{L}_i^*$ and $L^*(\theta) \equiv \sum_i L_i^*$ where from (D10)

$$\begin{aligned} \hat{L}_i^* &\equiv I\{\hat{V}_i \epsilon \hat{\Psi}_v\} \sum_{d_1 \le d_2} Y_i(d_1, d_2) \ln\left[\hat{P}_i^*(d_1, d_2; \theta)\right] \\ L_i^* &\equiv I\{V_{i0} \epsilon \Psi_v\} \sum_{d_1 \le d_2} Y_i(d_1, d_2) \ln\left[P_i^*(d_1, d_2; \theta)\right]. \end{aligned}$$

Then, assuming (A1-4)

$$\sup_\theta \frac{1}{N} |\hat{L}^*(\theta) - \sum_i I\{X_i \epsilon \mathfrak{B}_N^*\} L_i^*| = o_p(1).$$

Proof. First we show $\sup_\theta \frac{1}{N} |\hat{L}^*(\theta) - L^*(\theta)| = o_p(1)$. Since $\frac{1}{N} |\hat{L}^*(\theta) - L^*(\theta)| \le D_1 + D_2$, where

$$\begin{aligned} D_1 &\equiv \frac{1}{N} \sum_i \sum_{d_1 \le d_2} \left| I\{\hat{V}_i \epsilon \hat{\Psi}_v\} - I\{V_{i0} \epsilon \Psi_v\} \right| \left| \ln\left[P_i^*(d_1, d_2; \theta)\right] \right| \\ D_2 &\equiv \frac{1}{N} \sum_i \sum_{d_1 \le d_2} \left| \ln\left(\hat{P}_i^*(d_1, d_2; \theta) / P_i^*(d_1, d_2; \theta)\right) \right|, \end{aligned}$$

it suffices to show that each of the above terms uniformly converges in probability to zero. For D_1:

$$D_1 \le \sum_{d_1 \le d_2} \sup_{i,\theta} \left| \ln\left[P_i^*(d_1, d_2; \theta)\right] \right| \frac{1}{N} \sum_i \sum_{d_1 \le d_2} \left| I\{\hat{V}_i \epsilon \hat{\Psi}_v\} - I\{V_{i0} \epsilon \Psi_v\} \right|$$

For the first factor of D_1, define $P_c^*(V_i(\theta); d_1)$ and $P_2^*(V_{2i}(\theta); d_2)$ as the probability limit of $\widehat{\Pr}^*(Y_{1i} = d_1 | Y_{2i} = d_2, V_i = t)$

and $\widehat{\Pr}^{*}\left(Y_{2i}=d_2|V_{2i}=t_2\right)$ holding $\Delta's$ fixed, then

$$P_i^*\left(d_1,d_2;\theta\right)\equiv P_c^*\left(V_i(\theta);d_1\right)P_2^*\left(V_{2i}(\theta);d_2\right)$$

As the analysis for P_c^* and P_2^* components is identical, here we consider the former.

$$\begin{aligned}
P_c^*\left(V_i(\theta);d_1\right) &= \frac{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)+\Delta(\tau_I(t),q_L^*(\theta,1))}{g(V|Y_2=1)\Pr(Y_2=1)+\sum_{d_1=0}^{1}\Delta(\tau_I(t),q_L^*(\theta,1))} \\
&\geq \frac{\min\left[\begin{array}{c}\inf\limits_{t\epsilon\mathfrak{B}_N^*,\theta}P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)+\Delta(\tau_I(t),q_L^*(\theta,1)),\\ \inf\limits_{t\epsilon\mathfrak{B}_N^*,\theta}P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)+\Delta(\tau_I(t),q_L^*(\theta,1))\end{array}\right]}{\sup\limits_{t,\theta}g(V|Y_2=1)\Pr(Y_2=1)+\sup\sum_{d_1=0}^{1}\Delta(\tau_I(t),q_L^*(\theta,1))} \\
&\geq \frac{\min\left[\inf\limits_{t\epsilon\mathfrak{B}_N,\theta}\Delta(\tau_I(t),q_L^*(\theta,1)),\inf\limits_{t\epsilon\mathfrak{B}_N^*,\theta}P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)\right]}{\sup\limits_{t,\theta}g(V|Y_2=1)\Pr(Y_2=1)+\sup\sum_{d_1=0}^{1}\Delta(\tau_I(t),q_L^*(\theta,1))}.
\end{aligned}$$

From the definition of $\mathfrak{B}_N^*$,

$$\inf_{t\epsilon\mathfrak{B}_N^*,\theta}P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)>N^{-r_c/2}$$

and from the definition of $\Delta(\tau_I(t),q_L^*(\theta,1)$:

$$\inf_{t\epsilon\mathfrak{B}_N,\theta}\Delta(\tau_I(t),q_L^*(\theta,1))>N^{-r_c/2}$$

From Assumption (A4) and the definition of Δ we have both $\sup\limits_{t,\theta}g(V|Y_2=1)\Pr(Y_2=1)$ and $\sup\sum_{d_1=0}^{1}\Delta(\tau_I(t),q_L^*(\theta,1)$ bounded. It follows that $\sup_{\theta,i}\left|\ln\left[P_c^*\left(V_i(\theta);d_1\right)\right]\right|=O_p\left[\ln\left(N^{r_c/2}\right)\right]=O_p(\ln N)$. From Klein (1993, Lemmas 1-2), there exists a $\varepsilon>0$ such that the indicator difference converges much faster than $o_p\left(N^{-\varepsilon}\right)$, which completes the argument for D_1.

Lemma 11 will now follow by showing uniform convergence for D_2 on $\mathfrak{B}_N^*$ and on $\mathfrak{B}_N$. Similar to the analysis for D_1,it will suffice to analyze the P_c^*-components of D_2. To this end, let

$$D_{2c}\left(\mathfrak{B}_N^*\right)\equiv\frac{1}{N}\sum_i\sum_{d_1\leq d_2}I\{X_i\epsilon\mathfrak{B}_N^*\}\left|\ln\left[\hat{P}_c^*\left(V_i(\theta);d_1\right)/P_c^*\left(V_i(\theta);d_1\right)\right]\right|,$$

and define $D_{2c}(\mathfrak{B}_N)$ analogously for $X_i \epsilon \mathfrak{B}_N$. We then need to show that $D_{2c}(\mathfrak{B}_N^*)$ and $D_{2c}(\mathfrak{B}_N)$ converge in probability to 0, uniformly in θ. Beginning with $D_{2c}(\mathfrak{B}_N^*)$, from a Taylor series expansion of $\ln\left(\hat{P}_c^*(V_i(\theta); d_1)\right)$ about $P_c^*(V_i(\theta); d_1)$, uniform convergence in probability to 0 follows from Lemma 1. For $D_{2c}(\mathfrak{B}_N)$,

$$\sup_\theta D_{2c}(\mathfrak{B}_N) \leq \left[\sup_{X_i \epsilon \mathfrak{B}_N^*, \theta} \sum_{d_1 \leq d_2} \left| \ln\left[\hat{P}_c^*(V_i(\theta); d_1) / P_c^*(V_i(\theta); d_1)\right] \right|\right] \left[\frac{1}{N} \sum_i I\{X_i \epsilon \mathfrak{B}_N^*\}\right].$$

Similar to the analysis for D_1, it can be shown that the first component increases at a $\ln N$ rate, while from Lemma 1 the second term converges to zero at a rate given by N raised to a negative power. Hence $\sup_\theta \frac{1}{N} |\hat{L}^*(\theta) - L^*(\theta)| = o_p(1)$.

Now we show $\sup_\theta \frac{1}{N} |L^*(\theta) - \sum_i I\{X_i \epsilon \mathfrak{B}_N^*\} L_i^*| = o_p(1)$ to complete the proof. Note that

$$\begin{aligned}
& \sup_\theta \left| \frac{1}{N} \sum_i I\{V_i(\theta_0) \epsilon \Psi_v\} \sum_{d_1 \leq d_2} I\{V_i(\theta) \epsilon \mathfrak{B}_N\} Y_i(d_1, d_2) \ln[P_i^*(d_1, d_2; \theta)] \right| \\
\leq & \sup_\theta \frac{1}{N} \sum_i \sum_{d_1 \leq d_2} I\{X_i \epsilon \mathfrak{B}_N\} |\ln[P_i^*(d_1, d_2; \theta)]| \\
\leq & \sup_{X_i \epsilon \mathfrak{B}_N, \theta} |\ln[P_i^*(d_1, d_2; \theta)] \sup_\theta \frac{1}{N} \sum_i \sum_{d_1 \leq d_2} I\{X_i \epsilon \mathfrak{B}_N\}.
\end{aligned}$$

The first term above explodes slower than $O_p(\ln N)$ as above. From the conditions on $\mathfrak{B}_N$, the second term decreases at a rate given by N raised to a negative power. The result then follows.

To establish consistency for the proposed estimator, we will need to show that the objective function underlying the estimator is uniformly close to one that is continuous and that is uniquely maximized at θ_0.

Lemma 12. Under (A1-4, A8), let:

$$L_i \equiv I\{V_{i0} \epsilon \Psi_v\} \sum_{d_1 \leq d_2} Y_i(d_1, d_2) \ln[P_i(d_1, d_2; \theta)]$$

and recall the definition of L_i^* in Lemma 11. Then,

$$\sup_\theta \frac{1}{N} \left| \sum I\{X_i \epsilon \mathfrak{B}_N^*\} L_i^* - E\left[\sum L_i\right] \right| = o_p(1)$$

and $E\frac{1}{N}\left[\sum L_i\right]$ is a bounded and continuous function of θ.[11].

Proof. First we show $\sup_\theta \frac{1}{N}|\sum I\left\{X_i \epsilon \mathfrak{B}_N^*\right\}\left[L_i^* - L_i\right]| = o_p(1)$. Note that:

$$|\ln P_i^* - \ln P_i| \leq \left|\ln \frac{P_c^*\left(V_i(\theta); d_1\right)}{P_{ci}}\right| + \left|\ln \frac{P_2^*\left(V_{2i}(\theta); d_2\right)}{P_{2i}}\right|$$

Following the same strategy as in Lemma 11, as arguments for P_c^* and P_2^* components are identical, here we consider the former.

$$\begin{aligned}
\left|\ln \frac{P_c^*\left(V_i(\theta); d_1\right)}{P_{ci}}\right| &= \ln\left[\begin{array}{c} \frac{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)+\Delta(\tau_I(t),q_L^*(\theta,1))}{g(V|Y_2=1)\Pr(Y_2=1)+\sum_{d_1=0}^{1}\Delta(\tau_I(t),q_L^*(\theta,1))} \\ *\frac{g(V|Y_2=1)\Pr(Y_2=1)}{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)} \end{array}\right] \\
&\leq \left|\ln\left(1+\frac{\Delta(\tau_I(t),q_L^*(\theta,1))}{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)}\right)\right| \\
&\quad + \left|\ln\left(1+\frac{\sum_{d_1=0}^{1}\Delta(\tau_I(t),q_L^*(\theta,1))}{g(V|Y_2=1)\Pr(Y_2=1)}\right)\right|.
\end{aligned}$$

As the arguments for these terms are identical, here we consider the first term. From a Taylor series expansion in $\frac{\Delta(\tau_I(t),q_L^*(\theta,1))}{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)}$ about zero

$$\left|\ln\left(1+\frac{\Delta(\tau_I(t),q_L^*(\theta,1))}{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)}\right)\right| = \left|\begin{array}{c} \frac{1}{1+\left(\frac{\Delta(\tau_I(t),q_L^*(\theta,1))}{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)}\right)^+} \\ *\left[\frac{\Delta(\tau_I(t),q_L^*(\theta,1))}{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)}\right] \end{array}\right|,$$

where $+$ indicates an intermediate point. On $\mathfrak{B}_N^*$, $\frac{\Delta(\tau_I(t),q_L^*(\theta,1))}{P(Y_1=1,Y_2=1)g(V|Y_1=1,Y_2=1)}$ converges to 0. Hence $\sup_\theta \frac{1}{N}|\sum I\left\{X_i \epsilon \mathfrak{B}_N^*\right\}|$ $o_p(1)$. From Lemma 1,

$$\sup_\theta \frac{1}{N}|\sum I\left\{X_i \epsilon \mathfrak{B}_N^*\right\} L_i - E\left[\sum I\left\{X_i \epsilon \mathfrak{B}_N^*\right\} L_i\right]| = o_p(1).$$

Now we show $\sup_\theta \frac{1}{N}\left|E\left[\sum I\left\{X_i \epsilon \mathfrak{B}_N^*\right\} L_i\right] - E\left[\sum L_i\right]\right| = o_p(1)$ to complete the proof. From the definition of

[11] If there is only one unbounded X with non-zero true coefficient, then this dominance condition is not required as we are trimming on the true index.

L_i :

$$\begin{aligned} |L_i| &\leq \left| \sum_{d_1 \leq d_2} \ln \left[P_i \left(d_1, d_2; \theta \right) \right] \right| \\ &\leq \sum_{d_1 \leq d_2} \sup_{\theta} \left| \ln \left[P_i \left(d_1, d_2; \theta \right) \right] \right|, \end{aligned}$$

which is bounded under (A8).

$$\begin{aligned} & \sup_{\theta} \frac{1}{N} \left| E \left[\sum I \left\{ X_i \epsilon \mathfrak{B}_N^* \right\} L_i \right] - E \left[\sum L_i \right] \right| \\ = & \sup_{\theta} \frac{1}{N} \left| E \left[\sum I \left\{ X_i \epsilon \mathfrak{B}_N \right\} L_i \right] \right| \\ = & \sup_{\theta} \left| E \left[I \left\{ X_i \epsilon \mathfrak{B}_N \right\} L_i \right] \right|. \end{aligned}$$

From Cauchy-Schwarz,

$$\sup_{\theta} \left| E \left[I \left\{ X_i \epsilon \mathfrak{B}_N \right\} L_i \right] \right| \leq \sqrt{\left| E \left[I \left\{ X_i \epsilon \mathfrak{B}_N \right\} \right] \right|} \sqrt{\sup_{\theta} \left| E \left[L_i^2 \right] \right|}$$

The first term converges to 0, while the second is bounded under (A8). Hence $\sup_{\theta} \frac{1}{N} \left| \sum I \left\{ X_i \epsilon \mathfrak{B}_N^* \right\} L_i^* - E \left[\sum L_i \right] \right| = o_p(1)$. Lastly, from (A8), $E \left[\frac{1}{N} \sum L_i \right]$ is a bounded and continuous function of θ.

Lemma 13. Initial Index Estimator. Recall the definition of $\hat{P}_i(d_1, d_2; \theta)$ in (D5) and that

$$\hat{\delta}_i \left(d_1, d_2; \theta \right) \equiv \nabla_\theta \hat{P}_i(d_1, d_2; \theta) / \hat{P}_i(d_1, d_2; \theta).$$

Define:

$$\begin{aligned}
A &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1\le d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_0\right)I\{X_i\epsilon\hat{\Psi}_x\}\\
A_0 &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1\le d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\delta_i\left(d_1,d_2;\theta_0\right)I\{X_i\epsilon\Psi_x\}\\
B &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1\le d_2}\left[\hat{P}_i\left(d_1,d_2;\theta_0\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_0\right)I\{X_i\epsilon\hat{\Psi}_x\}\\
B_0 &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1\le d_2}\left[\hat{P}_i\left(d_1,d_2;\theta_0\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\delta_i\left(d_1,d_2;\theta_0\right)I\{X_i\epsilon\Psi_x\}.
\end{aligned}$$

Then, under (A1-4),

$$\begin{aligned}
a) &: \sqrt{N}\left[A-A_0\right]=o_p(1)\\
b) &: \sqrt{N}\left[B-B_0\right]=o_p(1)\\
c) &: \left[\hat{\theta}_I-\theta_0\right]=O_p\left(N^{-\frac{2}{8+\varepsilon}}\right)
\end{aligned}$$

Proof. For (a), with $\varepsilon_i \equiv Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_0\right)$ and $\hat{\delta}_i \equiv \hat{\delta}_i\left(d_1,d_2;\theta_0\right)$, define

$$\begin{aligned}
\Delta_{A1} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\left[\hat{\delta}_i-\delta_i\right]I\{X_i\epsilon\Psi_x\};\\
\Delta_{A2} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\left[I\{X_i\epsilon\hat{\Psi}_x\}-I\{X_i\epsilon\Psi_x\}\right]\delta_i;\\
\Delta_{A3} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\left[I\{X_i\epsilon\hat{\Psi}_x\}-I\{X_i\epsilon\Psi_x\}\right]\left[\hat{\delta}_i-\delta_i\right].
\end{aligned}$$

then

$$\Delta_{A1}+\Delta_{A2}+\Delta_{A3}=o_p(1)\Rightarrow A-A_0=o_p(1).$$

Below we prove that these $\Delta-$ terms are $o_p(1)$.

To analyze Δ_{A1}, referring to (D5) the following notation is convenient:

$$\begin{aligned}\hat{f}_{ic} &\equiv \hat{f}\left(t; d_1, h_{1c}, h_{2c}\right), \ \hat{g}_{ic} \equiv \sum_{d_1} \hat{f}\left(t; d_1, h_{1c}, h_{2c}\right), \ \hat{P}_{ic} \equiv \hat{f}_{ic}/\hat{g}_{ic} \\ \hat{f}_{im} &\equiv \hat{f}_2\left(t_2; d_2, h\right), \ \hat{g}_{im} \equiv \sum_{d_2} \hat{f}_2\left(t_2; d_2, h\right), \ \hat{P}_{im} \equiv \hat{f}_{im}/\hat{g}_{im}.\end{aligned}$$

Then, recalling the definitions of $\hat{\delta}_i$ and δ_i, let:

$$\begin{aligned}\hat{\delta}_i\left(d_1, d_2; \theta_0\right) &\equiv \nabla_\theta \hat{P}_i(d_1, d_2; \theta_0)/\hat{P}_i(d_1, d_2; \theta_0) \\ &\equiv \nabla_\theta \hat{P}_{ic}/\hat{P}_{ic} + \nabla_\theta \hat{P}_{im}/\hat{P}_{im} \equiv \hat{\delta}_{ic} + \hat{\delta}_{im}. \\ \delta_i\left(d_1, d_2; \theta_0\right) &\equiv \nabla_\theta P_{ic}/P_{ic} + \nabla_\theta P_{im}/P_{im} \equiv \delta_{ic} + \delta_{im}.\end{aligned} \tag{26}$$

As the arguments for $\hat{\delta}_{ic}$and $\hat{\delta}_{im}$ are similar, with $\hat{\delta}_{ic}$ having a slower convergence rate, it suffices to provide the argument for

$$\Delta_{A1c} \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \left[\hat{\delta}_{ic} - \delta_{ic}\right] I\{X_i \epsilon \Psi_x\}.$$

To deal with the estimated denominators of $\hat{\delta}_{ic}$, by definition

$$\begin{aligned}(\hat{\delta}_{ic} - \delta_{ic})\hat{f}_{ic}\hat{g}_{ic} &= \left(\left[\hat{f}_{ic}\nabla_\theta \hat{g}_{ic} - \hat{g}_{ic}\nabla_\theta \hat{f}_{ic}\right] - \delta_{ic}\hat{f}_{ic}\hat{g}_{ic}\right) \\ \delta_{ic} f_{ic} g_{ic} &= f_{ic}\nabla_\theta g_{ic} - g_{ic}\nabla_\theta f_{ic}\end{aligned}$$

Therefore,

$$\begin{aligned}(\hat{\delta}_{ic} - \delta_{ic})\hat{f}_{ic}\hat{g}_{ic} &= \left[\begin{array}{c}\left(\left[f_{ic}\nabla_\theta \hat{g}_{ic} - g_{ic}\nabla_\theta \hat{f}_{ic}\right] - \delta_{ic} f_{ic} g_{ic}\right) + \\ (\hat{f}_{ic} - f_{ic})\nabla g_{ic} + (\hat{f}_{ic} - f_{ic})(\nabla g_{ic} - \nabla \hat{g}_{ic}) + ... +\end{array}\right] \\ &= \left[\begin{array}{c}\left(f_{ic}\left(\nabla_\theta \hat{g}_{ic} - \nabla_\theta g_{ic}\right) - g_{ic}\left(\nabla_\theta \hat{f}_{ic} - \nabla_\theta f_{ic}\right)\right) + \\ (\hat{f}_{ic} - f_{ic})\nabla g_{ic} + (\hat{f}_{ic} - f_{ic})(\nabla g_{ic} - \nabla \hat{g}_{ic}) + ... +\end{array}\right],\end{aligned} \tag{27}$$

where all of the terms either involve the difference between estimated and true functions such as $f_{ic}\left(\nabla_\theta \hat{g}_{ic} - \nabla_\theta g_{ic}\right)$ or cross-product differences such as $(\hat{f}_i - f_i)(\nabla g - \nabla \hat{g})$. Employing (27), Cauchy-Schwarz, and Lemma 2,

it can be shown that:

$$\Delta_{A1c} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \left[\hat{\delta}_{ic} - \delta_{ic}\right] \left[\frac{\hat{f}_{ic}\hat{g}_{ic}}{f_{ic}g_{ic}}\right]] I\{X_i \epsilon \Psi_x\} + o_p(1) \tag{28}$$

Next, substituting (27) into the simplified form of Δ_{A1c}, in (28) we obtain a collection of terms involving the difference between estimated and true functions. Employing Cauchy-Schwarz and Lemma 2, we may ignore terms involving cross product terms such as $(\hat{f}_i - f_i)(\nabla g - \nabla \hat{g})$. From a mean-square convergence argument similar to that in Klein and Shen (2010, Lemma 8), it can be shown that we may also ignore terms that depend on linear combinations of the difference between estimated and true functions as for example the first two components in (27).

For Δ_{A2}, recall from (D9)

$$\hat{\Psi}_x \equiv \{x : \hat{q}_{Lx} < x < \hat{q}_{Ux}\}$$

write it as a function of $\hat{q}$:

$$\hat{\Psi}_x \equiv \Psi_x(\hat{q}) \equiv \{x : \hat{q}_{Lx} < x < \hat{q}_{Ux}\}$$

Then, with q_0 as a vector of true quantiles, $\sqrt{N}\Delta_2 = o_p(1)$ if

$$\sup_{|q-q_o|<\xi} \frac{1}{\sqrt{N}} \sum \left[I\{X_i \epsilon \Psi_x(q)\} - I\{X_i \epsilon \Psi_x(q_0)\}\right] \varepsilon_i \delta_i = o_p(1)$$

for all $\xi = o(1)$.[12] The result then follows from Pakes and Pollard (1989, Lemma 2.17, p. 1037).

Turning to Δ_{A3}, and letting

$$\Psi_x^*(\hat{q}) \equiv \Psi_x(\hat{q}) \cup \Psi_x(q_0), \tag{29}$$

we have:

[12] If uniformity holds for $\alpha \, \epsilon \, \mathcal{N}_\xi$ for all $_\xi = o(1)$, then uniformity holds over $o_p\,(1)$ neighborhoods of α.

$$
\begin{aligned}
|\Delta_{A3}| &\leq \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i \mid I\{X_i\epsilon\Psi_x(\hat{q})\} - I\{X_i\epsilon\Psi_x(q_0)\}\mid I\{X_i\epsilon\Psi_x^*(\hat{q}))\}\left|\hat{\delta}_i - \delta_i\right| \leq \Delta_{31}\Delta_{32} \\
\Delta_{31} &\equiv \left[\sum \varepsilon_i^2 \left[I\{X_i\epsilon\Psi_x(\hat{q})\} - I\{X_i\epsilon\Psi_x(q_0)\}\right]^2/N\right]^{1/2}, \qquad (30) \\
\Delta_{32} &= \left[\sup_{|q-q_o|<\xi}\sum I\{X_i\epsilon\Psi_x^*(q)\}2\left(\left[\hat{\delta}_{im} - \delta_{im}\right]^2 + \left[\hat{\delta}_{ic} - \delta_{ic}\right]^2\right)/N\right]^{1/2}, \qquad (31)
\end{aligned}
$$

$\Delta_{31} = o_p(N^{-\frac{1}{2}+\varepsilon})$. For Δ_{32}, the convergence rate is determined by the $\hat{\delta}_{ic}$ term. It then follows from Lemma 1 that $|\Delta_{32}|^2 = O_p(N^{-c})$, $c > 2/9$. The result (a) follows.

To prove (b), let $\hat{P}_i - P_i \equiv \hat{P}_i(d_1, d_2; \theta_0) - P_i(d_1, d_2; \theta_0)$ and define:

$$
\begin{aligned}
\Delta_{B1} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{P}_i - P_i\right]\left[\hat{\delta}_i - \delta_i\right] I\{X_i\epsilon\Psi_x\} \\
\Delta_{B2} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{P}_i - P_i\right]\left[I\{X_i\epsilon\hat{\Psi}_x\} - I\{X_i\epsilon\Psi_x\}\right]\delta_i \\
\Delta_{B3} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{P}_i - P_i\right]\left[I\{X_i\epsilon\hat{\Psi}_x\} - I\{X_i\epsilon\Psi_x\}\right]\left[\hat{\delta}_i - \delta_i\right]
\end{aligned}
$$

Similar to (a), it now suffices to show that each of these terms is $o_p(1)$.

To analyze Δ_{B1}, recall the definitions of $\hat{P}_i$ and P_i and notice that

$$
\hat{P}_i - P_i \equiv \hat{P}_{im}\hat{P}_{ic} - P_{im}P_{ic} = \left[\hat{P}_{im} - P_{im}\right]\hat{P}_{ic} + P_{im}\left[\hat{P}_{ic} - P_{ic}\right].
$$

Because of its slower convergence rate it suffices to analyze the term involving $P_{im}\left[\hat{P}_{ic} - P_{ic}\right]$. Recalling the characterization of $\hat{\delta}_i$ and δ_i in (26) it also suffices to analyze the component involving $\hat{\delta}_{ic} - \delta_{ic}$. Accordingly, we analyze:

$$
\Delta_{B1c} \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{P}_{ic} - P_{ic}\right]\left[\hat{\delta}_{ic} - \delta_{ic}\right] I\{X_i\epsilon\Psi_x\};
$$

By definition

$$
\left[\hat{P}_{ic} - P_{ic}\right]\hat{g}_{ic} = (f_{ic} - \hat{g}_{ic}P_{ic}) \qquad (32)
$$

Employing the decomposition in (27), Cauchy-Schwarz, and Lemma 2:

$$\Delta_{B1c} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\hat{P}_{ic} - P_{ic} \right) \left(\hat{\delta}_{ic} - \delta_{ic} \right) \left[\frac{\hat{f}_{ic} \hat{g}_{ic}^2}{f_{ic} g_{ic}^2} \right] I\{X_i \epsilon \Psi_x\} + o_p(1)$$

Employing an argument similar to that for Δ_{1A}^c, from(32), (27),Cauchy-Schwarz, and Lemma 2, $\Delta_{B1c} = o_p(1)$.

Turning to Δ_{B2}, it suffices to analyze:

$$\Delta_{B2c} \equiv \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[\hat{P}_{ic} - P_{ic} \right] \left[I\{X_i \epsilon \hat{\Psi}_x\} - I\{X_i \epsilon \Psi_x\} \right] I\{X_i \epsilon \Psi_x(\hat{q}) \cup \Psi_x(q_0)\} \delta_i$$

From Cauchy-Schwarz and the definition of the set Ψ_x^* in (29)

$$|\Delta_{B2c}| \leq \sqrt{N} \sqrt{\frac{\sup_{|q-q_o|<\xi} \sum_{i=1}^{N} \frac{1}{N} \left[\hat{P}_{ic} - P_{ic} \right]^2 I\{X_i \epsilon \Psi_x^*(q)\} \delta_i^2 *}{\sum_{i=1}^{N} \frac{1}{N} \left[I\{X_i \epsilon \hat{\Psi}_x\} - I\{X_i \epsilon \Psi_x\} \right]^2}}. \tag{33}$$

For the first term under the square-root, from the definition of $\hat{P}_{ic} \equiv \hat{f}_{ic}/\hat{g}_{ic}$, (32), and Lemma 2, it can be shown that it is bounded above by:

$$\sup_{|q-q_o|<\xi} \sum_{i=1}^{N} \frac{1}{N} \left[\left(\hat{f}_{ic} - \hat{g}_{ic} P_{ic} \right) \frac{1}{g_{ic}} \right]^2 I\{X_i \epsilon \Psi_x^*(q)\} \delta_i^2 + o_p(N^{-1/2}.)$$

From Lemma 2, this term is $O_p\left(N^{-\frac{4}{8+\varepsilon}}\right)$. In an analysis almost identical to that for Δ_{31} in (31), the second component of (33) is $O_p(N^{-c})$, $c \geq 2/3$, from which it now follows that $\Delta_{B2c} = o_p(1)$.[13] Employing an analysis very similar to that for Δ_{B3}, it can be shown Δ_{B3} converges in probability to 0 at a rate faster than that for Δ_{B2}. Part (b) now follows.

For (c), from a standard Taylor series expansion and results (a-b) above:

$$\left[\hat{\theta}_I - \theta_0 \right] = -\hat{H}^{-1} \left[A - B \right] = \hat{H}^{-1} \left[A_0 - B_0 \right] + o_p(N^{-\frac{1}{2}}), \tag{34}$$

where $\hat{H}$ is the estimated Hessian to the first-stage likelihood. Using Lemma 1 and recalling the windows

[13] From Klein (1993, LemmasA1-2), the second component is actually $o_p(N^{-1/2+\varepsilon}$. To facilitate the analysis in subsequent lemmas, we are emphasizing here that a much slower convergence rate suffices.

defined in (D5), it follows that $\hat{H} = O_p(1)$. Since $\sqrt{N}A_0$ converges in distribution to a normal random variable, $A_0 = O_p(1/\sqrt{N})$, the proof will now follow from the convergence rate for B_0. Recalling that:

$$\hat{P}_i - P_i \equiv \hat{P}_{im}\hat{P}_{ic} - P_{im}P_{ic} = \left[\hat{P}_{im} - P_{im}\right]\hat{P}_{ic} + P_{im}\left[\hat{P}_{ic} - P_{ic}\right],$$

it suffices to examine:

$$B_{0c} \equiv \frac{1}{N}\sum_{i=1}^{N}\left[\hat{P}_c - P_c\right]\delta_i I\{X_i \epsilon \Psi_x\}. \tag{35}$$

Recalling that $\left[\hat{P}_{ic} - P_{ic}\right]\hat{g}_{ic} = (f_{ic} - \hat{g}_{ic}P_{ic})$, from Cauchy-Schwarz and Lemma 2:

$$\begin{aligned} B_{0c} &\equiv \frac{1}{N}\sum_{i=1}^{N}\left[\hat{P}_c - P_c\right]\frac{\hat{g}_{ic}}{\hat{g}_{ic}}\delta_i\left(d_1, d_2; \theta_0\right) I\{X_i \epsilon \Psi_x\} + o_p(1) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[(f_{ic} - \hat{g}_{ic}P_{ic})\right]\frac{1}{\hat{g}_{ic}}\delta_i\left(d_1, d_2; \theta_0\right) I\{X_i \epsilon \Psi_x\} + o_p(1). \end{aligned} \tag{36}$$

From Cauchy-Schwarz and Lemma 2, it can be shown that $B_{0c} = O_p(N^{-\frac{2}{8+\varepsilon}})$. Part (c) now follows.

The second stage estimator requires that trimming be based on an estimated index, and the theory requires that this index converge sufficiently fast to the true index. As the initial index in Lemma 13 does not satisfy the required rate condition, Lemma 14 below makes a bias correction to the initial estimator and shows that the resulting correction significantly increases the convergence rate. We will find that this rate is sufficient for our purposes.

Lemma 14. The Initial Bias-Corrected Estimator. Let:

$$B_I^o \equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{d_1 \le d_2}\left[\hat{P}_i^o\left(d_1, d_2; \theta_0\right) - P_i\left(d_1, d_2; \theta_0\right)\right]\hat{\delta}_i\left(d_1, d_2; \hat{\theta}_I\right) I\{X_i \epsilon \hat{\Psi}_x\}$$

and recall the definition of $\hat{\theta}_I^o$ in (D9). Then, under (A1-4),

$$\begin{aligned} a) \quad &: \quad B_I^o = O_p(N^{-1/3}) \\ b) \quad &: \quad (\hat{\theta}_I^o - \theta_0) = O_p(N^{-1/3}). \end{aligned}$$

Proof. For part (a), employing an analysis very similar to in Lemma 13b, it can be easily shown that:

$$B_I^o = \sum_{i=1}^{N} \sum_{d_1 \leq d_2} \left[\hat{P}_i^o \left(d_1, d_2; \theta_0 \right) - P_i \left(d_1, d_2; \theta_0 \right) \right] \delta_i \left(d_1, d_2; \theta_0 \right) I\{X_i \epsilon \Psi_x\} + o_p(N^{-1/2}).$$

Employing the same argument in (36), the rate of convergence basically comes from $\hat{P}_i^o \left(d_1, d_2; \theta_0 \right) - P_i \left(d_1, d_2; \theta_0 \right)$, which is $O_p(N^{-1/3})$.

Turning to (b), Employing Lemma 13a-b, we have

$$\left(\hat{\theta}_I - \theta_0 \right) = -\hat{H} \left(\theta^+ \right)^{-1} \left[A_0 - B_0 \right] + o_p(N^{-1/2}),$$

where θ^+ is an intermediate point. To simplify the adjustment to this estimator, referring to (D10) we will show below:

$$\Delta \equiv \hat{H} \left(\hat{\theta}_I \right)^{-1} \hat{C}_X \left(\hat{\theta}_I \right) - \hat{H} \left(\theta^+ \right)^{-1} \hat{C}_X \left(\theta_0 \right) = o_p(N^{-1/3}).$$

Rewriting the above expression, $\Delta = \Delta_1 + \Delta_2$, where

$$\begin{aligned} \Delta_1 &\equiv \hat{H} \left(\hat{\theta}_I \right)^{-1} \left[\hat{H} \left(\theta^+ \right) - \hat{H} \left(\hat{\theta}_I \right) \right] \hat{H} \left(\theta^+ \right)^{-1} \hat{C}_X \left(\hat{\theta}_I \right) \\ \Delta_2 &\equiv \hat{H} \left(\theta^+ \right)^{-1} \left[\hat{C}_X \left(\hat{\theta}_I \right) - \hat{C}_X \left(\theta_0 \right) \right]. \end{aligned}$$

To study Δ_1, note that Lemma 13 gives a convergence rate for $\hat{\theta}_I - \theta^+$. Then, using a Taylor series expansion on $\left[\hat{H} \left(\theta^+ \right) - \hat{H} \left(\hat{\theta}_I \right) \right]$ and Lemmas 1, it can be shown that $\left[\hat{H} \left(\theta^+ \right) - \hat{H} \left(\hat{\theta}_I \right) \right]$ and $\hat{C} \left(\hat{\theta}_I \right)$ converge to zero sufficiently fast that $\Delta_1 = o_p(N^{-1/3})$.

For Δ_2, Taylor expand the second component to obtain

$$\Delta_2 = \hat{H} \left(\theta^+ \right)^{-1} \left(\nabla \hat{C} \right) \left(\hat{\theta}_I - \theta_0 \right),$$

where $\nabla \hat{C}$ is evaluated at an intermediate point. The first component is $O_p(1)$ from Lemma 1; the second component is $O_p(\frac{1}{\sqrt{N}h^3})$ from Lemma 1; and the third component is $O_p\left(h^2\right)$, hence we have $\Delta_2 = o_p\left(N^{-1/3}\right)$.

From the definition of $\hat{\theta}_I^o$ in (D9) and employing the result above:

$$\left(\hat{\theta}_I^o - \theta_0\right) = \left(\hat{\theta}_I - \theta_0 + \hat{H}\left(\theta^+\right)^{-1}\hat{C}\left(\theta_0\right)\right) + o_p(1) = -\hat{H}\left(\theta^+\right)^{-1}\left(A_0 - B_I^o\right) + o_p(1),$$

Since $\hat{H}\left(\theta^+\right)$ converges to a positive definite matrix, and $A_0 = O_p(N^{-1/2})$, the proof then follows from part (a).

Lemma 15 provides a characterization for the gradient components of the second stage estimator, which are in turn used to obtain its convergence rate.

Lemma 15. The Second Stage Estimator. Define

$$\hat{\delta}_i^*\left(d_1, d_2; \theta\right) \equiv \nabla_\theta \hat{P}_i^*(d_1, d_2; \theta) / \hat{P}_i^*(d_1, d_2; \theta)$$

and let:

$$\begin{aligned}
A^* &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1 \le d_2}\left[Y_i\left(d_1, d_2\right) - P_i\left(d_1, d_2; \theta_0\right)\right]\hat{\delta}_i^*\left(d_1, d_2; \theta_0\right) I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\} \\
A_0^* &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1 \le d_2}\left[Y_i\left(d_1, d_2\right) - P_i\left(d_1, d_2; \theta_0\right)\right]\delta_i\left(d_1, d_2; \theta_0\right) I\{V_i\epsilon\Psi_v\} \\
B^* &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1 \le d_2}\left[\hat{P}_i^*\left(d_1, d_2; \theta_0\right) - P_i\left(d_1, d_2; \theta_0\right)\right]\hat{\delta}_i^*\left(d_1, d_2; \theta_0\right) I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\} \\
B_0^* &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1 \le d_2}\left[\hat{P}_i^*\left(d_1, d_2; \theta_0\right) - P_i\left(d_1, d_2; \theta_0\right)\right]\delta_i\left(d_1, d_2; \theta_0\right) I\{V_i\epsilon\Psi_v\}.
\end{aligned}$$

Then, under A1-4 and recalling the definition of the second stage estimator, $\hat{\theta}^*$:

$$\begin{aligned}
a) &: \sqrt{N}\left[A^* - A_0^*\right] = o_p(1) \\
b) &: \sqrt{N}\left[B^* - B_0^*\right] = o_p(1) \\
c) &: \left[\hat{\theta}^* - \theta_0\right] = O_p(N^{-1/3})
\end{aligned}$$

Proof. For (a), with $\varepsilon_i \equiv Y_i(d_1,d_2) - P_i(d_1,d_2;\theta_0)$ and $\hat{\delta}_i^* \equiv \hat{\delta}_i^*(d_1,d_2;\theta_0)$, define

$$\begin{aligned}
\Delta_{A1}^* &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\left[\hat{\delta}_i^* - \delta_i\right] I\{V_i \epsilon \Psi_v\};\\
\Delta_{A2}^* &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\left[I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\} - I\{V_i\epsilon\Psi_v\}\right]\delta_i;\\
\Delta_{A3}^* &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\varepsilon_i\left[I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\} - I\{V_i\epsilon\Psi_v\}\right]\left[\hat{\delta}_i^* - \delta_i\right].
\end{aligned}$$

Then

$$\Delta_{A1}^* + \Delta_{A2}^* + \Delta_{A3}^* = o_p(1) \Rightarrow A^* - A_0^* = o_p(1).$$

Using a strategy very similar to Lemma 13, we can prove that these Δ terms are $o_p(1)$.

To prove (b), notice that for $V_i\epsilon\Psi_v$, $\hat{P}_i^* - \hat{P}_i$ goes to zero at an exponential rate. Let $\hat{P}_i^* - P_i \equiv \hat{P}_i^*(d_1,d_2;\theta_0) - P_i(d_1,d_2;\theta_0)$ and define:

$$\begin{aligned}
\Delta_{B1} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{P}_i^* - P_i\right]\left[\hat{\delta}_i^* - \delta_i\right] I\{V_i\epsilon\Psi_v\}\\
\Delta_{B2} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{P}_i^* - P_i\right]\left[I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\} - I\{V_i\epsilon\Psi_v\}\right]\delta_i\\
\Delta_{B3} &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[\hat{P}_i^* - P_i\right]\left[I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\} - I\{V_i\epsilon\Psi_v\}\right]\left[\hat{\delta}_i^* - \delta_i\right].
\end{aligned}$$

The proof is very similar to that in Lemma 13. Finally, using a similar Taylor series argument and proof strategy as in Lemma 13, part (c) follows from (a) and (b)**.**

Lemma 16 below provides the key required result to prove asymptotic normality for the final adjusted estimator.

Lemma 16. Assuming (A1-4),

$$\sqrt{N}B^o = o_p(1),$$

where

$$B^o \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1\le d_2}\left[\hat{P}_i^o(d_1,d_2;\theta_0) - P_i(d_1,d_2;\theta_0)\right]\hat{\delta}_i^*(d_1,d_2;\theta_0)\, I\{V_i\left(\hat{\theta}_I^o\right)\epsilon\hat{\Psi}_v\}.$$

Proof. Using a similar strategy as for Lemma 15b, we can show $\sqrt{N}\left(B^o - B_0^o\right) = o_p(1)$, where

$$B_0^o \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1 \leq d_2}\left[\hat{P}_i^o\left(d_1, d_2; \theta_0\right) - P_i\left(d_1, d_2; \theta_0\right)\right]\delta_i\left(d_1, d_2; \theta_0\right) I\{V_i \epsilon \Psi_v\}.$$

Next, note that:

$$\hat{P}_i^o - P_i \equiv \hat{P}_{im}^o \hat{P}_{ic}^o - P_{im}P_{ic} = \left[\hat{P}_{im}^o - P_{im}\right]\hat{P}_{ic}^o + P_{im}\left[\hat{P}_{ic}^o - P_{ic}\right],$$

where for notational simplicity we have suppressed the fact that every probability function depends on $(d_1, d_2; \theta_0)$. Here, we provide the analysis for the last term as it converges to zero slower than the first. In other words, we study

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{d_1 \leq d_2}\left[P_{im}\left(\hat{P}_{ic}^o - P_{ic}\right)\right]\delta_i\left(d_1, d_2; \theta_0\right) I\{V_i \epsilon \Psi_v\} = U + o_p(1),$$

where the U-statistic satisfies:

$$\begin{aligned} U &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{d_1 \leq d_2}\left[\left(\frac{\hat{f}_c^o\left(t_2; d_2\right)}{\hat{g}_c^o\left(t_2; d_2\right)} - P_{ic}\right)P_{im}\right]\left[\frac{\hat{g}_c^o\left(t_2; d_2\right)}{g_c\left(t_2; d_2\right)}\right]\delta_i\left(d_1, d_2; \theta_0\right) I\{V_i \epsilon \Psi_v\} \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{d_1 \leq d_2}\left[\left(\hat{f}_c^o\left(t_2; d_2\right) - \hat{g}_c^o\left(t_2; d_2\right)P_{ic}\right)P_{im}\right]\left[\frac{\delta_i\left(d_1, d_2; \theta_0\right) I\{V_i \epsilon \Psi_v\}}{g_c\left(t_2; d_2\right)}\right]. \end{aligned}$$

Notice that from Newey's result (8), $E[\delta_i | V_i] = 0$, which implies that the expectation of U is 0. It then follows from standard projection arguments that the U-statistic vanishes in probability.