

# Recursive Differencing for Estimating Semiparametric Models\*

Chan Shen  
Pennsylvania State University

Roger Klein  
Rutgers University

September, 2019

## Abstract

Controlling the bias is central to estimating semiparametric models. Many methods have been developed to control bias in estimating conditional expectations while maintaining a desirable variance order. However, these methods typically do not perform well at moderate sample sizes. Moreover, and perhaps related to their performance, non-optimal windows are selected with undersmoothing needed to ensure the appropriate bias order. In this paper, we propose a recursive differencing estimator for conditional expectations. When this method is combined with a bias control targeting the derivative of the semiparametric expectation, we are able to obtain asymptotic normality under optimal windows.

As suggested by the structure of the recursion, in a wide variety of triple index designs, the proposed bias control performs much better at moderate sample sizes than regular or higher order kernels and local polynomials.

**Keywords:** semiparametric model, bias reduction, conditional expectation.

## 1 Introduction

In this paper, our primary emphasis is on semiparametric index models, which perform well at moderate sample sizes. Often, such models require estimating an expectation conditioned on a vector of indices, where each index is a parametric function of observables and an unknown finite dimensional parameter vector. We will term such an expectation as semiparametric due to the index structure of the conditioning variables. For models with an index structure, see for example, Robinson (1988), Powell et. al. (1989), Ichimura (1993), Ichimura and Lee (1991), Klein and Spady (1993), Horowitz (1996), Li et. al. (2014), and Klein, Shen and Vella (2015). The first objective of this paper is to develop a recursive estimator for a semiparametric expectation that can deliver a bias of any order while maintaining desirable variance properties and finite sample performance. Second, we combine the properties of this recursive estimator with a residual property of semiparametric derivatives to obtain asymptotic normality under optimal windows.

---

\*We thank the seminar participants at Columbia University and New York University for helpful comments and suggestions.

To obtain asymptotic normality at a  $\sqrt{N}$  rate for a finite dimensional parameter vector in a semiparametric model, the bias in the estimator must vanish faster than  $N^{-1/2}$  while the variance must converge to zero at a sufficiently fast rate. Methods have been developed in the literature to control for the bias while maintaining a desirable variance order. In some cases, an estimate of the bias in the parameter estimator can be removed from the estimator as in Honore and Powell (2005). In other cases, it is possible to employ different estimators for conditional expectations. Higher order kernels (e.g. Muller (1984)) achieve this bias order by increasing the degree of the kernel. An extension of Newey (2004) can similarly control the bias by increasing the convolution degree. For estimating semiparametric models, both approaches require suboptimal windows and hence leads to much larger variability. Local polynomials (e.g. Fan and Gijbels (1995, 1996), Ruppert and Wand (1994), Lu (1994), Masry (1995) and Gu, Li, and Yang (2015)) obtain this bias order by increasing the degree of the local polynomial. While the performance of the local linear estimator is quite good, higher degrees require introducing more local parameters and hence also leads to higher variability. In this paper, we propose alternative approaches for bias reduction that enables us to obtain  $\sqrt{N}$  normality in semiparametric index models. In finite samples, we find much lower variability for the recursive differencing estimator in simulation studies.

The estimator proposed here has a recursive differencing structure with a local linear estimator providing the basis for the first stage of the recursion. The bias in the first stage estimator depends on a localization error defined as the difference between the expectation at a point of interest and a nearby point. Accordingly, in the second stage of the recursion, we remove an estimator of this localization error from the previous stage. Continuing in this manner, we show that the bias declines at each stage of the recursion, with the variance order being unchanged.

Taking advantage of this recursive mechanism and a residual property of semiparametric derivatives, we propose an approach to estimate semiparametric multiple index models of arbitrary dimension under optimal windows. In a Monte Carlo study, we considered four semiparametric triple index models. Employing a three stage recursion, which is appropriate in this context (Theorem 2), we found that the resulting estimator had very good finite sample performance in terms of both bias and variance. In all of the cases the RMSE decreased, often substantially, with the stage of the recursion. The resulting estimator also performed much better than either a higher order kernel estimator or a local polynomial whose bias order is below  $N^{-1/2}$ .

To develop the proposed parameter estimator for a semiparametric model and the underlying conditional expectation estimator, Section 2 provides the intuition for these two estimators and their theoretical properties. Section 3 formally defines the estimators and obtains their large sample properties. Section 4 provides Monte Carlo results that demonstrate very good finite sample properties of the estimators in triple index models, exhibiting a substantial improvement over regular, higher order kernels, and local polynomial estimators. We note that the local linear estimator forms the basis for the first stage of the recursion and does significantly improve its performance. Section 5 contains our conclusions. The appendix contains proofs of all theorems and supporting intermediate lemmas.

## 2 Estimators

### 2.1 Estimating Expectations under Recursive Differencing

The semiparametric model that we study assumes:

$$E(Y_i|W_i) = E(Y_i|V(W_i; \theta_0)) \equiv M(V_i), \quad (1)$$

where the vector  $\{Y_i, W_i\}$  is i.i.d. over  $i = 1, \dots, N$ , and takes on values in  $\mathfrak{X}_Y \times \mathfrak{X}_W \subset \mathfrak{R}^{1+d_w}$  with  $d_w$  the dimension of  $W_i$ . Here,  $V_i \equiv V(W_i; \theta_0)$  is a vector of  $d < d_w$  continuous indices that depend on a finite dimensional parameter vector,  $\theta_0$ . For expositional purposes, in this section we take the parameter vector as known and discuss its estimation in the next section.

To motivate the form of the bias reduction, we defer discussions of trimming considerations to the next section. Denote  $M(V_i) \equiv E(Y_i|W_i)$  with  $V_i \equiv V(W_i; \theta_0)$  and consider the model in localized form:

$$Y_i = M(v) + [M(V_i) - M(v)] + \varepsilon_i,$$

where  $\varepsilon_i$  is an error satisfying  $E(\varepsilon_i|W_i) = 0$ . With  $K_i(v)$  as a kernel weight that controls the localization error  $[M(V_i) - M(v)]$  by downweighting observations  $V_i$  not close to  $v$ , an often employed conditional expectation estimator is given as:

$$\hat{M}(v) \equiv \frac{\frac{1}{N} \sum_i Y_i K_i(v)}{\hat{g}(v)}, \quad \hat{g}(v) \equiv \frac{1}{N} \sum_i K_i(v).$$

Substituting the local model for  $Y_i$ :

$$\hat{M}(v) - M(v) = \frac{\frac{1}{N} \sum_i [M(V_i) - M(v) + \varepsilon_i] K_i(v)}{\hat{g}(v)}.$$

As the localization error,  $M(V_i) - M(v)$ , is responsible for the bias, we propose a recursive mechanism that removes the estimated localization error. For  $s > 1$ , define the stage  $s$  estimator as:

$$\hat{M}_s(v) \equiv \frac{\frac{1}{N} \sum_i [Y_i - (\hat{M}_{s-1}(V_i) - \hat{M}_{s-1}(v))] K_i(v)}{\hat{g}(v)} \quad (2)$$

where  $\hat{M}_{s-1}(V_i) - \hat{M}_{s-1}(v)$  is the estimated localization error from the previous stage. It then follows that:

$$\begin{aligned} & \hat{g}(v) [\hat{M}_s(v) - M(v)] \\ &= \frac{1}{N} \sum_i \left\{ [\hat{M}_{s-1}(v) - M(v)] - [\hat{M}_{s-1}(V_i) - M(V_i)] + \varepsilon_i \right\} K_i(v) \end{aligned}$$

Note that the difference  $\hat{M}_s(v) - M(v)$  has been scaled by a density estimator, which is convenient for obtaining asymptotic results. With the kernel ensuring that  $V_i$  is close to  $v$ , intuitively this differencing structure reduces the estimation error because the error at  $V_i$  will be close to that at  $v$ . In the appendix, we formally prove that there is a recursion uniformly within  $o_p(N^{-1/2})$  of that above for which the bias order diminishes with each stage of the

recursion. The variance order remains the same. To provide some intuition for this result, consider stage  $s = 2$  and write:

$$\hat{g}(v) \left[ \hat{M}_2(v) - M(v) \right] = \hat{g}(v) \left[ \hat{M}_1(v) - M(v) \right] - \frac{1}{N} \sum_i \left\{ \frac{\hat{g}(V_i) \left[ \hat{M}_1(V_i) - M(V_i) \right]}{\hat{g}(V_i)} + \varepsilon_i \right\} K_i(v)$$

To deal with estimated density denominators, in the appendix we provide a recursion within  $o_p(N^{-1/2})$  of the above and having the form:

$$\hat{g}(v) \left[ \hat{M}_2^*(v) - M(v) \right] = \hat{g}(v) \left[ \hat{M}_1(v) - M(v) \right] - \frac{1}{N} \sum_i \left\{ \frac{\hat{g}(V_i) \left[ \hat{M}_1(V_i) - M(V_i) \right]}{g(V_i)} [1 + \delta(V_i)] + \varepsilon_i \right\} K_i(v)$$

where  $\delta(V_i)$  contains the higher order terms from an approximating expansion. It can be shown that for the leading term in this expansion:

$$\begin{aligned} & E \left[ \hat{g}(v) \left[ \hat{M}_1(v) - M(v) \right] - \frac{1}{N} \sum_i E \left\{ E \left[ \frac{\hat{g}(V_i) \left[ \hat{M}_1(V_i) - M(V_i) \right]}{g(V_i)} + \varepsilon_i | V_i \right] K_i(v) \right\} \right] \\ &= h^2 B_2(v) + h^4 B_4(v) + o(h^4) - \frac{1}{N} \sum_i E \left\{ \frac{h^2 B_2(V_i) + h^4 B_4(V_i) + o(h^4)}{g(V_i)} K_i(v) \right\} \\ &= h^2 B_2(v) + h^4 B_4(v) + o(h^4) - \int [h^2 B_2(V_i) + h^4 B_4(V_i) + o(h^4)] K_i(v), \end{aligned}$$

where the  $B$ -functions are uniformly bounded. In the case of local constant estimators, these expansions are well known in the literature, but as shown below also hold in other cases. With the kernel ensuring that  $V_i$  is close to  $v$  the above differencing structure results in a canceling of bias terms resulting in the bias order decreasing from  $h^2$  to  $h^4$ .<sup>1</sup> We show that higher order terms in  $\delta(V_i)$  do not have a larger bias order. As shown in the appendix, higher stages have a differencing in differencing structure which further reduces the bias.

To start this recursion, the theory developed here will hold for any initial estimator that satisfies certain convergence properties and has an approximating bias expansion of the above form. These conditions are satisfied by both the local constant estimator and the local linear estimator. Here, we employ a modified local linear estimator since it performs better than the local constant estimator.<sup>2</sup>

With  $V_i$  as a row vector of the model's  $d$  continuous indices and  $v$  a conformable row vector of fixed values, let  $Z_i \equiv (V_i - v)/h$ , where  $h$  is a kernel window parameter. The local linear estimator solves:

$$\hat{M}_L, \hat{M}'_L \equiv \arg \min_{M, M'} \sum_i [Y_i - M(v) - h Z_i M'(v)]^2 K_i(v) \quad (3)$$

$$\Rightarrow \hat{M}_L = \bar{Y}(v) - h \bar{Z}(v) \hat{M}'_L, \quad (4)$$

<sup>1</sup>For a symmetric kernel:

$$\int [h^2 B_2(V_i)] K_i(v) = h^2 B_2(v) + O(h^4).$$

<sup>2</sup>We modified the local linear estimator as it simplifies the bias arguments without sacrificing finite-sample performance.

where  $\bar{Y}(v)$  and  $\bar{Z}(v)$  are kernel weighted averages:

$$\bar{Y}(v) \equiv \sum_i Y_i K_i(v) / \sum_i K_i(v); \quad \bar{Z}(v) \equiv \sum_i Z_i K_i(v) / \sum_i K_i(v)$$

and  $\hat{M}'_L$  is a local linear estimator of the derivative of  $M(v)$ . To simplify arguments, define a modified derivative estimator as

$$\begin{aligned} \hat{M}'_m &\equiv \arg \min_{M'} \sum_i [Y_i - \bar{Y}(v) - hZ_i M'(v)]^2 K_i(v) \\ &= \frac{1}{h} [Z' D(v) Z]^{-1} [Z' D(v)] [Y - \tilde{Y}(v)], \end{aligned}$$

where  $\tilde{Y}(v)$  is a vector with each element equal to the local constant estimator,  $\bar{Y}(v)$ , and  $D(v) \equiv \text{diag}(K_i(v))$ . The modified local linear estimator is then defined as:

$$\hat{M}_1(v) \equiv \bar{Y}(v) - h\bar{Z}(v)\hat{M}'_m = \bar{Y} - \bar{Z}[Z' D(v) Z]^{-1} [Z' D(v)] [Y - \tilde{Y}(v)]. \quad (5)$$

In analyzing this estimator, we employ a regular kernel function, with the bias declining at each stage of the recursion. If other bias reducing kernels are employed at each stage, then the bias will be even smaller at each stage and continue to decline over the stages. The proofs for these alternative kernels are identical to those employed here. Using a modified local linear estimator as the start of the recursion, we have found that the recursion based on regular kernels performs very well.

## 2.2 Estimating Index Parameters in Semiparametric Models

Many semiparametric models depend on expectations conditioned on a vector of indices. Here, we employ recursive differencing to estimate conditional expectations. In addition to the recursive differencing structure, we also propose an extra mechanism that further reduces the bias in estimating index parameters in semiparametric models. Combining these mechanisms, we will be able to estimate a wide class of multiple index semiparametric models using optimal windows. We refer to this additional control as the residual property of semiparametric derivatives, which is given in the following proposition due to Whitney Newey.<sup>3</sup>

**Proposition 1** *Assume  $E(Y|W) = M[V(\theta_0)]$  from the index assumption in (1) and let  $M[V(\theta); \theta] \equiv E[Y|V(\theta)]$ .<sup>4</sup> Then,*

$$E \left\{ \frac{\partial M[V(\theta); \theta]}{\partial \theta} \Big|_{V(\theta_0)} \right\}_{\theta=\theta_0} = 0.$$

Exploiting this result as a bias control for single index models, Klein and Shen (2010) show that asymptotic normality can be obtained with regular kernels. This approach does not extend to higher dimensions under regular kernels, but does extend under recursive differencing. Before providing the theorem that covers this extension, here we briefly discuss

<sup>3</sup>See Klein and Shen (2010) for Newey's proof of this property.

<sup>4</sup>Notation for the dual dependence on  $\theta$  is employed because the conditional expectation function itself will generally be different when the conditioning variable changes via changes in  $\theta$ .

the estimation strategy. To exploit Newey’s residual result, it can be shown that trimming must be based on the index. Accordingly, we employ a two-step estimation strategy where the index parameters are estimated in the first step with trimming based on  $X$ , the continuous variables in  $W$ . In the second step, trimming is based on the estimated index from the first step. Throughout, we use the term step to distinguish the initial estimator based on  $X$ -trimming and the estimator based on index trimming; while we use the term stage to refer to the recursive differencing sequence. Under recursive differencing, it is possible to show that the semiparametric derivative component of the estimation gradient can be taken as known for any index dimension. We are then able to establish asymptotic normality under optimal windows. It should be noted that this argument requires an adjustment strategy, apart from trimming, to control density denominators in the second step. We defer discussion of this issue to the next section.

Theorem 1 in the next section provides the properties of the expectation estimator. For a class of semiparametric models, Theorem 2 then combines recursive differencing with a residual control introduced below so as to provide conditions on the number of stages and the kernel window to obtain  $\sqrt{N}$ -asymptotic normality under optimal windows.

### 3 Large Sample Results

To establish large sample results, we require the following definitions and notations.

#### 3.1 Definitions and Notations

**D1)** Index Functions. Let  $W_i$  be an *i.i.d.* vector of continuous and discrete variables,  $i = 1, \dots, N$ . Let  $\theta$  be a finite dimensional parameter vector and  $V_i = V(W_i; \theta)$  a vector of  $d$  continuous parametric index functions.

**D2)** Conditional Expectations.

$$M(v) \equiv E [Y_i | V(W_i; \theta) = v].$$

**D3)** Trimming. Denote  $\lambda_1$  and  $\lambda_2$  as lower and upper percentiles and let  $\lambda \equiv [\lambda_1, \lambda_2]$ . Define  $q_x(\lambda) \equiv [q_x(\lambda_1), q_x(\lambda_2)] : d_c \times 2$  as the corresponding matrix of lower and upper population quantiles for  $X_i$ , where  $X_i : d_c \times 1$  is the continuous subvector of  $W_i$ . Similarly, define  $q_v(\lambda) \equiv [q_v(\lambda_1), q_v(\lambda_2)] : d \times 2$  as a matrix of population quantiles for  $V_i : d \times 1$ . Then, define exterior trimming based on  $X_i$  or  $V_i$  as:

$$\begin{aligned} \tau_x(X_i) &\equiv 1 \{X_i \in \mathcal{C}_x(q_x)\}, \mathcal{C}_x(q_x) \equiv \{x : q_x(\lambda_1) < x < q_x(\lambda_2)\}; \\ \tau_v(V_i) &\equiv 1 \{V_i \in \mathcal{C}_v(q_v)\}, \mathcal{C}_v(q_v) \equiv \{v : q_v(\lambda_1) < v < q_v(\lambda_2)\}. \end{aligned}$$

With  $0 < \lambda_1 < \lambda_{1I} < \lambda_{2I} < \lambda_2 < 1$ , define corresponding interior trimming functions as:

$$\begin{aligned} \tau_{xI}(X_i) &\equiv 1 \{X_i \in \mathcal{C}_x(q_{xI})\}, \mathcal{C}_x(q_{xI}) \equiv \{x : q_x(\lambda_{1I}) < x < q_x(\lambda_{2I})\}; \\ \tau_{vI}(V_i) &\equiv 1 \{V_i \in \mathcal{C}_v(q_{vI})\}, \mathcal{C}_v(q_{vI}) \equiv \{v : q_v(\lambda_{1I}) < v < q_v(\lambda_{2I})\}. \end{aligned}$$

Let  $\hat{\tau}$  be the sample quantile version of  $\tau$ . For index trimming, the estimated index is constrained to being between lower and upper sample quantiles for the estimated index vector.

**D4)** Kernel. Let  $v$  and  $V_i$  be  $d$ -dimensional vectors with  $l^{th}$  elements as  $v_i(l)$  and  $V_i(l)$  respectively and denote  $s_l$  as the standard deviation of  $V_i(l)$ .<sup>5</sup> Define:

$$K_{il}(v) \equiv \frac{1}{s_l h} \phi \left( \frac{v_i(l) - V_i(l)}{s_l h} \right), K_i(v) \equiv \prod_{l=1}^d K_{il}(v), D_i(v) \equiv \text{diag}(K_i(v)),$$

$$k(z) \equiv \prod_{l=1}^d \phi(z(l)), K_i^*(v; \hat{\tau}) \equiv \hat{\tau}_i K_i(v),$$

where  $h = O(N^{-r})$ , and  $\phi(z(l))$  is a density symmetric about 0 with finite moments of all orders. The function  $\hat{\tau}_i$  is a trimming function that may be either  $\hat{\tau}_v(V_i)$  or  $\hat{\tau}_x(X_i)$ .

**D5)** Kernel Averages. Referring to D4), define:

$$\hat{g}_s(v) \equiv \begin{cases} \hat{g}_1(v) \equiv \frac{1}{N} \sum_{i=1}^N K_i(v), & s = 1 \\ \hat{g}(v) \equiv \frac{1}{N} \sum_{i=1}^N K_i^*(v; \hat{\tau}), & s > 1 \end{cases}.$$

When the estimators are evaluated at a data point, the average is taken over the  $N - 1$  observations excluding that data point observation.

**D6)** Conditional Expectation Estimator. Let  $\hat{D}^*$  and  $D^*$  be  $N \times N$  diagonal matrices with  $i^{th}$  element  $K_i^*(v; \hat{\tau})$  and  $K_i^*(v; \tau)$  respectively. Let  $Z$  be an  $N \times d$  matrix with  $i^{th}$  row  $(V_i - v)/h$ . Referring to (5), with  $\bar{Y}(v)$  and  $\bar{Z}(v)$  depending on  $\hat{g}_1(v)$ , for stage  $s = 1$ :

$$\hat{M}_1(v) \equiv \bar{Y}(v) - \bar{Z}(v) \left[ Z' \hat{D}^* Z \right]^{-1} Z' \hat{D}^* \left[ Y - \tilde{Y}(v) \right], \text{ where } E \left[ \frac{Z' D^* Z}{N} \right] \text{ is positive definite.}$$

For stage  $s > 1$ :

$$\hat{M}_s(v) \equiv \frac{\frac{1}{N} \sum_{i=1}^N \left\{ Y_i - \left[ \hat{M}_{s-1}(V_i) - \hat{M}_{s-1}(v) \right] \right\} \hat{\tau}_i K_i(v)}{\hat{g}(v)}.$$

**D7)** Adjusted Densities. With  $\hat{\gamma}_s$  as lower quantiles for the  $\hat{g}_s$  in D5), define adjustment factors:

$$A_s(v) = h^a \hat{\gamma}_s [1 - \hat{\tau}_\Delta(v)],$$

where  $h$  is the window and  $0 < a < 1$ . To smoothly restrict  $v$  to be between lower quantile  $q_v(\lambda_1)$  and upper quantile  $q_v(\lambda_2)$ , define

$$\tau_\Delta(v) = \left\{ 1 + \exp \left( Ln(N)^2 [v - q_v(\lambda_1)] \right) \right\}^{-1} \left\{ 1 + \exp \left( Ln(N)^2 [q_v(\lambda_2) - v] \right) \right\}^{-1}.$$

Let  $\hat{\tau}_\Delta(v)$  be the sample quantile version of  $\tau_\Delta(v)$ . Adjusted densities are then defined as:

$$\hat{g}_{sa}(v) \equiv \hat{g}_s(v) + A_s(v).$$

**D8)** Adjusted Recursive Differencing. Referring to D6) and D7), obtain the adjusted estimator  $\hat{M}_{sa}(v)$  from  $\hat{M}_s(v)$  by replacing  $\hat{g}(v)$  with  $\hat{g}_{sa}(v)$ .

---

<sup>5</sup>It can readily be shown that the standard deviation can be taken as known.

While the purpose of most of these definitions is clear, further discussion of the trimming and adjustment strategies can be useful. The estimators above depend on different types of trimming. Because of the recursive structure of the estimator in D6), the estimator at stage  $s$  will depend on a vector of estimated conditional mean functions from the previous stage. Trimming (exterior) is required to control this vector. To avoid boundary bias, this exterior trimming set must strictly contain the set over which the  $\hat{M}(v)$  functions are defined. We refer to the set of such  $v$ -values as an interior set.

The adjustment factor in D7) is employed to control density denominators in the second step for estimating a class of semiparametric models. Recall from the previous section that we employ a two-step estimator for index parameters in semiparametric models. In the first step, the estimated indices are recovered with trimming based on  $X$ . The estimated index from this step is then used for trimming in the second step. This procedure with expectations estimated under recursive differencing ensures that the gradient to the estimation problem has asymptotic bias of order  $o(N^{-1/2})$  under optimal windows. However in uniform convergence arguments underlying the consistency argument, density estimators must be controlled for  $\theta$  away from the truth  $\theta_0$ . The adjustment factor is employed for this purpose. The  $\tau_\Delta(v)$  function exponentially approaches 0 for  $v$  in the interior of its support. Near the support boundary where the density is allowed to approach 0 at an unrestricted rate, the adjustment factors vanish slowly at a rate of  $h^a$ ,  $0 < a < 1$ . In this manner, the rate at which the density denominator converges to zero is controlled.

In Theorem 1 below, we obtain convergence properties for the proposed expectation estimator. In Theorem 2 we prove asymptotic normality for parameter estimators in a class of semiparametric models under recursive differencing and the residual control. To this end, we make the following assumptions.

### 3.2 Assumptions

**A1)** The vector  $\{Y_i, W_i\}$  is *i.i.d.* over  $i = 1, \dots, N$ , and takes on values in  $\mathfrak{X}_Y \times \mathfrak{X}_W \subset \mathfrak{R}^{1+d_w}$ , where  $d_w$  is the dimension of  $W_i$ .

**A2)** The following index assumption holds:

$$E(Y_i|W_i) = E(Y_i|V(W_i; \theta_0)).$$

**A3)** Refer to D1) and assume that  $\theta \in \Phi$ , a compact set. Let  $g(v(w; \theta); \theta)$  be the density for  $V(W_i; \theta)$  evaluated at  $v(w; \theta)$  and let  $f(x|V_i = v)$  be the density for  $X$  conditioned on  $V_i = v$ . Let  $\alpha = 0, 1, 2$  and  $\beta = 0, 1, \dots, 2s + 1$ , where  $s$  is the stage. For  $v, w$ , and  $\theta$  in compact sets and with  $c > 0$ :

$$\begin{aligned} a) & : \inf_{v, \theta} |g(v; \theta)| > c \\ b) & : \sup_v |\nabla_v^\beta M(v; \theta_0)g(v; \theta_0)| \text{ and } \sup_v |\nabla_v^\beta g(v; \theta_0)| = O(1) \\ c) & : \sup_{w, \theta} |\nabla_\theta^\alpha \nabla_v^\beta M(v(w; \theta); \theta)| \text{ and } \sup_{w, \theta} |\nabla_\theta^\alpha \nabla_v^\beta g(v(w; \theta); \theta)| = O(1). \end{aligned}$$

**A4)** There exists  $m > 4$  such that  $E[|Y_i|^m] = O(1)$ .

**A5)** For the window parameter:  $r < \frac{1}{2d}$ .



**A6)** The vector of continuous random variables,  $X_i$ , is bounded.

The first three assumptions are standard for index models. Assumption A4) is useful for obtaining uniform convergence results for functions of unbounded random variables. Assumption A5) is needed to establish the bias properties of the proposed recursive differencing estimator (Theorem 1). Condition A6) is not required, but greatly simplifies the exposition. It should be noted that many and perhaps most models in applied work satisfy this condition. Moreover, when it does not hold, it is always possible to trim out a very small fraction of the observations to guarantee that A6) holds.

Convergence properties are important for obtaining  $\sqrt{N}$ -normality for a finite dimensional parameter vector in semiparametric models. Theorem 1 provides these properties for the conditional expectation estimator. It holds in the semiparametric case when conditioning on a vector of indices and in the nonparametric case.

**Theorem 1. The Recursive Differencing Estimator.** Assume A1)-A5), then for  $v$  and  $\theta$  in compact sets, with  $\Delta_s(v) \equiv \hat{g}_s(v) \left( \hat{M}_s(v) - M(v) \right)$ , there exists a  $\Delta_s^*(v)$  with its form provided in Lemma 2, that satisfies the following:

$$\begin{aligned}
 a) & : \sup_v |\Delta_s^*(v) - \Delta_s(v)| = o_p(N^{-1/2}) \\
 b) & : \sup_v |E\Delta_s^*(v)| = O(h^{2s}) + o(N^{-1/2}) \\
 c) & : \sup_v Var[\Delta_s^*(v)] = O(N^{-(1-rd)}) \\
 d) & : \left| \hat{M}_s(v) - M(v) \right| = O_p(h^{2s}) + O_p(N^{-(1-rd)/2}).
 \end{aligned}$$

For estimating semiparametric models, it can be readily shown that it is the bias in the scaled estimator  $\Delta_s(v)$  that is relevant. As this quantity is nonlinear, it is difficult to study directly. Lemmas 1-2 provide an approximating sequence  $\Delta_s^*(v)$  that from a) is uniformly within  $o_p(N^{-1/2})$  of  $\Delta_s(v)$ . The bias result is then given in b), followed by a variance result in c) and a convergence rate in d).

To obtain  $\sqrt{N}$ -normality for a finite dimensional parameter vector, we will require conditions on the stage  $s$  and the window parameter  $r$ . Theorem 2 provides these conditions for a class of multiple index models estimated by Semiparametric-Least-Squares (SLS) introduced by Ichimura (1992) and Ichimura and Lee (1991). The results of Theorem 2 readily extend to Quasi-Maximum-Likelihood estimators for semiparametric binary response and to ordered models. Though, such extensions require a more complicated trimming strategy than that employed here.

In the single index case, Klein and Shen (2010) exploit Newey's residual result to obtain  $\sqrt{N}$ -normality. Theorem 2 obtains this result for the multiple index case by combining recursive differencing and the residual bias controls. In so doing, it is possible to employ optimal windows.

**Theorem 2. Estimating Index Parameters with Recursive Differencing and Residual Bias Controls.** Set the stage  $s$  and window parameter  $r$  to satisfy:<sup>6</sup>

$$C1) : s > \frac{(d+2)(m+2)}{2m}, \quad \frac{1}{4s} < r < \frac{m}{2(d+2)(m+2)}$$

With all trimming based on sample quantiles for  $X$  as described in D3), define the first step estimator as:<sup>7</sup>

$$\hat{\theta} = \arg \min_{\theta} \sum_i \hat{\tau}_{vI}(X_i) \left\{ Y_i - \hat{M}_s[V_i(\theta)] \right\}^2. \quad (6)$$

For the second step estimator, set the stage  $s$ , and window parameter  $r$ , to satisfy:<sup>8</sup>

$$C2) : s \geq \frac{d+4}{4} + \frac{d+2}{m}, \quad r = r^* = \frac{1}{4s+d},$$

where  $r^*$  is the optimal window. Referring to D3) for the definitions of index trimming functions, base all trimming functions on sample quantile regions for the estimated index from the first stage. Referring to D7) -D8) and letting  $\hat{V}_i \equiv V_i(W_i; \hat{\theta})$  define the adjusted second stage estimator as:

$$\hat{\theta}_a \equiv \arg \min_{\theta} \left\{ \sum_i \hat{\tau}_{vI}(\hat{V}_i) \left\{ Y_i - \hat{M}_{sa}[V_i(\theta)] \right\}^2 \right\} \quad (7)$$

Then, under A1)-A5):

$$\begin{aligned} a) & : \left| \hat{\theta} - \theta_0 \right| = O_p(N^{-1/2}) \\ b) & : \left| \hat{\theta}_a - \theta_0 \right| = o_p(1) \\ c) & : \sqrt{N} \left( \hat{\theta}_a - \theta_0 \right) \xrightarrow{d} Z^* \sim N \left( 0, \sigma_{\varepsilon}^2 E \left[ G_i^*(\theta_0) G_i^*(\theta_0)' \right]^{-1} \right), \end{aligned}$$

where  $\sigma_{\varepsilon}^2$  is the variance of  $\varepsilon_i$  and  $G_i^*(\theta_0) \equiv \tau_{vI}(V_{0i}) \nabla_{\theta} [M(V(W_i; \theta); \theta)]_{\theta_0}$ .

It should be noted that recursive differencing makes it possible to take the estimated semiparametric derivative as known in the multiple index case. The gradient evaluated with a known semiparametric derivative and index trimming function has no bias from the residual property of the semiparametric derivative.

## 4 Monte Carlo Results

We conducted Monte Carlo experiments using four different designs: quadratic, cubic, exponential, and sin. In all designs, we constructed three indices:  $V_1 = X_1 + X_4$ ,  $V_2 = X_2 - X_4$ ,  $V_3 =$

<sup>6</sup>The upper bound for  $r$  follows from uniform convergence for second derivatives (Lemma 9). The lower bound for  $r$  follows because the bias order of  $O(N^{-2rs})$  must be  $o(N^{-1/2})$ . The condition on  $s$  is required for the interval on  $r$  to be non-empty.

<sup>7</sup>This first step estimator provides an estimated index that can serve as the basis for trimming at the second step.

<sup>8</sup>For uniform convergence of second derivatives to hold at the optimal window, we require  $r^* = \frac{1}{4s+d} < \frac{m}{2(d+2)(m+2)}$ . This condition will hold if  $s$  is set as in C2).

$X_3 + X_4$  where  $X_1, X_2, X_3$  and  $X_4$  follow standard normal distributions. We also generated an error term  $\varepsilon$  that follows a standard normal distribution. In all designs, the outcome has the form:

$$Y = \alpha V_1 + \beta T_2 + \gamma V_1 V_3 + \varepsilon,$$

where  $T_2$  was set respectively to be a quadratic, cubic, exponential, or sin function of  $V_2$  in the four designs. The  $\alpha, \beta, \gamma$  are standardizing constants selected so that in all designs, each of the three explanatory components has an approximate standard deviation (SD) of one.

As the focus of this paper is on the semiparametric case, we begin by reporting results for different estimators of the parameters in the four designs discussed above (Table 1). We will also compare results for different estimators of the conditional mean function (Table 2).

Five measures of estimator performance are provided in Table 1 including root mean squared error (RMSE), mean, standard deviation (SD), median, and median absolute deviation (MAD). The first three columns of results in Table 1 focus on the proposed recursive differencing estimator with a fixed window. We set the window parameter at  $r = \frac{1}{15}$  throughout these three stages so as to facilitate the comparison between stages.<sup>9</sup> The next three columns show the results for the proposed recursive differencing estimator with optimal windows over the stages. Stage 1 had a window size  $r = \frac{1}{7}$ ; stage 2 had a window size  $r = \frac{1}{11}$ ; while stage 3 had a window size  $r = \frac{1}{15}$ . The third stage estimator also has the residual control to ensure  $\sqrt{N}$  normality. The last three columns of Table 1 include three estimators that have a  $\sqrt{N}$ -asymptotic normality property due to undersmoothing without the residual control. The first one of them (RDS) employs recursive differencing as the sole bias control with undersmoothing window parameter  $r = \frac{1}{11.99}$  to ensure  $\sqrt{N}$  asymptotic normality. The next column provides results for a higher order kernel estimator (HK), which is an extension of the twicing kernel (Newey et. al, 2004). The last column is for a fifth degree local polynomial estimator (LP) with bias  $O(h^6)$  as in Ruppert and Wand (1994). For these last two estimators, we also set  $r = \frac{1}{11.99}$  to ensure that the bias is  $o(N^{-1/2})$ .

The first three columns of results show that with fixed window, the proposed recursive differencing estimator has decreasing bias over the stages while the standard deviation remains stable. This finding is consistent with the theory behind the recursive differencing mechanism. In all four designs, there was a substantial decrease in RMSE over the stages. This decline is most pronounced for the sin design. For example, the RMSE of the first parameter estimate declines from .368 at the first stage to .219 at the second stage and finally to .127 at the third stage. The next three columns show the results when we set the optimal window for each stage of the recursive differencing estimator. As expected, we found that the RMSE monotonically declined over the stages in most cases; in all cases, the third stage achieved the smallest RMSE. The small third stage RMSE is due both to recursive differencing and the residual control. The value added by using the residual control can be seen by comparing recursive differencing estimators with optimal windows with (column6) and without (column3) the residual control. Across all four designs, RMSE decreased by 20%-35% when the additional residual bias control is employed.

The last three columns of Table 1 provide a comparison between three  $\sqrt{N}$ -asymptotically normal estimators. The recursive differencing estimator with undersmoothing (RDS) strongly dominates the other two estimators (HK and LP) with substantially smaller RMSE. Its advantage is most pronounced in terms of variation (i.e. standard deviation and MAD). Similarly, the proposed recursive differencing estimator with residual control also strongly

---

<sup>9</sup>We chose  $r = 1/15$  as it is the optimal window for the third stage estimator.

dominates the HK and LP estimators. The performance of the proposed recursive differencing estimator with residual control (column6) is overall better than the undersmoothing version (RDS), with smaller RMSE in 8 out of 12 parameters from 4 designs, and the difference in RMSE is relatively small in the 4 cases where it did not dominate the undersmoothing version. In summary, recursive differencing estimators dominate the other estimators and there is value added to employing the residual bias control with recursive differencing.

In addition to comparing parameter estimators, we also compare results for estimating the conditional mean functions. The estimation of these functions plays a fundamental role in estimating parameters. Further, it is essential in estimation of marginal effects, which is an important object of interest in empirical studies. Therefore, we provide results on the estimation of conditional mean functions in Table 2. We investigated the performance of three conditional mean estimators underlying the parameter estimators we studied above: the recursive differencing estimator, the higher order kernel estimator, and the local polynomial estimator. To avoid confounding performance of the parameter estimators with those for conditional mean functions, all estimators for the conditional mean functions are reported at the true parameter values. Since marginal effects are often evaluated at different regions of the distribution, we calculated the conditional mean function estimators at every point in a trimmed set<sup>10</sup> and then averaged over decile intervals so as to obtain  $\sqrt{N}$ -asymptotic normality. In so doing, a window size of  $r = \frac{1}{11.99}$  was set for all three estimators.

In terms of RMSE, we found that the recursive differencing estimator was superior to the higher order kernel and local polynomial estimators across designs and deciles. The advantage was especially pronounced at the higher and lower deciles; and was relatively moderate at the middle deciles. To further investigate the source of such advantages, we examined the bias and standard deviations of the estimators. In terms of bias, we found that both the recursive differencing and the higher order kernel estimators perform better than the local polynomial estimator at the middle deciles. However, the recursive differencing estimator dominates both the higher order kernel estimator and the local polynomial estimator considerably at the higher and lower deciles.

The standard deviation of the recursive differencing estimator was smaller than that of the higher order kernel and local polynomial estimators across designs and deciles, with substantial advantage at the higher and lower deciles. Indeed, all three estimators had worse performance at extreme deciles. We remark that the performance of the local polynomial estimator near the boundary improves significantly when the sample size increases. We experimented with increasing the sample size to 10,000 for the cubic design. In that case, the RMSE of the local polynomial estimator for the first decile reduced to 0.275; bias reduced to -0.222; and standard deviation reduced to 0.162. However, the recursive differencing estimator continue to dominate it with first decile RMSE of 0.126, bias of 0.115 and standard deviation of 0.053. Results are similar for the 10th decile. There is a large and growing literature on boundary correction methods in kernel density estimation that could improve the performance of all three estimators at the boundaries. However, such a boundary correction is outside of the scope of this paper.

In summary, the Monte Carlo experiment showed that the proposed recursive differencing estimator performs much better under a moderate sample size than the other methods that were considered. Further, the behavior of the recursive differencing estimator is consistent with the theory underlying this estimator.

---

<sup>10</sup>Each index was trimmed at 3% and 2% from each tail based on interior and exterior trimming functions respectively.

## 5 Conclusions

In this paper, we propose recursive differencing estimators for estimating conditional expectations in semiparametric models with multiple indices. The order of the bias decreases with the stage of the recursion while the order of the variance remains the same.

While higher order kernels and local polynomials share the above properties, they differ from the proposed estimator in two important respects. First, the RMSE of the recursive differencing estimator became smaller over the stages. In contrast, higher order kernel or local polynomial estimators would require higher order terms to achieve the same bias order and often lead to higher RMSE. Second, in estimating index models, we show that with recursive differencing it is possible to exploit a residual property of semiparametric derivatives. In so doing, we obtain asymptotic normality without undersmoothing, regardless of the dimension of the index vector. This theoretical property contributes to the very good finite sample performance of the proposed estimator.

Table 1. Monte Carlo Results for Parameter Estimates

	Proposed Estimator with Fixed Window			Proposed Estimator with Optimal Window			Asymptotic Normality with Under Smoothing		
	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3	RDS	HK	LP
<b>Quadratic Design</b>									
RMSE	0.089	0.072	0.068	0.073	0.068	0.051	0.068	5.392	1.087
	0.112	0.101	0.088	0.099	0.088	0.070	0.086	5.621	1.323
	0.181	0.133	0.125	0.162	0.132	0.090	0.130	6.069	1.458
Mean	0.935	0.969	0.985	0.99	0.987	0.994	1.001	0.983	1.195
	-0.917	-0.940	-0.972	-0.992	-0.979	-0.968	-1.004	-1.986	-0.711
	1.125	1.044	1.033	1.066	1.049	1.010	1.059	1.141	1.471
SD	0.061	0.065	0.067	0.072	0.067	0.051	0.051	5.419	1.075
	0.076	0.081	0.083	0.099	0.086	0.062	0.062	5.562	1.297
	0.132	0.126	0.121	0.149	0.123	0.090	0.090	6.098	1.387
Median	0.933	0.970	0.984	0.992	0.982	0.994	0.997	1.036	1.087
	-0.921	-0.943	-0.973	-0.989	-0.981	-0.973	-1.006	-0.942	-0.989
	1.128	1.043	1.035	1.064	1.061	1.015	1.064	1.081	1.155
MAD	0.032	0.041	0.042	0.054	0.044	0.037	0.040	0.195	0.145
	0.045	0.056	0.062	0.071	0.060	0.038	0.056	0.142	0.134
	0.097	0.091	0.077	0.101	0.091	0.056	0.077	0.227	0.193
<b>Cubic Design</b>									
RMSE	0.211	0.136	0.090	0.074	0.085	0.062	0.056	6.618	1.263
	0.335	0.216	0.136	0.112	0.124	0.104	0.083	15.343	1.250
	0.121	0.086	0.081	0.092	0.083	0.058	0.081	7.894	1.327
Mean	0.793	0.872	0.923	0.950	0.930	0.954	0.977	0.595	1.147
	-0.673	-0.796	-0.885	-0.945	-0.903	-0.919	-0.990	-0.186	-1.249
	1.088	1.037	1.025	1.034	1.029	1.001	1.029	0.789	1.110
SD	0.039	0.043	0.047	0.055	0.048	0.042	0.052	6.639	1.260
	0.074	0.069	0.073	0.098	0.078	0.065	0.083	15.398	1.231
	0.084	0.078	0.078	0.086	0.079	0.058	0.076	7.931	1.329
Median	0.796	0.875	0.925	0.946	0.929	0.953	0.974	0.988	1.078
	-0.676	-0.794	-0.887	-0.942	-0.902	-0.918	-0.984	-0.997	-1.133
	1.076	1.037	1.034	1.031	1.032	1.007	1.034	1.047	1.082
MAD	0.025	0.026	0.025	0.034	0.026	0.028	0.026	0.189	0.127
	0.057	0.039	0.041	0.061	0.043	0.036	0.049	0.203	0.194
	0.053	0.058	0.054	0.058	0.053	0.042	0.051	0.223	0.144

	Proposed Estimator with Fixed Window			Proposed Estimator with Optimal Window			Asymptotic Normality with Under Smoothing		
	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3	RDS	HK	LP
<b>Exponential Design</b>									
RMSE	0.213	0.134	0.087	0.084	0.084	0.063	0.066	5.749	1.079
	0.275	0.176	0.114	0.122	0.111	0.093	0.101	6.444	1.989
	0.207	0.132	0.109	0.130	0.113	0.075	0.101	5.408	2.303
Mean	0.791	0.877	0.936	0.957	0.941	0.963	0.992	0.765	0.964
	-0.735	-0.844	-0.929	-0.964	-0.939	-0.950	-1.013	-1.982	-0.748
	1.178	1.088	1.050	1.059	1.057	1.014	1.038	1.125	1.581
SD	0.045	0.053	0.053	0.072	0.060	0.051	0.066	5.773	1.084
	0.075	0.081	0.081	0.117	0.094	0.078	0.101	6.401	1.983
	0.105	0.099	0.099	0.117	0.099	0.074	0.094	5.434	2.240
Median	0.798	0.882	0.940	0.955	0.945	0.962	0.990	1.001	1.076
	-0.733	-0.845	-0.932	-0.961	-0.944	-0.950	-1.017	-0.989	-1.059
	1.182	1.091	1.055	1.062	1.061	1.014	1.047	1.082	1.091
MAD	0.030	0.034	0.036	0.050	0.040	0.031	0.042	0.162	0.129
	0.055	0.063	0.057	0.078	0.057	0.050	0.069	0.190	0.178
	0.072	0.074	0.075	0.066	0.078	0.054	0.064	0.178	0.148
<b>Sin Design</b>									
RMSE	0.368	0.219	0.127	0.120	0.128	0.082	0.072	6.343	0.453
	0.367	0.218	0.127	0.116	0.127	0.083	0.073	28.820	0.854
	0.187	0.105	0.084	0.099	0.089	0.061	0.081	7.333	3.327
Mean	0.634	0.786	0.887	0.901	0.885	0.938	0.971	0.722	1.062
	-0.636	-0.787	-0.887	-0.907	-0.887	-0.939	-0.971	0.841	-1.044
	1.164	1.068	1.033	1.051	1.043	1.010	1.029	0.476	0.848
SD	0.041	0.048	0.048	0.068	0.057	0.053	0.066	6.368	0.451
	0.048	0.050	0.050	0.069	0.059	0.056	0.067	28.906	0.857
	0.090	0.081	0.081	0.085	0.078	0.061	0.076	7.351	3.340
Median	0.641	0.789	0.890	0.901	0.889	0.942	0.977	0.960	1.054
	-0.636	-0.787	-0.888	-0.911	-0.890	-0.939	-0.977	-0.966	-1.055
	1.156	1.067	1.024	1.050	1.035	1.009	1.019	1.012	1.096
MAD	0.029	0.031	0.034	0.047	0.039	0.031	0.040	0.140	0.134
	0.036	0.035	0.042	0.049	0.044	0.038	0.045	0.118	0.145
	0.066	0.054	0.048	0.051	0.045	0.042	0.048	0.162	0.126

Table 2: Monte Carlo Results for Conditional Expectation Estimators

	RMSE			Bias			SD		
	RD	HK	LP	RD	HK	LP	RD	HK	LP
<b>Quadratic Design</b>									
1st decile	0.113	4.990	1.027	0.077	-0.792	-0.848	0.083	4.951	0.583
2nd decile	0.047	0.053	0.087	0.001	0.014	0.068	0.047	0.051	0.054
3rd decile	0.035	0.037	0.107	0.004	0.002	0.098	0.035	0.037	0.041
4th decile	0.039	0.040	0.136	0.015	0.008	0.129	0.037	0.039	0.042
5th decile	0.043	0.043	0.128	0.012	-0.001	0.119	0.041	0.043	0.047
6th decile	0.048	0.054	0.100	0.004	-0.012	0.085	0.048	0.052	0.053
7th decile	0.055	0.066	0.073	-0.009	-0.022	0.041	0.055	0.063	0.061
8th decile	0.074	0.083	0.076	-0.023	-0.018	-0.016	0.070	0.081	0.075
9th decile	0.098	0.104	0.133	-0.042	0.001	-0.094	0.090	0.104	0.095
10th decile	0.155	0.643	1.748	-0.038	0.192	1.016	0.151	0.617	1.429
<b>Cubic Design</b>									
1st decile	0.212	4.922	1.123	0.188	-0.826	-0.891	0.099	4.877	0.686
2nd decile	0.067	0.090	0.079	0.025	0.054	0.036	0.063	0.072	0.070
3rd decile	0.044	0.055	0.079	0.000	0.030	0.058	0.044	0.047	0.054
4th decile	0.041	0.043	0.087	-0.002	0.008	0.074	0.041	0.043	0.046
5th decile	0.045	0.048	0.124	0.015	0.016	0.116	0.043	0.046	0.045
6th decile	0.050	0.054	0.126	0.017	0.012	0.116	0.048	0.053	0.050
7th decile	0.057	0.064	0.093	0.003	-0.005	0.072	0.057	0.064	0.059
8th decile	0.074	0.088	0.077	-0.013	-0.019	0.010	0.073	0.086	0.077
9th decile	0.103	0.114	0.122	-0.038	-0.030	-0.067	0.096	0.111	0.102
10th decile	0.172	1.016	1.570	-0.082	0.228	0.877	0.152	0.996	1.309
<b>Exponential Design</b>									
1st decile	0.146	4.947	0.923	0.117	-0.807	-0.772	0.087	4.905	0.510
2nd decile	0.057	0.068	0.076	0.002	0.025	0.040	0.057	0.064	0.064
3rd decile	0.047	0.058	0.099	0.003	0.026	0.086	0.047	0.052	0.051
4th decile	0.043	0.050	0.112	0.008	0.023	0.103	0.042	0.044	0.045
5th decile	0.046	0.048	0.107	0.014	0.016	0.096	0.044	0.046	0.047
6th decile	0.053	0.053	0.088	0.022	0.007	0.071	0.049	0.053	0.053
7th decile	0.061	0.065	0.074	0.022	-0.005	0.037	0.057	0.065	0.065
8th decile	0.066	0.080	0.074	0.004	-0.023	-0.013	0.066	0.077	0.073
9th decile	0.087	0.101	0.113	-0.026	-0.032	-0.074	0.084	0.096	0.086
10th decile	0.157	0.794	1.744	-0.101	0.211	1.028	0.121	0.769	1.416
<b>Sin Design</b>									
1st decile	0.179	5.092	0.819	0.155	-0.892	-0.641	0.091	5.039	0.513
2nd decile	0.080	0.106	0.165	0.050	0.073	0.145	0.063	0.077	0.078
3rd decile	0.073	0.112	0.131	0.028	0.080	0.109	0.068	0.078	0.074
4th decile	0.072	0.096	0.094	0.014	0.050	0.062	0.071	0.082	0.071
5th decile	0.072	0.082	0.074	0.013	0.019	0.018	0.071	0.080	0.072
6th decile	0.072	0.080	0.071	0.013	-0.011	-0.018	0.071	0.080	0.069
7th decile	0.068	0.089	0.087	0.004	-0.038	-0.052	0.068	0.081	0.070
8th decile	0.066	0.095	0.105	-0.007	-0.055	-0.076	0.066	0.078	0.072
9th decile	0.072	0.088	0.097	-0.022	-0.043	-0.064	0.069	0.077	0.073
10th decile	0.171	0.965	1.904	-0.123	0.304	1.098	0.119	0.921	1.563



## References

- [1] Bhattacharya, P. K. (1967), "Estimation of a Probability Density Function and its Derivatives," *The Indian Journal of Statistics, Series 4*, v. 29, 373-382.
- [2] Fan, J. and Gijbels, I. (1995), "Adaptive Order Polynomial Fitting: Bandwidth Robustification and Bias Reduction." *Journal of Computational and Graphical Statistics*, Vol. 4, No. 3 , 213-227.
- [3] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and its Applications*, Chapman & Hall.
- [4] Hansen, B. (2008), "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, Vol. 24, No. 3, 726-748.
- [5] Honore, B. E. and Powell, J. L. (2005), "Pairwise Difference Estimation of Nonlinear Models", in *Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg* ed. by Andrews, D. W. K. and Stock, J. H., Cambridge University Press, Chap. 22, 520-553.
- [6] Horowitz, J. (1996), "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable, *Econometrica*, 64, 103-137.
- [7] Ichimura, H. (1993), "Semiparametric least squares (SLS) and weighted SLS estimation of single index models," *Journal of Econometrics*, 58, 71-120.
- [8] Ichimura, H. and Lee, L. F. (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.
- [9] Jiang, Y. (2019), "Semiparametric Estimation of Financial Risk: Corporate Default, Credit Ratings, and Implied Volatility," Ph.D. Thesis, Economics Department, Rutgers University.
- [10] Klein, R. and Shen. C. (2010), "Bias Corrections in Testing and Estimating Semiparametric, Single Index Models," *Econometric Theory*, 1683-1718.
- [11] Klein, R., Shen, C., Vella, F. (2015), "Estimation of marginal effects in semiparametric selection models with binary outcomes," *Journal of Econometrics* 185(1), 82-94.
- [12] Klein, R. and Spady. R. (1993), "An Efficient Semiparametric Estimator for the Binary Response Model," *Econometrica*, 61, 387-421.
- [13] Klein, R. and Vella, F. (2010), "Estimating a Class of Triangular Simultaneous Equations Models Without Exclusion Restrictions," *Journal of Econometrics* 154 (2010), 154-164.
- [14] Li, Q. and Sun, Y. (2014), "Nonparametric and Semiparametric Estimation and Hypothesis Testing with Nonstationary Time Series," *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* (edited by J. Racine, L. Su, and A. Ullah), 444-482.

- [15] Gu, J., Li, Q. and Yang, J.C. (2015), "Multivariate Local Polynomial Kernel Estimators: Leading Bias and Asymptotic Distributions," *Econometric Reviews*, 34,979-1010.
- [16] Lu, Z. Q. (1996), "Multivariate Locally Weighted Polynomial Fitting and Partial Derivative Estimation," *Journal of Multivariate Analysis*, v. 59, 187-205.
- [17] Masry, E. (1996), "Multivariate Regression Estimation Local Polynomial Fitting for Time Series," *Stochastic Processes and their Applications*, 65, 81-101.
- [18] Muller, H. G. (1984), "Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes," *Annals of Statistics*, 12.
- [19] Newey, W., Hsieh, F. and Robins, J. (2004), "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947-962.
- [20] Pakes, A. and Pollard, D. (1989), "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1058.
- [21] Powell, J., Stock, J. and Stoker, T. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 51, 1403-1430.
- [22] Robinson, P. M. (1988), "Root N-Consistent Semiparametric Regression," *Econometrica*, 56, 931-954.
- [23] Ruppert, D. and Wand, M. P. (1994), "Multivariate Locally Weighted Least-Squares Regression," *Annals of Statistics*, v. 52, #3, 1346-1370.

## 6 Appendix

### 6.1 Proofs of Main Theorems

**Proof of Theorem 1.** The proof for part a) follows from Lemmas 1-2. Parts b-c) follow from Lemma 6. With  $\hat{g}(v; \theta)$  converging to  $g > 0$ , part d) follows from parts a-c).

**Proof of Theorem 2.** The proofs for a) and c) are very similar. To establish a), note that under  $X$ -trimming (see, e.g. Ichimura(1993)), it can readily be shown that  $|\hat{\theta} - \theta_0| = o_p(1)$ . From a standard Taylor series expansion:

$$\left(\hat{\theta} - \theta_0\right) = -\hat{H}(\theta^+)^{-1} \hat{G}(\theta_0),$$

where  $\hat{H}$  is the (estimated) Hessian matrix,  $\hat{G}$  is the estimated gradient to the SLS objective function, and  $\theta^+$  is between  $\hat{\theta}$  and  $\theta_0$ . Let  $H$  denote the Hessian matrix with all estimated functions replaced by the corresponding true ones. From Lemma 9,

$$\sup_{\theta} \left| \hat{H}(\theta) - H(\theta) \right| = o_p(1).$$

Under standard arguments,

$$\sup_{\theta} \left| \hat{H}(\theta) - E[H(\theta)] \right| = o_p(1).$$

Therefore, with  $\theta^+ \xrightarrow{p} \theta_0$ ,  $\hat{H}(\theta^+) \xrightarrow{p} E[H(\theta_0)]$ . Under an invertibility assumption, the convergence rate for  $|\hat{\theta} - \theta_0|$  is then determined by the convergence rate for  $\hat{G}(\theta_0)$  to 0.

With  $\varepsilon_i = Y_i - M_i$ , write the estimated gradient component as  $\hat{G}_A - \hat{G}_B$ , where:

$$\hat{G}_A \equiv \frac{1}{N} \sum \hat{\tau}_{xi} \varepsilon_i \nabla_{\theta} \hat{M}_i; \quad \hat{G}_B \equiv \frac{1}{N} \sum \hat{\tau}_{xi} \left[ \hat{M}_s(V_i) - M_i \right] \nabla_{\theta} \hat{M}_i$$

Write  $\hat{G}_A = \hat{G}_{A_1} + \hat{G}_{A_2} + \hat{G}_{A_3} + \hat{G}_{A_4}$ , where with  $\Delta_s(V_i) \equiv \hat{g}_i \left( \hat{M}_s(V_i) - M_i \right)$ ,

$$\begin{aligned} \hat{G}_{A_1} &\equiv \frac{1}{N} \sum \tau_{xi} \varepsilon_i \nabla_{\theta} M_i \\ \hat{G}_{A_2} &\equiv \frac{1}{N} \sum [\hat{\tau}_{xi} - \tau_{xi}] \varepsilon_i \nabla_{\theta} M_i \\ \hat{G}_{A_3} &= \frac{1}{N} \sum \tau_{xi} \varepsilon_i \nabla_{\theta} \left[ \frac{\Delta_s(V_i)}{\hat{g}_i} \right] \\ \hat{G}_{A_4} &\equiv \sum [\hat{\tau}_{xi} - \tau_{xi}] \varepsilon_i \nabla_{\theta} \left[ \frac{\Delta_s(V_i)}{\hat{g}_i} \right] \end{aligned}$$

From a standard central limit theorem,  $\hat{G}_{A_1} = O_p(N^{-1/2})$ . From Pakes and Pollard (1989; Lemma 2.18),  $\hat{G}_{A_2} = o_p(N^{-1/2})$ . From Lemmas 1-2, with  $\rho_{si} \equiv \nabla_{\theta} \left[ \frac{\Delta_s^*(V_i)}{g_i} (1 + \delta_{gi}) \right]$ :

$$\hat{G}_{A_3} = \hat{G}_{A_3}^* + o_p(N^{-1/2}), \quad \hat{G}_{A_3}^* = \frac{1}{N} \sum \tau_{xi} \varepsilon_i \rho_{si}$$

From Lemma 10, we may take exterior trimming as known. From a mean-square convergence argument and Lemma 7,  $\hat{G}_{A_3}^* = o_p(N^{-1/2})$ . An induction argument obtains this result for  $s > 1$ .

As above, for  $\hat{G}_{A_4}$ , replace  $\nabla_\theta \left[ \frac{\Delta_s(V_i)}{\hat{g}} \right]$  with  $\rho_{si} \equiv \nabla_\theta \left[ \frac{\Delta_s^*(V_i)}{g_i} (1 + \delta_{gi}) \right]$  to obtain  $\hat{G}_{A_4}^*$  and note that  $\hat{G}_{A_4}$  is uniformly within  $o_p(N^{-1/2})$  of  $\hat{G}_{A_4}^*$ . Write:

$$\left| \hat{G}_{A_4}^* \right| \leq \frac{1}{N} \sum |\hat{\tau}_{xi} - \tau_{xi}| |\varepsilon_i| |\tau_{xi}^* \rho_{si}|,$$

where  $\tau_{xi}^* = 1$  if either  $\hat{\tau}_{xi} = 1$  or  $\tau_{xi} = 1$  and is 0 otherwise. Approximating  $|\hat{\tau}_{xi} - \tau_{xi}|$  by a smooth function as in Klein and Shen (2010), from Cauchy-Schwarz and Lemma 6,  $\hat{G}_{A_4}^* = o_p(N^{-1/2})$ .

For  $\hat{G}_B$ , write it as  $\hat{G}_{B_1} + \hat{G}_{B_2} + \hat{G}_{B_3} + \hat{G}_{B_4}$ , where

$$\begin{aligned} \hat{G}_{B_1} &\equiv \frac{1}{N} \sum \tau_{xi} \left[ \hat{M}_i - M_i \right] \nabla_\theta M_i \\ \hat{G}_{B_2} &\equiv \frac{1}{N} \sum \tau_{xi} \left[ \hat{M}_i - M_i \right] \left[ \nabla_\theta \hat{M}_i - \nabla_\theta M_i \right] \\ \hat{G}_{B_3} &\equiv \frac{1}{N} \sum [\hat{\tau}_{xi} - \tau_{xi}] \left[ \hat{M}_i - M_i \right] \nabla_\theta M_i \\ \hat{G}_{B_4} &\equiv \frac{1}{N} \sum [\hat{\tau}_{xi} - \tau_{xi}] \left[ \hat{M}_i - M_i \right] \left[ \nabla_\theta \hat{M}_i - \nabla_\theta M_i \right] \end{aligned}$$

For  $\hat{G}_{B_1}$ , from Cauchy-Schwarz, Lemmas 1,6, and C1), it can be shown that

$$\frac{1}{\sqrt{N}} \sum \left\{ \tau_{xi} \left[ \hat{M}_i - M_i \right] \left( \frac{\hat{g}_i}{g_i} - 1 \right) \right\} \nabla_\theta M = o_p(1),$$

from which it follows that  $\sqrt{N} \left[ \hat{G}_{B_1} - \hat{G}_{B_1}^* \right] = o_p(1)$ , where

$$\hat{G}_{B_1}^* \equiv \frac{1}{N} \sum \left\{ \tau_{xi} \frac{\left[ \hat{M}_i - M_i \right] \hat{g}_i}{g_i} \right\} \nabla_\theta M.$$

Under the stage and window conditions,  $E \left[ N^{1/2} \hat{G}_{B_1}^* \right] = o(1)$ . Employing the arguments similar to those in Jiang (2019), it can then be shown that  $N^{1/2} [\hat{G}_{B_1}^* - U_N] = o_p(1)$  where  $U_N$  is a centered U-statistic with  $N^{1/2} U_N$  asymptotically distributed as normal. It follows that  $\hat{G}_{B_1}^* = O_p(N^{-1/2})$ . From Cauchy Schwarz, Lemmas 6-7, and C1),  $\hat{G}_{B_2} = o_p(N^{-1/2})$ . The analysis for  $\hat{G}_{B_3}$  and  $\hat{G}_{B_4}$  is similar to that for  $\hat{G}_{A_4}^*$  and  $\hat{G}_{B_2}$ .

To establish b), under index trimming at the second step, consistency follows from an extension of Lemma 5 in Klein and Shen (2010) to  $d$  multiple indices. To outline the argument, recall the definition of the adjusted estimator in D7) and D8):

$$\begin{aligned} \hat{M}_{sa}(v) &\equiv \frac{\frac{1}{N} \sum_{i=1}^N Y_i K_i(v)}{\hat{g}_{1a}(v)} - \frac{\frac{1}{N} \sum_{i=1}^N Z_i K_i(v)}{\hat{g}_{1a}(v)} \hat{d}_m, s = 1 \\ \hat{M}_{sa}(v) &\equiv \frac{\frac{1}{N} \sum_i \left[ Y_i - (\hat{M}_{(s-1)a}(V_i) - \hat{M}_{(s-1)a}(v)) \right] K_i(v)}{\hat{g}_a(v)}, s \geq 2 \end{aligned}$$

For  $s = 1$ , consider the first component of  $\hat{M}_{1a}(v)$ ,

$$\frac{\frac{1}{N} \sum_{i=1}^N Y_i K_i(v)}{\hat{g}_{1a}(v; \theta)} \equiv \frac{\frac{1}{N} \sum_{i=1}^N Y_i K_i(v)}{\hat{g}_{1a}(v; \theta)} \equiv \frac{\hat{f}(v; \theta)}{\hat{g}_{1a}(v; \theta)}.$$

With  $f(v; \theta) = M(v; \theta)g(v; \theta)$ ,

$$\left| \frac{\hat{f}(v; \theta)}{\hat{g}_{1a}(v; \theta)} - \frac{f(v; \theta)}{g(v; \theta)} \right| \leq \left| \frac{\hat{f}(v; \theta) - f(v; \theta)}{\hat{g}_{1a}(v; \theta)} \right| + \left| \frac{f(v; \theta)}{g(v; \theta)} \frac{\hat{g}_{1a}(v; \theta) - g(v; \theta)}{\hat{g}_{1a}(v; \theta)} \right|.$$

With  $\frac{f(v; \theta)}{g(v; \theta)}$  bounded, it suffices to consider  $\left| \frac{\hat{f}(v; \theta) - f(v; \theta)}{\hat{g}_{1a}(v; \theta)} \right|$  and  $\left| \frac{\hat{g}_{1a}(v; \theta) - g(v; \theta)}{\hat{g}_{1a}(v; \theta)} \right|$ . With the proofs for these terms being similar, consider the first term and write:

$$\sup_{v, \theta} \left| \frac{\hat{f}(v; \theta) - f(v; \theta)}{\hat{g}_{1a}(v; \theta)} \right| \leq \sup_{v, \theta} \left| \frac{\hat{f}(v; \theta) - E \left[ \hat{f}(v; \theta) \right]}{\hat{g}_{1a}(v; \theta)} \right|_A + \sup_{v, \theta} \left| \frac{E \left[ \hat{f}(v; \theta) \right] - f(v; \theta)}{\hat{g}_{1a}(v; \theta)} \right|_B.$$

To analyze the  $A$ -term, let

$$\Delta_A \equiv h^{-a} \sup_{v, \theta} \left| \hat{f}(v; \theta) - E \left[ \hat{f}(v; \theta) \right] \right|.$$

Recall that in the adjustment factor defined in D7),  $\tau_\Delta(v)$  is a smoothed indicator approaching 1 on the set  $\{v : q_v(\lambda_1) < v < q_v(\lambda_2)\}$ . Let  $\tau^*(v)$  be an indicator for  $\{v : q_v(\lambda_1^*) < v < q_v^*(\lambda_2^*)\}$  where  $\lambda_1^* < \lambda_1 < \lambda_2 < \lambda_2^*$ . Then,

$$A \leq \Delta_A \left\{ h^a \sup_{v, \theta} \left[ \frac{\tau^*}{\hat{g}(v; \theta)} \right] + h^a \sup_{v, \theta} \left[ \frac{1 - \tau^*}{\Delta} \right] \right\}.$$

When  $\tau^*(v) = 1$ ,  $\hat{g}(v; \theta)$  uniformly converges to  $g(v; \theta) > 0$ , while on its complement,  $\Delta = O(h^{-a})$ . Accordingly, it suffices to show that  $\Delta_A = o_p(1)$ . From Lemma 8,

$$\Delta_A = O(N^{-ra}) O_p \left( N^{-\left(\frac{m}{2(m+2)} - rd\right)} \right).$$

Under A5) with  $a < 1$ ,  $\Delta_A = o_p(1)$ .

For the  $B$ -term, as in Klein and Shen (2010), let  $\tau_s$  be an indicator for  $\{x : a_k + h^c < x_k < b_k - h^c\}$ , where  $0 < a < c < 1$ . Letting

$$\Delta_B \equiv h^{-a} \sup_{v, \theta} \left| E \left[ \hat{f}(v; \theta) \right] - f(v; \theta) \right|,$$

it follows that

$$B \leq \Delta_B \left\{ h^a \sup_{v, \theta} \left[ \frac{\tau^*}{\hat{g}(v; \theta)} \right] + h^a \sup_{v, \theta} \left[ \frac{1 - \tau^*}{\Delta} \right] \right\}.$$

Under the same arguments as above,  $B$  will be uniformly  $o_p(1)$  if  $\Delta_B$  is uniformly  $o_p(1)$ . Write  $B$  as:

$$B = \tau_s B + (1 - \tau_s) B.$$

The first term is uniformly  $o_p(1)$  as the boundary bias vanishes when  $\tau_s = 1$ . When  $\tau_s = 0$ , it can be shown that  $B = O_p(1)$ . With the probability of  $\tau_s = 0$  being  $O(N^{-c})$ ,  $0 < a < c < 1$ ,

the second term uniformly converges in probability to zero. Similar arguments apply for  $s > 1$ . From the above uniform convergence, the second stage objective function uniformly converges to its expectation, which has a unique maximum at the true parameter vector under standard arguments. Consistency for the second stage estimator in (7) of Theorem 2 follows.

To establish c), note that it has the following linear form:

$$\sqrt{N} \left( \hat{\theta}_a - \theta_0 \right) = -\hat{H}^{-1}(\theta^+) \sqrt{N} \hat{G}(\theta_0) + o_p(1), \quad \theta^+ \in \left[ \hat{\theta}_a, \theta_0 \right].$$

Employing arguments very similar to those in Part a), it can be shown under C2) that

$$\begin{aligned} i) & : \hat{H}^{-1}(\theta^+) \xrightarrow{p} \sigma^2 E[\tau_{vi} \nabla_{\theta} M_i \nabla_{\theta} M_i'] \\ ii) & : \sqrt{N} \hat{G}(\theta_0) = \frac{1}{\sqrt{N}} \sum \tau_{vi} \varepsilon_i \nabla_{\theta} M_i - \frac{1}{\sqrt{N}} \sum \tau_{vi} \left[ \hat{M}_i - M_i \right] \nabla_{\theta} M_i + o_p(1). \end{aligned}$$

The proof for part c) will then follow if

$$B \equiv \frac{1}{\sqrt{N}} \sum \tau_{vi} \left[ \hat{M}_i - M_i \right] \nabla_{\theta} M_i = o_p(1).$$

From Lemmas 1-2:

$$B = B^* + o_p(1), \quad B^* \equiv \frac{1}{\sqrt{N}} \sum \tau_{vi} \frac{\left[ \hat{M}_i^* - M_i \right] \hat{g}_i}{g_i} [1 + \delta_{g_i}] \nabla_{\theta} M_i.$$

From Newey's residual result,  $E[B^*] = 0$ . In the single index case, Klein and Shen (2010) show that  $B^*$  is a degenerate U-statistic under regular kernels. This argument had been extended to multiple indices under the recursive estimator, with the extension provided in Jiang (2019).

With Lemma 10 showing that exterior trimming can be taken as known, Lemmas 1-9 are provided for the known trimming case. For notational simplicity, let  $K_i^*(v) \equiv K_i^*(v; \tau)$  in Lemmas 1-9.

Lemmas 1-6 provide properties of the proposed expectations estimator while Lemmas 7-9 provide results for estimating semiparametric models. For ease of exposition, in Lemmas 1-6 we assume that the conditional expectation of the dependent variable,  $Y_i$ , is a function of the  $d$ -dimensional vector  $V_i$ . The vector  $V_i$  may be interpreted as a vector of indices in the semiparametric case or as a vector of exogenous variables in the nonparametric case.

We will obtain results for an approximating recursive sequence that is uniformly within  $N^{-1/2}$  of the original sequence. In constructing this sequence, we require the result on trimmed kernels in Lemma 1. Lemma 2 then provides the main result on the approximation.

## 6.2 Intermediate Lemmas

**Lemma 1. Density Estimators and Trimming.** Recall from D3) that index-trimming and  $X$ -trimming functions have the form:

$$\begin{aligned} \tau_v(V_i) & \equiv 1 \{V_i : q_v(\lambda_1) < V_i < q_v(\lambda_2)\}; \\ \tau_x(X_i) & \equiv 1 \{X_i : q_x(\lambda_1) < X_i < q_x(\lambda_2)\}. \end{aligned}$$

Let  $\mathcal{C}_v^*$  and  $\mathcal{C}_x^*$  be sets of the form:

$$\begin{aligned}\mathcal{C}_v^* &\equiv \{V_i : \underline{q} < V_i < \bar{q}\}, q_v(\lambda_1) < \underline{q} < \bar{q} < q_v(\lambda_2) \\ \mathcal{C}_x^* &\equiv \{X_i : \underline{q} < X_i < \bar{q}\}, q_x(\lambda_1) < \underline{q} < \bar{q} < q_x(\lambda_2)\end{aligned}$$

Then:

$$\begin{aligned}a) &: \sup_v \left| \frac{1}{N} \sum [1 - \tau_v(V_i)] K_i(v) \right| = o_p(N^{-1/2}) \text{ where } v \in \mathcal{C}_v^*; \\ b) &: \sup_v \left| \frac{1}{N} \sum [1 - \tau_x(X_i)] K_i(v) \right| = o_p(N^{-1/2}) \text{ where } x \in \mathcal{C}_x^*.\end{aligned}$$

**Proof.** Here we provide the proof for a), with the proof for b) being similar. Referring to a), denote  $V_i(m)$  as the  $m^{\text{th}}$  element of  $V_i$ ,  $m = 1, \dots, d$ . Let  $q_{1m}$  be the  $m^{\text{th}}$  element in the vector  $q_v(\lambda_1)$  and  $q_{2m}$  the  $m^{\text{th}}$  element in the vector  $q_v(\lambda_2)$ . Define an indicator:

$$\tau_m(V_i(m)) \equiv 1 \{V_i(m) : q_{1m} < V_i(m) < q_{2m}\}$$

and  $\tau_v(V_i) = \prod_{m=1}^d \tau_m(V_i(m))$ . It follows that  $[1 - \tau_v(V_i)]$  can be written as the sum of a finite number of terms, with a typical term given as:

$$\prod_{m=1}^{d_1} [1 - \tau_m(V_i(m))] \prod_{m=d_1+1}^d \tau_m(V_i(m)).$$

It then suffices to examine:

$$T(v) \equiv \frac{1}{N} \sum_{i=1}^N \prod_{m=1}^{d_1} [1 - \tau_m(V_i(m))] K_{im}(v(m)) \prod_{m=d_1+1}^d \tau_m(V_i(m)) K_{im}(v(m)),$$

where  $K_{im}(v)$  is the kernel component defined in D3). Under a), with  $q_{1m}^*$  and  $q_{2m}^*$  as the  $m^{\text{th}}$  elements in  $q_v(\lambda_{1I})$  and  $q_v(\lambda_{2I})$  respectively:

$$q_{1m} < q_{1m}^* < v(m) < q_{2m}^* < q_{2m}.$$

Therefore, noting that  $T_2(v) \equiv \prod_{m=d_1+1}^d \sup_{v(m)} \tau_m(V_i(m)) K_{im}(v(m)) = O(h^{-d_1})$ ,  $\sup_v T(v)$  is bounded above by:

$$\begin{aligned}& \frac{1}{N} \sup_v \sum_{i=1}^N \left[ \prod_{m=1}^{d_1} \sup_{v(m)} [1 - \tau_m(V_i(m))] K_{im}(v(m)) \right] \sup_v [T_2(v)] \\ &= O(h^{-d_1}) \frac{1}{N} \sum_{i=1}^N \prod_{m=1}^{d_1} [1 \{V_i(m) > q_{2m}\} K_{im}(q_{2m}^*) + 1 \{V_i(m) < q_{1m}\} K_{im}(q_{1m}^*)]. \quad (8)\end{aligned}$$

The result in (8) follows as the sup of each term in the sum is attained by selecting a value for  $v(m)$  that minimizes  $|v(m) - V_i(m)|$ . Hence, for each component in first product of terms, the sup is attained at either  $q_{1m}$  or  $q_{2m}$ . Under symmetry for the kernel, as the analysis for each of the two terms in (8) is essentially the same, here we consider the expectation

and variance of the first term and show that it has order smaller than  $o_p(N^{-1/2}h^{d_1})$ . For  $\bar{T}_1 \equiv \frac{1}{N} \sum_{i=1}^N \prod_{m=1}^{d_1} 1 \{V_i(m) > q_{2m}\} K_{im}(q_{2m}^*)$  and  $g$  the index density:

$$E [\bar{T}_1] \leq \sup_v g(v) \int \dots \int \prod_{m=1}^{d_1} [1 \{V_i(m) > q_{2m}\}] K_{im}(q_{2m}^*) dv_i \quad (9)$$

$$\leq \sup_v g(v) \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \prod_{m=1}^{d_1} 1 \{V_i(m) > q_{2m}\} K_{im}(q_{2m}^*) dv_i. \quad (10)$$

$$= \sup_v g(v) \prod_{m=1}^{d_1} \int_{-\infty}^{\infty} [1 \{V_i(m) > q_{2m}\}] K_{im}(q_{2m}^*) dv_i(m) \equiv P, \quad (11)$$

where in (9) the integration is over the support for  $V_i$ . From the definition of the indicator,

$$P = \prod_{m=1}^{d_1} \left[ \int_{q_{2m}}^{\infty} K_{im}(q_{2m}^*) dV_i(m) \right].$$

Making a change of variable with  $Z(m) \equiv \frac{q_{2m} - V_i(m)}{h}$ , as the kernel has moments of all orders under D4), from Markov's inequality:

$$P \leq \prod_{m=1}^{d_1} O(E [Z(m)]^{2t}) O(h^{2t}) = O(h^{2td_1}).$$

With  $t$  selected to be sufficiently large, the result follows. For the variance,

$$\begin{aligned} Var (\bar{T}_1) &= Var \left[ \prod_{m=1}^{d_1} 1 \{V_i(m) > q_{2m}\} K_{im}(q_{2m}^*) \right] / N \\ &\leq E \left[ \prod_{m=1}^{d_1} [1 \{V_i(m) > q_{2m}\}] K_{im}^2(q_{1m}) \right] \\ &= O \left( \frac{1}{h^{d_1}} \right) O \left( E \left[ \prod_{m=1}^{d_1} [1 \{V_i(m) > q_{2m}\}] h^{d_1} K_{im}^2(q_{1m}) \right] \right). \end{aligned}$$

From the analysis of the expectation above, the second component vanishes arbitrarily fast. The Lemma follows. ■

**Lemma 2. Recursion Approximation.** Recall the definition of the initial estimator in D6) and kernel functions in D4-5). With  $V_i \equiv V(W_i; \theta_0)$  and  $Z_i(v) \equiv [V_i - v] / h$  as the  $i^{th}$  row of  $Z$ , define:

$$\begin{aligned} \hat{A} &\equiv \sum \left[ \frac{1}{N} Z(v)' D^* Z(v) \right]; \quad A \equiv E \left[ \frac{1}{N} Z(v)' D^* Z(v) \right]; \quad \delta_A(v) \equiv \sum_{l=1}^{\mathcal{L}} \left[ (A - \hat{A}) A^{-1} \right]^l \\ \bar{g}_s(v) &= E(\hat{g}_s(v)); \quad \delta_{g_s}(V_i) \equiv \frac{1}{g_s(V_i)} \left[ \sum_{p=1}^{\mathcal{P}} \left[ \frac{\bar{g}_s(V_i) - \hat{g}_s(V_i)}{\bar{g}_s(V_i)} \right]^p \right] \\ \hat{d}^*(v) &\equiv A^{-1} [I + \delta_A(v)] \frac{1}{N} \sum_{i=1}^N \frac{Z_i(v) [Y_i - \hat{M}(v)] \hat{g}_1(v) K_i^*(v)}{\bar{g}_1(v)} [1 + \delta_{g_1}(v)] \\ \Delta_s^*(v) &\equiv \begin{cases} \Delta_1^*(v) \equiv \frac{1}{N} \sum_{i=1}^N [M(V_i) - M(v) + \varepsilon_i] K_i(v) - \frac{1}{N} \sum_{i=1}^N Z_i(v) K_i(v) \hat{d}^*(v), & s = 1 \\ \Delta_{s-1}^*(v) - \sum_{i=1}^N \left[ \frac{\Delta_{s-1}^*(V_i)}{\bar{g}_{s-1}(V_i)} \right] [1 + \delta_{g_{s-1}}(V_i)] K_i^*(v) + \frac{1}{N} \sum_{i=1}^N \varepsilon_i K_i^*(v), & s > 1 \end{cases} \end{aligned}$$



For  $\mathcal{L}, \mathcal{P}$  sufficiently large and finite:

$$\sup_v |\Delta_s^*(v) - \Delta_s(v)| = o_p(N^{-1/2}).$$

**Proof.** For  $s = 1$ , note that:

$$\begin{aligned} \hat{A}^{-1} &= A^{-1} + \hat{A}^{-1} [A - \hat{A}] A^{-1} \\ &= A^{-1} [I + \delta_A] + \hat{A}^{-1} [A - \hat{A}]^{-\mathcal{L}+1} A^{-\mathcal{L}+1}. \end{aligned}$$

With a similar expansion holding for  $1/\hat{g}(V_i)$ , for  $\mathcal{L}, \mathcal{P}$  sufficiently large:

$$\sup_v \left| \hat{A}^{-1} - A^{-1} [I + \delta_A] \right| = o_p(N^{-1/2}) \quad (12)$$

$$\sup_v \left| \frac{1}{\hat{g}(V_i)} - \frac{1}{\bar{g}(V_i)} [1 + \delta_{g_1}(V_i)] \right| = o_p(N^{-1/2}). \quad (13)$$

The result now follows for  $s = 1$ .

For  $s = 2$ ,

$$\Delta_2(v) \equiv \hat{g}(v) [\hat{M}_1(v) - M_1(v)] - \sum \left[ [\hat{M}_1(V_i) - M_1(V_i)] \right] K_i^*(v).$$

Therefore,  $\Delta_2^*(v) - \Delta_2(v) = T_1 - [T_{21} - T_{22}]$ , where

$$\begin{aligned} T_1 &\equiv \Delta_1^*(v) - \frac{\hat{g}(v)}{\hat{g}_1(v)} \Delta_1(v) \\ T_{21} &\equiv \frac{1}{N} \sum_{i=1}^N \left[ \left\{ \frac{\Delta_1^*}{\bar{g}_1(V_i)} \right\} [1 + \delta_{g_1}(V_i)] \right] K_i^*(v) \\ T_{22} &\equiv \frac{1}{N} \sum_{i=1}^N \left[ \hat{g}_1(V_i) \frac{\hat{M}_1(V_i) - M(V_i)}{\hat{g}_1(V_i)} \right] K_i^*(v). \end{aligned}$$

Because  $\sup_v |\Delta_1^*(v) - \Delta_1(v)| = o_p(N^{-1/2})$ , we have  $T_1$  uniformly  $o_p(N^{-1/2})$  from Lemma 1. From (12), (13), it follows that  $T_{21} - T_{22}$  is uniformly  $o_p(N^{-1/2})$ . An induction argument completes the proof. ■

**Lemma 3. Stage Characterization.** Define

$$KP_0(v) \equiv \frac{1}{N} \sum [M(V_i) - M(v)] \frac{1}{h^d k} \left[ \frac{(V_i - v)}{h} \right].$$

Recall the definitions of  $\delta_A(v)$  and  $\delta_{g_1}(V_i)$  in Lemma 2. With  $Z_i = (v - V_i)/h$ , let:

$$\begin{aligned} e_1(v, V_i) &\equiv Z_i K_i(v) A^{-1}(v) \\ e_2(v, V_i) &\equiv Z_i K_i^*(v) [M(V_i) - M(v)] \\ e_3(v, V_i) &\equiv Z_i K_i^*(v) \\ e_4(v, V_j) &\equiv \frac{[M(V_j) - M(v)] K_j^*(v)}{\bar{g}(v)}. \end{aligned}$$

Define  $\bar{e}_t(v) = \frac{1}{N} \sum_i e_t(v, V_i)$  and

$$KP_1(v) \equiv \bar{e}_1(v) [I + \delta_A(v)] \{ \bar{e}_2(v) + \bar{e}_3(v) \bar{e}_4(v) [1 + \delta_{g_1}(v)] \}.$$

Let  $U_1$  and  $U$  be random variables that have zero expectation conditioned on  $X$  and define:

$$KP_L(v) \equiv O\left(\frac{1}{N^L}\right) \sum_{i_1} \dots \sum_{i_L} T_{L, i_1, \dots, i_L},$$

$$T_{L, i_1, \dots, i_L} \equiv \frac{\Delta_1^*(V_{i_1})}{\prod_{l=1}^L \bar{g}_s(V_{i_l})} K_{i_L}^*(v) \prod_{l=1}^L \delta_{g_s}(V_{i_l}) \prod_{l=1}^{L-1} K_{i_l}^*(V_{i_{l+1}}).$$

Then, there exists integers  $C_1, \dots, C_{k-2}$  such that

$$a) : \Delta_1^*(v) = U_1 + KP_0(v) - KP_1(v), \quad s = 1 \quad (14)$$

$$b) : \Delta_{s-1}^*(v) - \Delta_1^*(v) = U + \sum_{l=1}^{s-1} C_l KP_l(v), \quad s > 1. \quad (15)$$

**Proof.** The proof for a) is immediate from Lemmas 1–2. For b), due to the form of  $\Delta_{s-1}^*(v) - \Delta_{s-2}^*(v)$ , the lemma follows because

$$\Delta_{s-1}^*(v) - \Delta_1^*(v) = [\Delta_{s-1}^*(v) - \Delta_{s-2}^*(v)] + [\Delta_{s-2}^*(v) - \Delta_{s-2}^*(v)] + \dots + [\Delta_2^*(v) - \Delta_1^*(v)].$$

■

To study the expectations of the estimators as characterized in Lemma 3, Lemma 4 provides conditional independence results for studying the expectation of products of averages, which is the structure of the stage  $s$  estimator. Part a) of this lemma is stated in a form more general than is required so as to illustrate a method of proof that applies to all other parts.

**Lemma 4. Kernel Products.** Let  $F(V_i)$  be a bounded function of  $V_i$  and  $A(v)$  a matrix that depends on  $v$ . For  $p$  a positive integer, define:

$$d1) : |A(v)| \text{ has } i, j \text{ element } |A_{ij}(v)|$$

$$d2) : \sup_v |A(v)| \text{ has } i, j \text{ element } \sup_v |A_{ij}(v)|$$

$$d3) : \delta(V_i) \equiv \left\{ \frac{1}{N} \sum_j [F(V_i) - K_j^*(V_i)] \right\}^p$$

$$d4) : \rho_i \equiv \left[ \frac{\Delta_{s-1}^*(V_i)}{g(V_i)} [1 + \delta_{g_s}(V_i)] \right].$$

Referring to the definitions in Lemmas 1-2, under A5):

$$\begin{aligned}
a) & : E[\delta(V_i)|V_i] = \left\{ \frac{1}{N} \sum_j E[F(V_i) - K_j^*(V_i) | V_i] \right\}^p + O\left(\frac{1}{Nh^d}\right) \\
b) & : E[\delta_A(v)] \equiv E\left\{[\bar{e}_2(v)]^l\right\} = [E(\bar{e}_2(v))]^l + O\left(\frac{1}{Nh^d}\right) \\
c) & : E[KP_1(v)] \equiv E\left\{\bar{e}_1(v) [I + \delta_A(v)] \{\bar{e}_2(v) + \bar{e}_3(v) \bar{e}_4(v) [1 + \delta_{g_1}(v)]\}\right\} = \\
& E[\bar{e}_1(v)] [I + E[\delta_A(v)]] \{E[\bar{e}_2(v)] + E[\bar{e}_3(v)] E[\bar{e}_4(v)] [1 + E[\delta_{g_1}(v)]]\} + O\left(\frac{1}{Nh^d}\right) \\
d) & : E\left\{\frac{\Delta_{s-1}^*(V_i)}{\bar{g}(V_i)} [1 + \delta_g(V_i)] | V_i\right\} = \frac{E[\Delta_{s-1}^*(V_i)|V_i]}{\bar{g}(V_i)} [1 + E[\delta_g(V_i)|V_i]] + O\left(\frac{1}{Nh^d}\right) \\
e) & : E\{\rho_i K_i(v) \rho_j K_j(v)\} = E\{E[\rho_i K_i(v)|V_i] E[\rho_j K_j(v)|V_j]\} + O\left(\frac{1}{Nh^d}\right).
\end{aligned}$$

**Proof.** For a), write the expectation of a typical term in  $\delta(V_i)$ :

$$E\left\{O\left(\frac{1}{N^p}\right) \left[\sum_{j \neq i} F(V_i) - K_j^*(V_i)\right]^p | V_i\right\} = E\left\{O\left(\frac{1}{N^p}\right) \sum_{j_1} \sum_{j_2} \dots \sum_{j_p} \prod_{l=1}^p [F(V_i) - K_{j_l}^*(V_i)] | V_i\right\}.$$

If all of the subscripts are distinct, the result is immediate from independence. Assume there are  $m+1 \geq 2$  identical subscripts and re-order terms so that these are at the end. The expectation is then:

$$\begin{aligned}
& O\left(\frac{1}{N^m}\right) E\left\{O\left(\frac{1}{N^{p-m}}\right) \left[\sum_{j_1} \sum_{j_2} \dots \sum_{j_{(p-m)}} \left(\prod_{l=1}^{p-m} [F(V_i) - K_{j_l}^*(V_i)]\right)\right] [F(V_i) - K_{j_p}^*(V_i)]^{m+1} | V_i\right\} \\
& = O\left[\left(\frac{1}{Nh^d}\right)^m\right] \prod_{l=1}^{p-m-1} [F(V_i) - E[K_{j_l}^*(V_i)|V_i]] E\left\{O(h^{md}) [F(V_i) - K_{j_p}^*(V_i)]^{m+1} | V_i\right\}.
\end{aligned}$$

For the three components above, from A5), the first term is  $O\left[\left(\frac{1}{Nh^d}\right)^m\right]$ ,  $m \geq 1$ . It can readily be shown the second component is  $O(1)$ . For the final component, let  $z$  be a  $d \times 1$  vector with  $l^{th}$  component  $\frac{V_i(l) - V_{j_p}(l)}{O(h)}$ . With this change of variable, and the integral being multi-dimensional, we can write the third component as:

$$\begin{aligned}
& \int O(h^{md}) \left[F(V_i + O(h)z) - \frac{1}{O(h^d)} k(z)\right]^{m+1} O(h^d) g(V_i + O(h)z) dz \\
& = \int O(h^{(m+1)d}) \left\{ \sum_{i=1}^{m+1} \binom{m+1}{i} F(V_i + O(h)z)^{m+1-i} \left[\frac{1}{O(h^d)} k(z)\right]^i \right\} g(V_i + O(h)z) dz.
\end{aligned}$$

With  $k(z)^p$  integrable under A5) for all finite  $p$ , the above integral is  $O(1)$  from which the result follows. The proofs for b-c) are very similar to that for a) in that they depend on matching subscripts as in the above argument and require definitions d1)-d2). For d), from the stage characterization in Lemma 4, it suffices to consider:

$$E[KP_L(V_j) \delta_g(V_j) | V_j].$$

Employing the same matching arguments as in a), it follows that:

$$E [KP_L(V_j)\delta_g(V_j) | V_j] = E [KP_L(V_j)|V_j] E [\delta_g(V_j) | V_j] + o(N^{-1/2}),$$

which completes the argument for d). ■

**Lemma 5. Bias Expansions for Kernel Expectations.** Let  $F(v)$  be a bounded function of  $v$  and  $g(v)$  the density for  $V_i$ . Assume that  $F(v)g(v)$  has uniformly bounded derivatives to order  $2m + 1$ . Then, with kernel functions defined in D4), and for  $v$  in a compact subset over which the trimming function (see D3) is defined:

$$\begin{aligned} E [F(V_i)K_i^*(v)] &\equiv E [F(V_i)\tau_v(V_i) K_i(v)] \\ &= F(v)g(v) + \sum_{i=1}^m h^{2i} B_i(v) + o(h^{2m}), \end{aligned}$$

where  $B_i$  is uniformly bounded in its argument.

**Proof.** Recalling D4):

$$E [F(V_i)K_i^*(v)] = \int_{\mathcal{C}_v(q_v)} F(v_i)K_i(v)g(v_i)dv_i,$$

where  $\mathcal{C}_v(q_v)$  as defined in D3) is the set over which the trimming function is one. Let  $z$  be a vector with  $l^{th}$  element  $\frac{v(l)-v_i(l)}{O(h)}$ . Then, making this change of variable and restricting all components of  $v$  to a strict subset of  $\mathcal{C}_v(q_v)$ , the expectation up to higher order terms is given as:

$$\int_{\mathcal{R}^d} F(v + hz)k(z)g(v + hz)dz,$$

where  $k$  is the standardized kernel in D4). From a standard Taylor expansion in  $h$  about 0 and A3), the result follows. ■

Employing the above lemmas, Lemma 6 obtains uniform bias and variance rates for the proposed estimator.

**Lemma 6. Stage Bias and Variance.** Assuming A5), with  $\Delta_s^*(v)$  defined as in Lemma 2, and  $B_s(v)$  a uniformly bounded function:

$$\begin{aligned} a) &: \sup_v E [\Delta_s^*(v)] = h^{2s} B_s(v) + h^{2(s+1)} B_{s+1}(v) + o(h^{2(s+1)}), \\ b) &: \sup_v Var [\Delta_s^*(v)] = O(N^{-(1-rd)}). \end{aligned}$$

**Proof.** For a), with  $KP_0$  and  $KP_1$  defined in Lemma 3:

$$E [\Delta_1^*(v)] = E [KP_0(v)] - E [KP_1(v)].$$

It then follows from Lemmas 4-5 that

$$\sup_v |E [\Delta_1^*(v)]| = h^2 B_1(v) + h^4 B_2(v) + o(h^4).$$

Continuing with an induction argument for a), with  $s > 1$ , assume that

$$\sup_v |E [\Delta_{s-1}^*(v)]| = h^{2(s-1)} B_{s-1}(v) + h^{2s} B_s(v) + o(h^{2s}), \quad (16)$$

Under Lemma 4d),  $E [\Delta_s^*(v)]$  is given as:

$$\begin{aligned} & E [\Delta_{s-1}^*(v)] - E \left[ \frac{\Delta_{s-1}^*(V_i)}{\bar{g}(V_i)} [1 + \delta_{g_s}(V_i)] K_i^*(v) \right] \\ = & E [\Delta_{s-1}^*(v)] - E \left[ \frac{E [\Delta_{s-1}^*(V_i)|V_i]}{\bar{g}(V_i)} [1 + E [\delta_{g_s}(V_i)|V_i]] K_i^*(v) \right]. \end{aligned}$$

From Lemma 5 it now follows that

$$E \{ E [\Delta_{s-1}^*(V_i)|V_i] K_i(v) \} = h^{2(s-1)} B_{s-1}(v) + h^{2s} B_s(v) + o(h^{2s}).$$

Part a) now follows because  $E [\delta_{g_s}(V_i)|V_i] = O(h^2)$ .

The proof for b) at stage  $s = 1$  is immediate. Assuming the result holds for stage  $s - 1$ , we show it holds for stage  $s > 1$ , recall that

$$\Delta_s^*(v) \equiv \Delta_{s-1}^*(v) - \frac{1}{N} \sum_i \left[ \frac{\Delta_{s-1}^*(V_i)}{g(V_i)} [1 + \delta_{g_s}(V_i)] \right] K_i(v) - \frac{1}{N} \sum_i \varepsilon_i K_i(v).$$

With  $T \equiv \Delta_s^*(v) - E [\Delta_s^*(v)]$  and  $\rho_i \equiv \frac{\Delta_{s-1}^*(V_i)}{g(V_i)} [1 + \delta_{g_s}(V_i)]$ , from the definition of  $\Delta_s^*(v)$ ,  $T = T_1 - T_2 - T_3 - T_4$  where

$$\begin{aligned} T_1 & \equiv \Delta_{s-1}^*(v) - E [\Delta_{s-1}^*(v)] \\ T_2 & \equiv \frac{1}{N} \sum_i \varepsilon_i K_i(v) \\ T_3 & \equiv \frac{1}{N} \sum_i \{ E (\rho_i | V_i) K_i(v) - E [\rho_i K_i(v)] \} \\ T_4 & \equiv \frac{1}{N} \sum_i [\rho_i - E (\rho_i | V_i)] K_i(v). \end{aligned}$$

Part b) will follow if  $E(T_l^2) = O(N^{-(1-rd)})$ ,  $l = 1, 2, 3, 4$ . The result is immediate for  $T_1$ . For  $T_2$  and any stage  $s$ :

$$\begin{aligned} \sup_v E(T_2^2) & = \left[ \frac{1}{Nh^d} \right] \sup_v \frac{1}{N} E \sum_i [\varepsilon_i^2 h^d K_i^2(v)] \\ & = \frac{1}{Nh^d} \sigma^2 \sup_v E [h^d K_i^2(v)] = O \left( \frac{1}{Nh^d} \right). \end{aligned}$$

For  $T_3$  and any stage  $s$ , since  $E \{ E (\rho_i | V_i) K_i(v) - E [\rho_i K_i(v)] \} = E \{ E (\rho_i | K_i(v) V_i) - E [\rho_i K_i(v)] \} = 0$ , from conditional independence, the expectation of cross products vanish, which implies that:

$$\sup_v E(T_3^2) = \left[ \frac{1}{Nh^d} \right] \sup_v \frac{1}{N} E \sum_i h^d \{ E (\rho_i | V_i) K_i(v) - E [\rho_i K_i(v)] \}^2 = o \left( \frac{1}{Nh^d} \right).$$

For  $T_4$ , write  $T_4^2 = ST + CPT$ , where

$$ST \equiv \frac{1}{Nh^d} \frac{1}{N} \sum_i [\rho_i - E(\rho_i|V_i)]^2 h^d K_i^2(v)$$

$$CPT \equiv \frac{1}{N^2} \sum_i \sum_{j \neq i} [\rho_i - E(\rho_i|V_i)] K_i(v) [\rho_j - E(\rho_j|V_j)] K_j(v).$$

For the squared terms (ST):

$$\sup_v E[ST] = \frac{1}{Nh^d} \sup_v E \{ [\rho_i - E(\rho_i|V_i)]^2 h^d K_i^2(v) \} = O\left(\frac{1}{Nh^d}\right)$$

For the cross-product terms (CPT), from Lemma 4, part e) :

$$\sup_v E[CPT] = \sup_v E \{ E[\rho_i - E(\rho_i|V_i)|V_i] E[\rho_j - E(\rho_j|V_j)|V_j] K_i(v) K_j(v) \} + O\left(\frac{1}{Nh^d}\right).$$

As the first component is 0, part b) of the lemma follows. ■

**Lemma 7. Mean-Square Convergence for First Derivatives.** For  $w$  in a compact subset of its support:

$$\sup_w E \left[ \left( \nabla_{\theta}^1 [\Delta_s^*(v(w; \theta))]_{\theta_0} \right)^2 \right] = O(N^{-4rs}) + O(N^{-(1-r(d+2))}).$$

**Proof.** Employing the same term decomposition as in Lemma 6, the proof for lemma 7 follows. ■

In providing results for a class of semiparametric models, we need uniform convergence results, where uniformity is taken with respect to index parameters. For this purpose, we now explicitly consider the dependence of index on the parameter vector  $\theta$ .

**Lemma 8. Uniform Convergence Rates.** Recall from D4) that  $K_i(v)$  depends on  $\theta$  via  $v(\theta) \equiv V(w; \theta)$  and  $V_i(\theta) \equiv V(W_i; \theta)$ . Under A4), with  $Z_i \equiv [v - V_i(\theta)]/h$ , for  $w, \theta$  restricted to compact sets and  $\alpha = 0, 1, 2$  :

$$a) : \sup_{w, \theta} \left| \frac{\frac{1}{N-1} \nabla_{\theta}^{\alpha} \sum_i Y_i K_i(v) - E \{ \nabla_{\theta}^{\alpha} [Y_i K_i(v)] \}}{E \{ \nabla_{\theta}^{\alpha} [Y_i K_i(v)] \}} \right| = O_p \left( N^{-\left(\frac{m}{2(m+2)} - r(\alpha+d)\right)} \right)$$

$$b) : \sup_{w, \theta} \left| \frac{\frac{1}{N-1} \nabla_{\theta}^{\alpha} \sum_i K_i(v) - E \{ \nabla_{\theta}^{\alpha} [K_i(v)] \}}{E \{ \nabla_{\theta}^{\alpha} [K_i(v)] \}} \right| = O_p \left( N^{-\left(\frac{1}{2} - r(\alpha+d)\right)} \right)$$

$$c) : \sup_{w, \theta} \left| \frac{\frac{1}{N-1} \nabla_{\theta}^{\alpha} \sum_i Z_i Z_i' K_i(v) - E \{ \nabla_{\theta}^{\alpha} [Z_i Z_i' K_i(v)] \}}{E \{ \nabla_{\theta}^{\alpha} [Z_i Z_i' K_i(v)] \}} \right| = O_p \left( N^{-\left(\frac{1}{2} - r(\alpha+d)\right)} \right)$$

$$d) : \sup_{w, \theta} \left| \frac{\frac{1}{N-1} \nabla_{\theta}^{\alpha} \sum_i Z_i' Y_i K_i(v) - E \{ \nabla_{\theta}^{\alpha} [Z_i' Y_i K_i(v)] \}}{E \{ \nabla_{\theta}^{\alpha} [Z_i' Y_i K_i(v)] \}} \right| = O_p \left( N^{-\left(\frac{1}{2} - r(\alpha+d)\right)} \right).$$

**Proof.** As the arguments for b), c), and d) are similar to but less complicated than those for a), here we provide the proof for a). Let:

$$b_i = \begin{cases} 1 & : |Y_i| > N^{\frac{1}{m+2}} \\ 0 & : \textit{otherwise} \end{cases} .$$

Write the term in a) as  $T_{1-b} + T_b$ , where:

$$\begin{aligned} T_{1-b} &\equiv \frac{1}{N} \sum_s (1 - b_i) Y_i \nabla_\theta^\alpha [K_i(v)] - E \{ Y_i \nabla_\theta^\alpha [K_i(v)] \} \\ T_b &\equiv \frac{1}{N} \sum_s b_i Y_i \nabla_\theta^\alpha [K_i(v)] - E \{ Y_i \nabla_\theta^\alpha [K_i(v)] \} . \end{aligned}$$

For  $T_{1-b}$ , from standard results in the literature:

$$\sup_\theta |T_{1-b} - E(T_{1-b})| = O_p \left( N^{-\left(\frac{1}{2} - r(\alpha+d) - \frac{1}{m+2}\right)} \right) = O_p \left( N^{-\left(\frac{m}{2(m+2)} - r(\alpha+d)\right)} \right) .$$

For  $T_b$ , the function  $K_i(V(\theta))$  incorporates trimming that can be taken as known (Lemma 1) and that bounds  $X_s$ . Then for  $x$  restricted to a compact set, it follows that

$$\sup_{\theta, i} |\nabla_\theta^\alpha [K_i(v)]| = O \left( \frac{1}{h^{\alpha+d}} \right) .$$

It then suffices to study

$$\frac{1}{h^{\alpha+d}} E[b_i | Y_i] \leq \frac{1}{h^{\alpha+d}} E[b_i]^{1/2} E[|Y_i^2|]^{1/2} ,$$

which follows from Cauchy-Schwarz. Note that:

$$\begin{aligned} E[b_i] &\leq \Pr \left( |Y_i|^m > N^{\frac{m}{m+2}} \right) \\ &\leq N^{-\left(\frac{m}{m+2}\right)} E[|Y_i|^m] = O \left( N^{-\frac{m}{m+2}} \right) . \end{aligned}$$

Therefore,

$$T_b = O \left( N^{-\left[\frac{m}{2(m+2)} - r(\alpha+d)\right]} \right) .$$

The result now follows.<sup>11</sup> ■

Finally, for establishing asymptotic normality, we need a uniform convergence lemma for the estimated mean functions and their first two derivatives. Lemma 9 provides the required results.

**Lemma 9. Uniform Convergence for Derivatives.** Under Assumption A4), for

$$0 < r < \frac{1}{2(d+\alpha)} \frac{m}{m+2}, \alpha = 0, 1, 2,$$

---

<sup>11</sup>The  $b_i$  indicator was defined to ensure that the  $T_b$ -terms and the  $T_{1-b}$ -terms converge to zero at the same rate.

we have

$$\sup_{w, \theta} |d(w, \theta)| \equiv o_p(1), \quad d(w, \theta) \equiv \nabla_{\theta}^{\alpha} \left[ \hat{M}_k^*(v(\theta); \theta) - M(v(\theta); \theta) \right].$$

**Proof.** From Lemma 8), the lemma is immediate for  $s = 1$ . For  $s \geq 2$ , the result follows from an induction argument. ■

For purposes of obtaining properties of the recursive estimator for  $M$ , above we have taken exterior  $X$ -trimming as known. For estimating index parameters in a semiparametric model, as described in Theorem 2, in the first stage we employ  $X$ -trimming based on sample quantiles. In the second stage, we trim based on the estimated index. To cover these cases, let  $\hat{\gamma}$  be either a vector of sample quantiles or a vector containing sample quantiles and the vector of index parameter estimates,  $\hat{\theta}$ .

**Lemma 10. Exterior Trimming.** Set the stage  $s$ , and window parameter  $r$ , to satisfy either condition C1) or C2) in Theorem 2. With  $\hat{q}$  as a vector of sample quantiles, denote  $\Delta_s^*(v; \hat{q})$  as the corresponding scaled estimator under estimated  $X$ -trimming and  $\Delta_s^*(v; q)$  as the corresponding estimator under known trimming. Then:

$$\begin{aligned} a) & : \sup_{v, \theta} |\{\Delta_s^*(v; \hat{q}) - \Delta_s^*(v; q)\}| = o_p(N^{-1/2}) \\ b) & : \sup_{v, \theta} |\nabla_{\theta}^1 \{\Delta_s^*(v; \hat{q}) - \Delta_s^*(v; q)\}| = o_p(N^{-1/2}) \\ c) & : \sup_{v, \theta} |\nabla_{\theta}^2 \{\Delta_s^*(v; \hat{q}) - \Delta_s^*(v; q)\}| = o_p(1). \end{aligned}$$

**Proof.** We prove the lemma first in the case of  $X$ -trimming and then show that the arguments immediately extend to index-trimming. To prove a) for  $s = 1$  and letting  $\hat{A} \equiv [Z'D^*(v, \hat{\tau})Z]/N$ ,  $A \equiv [Z'D^*(v, \tau)Z]/N$ ,  $\hat{B} \equiv Z'D^*(v, \hat{\tau})/N$ , and  $B \equiv Z'D^*(v, \tau)/N$ :

$$\begin{aligned} \Delta_1^*(v; \hat{q}) - \Delta_1^*(v; q) & = \bar{Z}(v) \left\{ \hat{A}^{-1} \hat{B} - A^{-1} B \right\} \left[ Y - \hat{M}(v) \right] \\ & = \bar{Z}(v) \left\{ \begin{array}{l} \hat{A}^{-1} \left[ A - \hat{A} \right] A^{-1} B + \\ A^{-1} \left[ \hat{B} - B \right] + \\ \hat{A}^{-1} \left[ A - \hat{A} \right] A^{-1} \left[ \hat{B} - B \right] \end{array} \right\}. \end{aligned}$$

As all terms are similar, with the third converging to zero faster than the first two, here we provide the argument for the first term given as:

$$\bar{Z}(v) \hat{A}^{-1} \left\{ \frac{1}{N} Z' [D^*(v, \hat{\tau}) - D^*(v, \tau)] Z \right\} A^{-1} \frac{1}{N} Z' B \left[ Y - \hat{M}(v) \right].$$

It can be shown that there exists  $\delta > 0$  such that  $\bar{Z}(v)$  is uniformly  $o_p(N^{-\delta})$ . Further, it can be shown that  $\hat{A}^{-1}$ ,  $A^{-1}$ , and  $\frac{1}{N} Z' B \left[ Y - \hat{M}(v) \right]$  are uniformly  $O_p(1)$ . The proof will then follow if  $\frac{1}{N} Z' [D^*(v, \hat{\tau}) - D^*(v, \tau)] Z$  is uniformly  $o_p(N^{-(1/2-\delta)})$ . Employing an inequality due to Jim Powell, Klein and Shen (2010, Proof of Theorem 2) provide a smooth upper bound to the difference in indicators. With the difference in sample and population quantiles being  $O_p(N^{-1/2})$ , a Taylor series then gives the result, with the smooth approximation slowing the



convergence rate to  $N^{-(1/2-e)}$ ,  $e > 0$  and arbitrarily small. The result then follows for  $s = 1$ . An induction argument then completes the proof for part a). The proofs for parts b) and c) are similar.

For exterior index trimming, let  $\hat{\gamma} \equiv (\hat{q}_1, \hat{q}_2, \hat{\theta})$ , where  $(\hat{q}_1, \hat{q}_2)$  are lower and upper sample quantiles for the estimated index,  $X_{1i} + X_{2i}\hat{\theta}$ . Write

$$\begin{aligned} \tau_i(\hat{\gamma}) &= 1 \left\{ \hat{q}_1 < X_{1i} + X_{2i}\hat{\theta} < \hat{q}_2 \right\} \\ &= 1 \left\{ \hat{q}_1 - X_{2i}(\hat{\theta} - \theta_0) < X_{1i} + X_{2i}\theta_0 < \hat{q}_2 - X_{2i}(\hat{\theta} - \theta_0) \right\} \\ &\equiv 1 \left\{ \hat{c}_1 < X_{1i} + X_{2i}\theta_0 < \hat{c}_2 \right\} \equiv 1 \left\{ \hat{c}_1 < V_{0i} < \hat{c}_2 \right\}. \end{aligned}$$

where  $V_{0i} \equiv X_{1i} + X_{2i}\theta_0$ . Recalling from A6) that  $X_{2i}$  is bounded and that  $|\hat{\theta} - \theta_0| = O_p(N^{1/2})$ , the proofs for index trimming are very similar to those for  $X$ -trimming. ■