Econometric Theory, **40**, 2024, 37–59.
doi:10.1017/S0266466622000329

RECURSIVE DIFFERENCING FOR ESTIMATING SEMIPARAMETRIC MODELS

CHAN SHEN
Penn State University

ROGER KLEIN
Rutgers University

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, nonoptimal 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.

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, Stock, and Stoker (1989), Ichimura and Lee (1991), Ichimura (1993), Klein and Spady (1993), Horowitz (1996), Li and Sun (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, for a class

We thank the seminar participants at Columbia University and New York University for helpful comments and suggestions. We also thank the Editor and referees for their insightful comments and suggestions. The authors are solely responsible for any errors. Address correspondence to Chan Shen, Departments of Surgery and Public Health Sciences, Penn State University, College of Medicine, Hershey, PA, USA; e-mail: chanshen@gmail.com.

of models estimated by semiparametric least squares (SLS) (Ichimura and Lee (1991); Ichimura (1993)), we employ recursive differencing to obtain asymptotic normality in multiple-index models. When we combine recursive differencing with an adjustment to utilize a residual property of semiparametric derivatives, we obtain asymptotic normality under optimal windows. Throughout this paper, we use the term optimal window to mean that we equate the orders of the squared bias and variance of the estimator of interest.

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 $N^{-1/2}$, whereas 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 Honoré and Powell (2005). In other cases, it is possible to employ different estimators for expectations conditioned on indices. To accommodate higher dimensions of the indices, higher-order kernels (HKs; e.g., Müller, 1984) increase the bias order by increasing the degree of the kernel; an extension of Newey et al. (2004) can similarly control the bias by increasing the convolution degree.

These methods perform reasonably well for single-index models. However, there are many instances when multiple-index models are required. For example, in many index models, one of the variables in a main index of interest is endogenous. Control estimators deal with endogeneity by conditioning not only on the model's index, but also on the control (e.g., Blundell and Powell (2003, 2004)). In such models, the control itself becomes a second index. The estimator for joint binary models, as in Klein et al. (2015), requires a multiple index formulation. Shen (2013) examines a multiple equation system for healthcare expenditures and related decisions, which results in a multiple index formulation. Maurer, Klein, and Vella (2011) examine a panel model where an unobserved individual effect is modeled as a separate and additional index to that in the main part of the model.

When the degree of bias reduction is increased to accommodate multiple index semiparametric models, the finite-sample variability of these estimators becomes large. Furthermore, both approaches require suboptimal windows. Local polynomials (LPs; e.g., Ruppert and Wand (1994); Fan and Gijbels (1995, 1996); Lu (1996); Masry (1996); Gu, Li, and Yang (2015)) obtain this bias order by increasing the degree of the LP. While the performance of the local linear estimator is quite good, higher degrees require more local parameters and hence higher variability. In this paper, we propose alternative approaches for bias reduction that enable us to obtain $\sqrt{N}$ normality in semiparametric multiple-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

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

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.

Employing the estimator for semiparametric expectations developed here, we provide two approaches for estimating a class of semiparametric models. One relies on recursive differencing as the sole bias control. Under additional assumptions, the other estimator takes advantage of this recursive mechanism and a residual property of semiparametric derivatives. Klein and Shen (2010) examine this second control for single-index models. Using recursive differencing, we are able to extend this approach to multiple-index models. In so doing, we obtain asymptotic normality 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 estimators had very good finite-sample performance in terms of both bias and variance. In all of the cases, the root-mean-squared error (RMSE) decreased, often substantially, with the recursion stage. Both estimators also performed much better than either an HK estimator or an LP whose bias order is below $N^{-1/2}$. Overall, we found a performance advantage to employing optimal windows.

To develop the proposed estimators for a conditional expectation and the parameters in a semiparametric model, 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 discusses how to implement the proposed estimators. Section 5 provides Monte Carlo results that demonstrate very good finite-sample properties of the estimators in triple-index models, exhibiting a substantial improvement over regular, HK, and LP 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 6 contains our conclusions. The Supplementary Material provides further details containing proofs of all theorems and supporting intermediate lemmas.

2. ESTIMATORS

2.1. Estimating Expectations Under Recursive Differencing

The semiparametric model that we study assumes that

$$E(Y_i|W_i) = E\left(Y_i|V\left(W_i;\theta_0\right)\right), \tag{1}$$

where the vector $\{Y_i, W_i\}$ is i.i.d. over $i = 1, \ldots, N$, and takes on values in R^{1+d_w} with d_w the dimension of W_i. Here, $V\left(W_i;\theta_0\right)$ is a vector of $d < d_w$ continuous indices that depend on a finite-dimensional parameter vector, θ_0.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

To motivate the form of the bias reduction, assuming that ε_i is an error satisfying $E(\varepsilon_i|W_i) = 0$, consider the model

$$Y_i = E\left(Y_i|V\left(W_i;\theta_0\right)\right) + \varepsilon_i.$$

For expositional purposes, in this section, we take the parameter vector as known and discuss its estimation in the next section. Let $V_i = V(W_i;\theta_0)$ and $M(V_i) = E(Y_i|V_i)$. Then, an often employed conditional expectation estimator is given as

$$\hat{M}(v) \equiv \frac{\frac{1}{N}\sum_i Y_i K_i(v)}{\hat{g}(v)},\ \hat{g}(v) \equiv \frac{1}{N}\sum_i K_i(v), \tag{2}$$

where $K_i(v)$ is a kernel weight; for example, in the single-index case, $K_i(v) \equiv \frac{1}{h}\phi\left(\frac{v-V_i}{h}\right)$, where $h = N^{-r}$ is the bandwidth and ϕ is a standard normal density function. Define the localization error as

$$e^*(v) \equiv \frac{\frac{1}{N}\sum_i \left[M\left(V_i\right) - M\left(v\right)\right] K_i(v)}{\hat{g}(v)}.$$

An infeasible bias-corrected estimator would then be given as[1]

$$\begin{aligned}
\tilde{M}^*(v) &\equiv \hat{M}(v) - e^*(v) \\
&= \frac{\frac{1}{N}\sum_i \left\{Y_i - \left[M\left(V_i\right) - M\left(v\right)\right]\right\} K_i(v)}{\hat{g}(v)} \\
&= \frac{\frac{1}{N}\sum_i \left\{M(v) + \left[M(V_i) - M(v)\right] + \varepsilon_i - \left[M\left(V_i\right) - M\left(v\right)\right]\right\} K_i(v)}{\hat{g}(v)} \\
&= M(v) + \frac{\frac{1}{N}\sum_i \varepsilon_i K_i(v)}{\hat{g}(v)}.
\end{aligned}$$

Since $E(\varepsilon_i|W_i) = 0$, the last term has zero expectation and the estimator is unbiased. As $e^*(v)$ is unknown, define the prediction error as

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

Then, a feasible bias-corrected estimator is given as

$$\begin{aligned}
\tilde{M}(v) &\equiv \hat{M}(v) - e(v). \\
&= \frac{\frac{1}{N}\sum_i \left\{Y_i - \left[\hat{M}\left(V_i\right) - \hat{M}\left(v\right)\right]\right\} K_i(v)}{\hat{g}(v)}.
\end{aligned}$$

[1] We thank a referee for suggesting this intuition.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

Replacing $\hat{M}(V_i) - \hat{M}(v)$ by $\hat{M}_{s-1}(V_i) - \hat{M}_{s-1}(v)$, then, leads to the recursive estimator at stage s where $s \geqq 2$:

$$\hat{M}_s(v) \equiv \frac{\frac{1}{N}\sum_i \left[Y_i - (\hat{M}_{s-1}(V_i) - \hat{M}_{s-1}(v))\right] K_i(v)}{\hat{g}(v)}. \tag{3}$$

From the above discussion, it is intuitive that this adjustment lowers the bias. We prove that the bias order declines over stages with order at stage s given by $O(N^{-2rs})$, where r is the kernel window parameter.

To start this recursion, we note that the theory will hold for any initial estimator that satisfies certain convergence properties. Here, we employ a modified local linear estimator as it simplifies the bias arguments; furthermore, in Monte Carlo simulations, it performed noticeably better than the local constant estimator in (2) and similar to the local linear estimator.

To describe the initial estimator, for observation i, let V_i be a row vector of the model's d continuous indices and let v be a conformable row vector of fixed values. With $Z_i \equiv \frac{V_i - v}{h}$, the local linear estimator solves

$$\hat{M}_L, \hat{M}'_L \equiv \arg\min_{M, M'} \sum_i \left[Y_i - M(v) - hZ_i M'(v)\right]^2 K_i(v) \tag{4}$$

$$\Rightarrow \hat{M}_L = \bar{Y}(v) - h\bar{Z}(v)\hat{M}'_L, \tag{5}$$

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

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

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

$$\hat{M}'_m \equiv \arg\min_{M'} \sum_i \left[Y_i - \bar{Y}(v) - hZ_i M'(v)\right]^2 K_i(v)$$

$$= \frac{1}{h}[Z'D(v)Z]^{-1}\left[Z'D(v)\right]\left[Y - \underline{1} \cdot \bar{Y}(v)\right],$$

where 1 is a vector of ones, and $D(v) \equiv diag(K_i(v))$. The modified local linear estimator is, then, defined as

$$\hat{M}_1(v) \equiv \bar{Y}(v) - h\bar{Z}(v)\hat{M}'_m = \bar{Y} - \bar{Z}[Z'D(v)Z]^{-1}\left[Z'D(v)\right]\left[Y - \underline{1} \cdot \bar{Y}(v)\right]. \tag{6}$$

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

In addition to the recursive differencing structure, we also propose an extra mechanism to reduce the bias in estimating index parameters in semiparametric models. Combining these mechanisms, we will be able to estimate a wide class

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

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.[2]

PROPOSITION 1 *Assume that* $E(Y|W) = M[V(\theta_0); \theta_0]$ *from the index assumption in (1), and let* $\delta_i(\theta) \equiv \nabla_\theta E(Y|V(\theta))$. *Then,*

$$E[\delta_i(\theta_0)|V(\theta_0)] = 0.$$

The estimators that we consider here are variants of SLS, with conditional expectations estimated under recursive differencing. We select estimates so as to minimize an SLS objective function of the form

$$\hat{Q}(\theta) \equiv \frac{1}{N}\sum_i \tau_i \left[Y_i - \hat{M}_s(V_i(\theta))\right]^2,$$

where τ_i is a trimming function and $\hat{M}_s$ is the stage-s recursive differencing estimator for the conditional expectation of Y_i given in (3). Under recursive differencing, we show that the gradient to the objective function asymptotically can be written as $A - B + o_p(N^{-1/2})$, where

$$\sqrt{N}A \equiv \frac{1}{\sqrt{N}}\sum_i \{Y_i - M[V_i(\theta_0)]\}\tau_i\delta_i(\theta_0),$$

$$\sqrt{N}B \equiv \frac{1}{\sqrt{N}}\sum_i \left\{\hat{M}_s[V_i(\theta_0)] - M[V_i(\theta_0)]\right\}\tau_i\delta_i(\theta_0).$$

The A-term is straightforward as it is an i.i.d. normalized sum with expectation 0. The B-term is more difficult to analyze and is the source of the bias. Employing an index assumption and omitting technical details, in essence, we let

$$F(V_0) \equiv E\left\{\hat{M}_s[V_i(\theta_0)] - M[V_i(\theta_0)]|W\right\}.$$

Then,

$$E(B) = EE(B|W) = E\{F(V_0)\tau_i\delta_i(\theta_0)\} = E\{F(V_0)E[\tau_i\delta_i(\theta_0)|V_i(\theta_0)]\}.$$

Although $E(\delta_i(\theta_0)|V_i(\theta_0)) = 0$, it is not possible to exploit this residual property with regular trimming. Here, we employ trimming that is asymptotically equivalent to trimming on the true index, $V_i(\theta_0)$.

Recursive differencing makes it possible to take the derivative $\delta_i(\theta_0)$ as known, which is required for implementing the residual control. Given an initial $\sqrt{N}$-normally distributed estimator $\hat{\theta}_1$,[3] we can exploit the residual property by using a similar strategy as in Klein and Shen (2010) combined with recursive differencing.

[2] See Klein and Shen (2010) for Newey's proof of this property.

[3] A specific initial estimator is provided in (D7) in Section 3.1).

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

Specifically, we can trim based on $V_i(\hat{\theta}_1)$, which we show is asymptotically equivalent to trimming based on $V_i(\theta_0)$. Then, $E(B) = 0$.

Consistency requires that the density for $V_i(\theta)$ does not converge to 0 "too fast" for θ away from θ_0. Unfortunately, the above trimming only controls $V_i(\theta)$ when θ is close to θ_0. To solve this problem, we follow the strategy in Klein and Shen (2010) and Klein et al. (2015) in which we adjust density denominators so as to control the rate at which they converge to 0. The adjusted density has the form

$$\hat{g}_a(v;\theta) = \hat{g}(v;\theta) + \hat{A}(v;\theta).$$

The adjustment term $\hat{A}(v;\theta)$ is set to vanish very slowly when v is close to its boundary where the density is going to zero and to rapidly vanish for v in the interior. In this manner, the rate at which the density tends to zero is controlled, and we preserve the consistency argument without impacting the asymptotic normality argument for the gradient.

By exploiting the residual control, we are able to establish asymptotic normality for the adjusted estimator, $\hat{\theta}_2$, with optimal windows. Here, it should be noted that the optimal window, r^*, depends on the stage of the recursion and is obtained by equating squared-bias and variance orders. In particular, with d as the number of indices and s as the recursion stage, $r^* = \frac{1}{4s+d}$. We discuss selection of the number of stages in Section 4 on implementation.

3. LARGE-SAMPLE RESULTS

3.1. Definitions and Notations

To establish large-sample results for the recursive differencing estimator and for estimators of a class of semiparametric models, we require definitions and notations. Of particular note, we will provide both the initial ($\hat{\theta}_1$) and the adjusted ($\hat{\theta}_2$) estimators for the index parameters, each of which will require separate definitions for the conditional expectation estimators (D6 and D9) and separate corresponding trimming (D3).

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

(D2) Conditional Expectations. $M(v) \equiv E[Y_i | V(W_i;\theta) = v]$.

(D3) Trimming. For a continuous random variable T_k, denote q_{1k} and q_{2k} as its lower and upper population quantiles. With trimming based on a set that slowly expands to the full support for T_k, we use the notation B for the case where the lower support point a_k is bounded and $\bar{B}$ when the upper support point b_k is bounded. Let U refer to the unbounded support case. Then, for

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

an arbitrarily small $\vartheta > 0$ and with c_N expanding slowly, let

$$q_{1k}(N) \equiv \begin{cases} \frac{q_{1k}+a_k\vartheta\ln(N)}{\vartheta\ln(N)+1} & : \underline{B}, \\ \frac{q_{1k}-c_N\vartheta\ln(N)}{\vartheta\ln(N)+1} & : U, \end{cases} \quad q_{2k}(N) \equiv \begin{cases} \frac{q_{2k}+b_k\vartheta\ln(N)}{\vartheta\ln(N)+1} & : \bar{B}, \\ \frac{q_{2k}+c_N\vartheta\ln(N)}{\vartheta\ln(N)+1} & : U. \end{cases}$$

With T_{ik} as the ith observation on T_k, define indicator trimming for T_{ik} as: $\tau(T_{ik}; q_k(N)) \equiv 1\{T_k : q_{1k}(N) < T_{ik} < q_{2k}(N)\}$. Define a corresponding smooth trimming function approximating the above as

$$\tau_{sm}(T_{ik}; q_k(N)) \equiv \left\{1 + \exp\left(-(\ln N)^2[T_{ik} - q_{1k}(N)]\right)\right\}^{-1} \left\{1 + \exp\left(-(\ln N)^2[q_{2k}(N) - T_{ik}]\right)\right\}^{-1}.$$

Let $\tau(T_i; q(N)) \equiv \prod_k \tau(T_{ik}; q_k(N))$ and $\tau_{sm}(T_i; q(N)) \equiv \prod_k \tau_{sm}(T_{ik}; q_k(N))$. For $\vartheta = 0$, we define fixed indicator trimming as $\tau(T_{ik}; q_k) \equiv 1\{T_k : q_{1k} < T_{ik} < q_{2k}\}$, and let $\tau(T_i; q) \equiv \prod_k \tau(T_{ik}; q_k)$.

(D4) Kernel. Let v and V_i be d-dimensional vectors with lth elements as $v(l)$ and $V_i(l)$, respectively, and denote s_l as the standard deviation (SD) of $V_i(l)$.[4] Referring to (D3), define

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

$$k(z) \equiv \prod_{l=1}^{d} \phi(z(l)),\ K_i^*(v) \equiv \tau_{sm}(T_i, \hat{q}(N))K_i(v),$$

where z is an arbitrary vector with lth component $z(l)$, $h = N^{-r}$, $0 < r < \frac{1}{2d}$, $\phi(\cdot)$ is a density symmetric about 0 with finite moments of all orders, and $\hat{q}(N)$ depends on sample quantiles. When evaluated at a data point V_i, we set $K_{jl}(V_i)$ to zero for $i = j$.

(D5) Kernel Averages. Referring to (D4), for stage s, define

$$\hat{g}_s(v) \equiv \begin{cases} \hat{g}_1(v) \equiv \frac{1}{N}\sum_{i=1}^{N} K_i(v), & s = 1, \\ \hat{g}_2(v) \equiv \frac{1}{N}\sum_{i=1}^{N} K_i^*(v), & s > 1. \end{cases}$$

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

(D6) Unadjusted Conditional Expectation Estimator. Let D be the $N \times N$ diagonal matrix with jth element $K_j(V_i)$. Let Z be an $N \times d$ matrix with jth row $(V_i - V_j)/h$. Referring to (6), with $\bar{Y}(V_i)$ and $\bar{Z}(V_i)$ depending on $\hat{g}_1(V_i)$, for stage $s = 1$,

$$\hat{M}_1(V_i) \equiv \bar{Y}(V_i) - \bar{Z}(V_i)\left[Z'DZ\right]^{-1} Z'D\left[Y - \underline{1} \cdot \bar{Y}(V_i)\right],$$

[4] It can be shown that the SD can be taken as known.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

where $E\left[\frac{Z'DZ}{N-1}\right]$ is positive definite. For stage $s > 1$, refer to (D3) and set $T_k \equiv X_k$ so that τ_{sm} represents smooth X-trimming here. Then,

$$\hat{M}_s(V_i) \equiv \frac{\frac{1}{N-1}\sum_{j\neq i}\left\{Y_j - \left[\hat{M}_{s-1}(V_j) - \hat{M}_{s-1}(V_i)\right]\right\}\tau_{sm}(X_j;\hat{q}(N))K_j(V_i)}{\hat{g}_s(V_i)}.$$

(D7) Initial Index Parameter Estimator. Let $(\hat{q}'_{1k}, \hat{q}'_{2k})$ and $(\hat{q}_{1k}, \hat{q}_{2k})$ be vectors of sample quantiles for $X_{ik}, \hat{q}_{1k} < \hat{q}'_{1k} < \hat{q}'_{2k} < \hat{q}_{2k}$. Then, with $\hat{q}'_{1k}, \hat{q}'_{2k}$ replacing q_{1k}, q_{2k}, define

$$\hat{\theta}_1 \equiv \arg\min_{\theta} \hat{Q}_1(\theta), \tag{7}$$

$$\hat{Q}_1(\theta) \equiv \frac{1}{N}\sum_{i=1}^{N}\tau(X_i;\hat{q}')\left\{Y_i - \hat{M}_s[V_i(\theta)]\right\}^2.$$

Referring to (D3), set $T_i \equiv X_i$ so that τ represents indicator X-trimming here.

(D8) Adjusted Densities. To adjust the density estimators in (D4), let $\hat{\gamma}_s$ be a lower sample quantile for $\hat{g}_s(V_j(\hat{\theta}_1))$ and $\tau_\Delta(\hat{g}_s(v))$ a smooth trimming function with the following form:

$$\tau_\Delta(\hat{g}_s(v)) \equiv \left\{1 + \exp\left(-(\ln N)^2[\hat{g}_s(v) - N^{-ar}]\right)\right\}^{-1}.$$

Then, define an estimated adjustment factor as

$$\hat{A}_s(v) \equiv \hat{\gamma}_s N^{-ar}\left[1 - \tau_\Delta(\hat{g}_s(v))\right].$$

Referring to the definition of ϑ in (D3), set the adjustment parameter a to satisfy $\vartheta < a < \frac{2}{s}$. With $1 - \tau_\Delta(\hat{g}_s(v))$ approaching 1 in the tails for $\hat{g}_s(v)$, adjusted estimated densities are then defined as

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

(D9) Adjusted Conditional Expectations Estimator. Let

$$\bar{Y}_a = \frac{\hat{g}_1(V_i)}{\hat{g}_{1a}(V_i)}\bar{Y},\ \bar{Z}_a = \frac{\hat{g}_1(V_i)}{\hat{g}_{1a}(V_i)}\bar{Z}.$$

Refer to (D3) and set $T_i \equiv V_i(\hat{\theta}_1)$, where $\hat{\theta}_1$ is the estimator in (D7), so that τ_{sm} represents smooth index trimming here. Define

$$\hat{M}_{1a}(V_i) \equiv \bar{Y}_a(V_i) - \bar{Z}_a(V_i)\left[Z'\hat{D}_a Z\right]^{-1} Z'D\left[Y - \underline{1}\cdot\bar{Y}(V_i)\right],$$

$$\hat{D}_a = D + \hat{A}_1(v).$$

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

For $s > 1$, define $\hat{M}_{sa}(V_i)$ as:

$$\frac{\frac{1}{N-1}\sum_{j\neq i}\left\{Y_j - \left[\hat{M}_{(s-1)a}(V_j(\theta)) - \hat{M}_{(s-1)a}(V_i(\theta))\right]\right\}\tau_{sm}(V_j(\hat{\theta}_1);\hat{q}(N))K_j(V_i(\theta))}{\hat{g}_{sa}(V_i(\theta))}.$$

(D10) Adjusted Index Parameter Estimator. Referring to (D7), let $(\hat{q}'_{1k},\hat{q}'_{2k})$ and $(\hat{q}_{1k},\hat{q}_{2k})$ be vectors of sample quantiles for $V_i(\hat{\theta}), \hat{q}_{1k} < \hat{q}'_{1k} < \hat{q}'_{2k} < \hat{q}_{2k}$. Define

$$\hat{\theta}_2 \equiv \arg\min_{\theta} \hat{Q}_2\left(\theta;\hat{\theta}_1\right), \tag{8}$$

$$\hat{Q}_2\left(\theta;\hat{\theta}_1\right) \equiv \frac{1}{N}\sum_{i=1}^{N}\tau\left(V_i\left(\hat{\theta}_1\right),\hat{q}'\right)\left\{Y_i - \hat{M}_{sa}\left[V_i(\theta)\right]\right\}^2.$$

Referring to (D3), set $T_i \equiv V_i(\hat{\theta}_1)$ so that τ represents indicator index trimming.

While the purpose of most of these definitions is clear, further discussion of kernel windows, trimming, and adjustment strategies can be useful. Throughout, we take the SDs of the indices as known, because consistency for our index parameter estimator does not depend on the consistency of the estimator for this SD (see Newey and McFadden, 1994).

The estimators above depend on different types of trimming, which we provide in (D3). Within $\hat{M}_{sa}$ and $\hat{M}_s$, for $s > 1$, trimming sequences are on sets that expand to the full support of the continuous variables as the sample size increases. For the bounded cases (B and $\bar{B}$), we can interpret the trimming as a weighted average of the population quantiles and support points. For example, in the lower bound (B) case, we could write it as $q_1 \cdot (1 - w_N) + a_k \cdot w_N$, where the weight $w_N \equiv \frac{\vartheta \ln(N)}{[\vartheta \ln(N)+1]}$ slowly approaches one ensuring that the density slowly converge to 0. The $\ln(N)$ function is a commonly employed slowly increasing function, increasing slower than N^{α} for any $\alpha > 0$. For the unbounded case, a slowly expanding sequence c_N plays a similar role as the support point.[5] We view the quantiles as the reference or anchoring points from which we slowly depart as the sample size increases.

In (D4), we provide conditions for the regular kernel that we employ. Note the restriction on the window parameter, r: $0 < r < \frac{1}{2d}$. This condition, which is commonly satisfied in the literature on semiparametric models, is important for Lemma 3 on kernel products. It ensures an asymptotic independence property which facilitates the calculation of an expectation of a product of terms involving averages of kernel functions. Definitions of kernel averages are provided in (D5), which are needed in the construction of the estimators.

[5]For example, with $c_N = \sqrt{\delta \ln(N)}$, we can accommodate any density with tails no thinner than that for a normal density.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

Because of the recursive structure of the estimator for conditional expectations in (D6), the estimator at stage s will depend on the previous stage. Namely, $\hat{M}_s(V_i)$ depends on the bias correction term $\hat{M}_{s-1}(V_j) - \hat{M}_{s-1}(V_i)$ from stage $s-1$. Notice that this bias correction term is evaluated at both observations i and j. Therefore, we need smooth trimming within the $\hat{M}$-function to control the density denominator at j in addition to the indicator trimming controlling i in the objective $\hat{Q}$-function. We need indicator trimming in the $\hat{Q}$-function to be on a narrower set compared to the smooth trimming in the $\hat{M}$-function to ensure enough V_j observations in a neighborhood of V_i to avoid a bias.

As discussed in Section 2.2, the second estimator, $\hat{\theta}_2$, is implemented so as to take advantage of Newey's residual result for bias control. Trimming sequences based on X have the desirable property that they do not depend on the unknown parameters. However, as discussed earlier, there is a problem in exploiting Newey's residual result when trimming is not based on the true index. Accordingly, for $\hat{\theta}_2$, all trimming is based on the estimated index obtained from the initial estimator $\hat{\theta}_1$. We show that such trimming is asymptotically equivalent to trimming on the true index.

This strategy does not control density denominators when evaluated at θ not in a neighborhood of θ_0, which is important for consistency. As previously explained we adjust density denominators so as to control how fast they converge to 0. (D8) defines this adjustment to have the following desirable properties. When the density for $V(\theta)$ is N^{-ar} or smaller, the adjustment ensures that the density reciprocal is $O(N^{ar})$, with $a < \frac{2}{s}$. When studying the gradient evaluated at θ_0, index trimming ensures that the adjustment rapidly approaches 0 at a rate faster than $O(1/N)$ as is required for normality. Employing this adjustment mechanism, (D9) defines a corresponding adjusted expectations estimator, and the adjusted parameter estimator is defined in (D10).

3.2. Assumptions

The main assumptions underlying theoretical results are shown below. Assumptions that may be required for one estimator but not for another are provided directly in theorems.

(A1) The vector $\{Y_i, W_i\}$ is *i.i.d.* over $i = 1, \ldots, N$, and takes on values in R^{1+d_w}, where d_w is the dimension of W_i.

(A2) The following index restriction holds:

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

With $\varepsilon_i \equiv Y_i - E(Y_i|W_i)$, assume that $\sigma_\varepsilon^2 \equiv Variance(\varepsilon_i|W_i)$ is constant.[6]

(A3) Refer to (D1), and assume that $\theta \epsilon \Phi$, a compact set. Let $g(v(w;\theta);\theta)$ be the density for $V(W_i;\theta)$ evaluated at $v(w;\theta)$. Let $\alpha = 0, 1, 2$ and

[6]Note that this is a restrictive homoscedasticity assumption.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

$\beta = 0, 1, \ldots, 2s+1$, where s is the stage. Let $\varepsilon^* \equiv 0$ under fixed trimming and $0 < \varepsilon^* < ar$ from (D8) under expanding trimming. Recall from (D3) that with $T_i \equiv V_i$ or X_i, $\tau^*(T_i; q(N))$ constrains T_i to a slowly expanding set. In either case, for $\theta \epsilon \Phi$, V_i is constrained to a slowly expanding set $\mathfrak{C}_N^*$. Define $\mathfrak{A}_v$ and $\mathfrak{A}_x$ as compact subsets of the supports for V and X, respectively, and let $\mathfrak{B}$ be $\mathfrak{C}_N^*$ or $\mathfrak{A}_v$ Then, with $\mathfrak{S}$ as the support for V,

$$(a): \inf_{v \epsilon \mathfrak{B}} |g(v;\theta_0)| \text{ and } \inf_{v \epsilon \mathfrak{B}, \theta \epsilon \Phi} |g(v;\theta)| = O(N^{-\varepsilon^*}),$$

$$(b): \sup_{v \epsilon \mathfrak{S}} \left|\nabla_v^\beta [M(v;\theta_0) g(v;\theta_0)]\right| \text{ and } \sup_{v \epsilon \mathfrak{S}} \left|\nabla_v^\beta g(v;\theta_0)\right| = O(1),$$

$$(c): \sup_{w(x \epsilon \mathfrak{A}_x), \theta \epsilon \Phi} \left|\nabla_\theta^\alpha \nabla_v^\beta [M(v(w;\theta);\theta)]\right| \text{ and}$$

$$\sup_{w(x \epsilon \mathfrak{A}_x), \theta \epsilon \Phi} \left|\nabla_\theta^\alpha \nabla_v^\beta g(v(w;\theta);\theta)\right| = O(1).$$

(A4) We assume that Y and each variable in X have a density whose tails are smaller than that of a t-distribution with, respectively, $m_y + 1$ and $m_x + 1$ degrees of freedom. With these variables having m_y and m_x moments, we assume that $m_y, m_x > 4$.

The first two assumptions are standard for index models. When trimming expands to the full support of the continuous variables, we need to construct a trimming sequence that controls how fast the density approaches. For example, when the continuous X's are jointly distributed as normal, the index will follow a normal density. Assumption (A3) ensures that density denominators do not converge to zero too fast. Assumption (A4) is useful for obtaining uniform convergence results for functions of unbounded random variables.

3.3. Theorems

Theorem 1 provides the properties of the expectation estimator whereas Theorem 2 obtains the asymptotic properties of the two proposed parameter estimators in a class of multiple-index models.

Convergence properties for the recursive differencing estimator are important for obtaining $\sqrt{N}$-normality for a finite-dimensional parameter vector in semiparametric models. These properties are also useful for obtaining asymptotic properties of marginal effects. To provide these results, we show that an approximating recursion is close to the original one and then focus on the properties of the approximation. The definition below provides the approximation.

Definition 1 (Recursion approximation). Recall the definition of the conditional expectation estimator in (D6) and kernel functions in (D4) and (D5). With $D \equiv diag(K(v))$, $\mathcal{L}$, and $\mathcal{P}$ positive integers, $V_i \equiv V(W_i;\theta_0)$, and $Z_i(v) \equiv \frac{V_i - v}{h}$ as the ith

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

row of Z, define

$$\hat{A} \equiv \frac{1}{N}\left[Z(v)' DZ(v)\right];\ A \equiv E(\hat{A});\ \delta_A(v) \equiv \sum_{l=1}^{\mathcal{L}}\left[(A-\hat{A})A^{-1}\right]^l;$$

$$\bar{g}_s(v) = E(\hat{g}_s(v)); \delta_{g_s} \equiv \sum_{p=1}^{\mathcal{P}}\left[\frac{\bar{g}_s(V_i)-\hat{g}_s(V_i)}{\bar{g}_s(V_i)}\right]^p;$$

$$\hat{d}(v) \equiv \left[Z'DZ\right]^{-1} Z'D\left[Y - \underline{1}\cdot\bar{Y}(v)\right];$$

$$\hat{d}^*(v) \equiv A^{-1}[I+\delta_A(v)]\frac{1}{N}\sum_{i=1}^{N}\frac{Z_i(v)\left[Y_i-\bar{Y}(v)\right]\hat{g}_1(v)K_i^*(v)}{\bar{g}_1(v)}\left[1+\delta_{g_1}(v)\right].$$

With $\Delta_s(v) \equiv \hat{g}_s(v)[\hat{M}_s(v) - M(v)]$, we have

$$\Delta_1(v) = \frac{\hat{g}_2(v)}{\hat{g}_1(v)}\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}[M(V_i)-M(v)+\varepsilon_i]K_i(v) \\ -\frac{1}{N}\sum_{i=1}^{N} Z_i(v)K_i(v)\hat{d}(v) \end{bmatrix},$$

$$\Delta_s(v) \equiv \Delta_{s-1}(v) - \sum_{i=1}^{N}\frac{\Delta_{s-1}(V_i)}{\hat{g}_2(V_i)}K_i^*(v) + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i K_i^*(v),\ s>1.$$

The approximating recursion is given as

$$\Delta_1^*(v) \equiv \frac{\hat{g}_2(v)}{\bar{g}_1(v)}[1+\delta_{g_{s-1}}(V_i)]\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}[M_1(V_i)-M_1(v)+\varepsilon_i]K_i(v) \\ -\frac{1}{N}\sum_{i=1}^{N} Z_i(v)K_i(v)\hat{d}^*(v) \end{bmatrix},$$

$$\Delta_s^*(v) \equiv \Delta_{s-1}^*(v) - \sum_{i=1}^{N}\left[\frac{\Delta_{s-1}^*(V_i)}{\bar{g}_2(V_i)}\right][1+\delta_{g_2}(V_i)]K_i^*(v) + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i K_i^*(v), s>1.$$

Employing this approximating recursion, Theorem 1 provides properties of the recursive differencing estimator for conditional expectations.

THEOREM 1 (The recursive differencing estimator). *Assume (A1)–(A4), with window parameter, $r, 0 < r < \frac{1}{2d}$, v and θ in compact sets, and $\Delta_s^*(v)$ defined above:*

$$(a): \sup_v\left|\Delta_s^*(v) - \Delta_s(v)\right| = o_p(N^{-\frac{1}{2}}),$$

$$(b): \sup_v\left|E\Delta_s^*(v)\right| = O(h^{2s}) + o(N^{-\frac{1}{2}}),$$

$$(c): \sup_v Var\left[\Delta_s^*(v)\right] = O\left((Nh^d)^{-1}\right),$$

$$(d): \left|\hat{M}_s(v) - M(v)\right| = O_p\left(h^{2s}\right) + O_p((Nh^d)^{-\frac{1}{2}}).$$

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

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

difficult to study directly. We provide an approximating sequence $\Delta_s^*(v)$ that from (a) is uniformly within $o_p(N^{-\frac{1}{2}})$ 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. Employing variants of SLS introduced by Ichimura and Lee (1991) and Ichimura (1993), Theorem 2 provides these conditions for two estimators in a class of multiple-index models. The first estimator employs recursive differencing as the sole bias control under X-trimming and makes no assumptions on the X-variables being bounded. This estimator is important in part because it makes weaker assumptions on the existence of X-moments. Furthermore, the other estimator in Theorem 2 depends on it. There are also moment-based estimators where the residual control is not applicable (e.g., semiparametric instrumental variable estimators). The second estimator is asymptotically distributed as normal under optimal windows when recursive differencing is combined with the residual control.

THEOREM 2 (Estimating index parameters). *Under assumptions (A1)–(A4), with $\gamma \equiv \frac{m_y}{m_y+1} - \frac{2}{m_x+1}$, set the stage s and window parameter r to satisfy*[7]

$$(C1): s > \frac{(d+2)}{2\gamma},\ \frac{1}{4s} < r < \frac{\gamma}{2(d+2)}. \tag{9}$$

With $\delta_i(\theta_0) \equiv \nabla_\theta [M(V(W_i;\theta);\theta)]_{\theta_0}$ and trimming function τ in (D3), let

$$H_1 \equiv E\left[\tau(X_i;q')\delta_i\delta_i'\right],$$
$$G_{1i} \equiv \tau(X_i;q')\delta_i - E(\tau(X_i;q')\delta_i|V_i).$$

With $\Sigma_1 \equiv \sigma_\varepsilon^2 H_1^{-1} E\left(G_1 G_1'\right) H_1^{-1}$ and $\sigma_\varepsilon^2 \equiv Var(\varepsilon_i)$,

$$(a): \hat{\theta}_1 \xrightarrow{p} \theta_0,\ \sqrt{N}\left[\hat{\theta}_1 - \theta_0\right] \xrightarrow{d} Z_1 \sim N(0, \Sigma_1).$$

For the second-step estimator $\hat{\theta}_2$, assume that each variable in X is bounded. Set the stage s, window parameter r, and adjustment parameter a to satisfy[8]

$$(C2): s > \frac{d+4}{4} + \frac{d+2}{2m_y}, \tag{10}$$

$$r = \frac{1}{4s+d} < \frac{1}{2[d+2]}\frac{m_y}{m_y+1},\ 0 < a < \frac{2}{s}.$$

[7] 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 nonempty.

[8] The upper bound for r follows from uniform convergence for second derivatives. We set s to ensure that this condition on r can hold.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

With $\Sigma_2 \equiv \sigma_\varepsilon^2 E\left[\tau\left(V_i\left(\theta_0\right); q\right)\delta_i\delta_i'\right]^{-1}$,

$$(b): \hat{\theta}_2 \xrightarrow{p} \theta_0;\ \sqrt{N}\left(\hat{\theta}_2 - \theta_0\right) \xrightarrow{d} Z_2 \sim N\left(0, \Sigma_2\right).$$

With the proof of this theorem being in the Supplementary Material, here we make a few brief remarks on the proof strategy. For part (a), consistency for the first estimator is established under identification conditions by showing that $\hat{M}_s\left[V_i\left(\theta\right)\right]$ converges uniformly (w.r.t. θ) in probability to the true conditional mean function. Asymptotic normality follows largely from a U-statistic result based on a low-order bias obtained in Theorem 1. In part (b), we employ both residual and recursive differencing controls, which makes it possible to set optimal window parameter $r = \frac{1}{4s+d}$ in (10). As described below in 1(a) and 1(b) of Section 4, the window parameter was chosen at the beginning of implementation based on the final stage.

It suffices to adjust the expectations so as to ensure that density denominators vanish slowly when the index is away from the truth and rapidly at the truth. This strategy, which was developed in Klein and Shen (2010) for single-index models under bounded X's, is extended here to the multiple-index case. To take advantage of the residual control, trimming at this second step is based on $\hat{\theta}_1$ in part (a).

4. IMPLEMENTATION

In this section, we describe the steps needed to implement the estimators in Theorem 2. One of these estimators employs recursive differencing as the sole bias control, whereas the other controls for the bias with recursive differencing and a residual control. As a necessary part of this discussion, we also provide the details for constructing the recursive differencing estimator for a conditional expectation. To guarantee that the estimators work well in practice as well as in theory, there are several different bias correcting mechanisms, trimming levels, and window parameters that must be selected. We provide these choices below. For the case of three indices, Gauss code for implementing the estimators is available at https://economics.rutgers.edu/people/faculty/people/86-faculty/220-klein-roger.

4.1. Recursive Differencing Bias Control

In this subsection, we describe the steps required to implement the estimator in (D7).

1(a) Window parameter r, and final stage s^*. Set r and s^* to satisfy conditions in (C1) of Theorem 2. We set s^* to be the smallest stage value satisfying (C1) and r as close as possible to its optimal value $\frac{1}{4s^*+d}$, which equates the orders of squared bias and variance.

2(a) Initial (stage $s = 1$) estimated expectation. Employing the kernel function in (D4), refer to (D6) and calculate the initial stage 1 estimator $\hat{M}_1$ for the conditional expectation.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

3(a) Stage s estimated expectation. For $s \geqq 2$, recursively calculate the stage s estimator $\hat{M}_s$ as in (D6) until the final stage s^* is reached. Note that this estimator uses $\tau_{sm}(X_i; \hat{q}(N))$ in (D3). This smooth trimming function depends on lower and upper sample quantiles (e.g., 0.02 and 0.98).

4(a) The estimator for index parameters $\hat{\theta}_1$. Referring to (D7), calculate the index parameter estimator in Theorem 2(a). Here, the indicator trimming $\tau(X_i; \hat{q}')$ is based on sample quantiles (e.g., 0.03 and 0.97).

While the estimator in Step 4(a) performs very well in simulations, the performance generally improves when we combine recursive differencing and residual controls. The implementation of this estimator is discussed in the next section.

4.2. Recursive Differencing and Residual Bias Controls

To combine the residual control with recursive differencing for estimating index parameters, we require a somewhat different estimation strategy than that above. For the estimator in (D10), the required steps are as follows:

1(b) Window parameter r, final stage s^*, and adjustment parameter a. Set r, s^*, and a as in (C2) of Theorem 2(b), with s^* set to be the smallest stage value satisfying (C2). For example, with $s^* = 3$ in the Monte Carlo, we set $a = 1/2$ and $r = 1/15$.

2(b) Stage 1 adjusted expectation estimator. Calculate the stage 1 adjusted expectation $\hat{M}_{1a}$ as in (D9), using the kernel function in (D4) and (D5). This adjusted expectation is obtained by replacing estimated densities with adjusted ones as provided in (D8). There are three multiplicative components in this adjustment. First, it depends on $\hat{\gamma}_s$, a lower sample quantile for $\hat{g}_1(V_j(\hat{\theta}_1); \hat{\theta}_1)$, which we suggest setting at the 0.01 level. Second, it depends on N^{-ar}. Finally, it depends on the smooth trimming function in (D8). Note that only the last component depends on θ.

3(b) Stage s adjusted expectation estimator. For $s \geqq 2$, recursively calculate the stage s estimator $\hat{M}_{sa}$ as in (D9) until the final stage s^* is reached. Note that this estimator uses $\tau_{sm}(V_j(\hat{\theta}_1); \hat{q}(N))$ in (D3). This smooth trimming function depends on lower and upper sample quantiles (e.g., 0.02 and 0.98).

4(b) The estimator for index parameters, $\hat{\theta}_2$. Referring to (D10), calculate the index parameter estimator with trimming based on $\tau(V_i(\hat{\theta}_1); \hat{q}')$, which restricts the estimated index to a quantile region that must be a subset of that in Step 3(b). We suggest setting these sample quantiles at the 0.03 and 0.97 levels, respectively.

Inferences are then conducted as if this were a parametric problem using an estimator for the covariance matrix. The estimator is given by replacing all components of the covariance matrices in Theorem 2 by their sample counterparts, with the estimated error variance $\hat{\sigma}_\varepsilon^2$ given as an average of squared residuals. For the second estimator, if maximum likelihood software is available (e.g., maxlik in

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

Gauss), then the calculation can be simplified as follows. Define

$$\hat{\theta}_2 \equiv \arg\max_{\theta} \left(-\frac{1}{2}\right) \hat{Q}_2(\theta; \hat{\theta}_1).$$

Let $\hat{C}$ be the returned covariance matrix obtained from maximum likelihood software. Then, a consistent estimator for the covariance matrix is given as $\hat{\sigma}_\varepsilon^2 \hat{C}$.

5. MONTE CARLO RESULTS

We conducted Monte Carlo experiments using four different designs: quadratic, cubic, exponential, and sin. In all designs, we constructed three indices: $V_1 = X_1 + X_4, V_2 = X_2 - X_4$, and $V_3 = X_3 + X_4$, where X_1, X_2, X_3, and X_4 follow truncated standard normal distributions with $|X_k| < 3$. We also generated an error term ε 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 α, β, γ are standardizing constants selected so that, in all designs, each of the three explanatory components has an approximate 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 (Figure 1). We will also compare results for different estimators of the conditional mean function (Figure 2). We set the sample size at 2,000 and the number of replications at 100 for both Figures 1 and 2.

Figure 1 provides a comparison between four $\sqrt{N}$-asymptotically normal estimators and results to check theoretical predictions on the recursive differencing estimator. Each design has three estimated parameters, and the plots are organized by design and parameter. In a series of box plots, each provides distributional results for comparing estimators. With deviations from the truth shown on the vertical axis, the figure indicates the types of estimators on the horizontal axis. The estimators have been shifted so that 0 corresponds to the true value. The length of each box is the interquartile range, with the median of each estimator shown by the bold line within the box.

The first box plot provides results for an HK estimator, which is an extension of the twicing kernel (Newey, Hsieh, and Robins, 2004). The second box plot is for a fifth degree LP estimator with bias $O(h^6)$ as in Ruppert and Wand (1994). For these two estimators, we set $r = \frac{1}{11.99}$ to ensure that the bias is $o(N^{-1/2})$. For the estimator with recursive differencing as the sole bias control (RDS), we employ the undersmoothing window parameter $r = \frac{1}{11.99}$ to ensure $\sqrt{N}$ asymptotic normality. The next three estimators (WF1, WF2, and WF3) provide results for three stages of the recursive differencing estimator without residual control. We fix the window

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

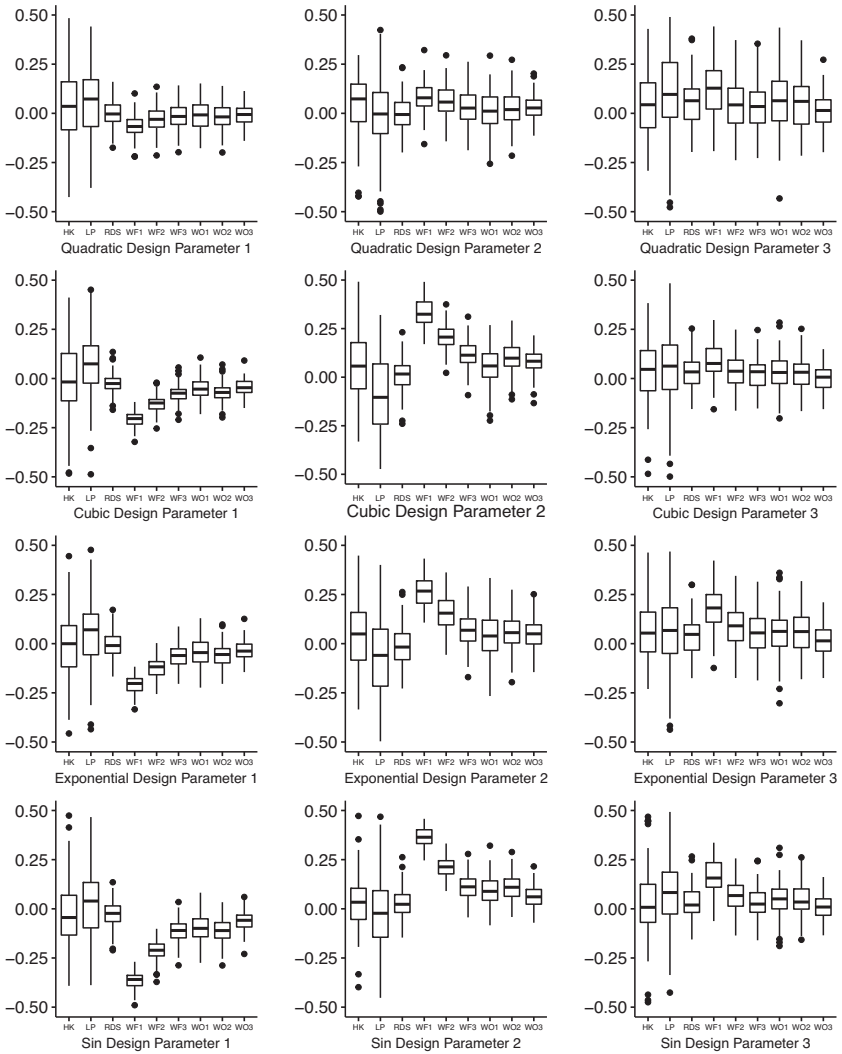

FIGURE 1. Monte Carlo results: Parameter estimators.

parameter at $r = \frac{1}{15}$ throughout these three stages so as to facilitate the comparison between stages.[9]

The last three box plots show the results for the proposed recursive differencing estimator with optimal windows set for single, double, and triple stages (WO1,

[9]We chose $r = 1/15$ as it is the optimal window for the third-stage estimator.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

WO2, and WO3). The optimal windows were set by equating the orders of squared bias and variance. More specifically, WO1 had a window size $r = \frac{1}{7}$; WO2 had a window size $r = \frac{1}{11}$; whereas WO3 had a window size $r = \frac{1}{15}$. The triple stage estimator also has the residual control to ensure $\sqrt{N}$ normality.

Across all designs and parameters, the recursive differencing estimator with undersmoothing (RDS) strongly dominates the other two estimators (HK and LP) with much smaller variation and low median bias. As a result, the RDS estimator had substantially smaller RMSE. Similarly, the proposed recursive differencing estimator with residual control also strongly dominates the HK and LP estimators. We found that the performance of the proposed recursive differencing estimator with residual control (WO3) is overall better than the undersmoothing version (RDS). The average reduction in RMSE was 0.013, which is an improvement of at least 10%. In summary, recursive differencing estimators dominate the other estimators, and there is a value added to employing the residual bias control with recursive differencing.

We also made a number of comparisons to check predictions made by the theory. Turning to recursive differencing under fixed windows, comparing WF1, WF2, and WF3, the proposed recursive differencing estimator for the most part has decreasing bias over the stages, whereas the variation as reflected in the interquartile range remains stable. This finding is consistent with the theory behind the recursive differencing mechanism. The decline is most pronounced for the sin design. Similarly, we made comparisons between WO1, WO2, and WO3. As predicted by the theory, 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 and residual control (WO3) to the recursive differencing estimator without residual control (WF3). Across all four designs, we found that RMSE decreased substantially when the additional residual bias control is employed. For the first design (quadratic), we also compared the performance of our proposed recursive differencing estimator (WO3) at different sample sizes: $N = 500$, 1,000, and 2,000. We found that the bias in the estimators remained reasonably small at the smaller sample sizes (ranging $00.02 \sim 0.06$ at $N = 500$, $0.01 \sim 0.05$ at $N = 1,000$, and $0.01 \sim 0.03$ at $N = 2,000$), whereas the SDs were larger (ranging $0.13 \sim 0.24$ at $N = 500$, $0.08 \sim 0.14$ at $N = 1,000$, and $0.05 \sim 0.09$ at $N = 2,000$). The theory suggests that the doubling of the sample size would reduce standard errors by $1/\sqrt{2}$ if the sample size is sufficiently large. We found this to be approximately the case when going from 1,000 to 2,000 observations, but not from 500 to 1,000 observations. Therefore, our experiment suggests that the sample size of 500 is probably too small.

In addition to comparing parameter estimators, we also compared results for estimating the conditional mean functions. The estimation of these functions

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

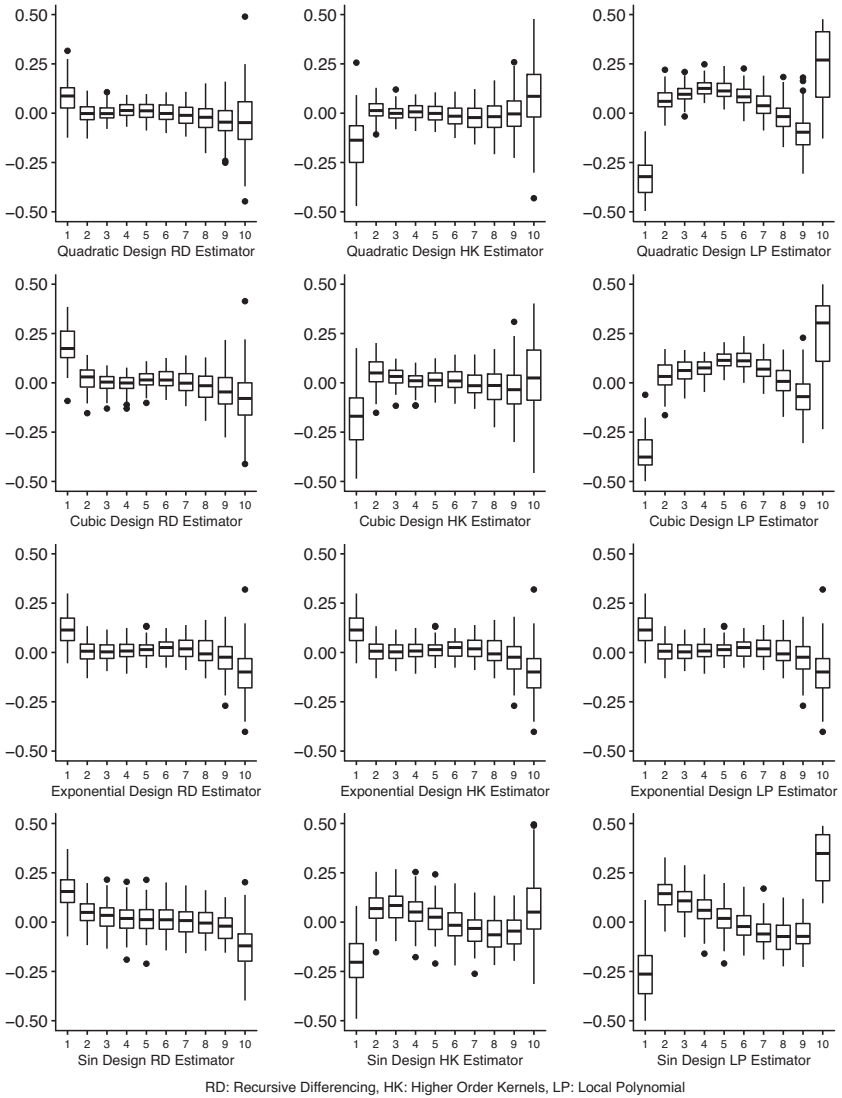

FIGURE 2. Monte Carlo results: Conditional expectation estimators by deciles.

plays a fundamental role in estimating parameters. Furthermore, changes in these conditional mean functions are important objects of interest in empirical studies. Therefore, we provide results on the estimation of conditional mean functions in Figure 2. We investigated the performance of three conditional mean estimators underlying the parameter estimators we studied above: the recursive differencing (RD) estimator, the HK estimator, and the LP estimator. We organized the plots by design and estimator. We calculated the conditional mean function estimators

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

at every point in a trimmed set[10] and then averaged over decile intervals. In so doing, a window size of $r = \frac{1}{11.99}$ was set for all three estimators. In each plot, we provided the results by deciles. 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.

From the box plots in Figure 2, the bias for the recursive differencing and the higher-order kernel estimators are similar to each other with both having smaller bias than the LP estimator. The variation of the recursive differencing estimator is smaller than that of the higher-order kernel and LP estimators across designs and deciles, with substantial advantage at the higher and lower deciles. As expected from the bias and variation results, we found that overall the recursive differencing estimator had smaller RMSE than the higher-order kernel and LP estimators across designs and deciles. The advantage was especially pronounced at the higher and lower deciles.

We remark that the performance of the LP 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, which is the most challenging design for LP estimator. In that case, the RMSE of the LP estimator for the first decile reduced from 1.027 to 0.275; bias reduced from −0.848 to −0.222; and SD reduced from 0.583 to 0.162. However, the recursive differencing estimator continues to dominate it with first decile RMSE of 0.126, bias of 0.115, and SD of 0.053. Results are similar for the 10th decile.

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. Furthermore, the behavior of the recursive differencing estimator is consistent with the underlying theory.

6. CONCLUSIONS

In this paper, we propose recursive differencing estimators for estimating conditional expectations and parameters in semiparametric models with multiple indices. The most important feature that we want to highlight is that the order of the bias decreases with the stage of the recursion, whereas the order of the variance remains the same.

While HKs and LPs share the above properties, they differ from the proposed estimator in two important respects. First, the RMSE of the recursive differencing estimator becomes smaller over the stages. In contrast, HK or LP estimators would require higher-order terms to achieve the same bias order, which

[10]Each index was trimmed at 3% and 1% from each tail based on indicator and smooth trimming functions, respectively.

https://doi.org/10.1017/S0266466622000329 Published online by Cambridge University Press

often leads to higher RMSEs.[11] 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.

SUPPLEMENTARY MATERIAL

Shen, C. and Klein R. (2022): Supplement to "Recursive Differencing for Estimating Semiparametric Models", Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466622000329

REFERENCES

Blundell, R.W. and J.L. Powell (2003) Endogeneity in nonparametric and semiparametric regression models. *Econometric Society Monographs* 36, 312–357.

Blundell, R.W. and J.L. Powell (2004) Endogeneity in semiparametric binary response models. *The Review of Economic Studies* 71(3), 655–679.

Fan, J. and I. Gijbels (1995) Adaptive order polynomial fitting: Bandwidth Robustification and bias reduction. *Journal of Computational and Graphical Statistics* 4(3), 213–227.

Fan, J. and I. Gijbels (1996) *Local Polynomial Modeling and Its Applications*. Chapman & Hall.

Gu, J., Q. Li, and J.C. Yang (2015) Multivariate local polynomial kernel estimators: Leading bias and asymptotic distributions. *Econometric Reviews* 34, 979–1010.

Honoré, B.E. and J.L. Powell (2005) Chapter 22: Pairwise difference estimation of nonlinear models. In D.W.K. Andrews and J.H. Stock (eds.), *Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg*, pp. 520–553. Cambridge University Press.

Horowitz, J. (1996) Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica* 64, 103–137.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *Journal of Econometrics* 58, 71–120.

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

Klein, R. and C. Shen (2010) Bias corrections in testing and estimating semiparametric, single index models. *Econometric Theory* 26, 1683–1718.

Klein, R., C. Shen, and F. Vella (2015) Estimation of marginal effects in semiparametric selection models with binary outcomes. *Journal of Econometrics* 185(1), 82–94.

Klein, R. and R. Spady (1993) An efficient semiparametric estimator for the binary response model. *Econometrica* 61, 387–421.

Li, Q. and Y. Sun (2014) Nonparametric and semiparametric estimation and hypothesis testing with nonstationary time series. In J. Racine, L. Su, and A. Ullah (eds.), *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, pp. 444–482. Oxford University Press, New York, NY.

[11] We note that a local linear estimator provides the basis for the start of the recursion sequence, which does contribute to its performance.

Lu, Z.Q. (1996) Multivariate locally weighted polynomial fitting and partial derivative estimation. *Journal of Multivariate Analysis* 59, 187–205.

Masry, E. (1996) Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and Their Applications* 65, 81–101.

Maurer, J., R. Klein, and F. Vella (2011) Subjective health assessments and active labor market participation of older men: Evidence from a semiparametric binary choice model with nonadditive correlated individual-specific effects. *The Review of Economics and Statistics* 93(3), 764–774.

Müller, H.G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Annals of Statistics* 12, 766–774.

Newey, W.K., F. Hsieh, and J. Robins (2004) Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* 72, 947–962.

Newey, W.K. and D. McFadden (1994) Large sample estimation and hypothesis testing. In: *Handbook of Econometrics*, vol. 4, pp. 2111–2245. North-Holland.

Powell, J., J. Stock, and T. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 51, 1403–1430.

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

Ruppert, D. and M.P. Wand (1994) Multivariate locally weighted least-squares regression. *Annals of Statistics* 52(3), 1346–1370.

Shen, C. (2013) Determinants of health care decisions: Insurance, utilization, and expenditures. *The Review of Economics and Statistics* 95(1), 142–153.