Bias Corrections in Testing and Estimating Semiparametric, Single Index Models

Roger Klein and Chan Shen*

July, 2009

Abstract

Semiparametric methods are widely employed in applied work where the ability to conduct inferences is important. To establish asymptotic normality for making inferences, bias control mechanisms are often used in implementing semiparametric estimators. The first contribution of this paper is to propose a mechanism that enables us to establish asymptotic normality with regular kernels. In so doing, we argue below that the resulting estimator performs very well in finite samples.

Semiparametric models are commonly estimated under a single index assumption. Since the consistency of the estimator critically depends on this assumption being correct, our second objective is to develop a test for it. To ensure that the test statistic has good size and power properties in finite samples, we employ a bias control mechanism similar to that underlying the estimator. Furthermore, we structure the test so that its form adapts to the model under the alternative hypothesis. Monte Carlo results confirm that the bias control and the adaptive feature significantly improve the performance of the test statistic in finite samples.

1 Introduction

Semiparametric models are commonly estimated under a single index assumption (e.g., Ahn (1997), Climov, Delecroix and Simar (2002), Fraga and Martins (2001), Gerfin (1996), Gorgens (2000), Gorgens and Horowitz (1999), Ichimura (1993), Klein and Sherman (2002), Klein and Spady (1993)). In this paper, we focus on developing bias controls for estimating and testing semiparametric models in such a setting. Before discussing bias controls, we begin with a brief description of these single index models

*This paper is based on a portion of Chan Shen's Ph.D. dissertation, Economics Department, Rutgers University. We would like to thank the editor and anonymous referees for helpful comments. Any errors are the sole responsibility of the authors.

with which this paper is concerned. To this end, let Y be a dependent variable of interest, X a vector of explanatory variable, and $V(X;\theta_o)$ a parametric function of X where θ_o is a vector of parameter values. With V termed an index, these single index models are characterized by the assumption:

$$E(Y|X) = G(V(X;\theta_o)) = E(Y|V(X;\theta_o))$$

The linear index, where V is linear in observed variables, is the most commonly used. Such a structure can cover a wide class of models because it permits V to be a linear combination of higher order and interaction terms. To cover an even wider class of models, it may be necessary to consider multiple index models where the conditional expectation is an unknown function of two or more separate indices.

As above, let θ_o be a parameter vector of interest and denote $\hat{\theta}$ as its estimator. To ensure that this estimator has desirable properties, in large and finite samples, it is critical to have a good estimator for the conditional expectation, $E(Y|V(X;\theta_o))$. In estimating this expectation, higher order kernels[1] are often used in the literature as a bias control mechanism to obtain asymptotic normality for $\hat{\theta}$. This bias control method can fail to satisfy restrictions on the true conditional expectation of interest, and hence does not perform well in finite samples. For example, in binary response models, where expectations are probabilities, such a method can deliver estimated probabilities that fall outside of the interval [0,1].

One objective of this paper is to provide an estimator with desirable large sample properties and that also performs well in finite samples. To obtain these properties, we propose a bias control mechanism with two controls and regular kernels. The bias controls will ensure normality, while the use of regular kernels allows us to impose known restrictions on the estimated conditional expectation.[2] We find that the resulting

[1] With V_i i.i.d. distributed as g, a kernel density estimator for $g(t)$ is given as:

$$\hat{g}(t) = \sum \frac{1}{Nh} K\left[(t - V_i)/h\right].$$

When K is a density that is symmetric about zero (e.g., a standard normal), we refer to K as a regular kernel. In this case, it can be shown that the bias in $\hat{g}$ is $O(h^2)$, where h tends to zero at a rate given below. When K is a function that is symmetric about zero, integrates to one, and

$$\int z^{2p} K(z) dz = 0, \ p = 1, 2, ...,$$

then K is termed a higher order kernel. It can be shown that the bias in a density estimator based on this kernel is $O\left(h^{2(1+p)}\right)$. Notice that unlike regular kernels, higher order kernels must take on negative values.

[2] There are other alternative methods that control for the bias under regular kernels. For example, Powell and Honore (2005) employ the following jackknife approach. Let $\hat{\beta}(h)$ be an estimator based on the window parameter h. Then, under this approach the final estimator is a linear combination of such estimators using different windows. In contrast, here we employ a two-stage approach that exploits a result due to Whitney Newey (see Theorem 0 in section 2) to ensure asymptotic normality

estimator performs well in finite samples, an important feature for its use in applications. It should be remarked that similar bias control mechanisms can be used for likelihood-based estimators in double index models (see Klein, Shen, and Vella, 2009). Such models are important and naturally arise when one or more of the explanatory variables is endogenous. For an application in the context of healthcare decisions, see Shen (2009).

When, as in the examples just mentioned, the model is characterized by more than one index, the imposition of the single index assumption will result in an inconsistent estimator for the conditional expectation of interest.[3] Given the sensitivity of the estimator to a single index assumption and given its wide use, a second objective of this paper is to formulate a test for this assumption.

There have been papers in the literature on testing parametric against semiparametric models (e.g., Härdle, Mammen, and Müller (1998), Härdle, Spokoiny, and Sperlich (1997), Horowitz and Härdle(1994)). Related tests of parametric models are given by Newey (1985), Bierens (1990), and Härdle and Mammen (1993). This paper differs from those above in that it formulates a test for a main assumption in semiparametric models. We note that in a likelihood context with a parametric null hypothesis, Newey develops conditional moment tests that have optimal local power properties. It may be possible to extend these results to the present context, but this extension is beyond the scope of the current paper.

Some papers focus on testing single index restrictions. Escanciano and Song (2007) provide a test focusing on average marginal effects and show that it has a minimax property; Andrews (1993) provides high level conditions for testing moment restrictions. Our paper differs from these in that we provide primitive conditions for a conditional moment test and for the estimator on which it is based. Tripathy and Kitamura (2003) employ an empirical likelihood approach for testing moment conditions of the form: $E\left[G(z,\theta)|X\right] = 0$. With G a known function and X continuous, they establish an optimality property for their test. The test proposed here is also an orthogonality test in that we test whether a function G is correlated with functions $M(X)$. However, unlike the above test, here the function G will be unknown as we set $G = Y - E(Y|V)$, where V is an index and the conditional expectation function $E(Y|V)$ is unknown. Further, the M-functions will be unknown and we will require nonparametric estimates of them. This feature is needed to ensure that the form of test statistic adapts to the model under the alternative hypothesis. Most importantly, the proposed statistic differs from those in the literature in the bias control mechanism that it employs. This mechanism is similar to that underlying the estimator, and results

under regular kernels. In addition, we also implement a smoothing adjustment to the final estimator. We find that the resulting estimator performs quite well in finite samples. It is an open question as to whether or not a further improvement would be obtained if we jackknifed our estimator.

[3] As an alternative example of a double index model, return to the binary response model discussed earlier, and let the error term have a conditional variance that depends on another index. Namely, let $u = s(X\beta_o)\varepsilon$, where ε is independent of X.

in a test statistic that has good size and power properties in finite samples.

In organizing this paper, we begin by discussing the moment conditions that characterize the estimator and the test statistic in Section 2. These conditions incorporate methods for controlling their bias using regular kernels. Section 3 contains assumptions and asymptotic results. Here, we will also outline the basic proof strategy, with the Appendix containing all formal and complete proofs. In section 5, we carry out Monte Carlo studies, where we evaluate the performance of the estimators and test statistics in finite samples. To preview the results, we find that a bias corrected estimator based on regular kernels performs the best and the bias-corrected form of the test statistic has good size and power properties.

2 Moment Conditions and Bias Control

2.1 The Estimator

In describing these conditions and the nature of the bias controls, it will be useful to have simplified notation for sample averages of the quantities of interest. For this purpose, define:

$$\langle AB \rangle \equiv \sum_{i=1}^{N} \left[A_i B_i \right] / N; \;\; \langle A/B \rangle \equiv \sum_{i=1}^{N} \left[A_i / B_i \right] / N$$

Further, we use the " Λ " symbol above a quantity of interest to indicate an estimator for it. Then, letting $V(\theta_0) \equiv V(X; \theta_0)$ be a single index depending on explanatory variables, X, and on a vector of true parameter values, θ_0, assume that we are interested in an extremum estimator for θ_0. Let τ be a trimming function that controls for small denominators in a manner that we will make explicit below. In this section, for expositional simplicity, we take this trimming function as known. In the Appendix, we let this function depend on an estimated argument and show that it may be taken as known. Employing this trimming function, consider estimators whose gradients have the following structural form:

$$\hat{G}^* (\theta_0) \equiv \left\langle \left[Y - \hat{E}(Y|V(\theta_0) \right] \tau \hat{W}^* \right\rangle,$$

where the weighting function, $\hat{W}^*$, has the form:

$$\hat{W}_i^* \equiv \hat{\alpha} \left(V_i (\theta_0) \right) \nabla \hat{E}_i$$

For SLS estimators (Ichimura (1993)), $\hat{\alpha} \left(V_i (\theta_0) \right) = 1$. For a QMLE estimator for binary response models (Klein and Spady (1993)), $\hat{\alpha} \left(V_i (\theta_0) \right) = 1 / \hat{E}_i \left[\left(1 - \hat{E}_i \right) \right]$. For a QMLE estimator of ordered models (Klein and Sherman (2002)), the gradient consists of a number of components, all of which have the structure above. The weights differ

above, but all consist of a function of the index and the derivative of a nonparametric expectation estimator. If the gradient, when normalized by $\sqrt{N}$ is asymptotically distributed as normal, then it is not difficult to show that the underlying estimator of interest has an asymptotic normal distribution. Accordingly, in what follows, we focus on these gradient expressions.

For such estimators characterized by the gradient structure above, write the gradient as:

$$\begin{aligned}\sqrt{N}\hat{G}^*(\theta_0) &= \sqrt{N}\left[\hat{A}^*(\theta_0) - \hat{B}^*(\theta_0)\right] \\ \hat{A}^*(\theta_0) &\equiv \left\langle [Y - E(Y|V_0)]\,\tau\hat{W}^*\right\rangle \\ \hat{B}^*(\theta_0) &\equiv \left[\hat{E}(Y|V_0) - E(Y|V_0)\right]\tau\hat{W}^*\end{aligned}$$

For the first term, with the argument given in the Appendix, it can be shown that:

$$\sqrt{N}\left[\hat{A}^*(\theta_0) - A^*(\theta_0)\right] \xrightarrow{p} 0,\ A^*(\theta_0) \equiv \langle [Y - E(Y|V_0)]\,\tau W^*\rangle$$

The second or B-component above contributes a bias to the estimator that we need to control in order to show that the gradient has an asymptotic normal distribution.

Below we will define $\hat{E}(Y|V_0)$ as a ratio of estimated functions: $\hat{f}/\hat{g}$, each of which converges to its true limiting value. We will be able to show that:[4]

$$\begin{aligned}&\sqrt{N}\left[\hat{B}^*(\theta_0) - \hat{B}_S^*\right] \xrightarrow{p} 0, \\ \hat{B}_S^* &= \left\langle \left[\hat{f}/\hat{g} - E(Y|V_0)\right]\tau\hat{W}^*(\hat{g}/g)\right\rangle \\ &= \left\langle \left[\hat{f} - \hat{g}E(Y|V_0)\right]\tau W^*/g\right\rangle + o_p(1).\end{aligned}$$

Since the above quantity is linear in the estimated components $\hat{f}$ and $\hat{g}$, it is possible to control for the bias in $\hat{B}_S^*$ by controlling for the bias in these estimated functions. Higher order kernels are commonly employed for this purpose. In this case a standard U-statistic projection argument, which we provide in the Appendix, immediately provides the result:

$$\begin{aligned}&\sqrt{N}\left[\hat{B}_S^* - B_S^*\right] \xrightarrow{p} 0, \\ B_S^* &= \langle [Y - E(Y|V_0)]\,\tau E[W^*|V_0]\rangle\end{aligned}$$

[4]Note that:

$$\left\langle \left[\hat{f}/\hat{g} - E(Y|V_0)\right]\tau\hat{W}^*[(\hat{g}/g) - 1]\right\rangle \xrightarrow{p} 0$$

with the first and third term each converging to zero at a rate somewhat below $N^{-1/2}$. We will show in the Appendix that the overall or combined rate of the product is sufficient to provide the desired result.

Asymptotic normality for the normalized gradient now follows from a standard central limit theorem. In using higher order kernels to control for the bias and deliver this result, it should be noted that such kernels can result in negative density estimates and (as is the case here) often do not perform as well as methods based on regular kernels that do not deliver the desired large sample properties. Here, we seek alternative bias controls that deliver the desired large sample results with regular kernels.

Recalling that the weight function contains the derivative of the expectation function, we exploit a property of this derivative due to Whitney Newey in the following theorem:[5]

Theorem 0: With $V(\theta_0) \equiv V(X;\theta_0)$ as a single index, assume the following single index restriction holds:

$$E(Y|X) = E(Y|V(\theta_0)) \equiv F(\ V(\theta_0))$$

Then:

$$E[\ \nabla_\theta E(Y|V(\theta))\ \mid V(\theta_0)]_{\theta=\theta_0} = 0\ .$$

Proof: Let $\delta(\theta) \equiv V(\theta_0) - V(\theta)$ and observe that $\delta(\theta_0) = 0$ and that $\nabla_\theta \delta(\theta) = -\nabla_\theta V(\theta)$. Then, employing the index restriction and using iterated expectations:

$$\begin{aligned} E(Y|V(\theta)) &= E_X[E(Y|V(\theta_0))\ \mid V(\theta)] \\ &\equiv E_X[F[V(\theta_0)] \mid V(\theta)] \\ &= E_X[F[V(\theta)+\delta(\theta)] \mid V(\theta)] \\ &\equiv G(V(\theta),\delta(\theta)) \end{aligned}$$

Let G_k be the partial derivative of G taken w.r.t. θ in the k^{th} argument of G, $k = 1,2$. From the chain rule:

$$\begin{aligned} \nabla_\theta G(V(\theta),\delta(\theta))|_{\theta=\theta_0} &= G_1(V(\theta),0)|_{\theta=\theta_0} + G_2(V(\theta_0),\delta(\theta))|_{\theta=\theta_0} \\ &= \nabla_\theta F(\ V(\theta))|_{\theta=\theta_0} - E[\nabla_\theta F(\ V(\theta))\ \mid V(\theta_0)]_{\theta=\theta_0} \end{aligned}$$

The proof now follows.

From above, $\nabla_\theta E[Y|V(\theta)]_{\theta=\theta_0}$ behaves as an error component with conditional expectation 0. As this component enters multiplicatively into the gradient, we exploit its residual-like properties as a bias control. To utilize Newey's result, return to the gradient discussed above and let

$$H(V) \equiv E\left(\left[\hat{f} - \hat{g}E(Y|V_0)\right]\ \mid X\right)$$

[5]This result and its proof were provided to one of the authors in a private communication. The proof, which is very short and can be found in Klein and Sherman (2002), is also provided here.

Then, take an iterated expectation to obtain:

$$\begin{aligned} E\left[\hat{B}_S^*\right] &= \left\langle E_X E\left[\hat{f} - \hat{g}E\left(Y|V_0\right)\right]\tau W^*/g \mid X\right\rangle + o(1) \\ &= \langle E_X\left(\tau H(V)W^*/g\right)\rangle \\ &= \langle E_V\{[H(V)/g]E\left[(\tau W^*|V)\right]\}\rangle \end{aligned}$$

If the trimming function, τ, depends on X, it is not possible to employ Newey's result and obtain 0 for this expectation. If the trimming depends on V, then this expectation would be zero by construction.

Based on the above observation, we consider a multi-stage estimation method. In the first stage, we trim on X and obtain consistent estimates for the index parameters. Using these parameter estimates, we construct an estimated index upon which to base trimming. In the second stage, we then trim on the basis of the (estimated) index rather than X. In so doing, the expected value of the gradient would be zero. However, such trimming upsets the consistency argument because it provides no protection for small denominators outside of a small neighborhood of the truth. To resolve this problem, we adjust expectations as follows. Recalling that $\hat{E} = \hat{f}/\hat{g}$, define an adjusted expectation as:

$$\hat{E}_a = \frac{\hat{f}}{\hat{g} + \Delta}$$

Below, we will define Δ such that it vanishes rapidly in regions where g is bounded away from zero. In regions where g tends to zero, Δ tends to zero very slowly. In this manner, we are able to preserve the consistency argument and establish asymptotic normality for the gradient.[6] It is possible to further improve the performance of the estimator in finite samples under a smoothing adjustment, but we defer discussion of this issue until Section 3.

2.2 Test Statistics

In what follows, we will first consider a general test of moment conditions and then specialize it to the test for the single index restriction. Consider test statistics based on "residuals"of the form:

$$G_{kT}\left(\theta_0\right) \equiv \langle\left[Y - E(Y|V\left(\theta_0\right)\right]W_{Tk}\rangle\ ,\ k = 1, .., K$$

where $W_{Tk} = W_{Tk}\left(X_k\right)$ is a vector of observations on a function of the k^{th} exogenous variable, X_k. Define G_T as a column vector with G_{kT} as the k^{th} element. Under a null hypothesis of interest, H_0, we assume that the following orthogonality condition holds:

$$E\left[G_T\left(\theta_0\right)\right] = 0,$$

[6] A similar strategy is employed in Klein and Spady (1993) so as to let trimming depend on an estimated density. That paper, however, relies on higher order kernels or local smoothing to obtain large sample results.

In testing whether or not these conditions hold, we allow the conditional expectation $E(Y|V)$ and the weight W_k to be unknown functions that can be estimated nonparametrically. Accordingly, write the estimated k^{th} moment as:

$$\hat{G}_k\left(\hat{\theta}\right) \equiv \left\langle \left[Y - \hat{E}(Y|V\left(\hat{\theta}\right)\right] \hat{W}_{Tk} \right\rangle .$$

With a test statistic based on these estimated moments, we will need to show that $\sqrt{N}\hat{G}_{kT}\left(\hat{\theta}\right)$ has an asymptotic normal distribution under the null hypothesis of interest. Employing a standard Taylor expansion and with $\hat{\theta}$ as a $\sqrt{N}-$consistent estimator that has an asymptotic linear representation[7], we will be able to write

$$\sqrt{N}\hat{G}_{kT}\left(\hat{\theta}\right) = \sqrt{N}\hat{G}_{kT}\left(\theta_0\right) + \nabla E\left[G_{kT}\left(\theta_0\right)\right]\sqrt{N}\left[\hat{\theta} - \theta_0\right] + o_p\left(1\right).$$

As the second or parameter-uncertainty component poses no difficulty, here we focus on the first component and discuss the nature of the bias control that we employ.

With $V_o \equiv V\left(\theta_0\right)$, we will be able to decompose these moment conditions in the same form as the gradient for the estimator above and write

$$\begin{aligned}
\sqrt{N}\hat{G}_{kT}\left(\theta_0\right) &= \sqrt{N}\left[A_T\left(\theta_0\right) - \hat{B}_T\left(\theta_0\right)\right] + o_p(1) \\
A_T\left(\theta_0\right) &\equiv \left[Y - E\left(Y|V_0\right)\right] W_{Tk} \\
\hat{B}_T\left(\theta_0\right) &\equiv \left\langle \left[\hat{E}\left(Y|V_0\right) - E\left(Y|V_0\right)\right] \hat{W}_{Tk} \right\rangle
\end{aligned}$$

As for the estimator, here the second or B-component contributes a bias that we seek to control. As above, we will be able to show:

$$\begin{aligned}
&\sqrt{N}\left[\hat{B}_T\left(\theta_0\right) - \hat{B}_S\right] \xrightarrow{p} 0, \\
\hat{B}_S &= \left\langle \left[\hat{f}/\hat{g} - E\left(Y|V_0\right)\right] \hat{W}_{Tk}\left(\hat{g}/g\right) \right\rangle \\
&= \left\langle \left[\hat{f} - \hat{g}E\left(Y|V_0\right)\right] W_{Tk}/g \right\rangle + o_p(1)
\end{aligned}$$

Using higher order kernels to control for the bias in $\hat{f}$ and $\hat{g}$, in a standard U-statistic argument, which is provided in the Appendix, we can show:

$$\begin{aligned}
&\sqrt{N}\left[\hat{B}_S - B\right] \xrightarrow{p} 0, \\
B &= \left\langle \left[Y - E\left(Y|V_0\right)\right] E\left[W_k|V_0\right] \right\rangle
\end{aligned}$$

[7]The estimators we consider are all of the form:

$$\sqrt{N}\left[\hat{\theta} - \theta_0\right] = -H_o^{-1}\sqrt{N}\left\langle G\right\rangle,$$

where H_0 is the Hessian matrix, and $\sqrt{N}\left\langle G\right\rangle$ is asymptotically distributed as $N(0, \Sigma)$.

The moment condition in large samples now has a form to which a central limit theorem would apply under the null hypothesis. Namely:

$$\sqrt{N}\hat{G}_{kT}(\theta_0) = \sqrt{N}\left\langle [Y - E(Y|V_0)]\,[W_k - E[W_k|V_0]]\right\rangle + o_p(1)$$

Taking estimation uncertainty into account, the "full" gradient has the form:

$$\sqrt{N}\hat{G}_{kT}\left(\hat{\theta}\right) = \sqrt{N}\hat{G}_{kT}(\theta_0) + \nabla E\left[G_{kT}(\theta_0)\right]\sqrt{N}\left[\hat{\theta} - \theta_0\right]$$

Letting $G\left(\hat{\theta}\right)$ be the vector with k^{th} element $\hat{G}_{kT}\left(\hat{\theta}\right)$, the test statistic is then given by a standard quadratic form:

$$T \equiv \sqrt{N}G\left(\hat{\theta}\right)'\hat{\Sigma}^{-1}\sqrt{N}G\left(\hat{\theta}\right),$$

where $\hat{\Sigma}$ is a consistent estimator for the covariance matrix of $\sqrt{N}G\left(\hat{\theta}\right)$. Various alternative estimators for this covariance matrix will be provided below and examined in the Monte Carlo section.

As in the case for the estimator, we find that the test statistic based on higher order kernels can be dominated by one based on an alternative bias control and regular kernels. Unfortunately, the weight need not and will not here have the residual property of the derivative weight entering the gradient for the estimator. Therefore, we propose to recenter the weight so that it has the same residual-like property as in the estimator case. Namely, with $\hat{V} \equiv V\left(\hat{\theta}\right)$ define $\hat{G}^*(\hat{\theta})$ as a vector with the k^{th} element being:

$$\begin{aligned}
\hat{G}^*_{kT}\left(\hat{\theta}\right) &\equiv \left\langle \left[Y - \hat{E}(Y|V\left(\hat{\theta}\right)\right]\hat{W}^*_k\right\rangle \\
\hat{W}^*_k &\equiv \hat{W}_k - E\left[\hat{W}^*_k|\hat{V}\right] \\
T^* &\equiv \sqrt{N}\hat{G}^*\left(\hat{\theta}\right)'\hat{\Sigma}^{-1}\sqrt{N}\hat{G}^*\left(\hat{\theta}\right)
\end{aligned}$$

For the test statistic proposed below, we will show that such recentering provides a bias control that makes it possible to employ regular kernels and still obtain the same large sample result obtained under higher order kernels. Namely, we will show that T^* is close in probability to T, with T^* having a χ^2 distribution. We find below that T^*, which is based on this alternative bias control, has much better finite sample properties than T.

To specialize the above moment conditions and develop a corresponding test statistic (a quadratic form in the moments) for the single index assumption, we need to specify the weight function. A natural choice for this function would not be a function of any particular exogenous variable, but rather the full conditional expectation:

$E(Y|X)$. In this case, the expected moment condition becomes:

$$\begin{aligned} & E\left(\ [Y - E(Y|V)]\, E(Y|X)\ \right) \\ = \ & E\left(\ [E(Y|X) - E(Y|V)]\, E(Y|X)\ \right) \\ = \ & E\left(\ [E(Y|X) - E(Y|V)]^2\right) \end{aligned}$$

Notice that this expected moment condition is zero iff $E(Y|X) = E(Y|V)$. The above weight would seem natural as the expected moment condition reduces to the distance between nonparametric and index expectations. However, it is difficult to obtain reasonable estimates of the full conditional expectation $W = E(Y|X)$ when the dimension of X is large. We are therefore motivated to seek low dimensional weights that are close to this full expectation. With "close" defined in a mean-squared error sense, low dimensional weights are given by:

$$W_k = \ \arg\min_{\omega} \left[E(W - \omega)^2 | X_k\right] = E(Y|X_k)$$

Notice that this weight depends on the actual form of the dependence of Y on X. In other words, it is adaptive to the alternative model. This property is desirable compared to fixed weights, because intuitively it yields better test power by being able to flexibly capture different violations of the null hypothesis. Our Monte Carlo study comparing one common fixed weight and our adaptive weight confirms the above observation. The fixed weight we use is the quadratic weight. Detailed discussions are in the Monte Carlo section.

3 Assumptions, Definitions, and Results

To obtain the above results, we require standard assumptions on the data generating process, smoothness conditions on unknown densities, and given sets over which densities are positive. Assume:

(A1) Observations. With (Y_i, X_i) as the i^{th} observation on the dependent and explanatory variables, assume that (Y_i, X_i) is i.i.d. With X as the NxK matrix of observations on the explanatory variables (including a column vector of ones), assume that X has full column rank with probability 1.

(A2) Model. Under the null hypothesis $E(Y_i|X_i) = E(Y_i|V_i)$, $V_i \equiv X_{1i} + X_{2i}\theta_0$, where X_{1i} is continuous and θ_0 is in the interior of a compact parameter space, Θ. Furthermore, to simplify arguments we assume that X is bounded.[8] In addition, $Var(Y_i|X_i)$ is bounded.

[8] The assumption on X being bounded is not necessary, but simplifies several of the arguments.

(A3) Estimator Characterization. Under the null hypothesis, with $-H_o$ positive definite and G_i being i.i.d., the estimator for θ_0 satisfies:

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta_0) &= -H_o^{-1} N^{-1/2} \sum_{i=1}^{n} G_i + o_p(1), \\ E(G_i) &= 0;\ Var(G_i) = O(1). \end{aligned}$$

(A4) Continuous Variable Density. With X_k as any of the continuous X variables, denote $g_k(\bullet|y)$ as its density conditioned on $Y = y$. Denote $\nabla^d g_k(t|y)$ as the d^{th} partial derivative with respect to t, with $\nabla^o g_k(t|y) \equiv g_k(t|y)$. With g_k supported on $[a_k^*, b_k^*]$:

$$\begin{aligned} g_k &> 0 \ \ on \ \ (a_k, b_k),\ a_k^* < a_k < b_k < b_k^* \\ |\nabla^d g_k| &= O(1) \ \ on \ \ [a_k, b_k], \quad d = 0, 1, 2, 3. \end{aligned}$$

(A5) Index Density. With $V_i \equiv X_{1i} + X_{2i}\theta_0$, let $g(x_1|y, x_2)$ be the indicated conditional density supported on $[a, b]$, Assume

$$\begin{aligned} g &> 0 \ \ on \ \ (a, b) \\ |\nabla^d g| &= O(1) \ \ on \ \ [a, b], \quad d = 0, 1, 2, 3. \end{aligned}$$

(A6) Tail Condition. With g_y as the density for the dependent variable, Y, assume that there exists T such that for $t > T$ and $df \geqslant 4$:

$$g_y(t) < 1/[\left(1 + t^2\right)^{(df+1)/2}].$$

The above assumptions are somewhat standard in the literature. Namely, the model must include a continuous variable (A2) and densities for continuous variables and the index must be sufficiently smooth, as implied by (A4-5). Notice that (A4-5) also specifies when density denominators become zero, which facilitate the trimming strategy. To establish uniform convergence results for estimated expectations, we require a tail condition on the density for the dependent variable, Y. While this assumption can be made in terms of the number of finite moments for Y, here we directly assume in (A6) that the density has tails that are no thinner than those for a t-distribution with

$d \geqslant 4$. Additional window conditions will be required and are stated directly in the Theorems for which they are needed. To define the estimators and test statistics, we will also require the definitions below.

(D1) Trimming. With Z_{ik} as the i^{th} observation on a continuous variable, Z_k, $k = 1, ...K$, let

$$\begin{aligned} \hat{\tau}_{ik} &\equiv \begin{cases} 1: & \hat{a}_k < Z_{ik} < \hat{b}_k \\ 0: & \text{otherwise,} \end{cases} \\ \hat{\tau}_i &\equiv \Pi_k \hat{\tau}_{ik} \end{aligned}$$

where $\hat{a}_k$ and $\hat{b}_k$ are respectively lower and upper sample quantiles for Z_k. With X_k as an exogenous variable, when $Z_{ik} = X_{ik}$, we refer to $\hat{\tau}_i$ as X-trimming and write $\hat{\tau}_{ix} = \hat{\tau}_i$; with $\hat{V}$ as the estimated index, when $Z_{ik} = \hat{V}, k = 1$, we refer to $\hat{\tau}_i$ as index-trimming and write $\hat{\tau}_{iv} = \hat{\tau}_i$.

In the case where a smooth trimming function is required, define:

$$\tau(z, \delta) \equiv \left[1 + \exp\left(-Ln(N)Ln(N)\left[z - \delta\right]\right)\right]^{-1}$$

as a smoothed approximation to an indicator on $z \geqslant \delta$. A smoothed indicator on $z \in [a, b]$ is then defined as $\tau(z, a) * \tau(b, z)$.

(D2) Kernels. The kernel function $K(z)$ is termed regular if $K(z) \succeq 0$, $\int K(z)dz = 1$, and $K(z) = K(-z)$. The function $K(z)$ will be termed a (normal) twicing kernel if $K(z) = 2\phi(z) - \phi\left(z/\sqrt{2}\right)/\sqrt{2}$.

(D3) Expectations. With $h = O\left(N^{-r}\right)$ and $K_{ij} \equiv K\left[\left(z_i - z_j\right)/h\right]$, the estimated conditional expectation with window parameter r is denoted as $\hat{E}_i \equiv \hat{E}\left(Y|Z = z_i\right)$ and is given by:

$$\hat{E}_i \equiv \left[\frac{1}{(N-1)\,h}\sum_{j \neq i} Y_j \hat{\tau}_j K_{ij}\right] / \left[\hat{\Delta}_i + \frac{1}{(N-1)\,h}\sum_{j \neq i} \hat{\tau}_j K_{ij}\right] \equiv \hat{f}_i / \hat{g}_i^*$$

The expectation is referred to as being:

a) regular $(\hat{E})$ if $\hat{\tau}_j = 1$, $\hat{\Delta}_i = 0$, and K is a regular kernel.

b) twicing if $\hat{\tau}_j = 1$, $\hat{\Delta}_i = 0$, and K is a (normal) twicing kernel (Newey, Hsieh, and Robins (2004)).

c) adjusted $(\hat{E}_a)$ if $\hat{\tau}_j = 1$, K is regular, and with $\hat{q}$ as a lower sample quantile, (e.g., 0.01) of $\hat{g}(z_i)$, $i = 1, ..., N$.

$$\hat{\Delta}_i \equiv h^\alpha \hat{q}\left[1 - \hat{\tau}_i\left(\hat{a}, \hat{b}\right)\right], 0 < \alpha < 1$$

(D4) First and Second Stage Estimators.[9]

$$\hat{\theta}_1 = \arg\max_{\theta} \hat{Q}_1, \quad \hat{Q}_1 \equiv -\frac{1}{2n}\sum_{i=1}^{n} \hat{\tau}_{ix}[Y_i - \tilde{E}\left(Y_i|v(X_i;\theta)\right)]^2,$$

$$\hat{\theta}_2 = \arg\max_{\theta} \hat{Q}_2, \quad \hat{Q}_2 \equiv -\frac{1}{2n}\sum_{i=1}^{n} \hat{\tau}_{iv}[Y_i - \hat{E}_a\left(Y_i|v(X_i;\theta)\right)]^2$$

(D5) Smoothing Adjustment. Letting $\hat{H}(\theta)$ be the Hessian w.r.t. $\hat{Q}_2$, and $\hat{E}^*$ be a regular expectation with window parameter $r^* = 1/5$, define:

$$\begin{aligned}
\hat{B}\left(\hat{\theta}_2\right) &= \sum_{i=1}^{n} \hat{\tau}_{iv}(\hat{E}_i\left(\hat{\theta}_2\right) - E_i\left(\hat{\theta}_2\right))\nabla \hat{E}_i\left(\hat{\theta}_2\right) \\
\hat{B}^*\left(\hat{\theta}_2\right) &= \sum_{i=1}^{n} \hat{\tau}_{iv}(\hat{E}_i^*\left(\hat{\theta}_2\right) - E_i\left(\hat{\theta}_2\right))\nabla \hat{E}_i\left(\hat{\theta}_2\right)
\end{aligned}$$

Then, define an adjusted estimator as:

$$\hat{\theta}^* = \hat{\theta}_2 - \hat{H}\left(\hat{\theta}_2\right)^{-1}\left[\hat{B}\left(\hat{\theta}_2\right) - \hat{B}^*\left(\hat{\theta}_2\right)\right]$$

(D6) Test Statistics. The test statistics, T and T^*, are defined as above.

As discussed earlier, we employ a two-stage estimator (D4) so as to utilize Newey's result as a bias control. The first stage of this estimator requires X-trimming (D1) and regular expectations (D3), while the second stage requires index-trimming (D1) and adjusted expectations (D3). We will compare results under regular and higher order kernels (D2). Notice that the twicing kernel in (D2) is a higher order kernel in that:

$$\begin{aligned}
&\int z^2\left[2\phi\left(z\right) - \phi\left(z/\sqrt{2}\right)/\sqrt{2}\right]dz \\
= \;& 2 - 2\int [z/\sqrt{2}]^2\phi\left(z/\sqrt{2}\right)/\sqrt{2}dz \\
= \;& 2 - 2\int w^2\phi(w)dw = 0, \; w \equiv z/\sqrt{2}
\end{aligned}$$

In examining the second stage estimator and the test statistic for various designs, we had one design where the finite sample bias for the estimator was significantly

[9] As discussed earlier, there are many different estimators to which this paper applies. We focus on variants of the SLS estimator so as to employ the same estimator over designs where the dependent variable is continuous or discrete.

larger than that for the other designs. As a result, we found that the test statistic had poor size properties in this case. The smoothing adjustment (D5) improved the size properties of our test statistic significantly in this case by reducing the bias in the estimator. To explain why this adjustment "works", recall the definition of $\hat{A}(\theta_0) - \hat{B}(\theta_0)$ in the previous section. Then, a standard Taylor expansion yields:

$$\hat{\theta}_2 - \theta_0 = -\hat{H}^{-1}(\theta^+)(\hat{A}(\theta_0) - \hat{B}(\theta_0)),\ \theta^+ \epsilon \left[\hat{\theta}_2, \theta_0\right].$$

Defining an estimator with an infeasible adjustment as:

$$\hat{\theta}_I = \hat{\theta}_2 - \hat{H}^{-1}\left(\theta^+\right)\left[\hat{B}\left(\theta_0\right) - \hat{B}^*\left(\theta_0\right)\right],$$

then it immediately follows that

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

This infeasible estimator is the same as $\hat{\theta}_2$, except the B-component now depends on an optimal expectation estimator. As a result, we would expect it to perform better in finite samples. Below, we show that this infeasible estimator can be approximated by the feasible estimator based on the adjustment in (D5) in that:

$$\sqrt{N}\left[\hat{\theta}_I - \hat{\theta}_2^*\right] \xrightarrow{p} 0.$$

Beginning with the estimator, Theorem 1 below establishes consistency at both stages.

Theorem 1: (Estimator Consistency). With $df \geqq 4$ given in (A6), set $\lambda \equiv df/\left(1 - df\right)$. Denote $\hat{\theta}_1$ and $\hat{\theta}_2$ as the first and second stage estimators respectively and assume (A1-6). Base the first-stage estimator on a regular expectation (D3) with window r_1:

$$1/8 < r_1 < 1/6;\ 0 < r_1 < \left[1/2 - \delta\right]/\left[\lambda + \varepsilon\right]$$

Base the second-stage estimator on an adjusted expectation (D3) with adjustment parameter $0 < \alpha < 1$ and window r_2 :

$$1/8 < r_2 < 1/6;\ \text{and } 0 < r_2 < \left[1/2 - \delta\right]/\left[\lambda\left(1 + a\right) + \varepsilon\right],$$

Then:

$$\left|\hat{\theta}_1 - \theta_o\right| = o_p\left(1\right);\quad \left|\hat{\theta}_2 - \theta_o\right| = o_p(1).$$

The normality arguments are very similar for the estimator and the test statistic as they are both based on similar moment conditions. After providing these results in Theorems 2-3 below, we will outline the common structure of the argument.

Theorem 2: (Estimator: Asymptotic Linearity and Normality). Assume (A1-6) and base the second stage estimator, $\hat{\theta}_2$, on an adjusted expectation (D3) with adjustment and window parameters as given in Theorem 1. Letting $G(\theta_0) \equiv \nabla_{\theta'} Q_2(\theta_0)$, $H_0 \equiv \nabla_{\theta\theta'} Q_2(\theta_0)$, and $\Sigma \equiv H_0^{-1} E\left[\sqrt{N} G_0 G_0' \sqrt{N}\right] H_0^{-1}$:

$$\begin{aligned}
a) &: \left|\hat{\theta}_1 - \theta_o\right| = o_p\left(N^{-1/4}\right) \\
b) &: \sqrt{N}\left(\hat{\theta}_2 - \theta_0\right) = -H_0^{-1}\sqrt{N} G(\theta_0) + o_p(1) \\
c) &: \sqrt{N}\left(\hat{\theta}_2^* - \hat{\theta}_2\right) = o_p(1) \\
d) &: \sqrt{N}\left(\hat{\theta}_2^* - \theta_0\right) \xrightarrow{d} Z \tilde{} N(0, \Sigma)
\end{aligned}$$

In the special case when $Var(Y_i|X_i) = \sigma_o^2$ is constant, $\Sigma = -\sigma_o^2 H_0^{-1}$.

Theorem 3. (Test Statistic: Asymptotic Null-Distribution): Let

$$\hat{M} \equiv \hat{E}\left(Y|\hat{V}\right);\ \hat{M}_k \equiv \hat{E}\left(Y|X_k\right);\ \hat{M}_T \equiv \hat{E}_T\left(Y|\hat{V}\right),$$

where the first two expectations are regular with window parameter $r : 1/6 < r < 1/4$ and the third is twicing with window parameter $r_T : 1/8 < r_T < 1/6$. Define:

$$\hat{w}_k \equiv \hat{\tau}_k \hat{M}_k;\ \hat{w}_k^* \equiv \hat{w}_k - \hat{E}\left(\hat{w}_k|\hat{V}\right),$$

where the above expectation is a regular with window parameter r^*: $1/6 < r^* < r < 1/4$. Denote $\hat{T}$ and $\hat{T}^*$ as the unscented and centered moments with respective k^{th} elements:

$$\hat{T}_k\left(\hat{\theta}\right) = \left\langle \hat{\tau}_v\left(Y - \hat{M}_T\right)\hat{w}_k \right\rangle;\ \hat{T}_k^*\left(\hat{\theta}\right) = \left\langle \left(Y - \hat{M}\right)\hat{w}_k^* \right\rangle$$

Then, with $\varepsilon \equiv (Y - M)$, under the null hypothesis of a single index:

$$\begin{aligned}
\mathbf{a}) &: \sqrt{N}\hat{T}_k^*\left(\hat{\theta}\right) = \sqrt{N} S_k + o_p(1),\ S_k = \langle \varepsilon w_k^* \rangle - \langle \nabla_\theta w_k^* \rangle H_0^{-1} G_o \\
\mathbf{b}) &: \sqrt{N}\left[\hat{T}_k^*\left(\hat{\theta}\right) - \hat{T}_k\left(\hat{\theta}\right)\right] = o_p(1) \\
\mathbf{c}) &: \sqrt{N} T' \Sigma^{-1} T \xrightarrow{d} \mathcal{X} \tilde{} \chi^2(K),\ T = \hat{T}^*\left(\hat{\theta}\right), \hat{T}\left(\hat{\theta}\right),
\end{aligned}$$

where with S_k as the k^{th} element of S: $\Sigma \equiv E[SS']$.

To outline the proof for Theorems 2b and 3a (other parts follow directly), note that the moment conditions underlying the estimator and the test statistic have the structure:

$$\sqrt{N}\left\langle \hat{\tau}(Y - \hat{M})\hat{\omega} \right\rangle = \sqrt{N}\left\langle \hat{\tau}(Y - M)\hat{\omega} \right\rangle - \sqrt{N}\left\langle \hat{\tau}(\hat{M} - M)\hat{\omega} \right\rangle.$$

Here $\hat{\omega}$ is an estimated weight vector whose form is given above, depending on whether the above moment conditions describe the estimator or the test statistic. Denote ω as the limiting value of the estimated weight. Then, for the first component above, a mean-squared convergence argument is employed in the Appendix together with a result from Pakes and Pollard (1989) to show that :

$$\sqrt{N}\left\langle \hat{\tau}(Y-M)\hat{\omega}\right\rangle = \sqrt{N}\left\langle \tau(Y-M)\omega\right\rangle + o_p\left(1\right).$$

With $\hat{M} = \hat{f}/\hat{g}$, in the Appendix we show that the second component is within $o_p\left(1\right)$ of

$$\sqrt{N}\left\langle \tau(\hat{M}-M)\omega\hat{g}/g\right\rangle = \sqrt{N}\left\langle \tau(\hat{f}-\hat{g}M)\tau\omega/g\right\rangle.$$

As a U-statistic, this last term can be analyzed by conventional projection arguments. Provided that its expectation tends to zero, this term vanishes for the estimator and the centered test statistic. For the uncentered test statistic, it contributes precisely the term that makes it asymptotically close to the centered form. As discussed in section 2, the above expression will have expectation tending to zero if appropriate higher kernels are employed or when the trimming function, τ, depends only on the index. For both the estimator and the centered test statistic, regular kernels can be employed as this last condition holds.

4 Monte Carlo Designs and Results

In this section Monte Carlo experiments are used to investigate different estimators and test variants. First, we use a Monte Carlo experiment to evaluate different estimators that are defined in Section 3. Our Monte Carlo study shows that the two-stage normal kernel estimator with bias correction has the smallest root mean-square error (RMSE). Therefore, we choose this estimator to study the empirical size and power of different variants of the test statistic.

Second, in evaluating the test statistics T and T^*, we will examine both bias-corrected and regular forms of the test statistics as defined in Section 3. Both test statistics depend on an estimated covariance matrix, and we provide results for two different estimates. With S defined as in Theorem 3, the covariance matrix is given by $\Sigma = E(SS')$, which may be estimated by a sample analogue. Alternatively, with $\varepsilon_i \equiv Y_i - E(Y_i|V_i)$, note that the k^{th} element of S has the form:

$$S_k = \sum_{i=1}^{N} w_{ik}^{*}\varepsilon_i - \left\langle \nabla_\theta w_k^{*}\right\rangle H_0^{-1}\sum_{i=1}^{N}\nabla_\theta E\left(Y_i|V_i\right)\varepsilon_i$$

Accordingly, elements of S will depend on ε_i^2 terms. Taking an iterated expectation and conditioning on X, write:

$$\Sigma = E\left[E(SS'|X)\right]$$

Given the form of S_k, it can be shown that the inner expectation depends on the variance of Y conditioned on the index. If this conditional variance is known to be constant, as it is in all but one of the designs below, then it can be factored out of the above expectation and directly estimated as an average of squared residuals. We will use the terms KCV (known constant conditional variance) and UCV (unknown conditional variance) to refer to these two covariance matrix estimates. Test statistics will be computed and compared under these two covariance matrix estimators.

Third, we compare the performance of our adaptive weight version of the test statistic and the fixed weight version. Recall that the test statistic depends on a weight that is a function of X_k, the k^{th} exogenous variable entering the model. The adaptive or predictive weight is given by $w(X_k) = E(Y|X_k)$, which is the optimal predictor of Y under quadratic loss. Notice that this weight has an unknown functional form that is model dependent. In contrast, a fixed weight has a known functional form that does not depend on the alternative. The fixed weight we use in our Monte Carlo study is the common quadratic weight $w(X_k) = X_k^2$. The Monte Carlo experiment confirms that the adaptive weight version of the test is robust. Namely, in some designs the two versions perform similarly, however, in other designs the adaptive weight strongly dominates the fixed one.

4.1 Designs

All of the designs have single index structures under the null hypothesis. For each design, the alternative does not satisfy a single index assumption. Under the alternative, the first design is a double index model with continuous dependent variables; the second design is a binary response model with index heteroscedasticity; the third design is a double index model with discrete dependent variables; the last two designs are general linear models with no index structure. In all the designs, we normalize such that $E(Y_i|X_i)$ has standard deviation 2 under the null and alternative models.

In the first (basic) design, we use the following data generating method for the null hypothesis:

$$Y_i = M_{0i} + \varepsilon_i,\ M_{0i} \propto (X_{1i} + X_{2i})^2,$$

where the $X\prime s$ ~ $\chi^2(1)$ and ε~$N(0,1)$. Under the alternative:

$$Y_i = M_{1i} + \varepsilon_i,\ M_{1i} \propto [(X_{1i} + X_{2i})^2 + (X_{1i} - X_{2i})^3].$$

The second design is a binary response design. With ε_i being i.i.d. $N(0,1)$, under the null of a single index model:

$$Y_i = \begin{cases} 1: & M_{0i} > \varepsilon_i \\ 0: & \text{otherwise} \end{cases},\ M_{oi} \propto X_{1i} + X_{2i} - 0.5$$

Unlike the previous two designs, here the two $X\prime s$ are correlated. In particular, the $X\prime s$ are linear combinations of the same χ^2 variable and different normal shocks.

The alternative model introduces heteroscedasticity, with $M_{1i}\varepsilon_i$ replacing ε_i above and with $M_{1i} \propto \sqrt{1+(X_{1i}-X_{2i})^2}$. We normalize M_{0i} and M_{0i}/M_{1i} so that they have expectation zero and standard deviation 2. This design is the only one that does not have a constant conditional variance.

Since discrete independent variables are very common in practice, the third design has a discrete regressor. The structure of the null and alternative are the same as in the basic design, however, here X_{2i} is a binary variable.

In the fourth (general linear model) design, under the null hypothesis we generate data by:

$$Y_i = M_{0i} + \varepsilon_i,\ M_{0i} \propto (X_{1i} + X_{2i} + 2X_{3i})^2,$$

where the $X\prime s$ ~ $\chi^2(1)$ and ε~$N(0,1)$. Under the alternative, which has no index structure:

$$Y_i = M_{1i} + \varepsilon_i,\ \ M_{1i} \propto [3{X_{1i}}^2 + 2{X_{2i}}^2 + {X_{3i}}^2 + 3].$$

A fifth design is constructed to compare the properties of our adaptive weight version of the test statistic and the fixed weight version. This design is different from the other four in a way that will be explained below. Under the null hypothesis we generate data by:

$$Y_i = M_{0i} + \varepsilon_i,\ M_{0i} \propto (X_{1i} + X_{2i} + X_{3i})^2,$$

where the $X\prime s$ ~ $N(0,1)$ and ε~$N(0,1)$. Under the alternative, which has no index structure:

$$Y_i = M_{1i} + \varepsilon_i,\ \ M_{1i} \propto [{X_{1i}}^3 + {X_{2i}}^3 + {X_{3i}}^3 + 1].$$

For all the designs, the sample size we use is n=1000, and the number of Monte Carlo replications is 1000. We provide results for theoretical sizes of 0.05 and 0.10.

For the estimator, there are a number of window and trimming choices that need to be specified. With windows having the form $h = O(N^{-r})$, for the stage1 and stage2 estimators, we set r at 1/6.1. Within the range of permissible values, the value gives the fastest point-wise convergence rate of the estimated expectation to the truth. For the smoothing adjustment, we select an optimal pointwise rate of 1/5. Finally, for the twicing kernel, we set this window at 1/7. In the case of trimming, all trimming is based on the .99 quantile for the relevant variables. Recall that in the second stage estimator, we adjust the denominator of estimated expectations. Here, we smoothly keep the index between the .005 and the .995 index quantiles. Recall also that this adjustment depends on a lower density quantile, and we select the .01 quantile for this purpose. Finally the adjustment depends multiplicatively on a window raised to the power of α, $0 < \alpha < 1$. In this case, we set α to be 1/2.

For the test statistics, with one exception given below, we set the window parameter r to be 1/5 for the expectations $E(Y|V)$ and $M_k = E(Y|X_k)$. The window parameter for $E(M_k|V)$ is 1/7. The index-trimming is set at .95, while the X-trimming is set at .99.

4.2 Monte Carlo Results

The first step is to use a Monte Carlo study to evaluate different estimators. The estimators studied are SLS estimators using twicing kernels (SLS-TW), using X-trimming in the first stage (S1SLS), and using index-trimming in the second stage (S2SLS). For each SLS variant there are two versions: having smoothing correction or not; the corrected ones have an extra "C" in front (e.g., CSLS-TW). Among the unadjusted estimators, in the general linear model and basic designs, S2SLS has an RMSE about 30% lower than that of the other estimators. The reduction in RMSE is smaller (8%) in the binary response case and close to zero in the discrete regressor design reported here. This last finding is design dependent and does not hold for other discrete designs.[10]

In terms of RMSE, bias adjusted estimators are quite close to uncorrected ones in all the designs except in the discrete regressor case, where it reduced RMSE by about 16% by cutting the bias in half.[11] Essentially, the bias correction makes little difference when the bias in the uncorrected estimator is very small, but can have a large impact when this bias is large. Hence our conclusion would be that the bias corrected two-stage normal kernel estimator with bias reducing structure is the best choice. Detailed results can be found in Table 1 Estimation Results. Note that with exception of the discrete regressor design, in all the other designs the twicing kernel design are not reported because there are severe outliers resulting in misleading bias and variance values.

After the CS2SLS estimator is chosen, we compare all the variants of test statistics we mentioned in our Monte Carlo study, involving known or unknown conditional variance (KCV or UCV) and different bias reducing mechanisms. The bias reducing mechanisms we employ are Twicing Kernel (TW), Regular Kernel using a window $r > \frac{1}{4}$(BRR); and Recentering. We investigate the empirical size, power, and adjusted power, which is the empirical power using bootstrap critical value adjusting the empirical size to be equal to the theoretical size. For reasons discussed below, our Monte Carlo results recommend the centered BRR as the best among all those variants. In addition, in all cases where the conditional variance is constant, it is better to impose this information.

In comparing different variants of the test statistic, note first that the recentered test statistics perform much better than uncentered ones in that the empirical sizes are much closer to theoretical value and power is also better. The uncentered test statistics in all the designs have highly inflated empirical sizes. As a result, the adjusted power

[10] Specifically, we interchanged quadratic and cubic components so that the conditional mean function was cubic under the null. For this case, we found that the gain is substantial as is shown in the detailed table below:

An Alternative Discrete Regressor Design

		SLS-TW	CSLS-TW	S1SLS	S2SLS	CS2SLS
Discrete Regressor (Flipped)	Bias	0.065	0.032	0.037	0.060	0.038
	Rvar	0.101	0.116	0.093	0.067	0.069
	Rmse	0.120	0.121	0.100	0.090	0.079

[11] The resulting bias reduction substantially improved the performance of the test statistic.

is substantially different from the unadjusted power. For example, turn to the general linear model design. The recentered KCV BRR gives empirical sizes of 0.049 and 0.097 for 5% and 10% theoretical sizes; while the uncentered version gives 0.153 and 0.231 respectively. The recentered test also has better power properties. Similar results occur for the other designs.

Second, we compare results that depend on whether or not a known constant conditional variance (KCV) assumption is correctly imposed in estimating the covariance matrix. In all three designs where this assumption holds, the performance of the test statistic is improved. The sizes are reasonable and similar, but the power of KCV is higher than UCV. For example, in the basic design, the power of the UCV version gives adjusted power of only 0.7 and 0.792 for 5% and 10% theoretical sizes; while the KCV version gives powers of 0.878 and 0.914 respectively. Not surprisingly, a better test statistic results from imposing correct (constant conditional variance) information when estimating the covariance matrix.

As for kernel selection, the results are quite close to one another. However, BRR is the most stable over designs. For the discrete regressor design, the recentered KCV with BRR gives empirical sizes of 0.053 and 0.089 for 5% and 10% theoretical sizes; while the corresponding ones for simple expectation are 0.22 and 0.353; twicing kernel yields 0.177 and 0.282. The power is also slightly better than the other two. The difference among them in other designs is often small. For example in the general linear model design our recentered KCV under BRR gives size power combinations of (0.049, 0.817) and (0.097,0.872) for 5% and 10% theoretical sizes; while corresponding expectation by index gives (0.045,0.85) and (0.089, 0.889); twicing provides (0.045, 0.806) and (0.088, 0.863).

As a conclusion, the recentered test statistic using BRR stands out among all the variations we tried. It performs well under all the designs. Furthermore, when it is known that the conditional variance is constant, such information should be imposed.

To compare fixed with adaptive weights, recall that the fixed weights are the squares of the exogenous variables that appear in the model, while the adaptive or predictive weights are the optimal (MSE) predictors of Y. With the exception of the fifth design, all of the other design have important quadratic elements. As a result, the fixed and adaptive weights explain a comparable proportions of the variation in the dependent variable in those designs. Not surprisingly, in these cases we find, but do not report, that fixed and adaptive weights perform similarly. In the fifth design, which is given above, quadratic elements are not important in the alternative model. Even collectively, such elements only explain 3.5% of the variation in Y. In contrast, collectively the adaptive weights explain 78.1%. Table 3 provides Monte Carlo results for the comparison between our adaptive weight version of the test and the fixed weight version. It is shown that our adaptive weight test statistic dominates the fixed weight version in this design by having much better power results. For example, at the 5% theoretical critical value panel, we find the adaptive weight version of the recentered BRR test with KCV has an empirical power of 0.962; while the number for the fixed weight

version is much lower at 0.706.

5 Conclusions

In summary, we have first developed an estimator that has desirable large sample properties (consistency and asymptotic normality), and that also performs well in finite samples. We have obtained these properties by employing bias controls that make it possible to base the estimator on regular kernels. These finite and large sample properties are important in applied work and are also central to the performance of test statistics that depend on the estimator.

Second, we have formulated a test statistic for testing the frequently made single index assumption in semiparametric models. We establish the large sample distribution of the test statistic under the null hypothesis and show that it performs well across a variety of designs in Monte Carlo experiments. This performance is obtained by an embedded bias control mechanism, the adaptive nature of the test statistic, and also the estimator upon which it is based.

Table 1. Estimation Results

Basic Design			
	S1SLS	S2SLS	CS2SLS
Bias	0.000	0.000	0.000
Rvar	0.042	0.031	0.031
Rmse	0.042	0.031	0.031

Binary Response Design			
	S1SLS	S2SLS	CS2SLS
Bias	-0.010	-0.002	0.000
Rvar	0.063	0.060	0.059
Rmse	0.064	0.060	0.059

Discrete Regressor Design					
	SLS-TW	CSLS-TW	S1SLS	S2SLS	CS2SLS
Bias	-0.019	-0.017	-0.023	-0.036	-0.019
Rvar	0.045	0.046	0.041	0.036	0.037
Rmse	0.049	0.049	0.047	0.050	0.042

General Linear Model Design			
	S1SLS	S2SLS	CS2SLS
Bias	-0.005	0.000	0.001
	0.043	-0.003	-0.004
Rvar	0.089	0.069	0.069
	0.132	0.109	0.109
Rmse	0.089	0.069	0.069
	0.139	0.109	0.109

Table 2. Test Results

Basic Design									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
	size			0.039	0.041			0.085	0.100
R	power			0.721	0.894			0.809	0.923
	adjusted power			0.746	0.903			0.827	0.923
	size	0.237	0.232	0.062	0.063	0.315	0.319	0.108	0.116
TW	power	0.774	0.899	0.729	0.890	0.841	0.930	0.810	0.921
	adjusted power	0.344	0.665	0.711	0.878	0.537	0.764	0.802	0.917
	size	0.219	0.232	0.063	0.065	0.319	0.309	0.110	0.118
BRR	power	0.768	0.905	0.723	0.889	0.842	0.936	0.809	0.920
	adjusted power	0.449	0.744	0.700	0.878	0.617	0.836	0.792	0.914

Binary Response									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
	size			0.049	0.035			0.097	0.067
R	power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	adjusted power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	size	0.147	0.125	0.044	0.022	0.253	0.206	0.098	0.057
TW	power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	adjusted power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	size	0.209	0.167	0.059	0.023	0.335	0.279	0.133	0.056
BRR	power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	adjusted power	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2. Test Results Continued

Discrete Regressor Design									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
	size			0.206	0.220			0.338	0.353
R	power			0.973	0.989			0.989	0.992
	adjusted power			0.911	0.967			0.943	0.978
	size	0.772	0.770	0.167	0.177	0.809	0.813	0.273	0.282
TW	power	0.996	0.996	0.99	0.994	0.996	0.998	0.994	0.995
	adjusted power	0.007	0.069	0.96	0.986	0.027	0.218	0.977	0.992
	size	0.864	0.874	0.051	0.053	0.909	0.901	0.096	0.089
BRR	power	0.996	0.999	0.996	0.997	1.000	1.000	0.996	0.997
	adjusted power	0.014	0.144	0.996	0.997	0.048	0.322	0.996	0.998

General Linear Model Design									
		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
	size			0.034	0.045			0.082	0.089
R	power			0.572	0.850			0.684	0.889
	adjusted power			0.624	0.856			0.717	0.892
	size	0.136	0.132	0.032	0.045	0.207	0.203	0.087	0.088
TW	power	0.565	0.822	0.511	0.806	0.675	0.861	0.629	0.863
	adjusted power	0.301	0.697	0.572	0.817	0.492	0.789	0.655	0.868
	size	0.152	0.153	0.038	0.049	0.227	0.231	0.090	0.097
BRR	power	0.695	0.909	0.522	0.817	0.800	0.931	0.647	0.872
	adjusted power	0.417	0.806	0.585	0.819	0.606	0.874	0.665	0.875

Table 3. Comparison of Fixed Weight and Adaptive Tests

Quadratic Weight Test

		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
R	size			0.036	0.043			0.076	0.091
	power			0.045	0.698			0.094	0.774
	adjusted power			0.062	0.713			0.119	0.783
TW	size	0.139	0.144	0.034	0.040	0.219	0.226	0.075	0.088
	power	0.055	0.716	0.043	0.695	0.109	0.780	0.094	0.774
	adjusted power	0.012	0.559	0.061	0.708	0.020	0.662	0.120	0.783
BRR	size	0.049	0.057	0.038	0.044	0.099	0.106	0.073	0.086
	power	0.044	0.704	0.044	0.697	0.094	0.774	0.095	0.771
	adjusted power	0.045	0.681	0.061	0.706	0.099	0.768	0.126	0.786

Adaptive Test

		5% theoretical critical value				10% theoretical critical value			
		Uncentered		Recentered		Uncentered		Recentered	
		UCV	KCV	UCV	KCV	UCV	KCV	UCV	KCV
R	size			0.043	0.048			0.089	0.095
	power			0.768	0.968			0.845	0.976
	adjusted power			0.786	0.968			0.852	0.976
TW	size	0.466	0.460	0.045	0.049	0.565	0.565	0.093	0.095
	power	0.813	0.975	0.757	0.966	0.877	0.981	0.833	0.976
	adjusted power	0.223	0.822	0.770	0.966	0.349	0.881	0.846	0.976
BRR	size	0.094	0.090	0.045	0.049	0.164	0.159	0.093	0.099
	power	0.796	0.971	0.757	0.962	0.860	0.978	0.830	0.975
	adjusted power	0.727	0.959	0.763	0.962	0.806	0.973	0.842	0.975

References

Ahn, H. (1997), Semiparametric Estimation of a Single-Index Model with Nonparametrically Generated Regressors, *Econometric Theory* 13, 3-31.

Bhattacharaya, P.K. (1967), Estimation of a Probability Density Function and its Derivatives. *Indian Journal of Statistics* Series A, 373-383

Bierens, Herman J., (1990), A consistent conditional moment test of functional form. *Econometrica* **58**, 1443-1458.

Climov, D. , M. Delecroix & L. Simar (2002), Semiparametric estimation in single index Poisson regression: a practical approach, *Journal of Applied Statistics* **29**, 1047-1070.

Delgado, M A. & J. Mora (1995), Nonparametric and semiparametric inference with discrete regressors. *Econometrica* **63**, 1477-1484.

Delgado, M A. & T. Stengos (1994), Semiparametric specification testing of nonnested econometric models. *Review of Economic Studies* **61**, 291-303.

Fraga, M. & O. Martins (2001), Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal, *Journal of Applied Econometrics* **16**, 23-39.

Gerfin, M. (1996), Parametric and Semi-parametric Estimation of the Binary Response Model of Labor Market Participation, *Journal of Applied Econometric* **11**, 321-39.

Gorgens, T. (2000), Semiparametric Estimation of Single-Index Transition Intensities, Econometric Society World Congress 2000 Contributed Papers 0596, Econometric Society

Gorgens,T. & J. L. Horowitz (1999), Semiparametric Estimation of a Censored Regression Model with an Unknown Transformation of the Dependent Variable, *Journal of Econometrics* **90**, 155-191.

Härdle, W. & E. Mammen (1993), Comparing nonparametric versus parametric regression fits. *Annals of Statistics* **21(4),** 1926-1947.

Härdle, W., E. Mammen & M. Müller (1998), Testing parametric versus semiparametric modelling in generalized linear models. *Journal of American Statistical Association* **93**, 1461-1474.

Härdle, W., V. Sponkoiny & S. Sperlich (1997), Semiparametric single index versus fixed link function modelling. *Annals of Statistics* **25**, 212-243.

Hoeffding, H. (1963), Probability Inequalities for Sums of Bounded Random Variables, *Journal of the American Statistical Association* **48**,13-30.

Honore, B. E. & J. L. Powell, 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* (Cambridge: Cambridge University Press, 2005), 520–53.

Horowitz, J. L. & V. G. Sponkoiny (2001), An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative, *Econometrica* **69**, 599-631.

Horowitz, J. L. & W. Härdle (1994), Testing a parametric model against a semiparametric alternative. *Econometric Theory* **10**, 821-848.

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

Klein, R. W. (1993), Specification tests for binary choice models based on index quantiles, *Journal of Econometrics* **59**, 343-375

Klein, R. W. & R. H. Spady (1993), An Efficient Semiparametric Estimator for Binary Response Models, *Econometrica* **61**, 387-421

Klein, R. W. & F. Vella (2007), Estimating a class of triangular simultaneous equations models without exclusion restrictions, manuscript.

Klein, R. W., C. Shen & F. Vella (2009), Joint Binary Selection and Treatment Models, manuscript.

Newey, W. K. (1985), Maximum likelihood specification testing and conditional moment tests, *Econometrica* **53**, 1047-1070.

Newey, W. K., F. Hsieh & J. Robins (2004), Twicing Kernels and a Small Bias Property of Semiparametric Estimators, *Econometrica* **72**, 947- 962.

Pakes, A. & D. Pollard, (1989), Simulation and the asymptotics of optimization estimators, *Econometrica* **57**, 1027-1058.

Powell, J. L. , J. H. Stock & T. M. Stoker (1989), Semiparametric Estimation of Weighted Average Derivatives, *Econometrica* **57**, 1403-1430.

Serfling, R. J. (1980), Approximation Theorems of Mathematical Statistics, New York: John Wiley & Sons.

Shen, C. (2009), Determinants of Healthcare Decisions: Insurance, Utilization and Expenditures, manuscript.

Tripathy, G. & Y. Kitamura (2003), Testing Conditional Moment Restrictions, *Annals of Statistic* 31, 2059-2095.

6 Appendix

6.1 Main Results

In the proofs of Theorems 1-2 below we provide proofs for the large sample properties of the second stage estimator, with the argument for the first-stage estimator being similar but shorter as it is based on a regular expectation. In so doing, we simplify notation by not subscripting objective functions, gradients, and hessian expressions.

Proof of Theorem 1. (Consistency: $\hat{\theta}_2$). Define:

$$\hat{Q}(\theta) \equiv \left\langle \hat{\tau}_v \left[Y - \hat{f}/\hat{g}^* \right]^2 \right\rangle ;\ Q(\theta) \equiv \left\langle \tau_v \left[Y - f/g \right]^2 \right\rangle$$

Then, recalling (D3), letting $\delta_i \equiv \hat{\tau}_v \left| \hat{f}_i/\hat{g}_i^* - f_i/g_i \right|$, and $\varepsilon_i \equiv Y_i - f_i/g_i$:

$$\left| \hat{Q}(\theta) - Q(\theta) \right| \leq C + S + T$$

$$C \equiv 2\left\langle |Y|\,\delta \right\rangle;\ S \equiv \left\langle \left(\hat{f}/\hat{g}^{*} + f/g \right) \delta \right\rangle;\ T \equiv \frac{1}{N}\sum |\hat{\tau}_{vi} - \tau_{vi}|\,\varepsilon_i^2$$

For C, with $\hat{C}_1^2 \equiv 4\left\langle \hat{\tau}_v Y^2 \right\rangle = O_p(1)$, from Cauchy's inequality and Lemma 5:

$$C \leq \hat{C}_1 \left\langle\ \delta^2 \right\rangle^{1/2} = O_p(1) \left\langle\ \delta^2 \right\rangle^{1/2} = o_p(1)$$

With a similar argument holding for S and T, $\hat{Q}(\theta)$ converges uniformly in θ to $Q(\theta)$ in probability. From standard arguments, $Q(\theta)$ converges uniformly to $E[Q(\theta)]$ in probability. Therefore:

$$\sup_{\theta} \left| \hat{Q}(\theta) - E[Q(\theta)] \right| \xrightarrow{p} 0.$$

From Ichimura (1993), $E[Q(\theta)]$ is uniquely maximized at θ_0, which completes the proof.

We provide the proof for the asymptotic linear characterization in Theorem 2(b); other results are similarly obtained or follow directly.

Proof of Theorem 2. (Asymptotic Normality: $\hat{\theta}_2$). With $\hat{H}(\theta) \equiv \nabla_{\theta\theta'}\hat{Q}(\theta)$ and $\hat{G}(\theta) \equiv \nabla_{\theta'}\hat{Q}(\theta)$, from a Taylor series expansion:

$$\sqrt{N}\left(\hat{\theta}_2 - \theta_0\right) = -\left[\hat{H}\left(\theta^{+}\right)^{-1}\right]\left[\sqrt{N}\hat{G}(\theta_0)\right],\ \ \theta^{+}\ \epsilon\ \left[\hat{\theta}_2, \theta_0\right].$$

For the Hessian, with $H(\theta) \equiv \nabla_{\theta\theta'}Q(\theta)$:

$$\sup_{\theta}\left|\hat{H}(\theta) - EH(\theta)\right| \leq \sup_{\theta}\left|\hat{H}(\theta) - H(\theta))\right| + \sup_{\theta}|H(\theta) - EH(\theta))|$$

From Lemma 5, the first term converges in probability to 0. From standard arguments, the second term also converges in probability to zero. Therefore, as $\theta^{+} \xrightarrow{p} \theta_0$: $\hat{H}\left(\theta^{+}\right) \xrightarrow{p} EH(\theta_0) \equiv H_0$

For the gradient, with $\hat{w} \equiv \nabla_{\theta}\hat{M}$:

$$\sqrt{N}\hat{G}(\theta_0) = \sqrt{N}\left[\left\langle [\ Y - M]\,\hat{\tau}\hat{w}\right\rangle - \left\langle \left[\hat{M} - M\right]\hat{\tau}\hat{w}\right\rangle\right] \equiv \sqrt{N}\left[\hat{G}_A - \hat{G}_B\right],$$

For $\hat{G}_A$, with $\varepsilon \equiv Y - M$, and $G_A \equiv \langle [Y - M]\tau w\rangle$,

$$\sqrt{N}\left[\hat{G}_A - G_A\right] = \sqrt{N}\left[\Delta_1 + \Delta_2 + \Delta_3\right],$$

$$\Delta_1 \equiv \left\langle \varepsilon\tau\left[\hat{w} - w\right]\right\rangle;\ \Delta_2 \equiv \left\langle \varepsilon\left[\hat{\tau} - \tau\right]w\right\rangle;\ \Delta_3 \equiv \left\langle \varepsilon\left[\hat{\tau} - \tau\right]\left[\hat{w} - w\right]\right\rangle$$

From Lemma 9, $\Delta_1 \xrightarrow{p} 0$. For Δ_2, let

$$\tau_i \equiv 1\left\{c_{1o} < v_i(\theta_o) < c_{2o}\right\} \equiv \tau_i(\alpha_o),\ \alpha_o \equiv [\theta_o, c_{1o}, c_{2o}]$$

Employing a similar strategy to that in Klein (1993), let $N_\varepsilon \equiv \langle \alpha : |\alpha - \alpha_o| < \varepsilon \rangle$, $\varepsilon = o(1)$. Then, $\sqrt{N}\Delta_2 = o_p(1)$ if

$$\sup_{N_\varepsilon} N^{1/2} \sum \left[\tau_i(\alpha) - \tau_i(\alpha_o)\right] \varepsilon_i w_i / N = o_p(1)$$

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

Turning to Δ_3, let $\tau^*(\hat{\alpha})$ be an indicator on the union of the sets over which $\tau(\hat{\alpha})$ and $\tau(\alpha_o)$ are defined. Then:

$$\sqrt{N}\,|\Delta_3| \;\leq\; \sqrt{N}\;\langle |\varepsilon|\,|\tau(\hat{\alpha}) - \tau(\alpha_o)|\,\tau^*(\hat{\alpha})\,|\hat{w} - w|\rangle \;\leq \sqrt{N}\,|\Delta_{31}|\,|\Delta_{32}|,$$

$$|\Delta_{31}| \;\equiv\; \left[\sum \varepsilon_i^2 \left[\tau_i(\hat{\alpha}) - \tau_i(\alpha_o)\right]^2 / N\right]^{1/2},\; |\Delta_{32}| = \left[\sup_{N_\varepsilon} \sum \tau_i^*(\alpha)\left[\hat{w}_i - w_i\right]^2 / N\right]^{1/2}$$

To analyze $|\Delta_{31}|$, for $k = 1, 2$ let:

$$\begin{aligned} S(z) &\equiv \left\{1 + \exp[-\left(\, N^{-(s-\varepsilon)} + z\right) / N^{-(s-\varepsilon)/2}\right\}^{-1},\; 0 < \varepsilon < s \\ S_k^* &\equiv S\left(|v(\theta) - v_0| + |c_k - c_{k0}| - |v_0 - c_{k0}|\right) + 1 - S(0),\; k = 1, 2. \end{aligned}$$

Then, from from Klein (1993, Lemma A.1):

$$|\tau_i(\alpha_o) - \tau(\hat{\alpha})| \leq S_1^* + S_2^*.$$

Let $\delta_N \equiv |v - v_0| + |c_1 - c_{10}|$, $w_k \equiv |v_0 - c_{k0}|$, and write $S_k^* \equiv S(\delta_N - w_k) + 1 - S(0)$. Note that $|\hat{\alpha} - \alpha_o| = N^{-(s+\varepsilon)}$, $s > 1/4$ and that $1 - S(0) = o_p(N^{-s})$. As in Klein (1993, Lemma A.2), Taylor expand $S(\delta_N - w_k)$ in δ_N about $\delta_N = 0$. Assuming that $\mathrm{E}(\varepsilon_i^2 | X_i)$ is bounded, it can then be shown that $|\Delta_{31}|^2 = o_p\left(N^{-1/2}\right)$ Since $|\Delta_{32}|^2 = O_p\left(N^{-1/2}\right)$, the result follows.

For $\hat{G}_B$, from Lemma 9:

$$\sqrt{N}\left[\hat{G}_B - \hat{G}_B^*\right] \xrightarrow{p} 0,\; \hat{G}_B^* \equiv \left\langle \left[\hat{M} - M\right] \tau w \right\rangle$$

Next, noting that $\hat{M}_i \equiv \hat{f}_i / \hat{g}_i$, recalling the definition of K_{ij} in (D3), letting $\rho_{ij} \equiv \frac{1}{h}\left[Y_j K_{ij} - M K_{ij}\right]\left[\tau_i w_i\right] / g_i$, and employing Lemma 10:

$$\begin{aligned} &\sqrt{N}\left[\hat{G}_B^* - U_N\right] \xrightarrow{p} 0, \\ U_N \;&\equiv\; \left\langle \left(\hat{f}/\hat{g} - M\right) w \frac{\hat{g}}{g} \right\rangle = \left\langle \left(\hat{f} - \hat{g}M\right) \frac{w}{g} \right\rangle = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \rho_{ij} / N. \end{aligned}$$

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

A U-statistic has the form:

$$\left(\begin{array}{c} N \\ 2 \end{array} \right)^{-1} \sum_i \sum_{j>i} \rho_{ij}^*,\ \rho_{ij}^* = \rho_{ji}^*.$$

For U_N above:

$$\begin{aligned} N(N-1)U_N &= \sum_i \sum_{j>i} \rho_{ij} + \sum_i \sum_{j<i} \rho_{ij} = \sum_i \sum_{j>i} \rho_{ij} + \sum_j \sum_{i>j} \rho_{ij} \\ &= \sum_i \sum_{j>i} \rho_{ij} + \sum_j \sum_{i>j} \rho_{ij} = \sum_i \sum_{j>i} \rho_{ij} + \sum_i \sum_{j>i} \rho_{ji}. \end{aligned}$$

Therefore, with $\rho_{ij}^* = \left[\rho_{ij} + \rho_{ji}\right]/2$, U_N has the conventional U-statistic form.

As discussed in section 2.1, with $U_N = B_S^*$, $\mathrm{E}\left(\rho_{ij}\right) = 0 \Longrightarrow E(U_N) = 0$ under index trimming and the residual property of w_i. Since U_N is a centered U-statistic and since it can be shown that $E(\rho_{ij}^{*\prime}\rho_{ij}^*) = o(N)$, then (see Serfling(1980) and Powell, Stock, and Stoker(1989)): $\sqrt{N}[U_N - \hat{U}_N] = o_p(1)$, where:

$$\sqrt{N}\hat{U}_N \equiv N^{-1/2} \sum_i \left[E\left(\rho_{ij} \mid X_i, Y_i\right) + E\left(\rho_{ji} \mid X_i, Y_i\right)\right] \equiv T_1 + T_2$$

For $T_1 : E\left(\rho_{ij}\right) = 0 \Rightarrow E(T_1) = 0$. As $E\left(\rho_{ij} \mid X_i, Y_i\right) = O\left(h^2\right)$, it may be shown that $Var(T_1) \to 0$. It follows that $T_1 = o_p(1)$. The T_2-term vanishes because $E(w_j|V_j) = 0$.

Therefore, $\sqrt{N}\hat{G}_B \xrightarrow{p} 0$, from which it follows that

$$\sqrt{N}\left(\hat{\theta}_2 - \theta_0\right) = -H_0^{-1}\left[\sqrt{N}\left\langle [Y - M]\tau w \right\rangle\right]$$

Asymptotic normality now follows from a standard central limit theorem.

Below we provide the proof for Theorem 3(a) which characterizes the k^{th}centered moment underlying the test statistic. Parts (b-c) of the theorem are either immediate or have arguments similar to (a).

Proof of Theorem 3. (Test Statistic: Asymptotic Null-Distribution) Define:

$$\sqrt{N}\hat{T}_k^*\left(\hat{\theta}\right) = \sqrt{N}\left\langle \left(Y - \hat{M}\left(\hat{\theta}\right)\right)\left(\hat{M}_k - \hat{E}(\hat{M}_k|V\left(\hat{\theta}\right)\right)\right\rangle$$

From a Taylor series expansion, Theorem 2, and Lemma 3:

$$\sqrt{N}\hat{T}_k^*\left(\hat{\theta}\right) = \sqrt{N}\hat{T}_k^*(\theta_0) - R_k,\ R_k \equiv \left\langle \nabla_\theta w_k^* \right\rangle H_0^{-1} G_o + o_p(1)$$

With $\hat{w}_k^* \equiv \hat{M}_k - \hat{E}\left(\hat{M}_k|V\right)$, write $\hat{T}_k^*(\theta_0) = \hat{T}_{Ak}^* + \hat{T}_{Bk}^*$, where:

$$\hat{T}_{Ak}^* \equiv \left\langle\ (Y - M)\right\rangle \hat{w}_k^*;\ \hat{T}_{Bk}^* \equiv -\left\langle \left(\hat{M} - M\right)\hat{w}_k^* \right\rangle$$

Analogous to the proof of Theorem 2, we show that $\sqrt{N}\hat{T}^*_{Ak} = \sqrt{N}T^*_k + o_p(1)$ and $\sqrt{N}\hat{T}_{Bk} = o_p(1)$, where:

$$T^*_k(\theta_0) = \langle (Y - M)(M_k - E(M_k|V)) \rangle$$

Write $\sqrt{N}\left[\hat{T}_{Ak} - T^*_k\right]$ as:

$$\sqrt{N}\left\langle (Y_i - M_i)\left\{\left[\hat{M}_k - M_k\right] - \left[\hat{E}(\hat{M}_k|V) - E(M|V)\right]\right\}\right\rangle$$

Convergence in probability to zero then follows for the first component from Lemma 7 and for the second component from Lemma 8. For $\hat{T}_{Bk}$, from lemmas 9 -10, $\sqrt{N}\hat{T}_B = o_p(1)$ by an argument similar to that for $\hat{G}_B$ in Theorem 2. Hence:

$$\sqrt{N}\hat{T}^*_k\left(\hat{\theta}\right) = \sqrt{N}T^*_k(\theta_0) - R_k(\theta_0) + o_p(1)$$

6.2 Intermediate Lemmas:

6.2.1 Convergence Rates

The proof of the following Lemma is due to Bhatacharaya (1967) and relies on an exponential bound due to Hoeffding (1963). A version of the proof is also contained in Klein (1993).

Lemma 1. (Uniform Convergence Rates for Bounded Functions). With z_i i.i.d., Let $m_i \equiv m(t; z_i, \theta)$ be random variables such that:

$$\left|m_i N^{-s}\right| = O(1)$$

Then, for θ and t in compact sets, m as the vector with i^{th} element m_i, and $\delta > 0$:

$$\sup_{t,\theta} |\langle m \rangle - E[\langle m \rangle]| = o_p\left(N^{-(1/2)+s+\delta}\right)$$

Lemma 2. Assume:

$$\langle \hat{a}\hat{a} \rangle = O_p\left(N^{-1}h^{-s}\right), \quad \left\langle \hat{b}\hat{b} \right\rangle = O_p\left(N^{-1}h^{-t}\right),$$

where $s + t < 6$. Then, with $h = O(N^{-r})$, $r < 1/6$: $\sqrt{N}\left\langle \hat{a}\hat{b} \right\rangle = o_p(1)$

Proof. The proof follows directly from Cauchy's inequality:

$$\left[\sqrt{N}\left\langle \hat{a}\hat{b} \right\rangle\right]^2 \leq N \langle \hat{a}\hat{a} \rangle \left\langle \hat{b}\hat{b} \right\rangle$$

Lemma 3. (Convergence Rates) For V a continuous random variable with density g_v, let $\nabla^d_\theta(g_v)$ be the d^{th} partial derivative of g with respect to θ, $\nabla^0_\theta(\hat{g}_v) \equiv \hat{g}_v$. Let $\hat{\psi}(t;\theta)$ refer to either $\hat{g}_v(t;v)$ or $\hat{f}_v(t;v)$ and let $\psi(t;\theta)$ refer to the corresponding true

functions, g_v or f_v. Then, for θ in a compact set and t in a compact subset of the support of V, the following rates hold for $d = 0, 1, 2$:

$$a) \quad : \quad \sup_{t,\theta} E\left\{\left[\nabla_\theta^d\left(\hat{\psi}\left(t;\theta\right)\right) - E\left(\ \nabla_\theta^d\left(\hat{\psi}\left(t;\theta\right)\right)\ \right)\right]^2\right\} = O\left(\frac{1}{Nh^{2d+1}}\right)$$

$$b) \quad : \quad \sup_{t,\theta}\left|E\left(\ \nabla_\theta^d\left(\hat{\psi}\left(t;\theta\right)\right)\ \right) - \nabla_\theta^d\left(\psi\left(t;\theta\right)\right)\right| = O\left(h^2\right)$$

Proof. As the proof is standard (e.g., see Klein, 1993), we outline it below. When $\psi\left(t;\theta\right) = g$ and $d = 1$. The variance calculation in (a) is immediate[13]. For the bias calculation (b), write $E\left[\nabla_\theta^1\left(\hat{g}\left(t;\theta\right)\right)\right]$ as:

$$\begin{aligned}\frac{1}{h}\int \nabla_\theta^1\left(\ K\left[\left(t-v\right)/h\right]\right) g_x(x)dx &= \frac{1}{h}\nabla_\theta^1\int K\left[\left(t-v\right)/h\right] g_x(x)dx = \\ \frac{1}{h}\nabla_\theta^1\int K\left[\left(t-v\right)/h\right] g_v(v)dv &= \nabla_\theta^1\int K\left(z\right) g_v(t+hz)dz\end{aligned}$$

The result now follows from a standard Taylor expansion in h, with t restricted to be away from the support boundary for V.

The test statistic depends on the marginal expectation of Y conditioned separately on each variable in the index. For a discrete variable Z, $E(Y|Z=t)$ can be estimated as the sample mean of Y for those observations at the support point or by using the same kernel representation employed for continuous random variables. Delgado and Mora (1995) provide a similar result using the nearest neighbor estimator. As the argument for kernels is very short, we provide it below.

Lemma 4. (Discrete Regressors). Let Z be a discrete random variable with support points $t_k : Pr\left(Z = t_k\right) > 0$. With t as one of these points, define the sample mean:

$$\bar{Y}\left(t\right) \equiv \sum_{Z_j\,=\,t} Y_j/N\left(t\right),$$

where $N(t)$ is the number of sample observations for which the random variable $Z = t$. Assuming $E\left|Y_j\right|$ is bounded,and that $\hat{E}$ is a regular expectation with window parameter $r > 0$ (D3), then:

$$\left|\hat{E}(Y|Z=t) - \bar{Y}\left(t\right)\right| = O_p\left(1/N\right)$$

[13]The estimator has the form:

$$\sum \frac{1}{h^2} k\left[\left(t-w_i\right)/h\right]/N$$

With the bias term vanishing faster than the second moment term, the order of the variance is given by:

$$\frac{E\left(k^2\left[\left(t-w_i\right)/h\right]\right)}{h^4 N}$$

Letting $z = \left(w-t\right)/h$, a factor of h disappears in the Jacobian; the result follows.

Proof. With $\{\bullet\}$ as an indicator on the indicated set, by definition $\hat{E}(Y|Z=t)$ is given as:

$$\frac{\sum\{Z_j=t\}Y_jK(0)+\sum\{Z_j\neq t\}Y_jK\left[(t-Z_j)/h\right]}{\sum\{Z_j=t\}K(0)+\sum\{Z_j\neq t\}K\left[(t-Z_j)/h\right]}\equiv\frac{\bar{Y}(t)+\Delta_1}{1+\Delta_0},$$

$$\Delta_d\equiv\sum_{Z_j\neq\ t}Y_j^dK\left(\left[t-Z_j\right]/h\right)/\left[N(t)K(0\right],\ d=0,1$$

Then,

$$\left|\bar{Y}(t)-\hat{E}(Y|Z=t)\right|=\left|\left[\Delta_0\bar{Y}(t)-\Delta_1\right]\right|/\left[1+\Delta_0\right]\leq|\Delta_1|+\left|\bar{Y}(t)\right||\Delta_0|$$

For $|t-Z_j|>c$, a fixed positive and finite constant, $K\left(\left[t-Z_j\right]/h\right)=o\left(1/N^2\right)$. For the normal-kernel case, this term vanishes at an exponential rate. The result follows by taking expectations of both sides.

To establish consistency for the estimator, we require the relative convergence results below.

Lemma 5. (Adjusted Expectations) Recalling that X is bounded and that θ lies in a compact set, assume that $E\equiv E(Y|V)$ is bounded, where V is the index. From the tail condition (A6), Y has tails thinner than a t-distribution with $df\geqq 4$ degrees of freedom. Define $\lambda\equiv df/(df-1)$ and let $\varepsilon,\delta>0$. Recalling the adjustment parameter α in (D3), let $\hat{E}_A$ be an adjusted expectation with adjustment parameter $\alpha:0<\alpha<1/2$ and window parameter r :

$$0<r<\frac{1/2-\delta}{\lambda(1+\alpha)+\varepsilon}$$

Then, with ∇_θ^k as the partial derivative operator as defined above and recalling the definition of $\hat{g}^*(t)$:

$$(1):\sup_\theta\left\langle\left[\hat{E}_A-E\right]^2\right\rangle=o_p(1)$$

From (D3), recall that $\hat{E}_a\equiv\hat{f}(x\theta;\theta)/\hat{g}^*(x\theta;\theta)$. Assume that ∇_θ^kE and ∇_θ^kg are $O(1),k=0,1,2$. From (D3), recall that $\hat{E}_a\equiv\hat{f}(x\theta;\theta)/\hat{g}^*(x\theta;\theta)$. Let the window parameter satisfy:

$$0<r<\frac{1/2-\delta}{\lambda(1+k)+\varepsilon}$$

Then, for θ in an $o_p(1)$ neighborhood of θ_0, $k=0,1,2$, and for $D=\nabla_\theta^k\left[\hat{f}(x\theta;\theta)-f(x\theta;\theta)\right]$ or $\nabla_\theta^k\left[\hat{g}^*(x\theta;\theta)-\hat{g}^*(x\theta;\theta)\right]$:

$$(2)\sup_{x,\theta}\hat{\tau}\left(x\hat{\theta}\right)D/\hat{g}^*(x\theta;\theta)^a=o_p(1)$$

Proof. For (1), since $f(t)/g(t)$ is by assumption bounded, it suffices to show T_f, $T_g = o_p(1)$:

$$T_f \equiv \sup_\theta \left\langle \left[\left(\hat{f} - f\right)/\hat{g}^*\right]^2 \right\rangle;\ T_g \equiv \sup_\theta \left\langle \left[(\hat{g} - g)/\hat{g}^* \right]^2 \right\rangle.$$

For T_f (the proof for T_g is similar), write $T_f \leq A + B$, where:

$$A \equiv \sup_\theta \left\langle \left|\left(\hat{f} - E\hat{f}\right)/\hat{g}^*\right| \right\rangle;\ B \equiv \sup_\theta \left\langle \left[\left(E\hat{f} - f\right)/\hat{g}^*\right]^2 \right\rangle$$

Each of these terms is examined below.

A: Relative Convergence to Expectation

With $b > 0$, let $b_j \equiv 1$ if $|Y_j| < h^{-b}$ and 0 otherwise. Following a strategy employed by Ichimura (1993), consider separately bounded and unbounded regions for Y_j. Letting $K_j \equiv K\left[(t - v_j(\theta))/h \right]$, define:

$$\hat{f}_b(t) \equiv \sum_{j=1}^N \frac{b_j Y_j}{hN} K_j; \quad \hat{f}_u(t) \equiv \sum_{j=1}^N \frac{(1 - b_j) Y_j}{hN} K_j$$

Then, $A \leq A_b + A_u$, where:

$$\begin{aligned} A_b &\equiv \sup_{\theta,t} \left| \left(\hat{f}_b(t) - E\hat{f}_b(t)\right)/\hat{g}^*(t) \right| \\ A_u &\equiv \sup_{\theta,t} \left| \left(\hat{f}_u(t) - E\hat{f}_u(t)\right)/\hat{g}^*(t) \right|. \end{aligned}$$

Recall that $\hat{g}^*(t) \equiv \hat{g}(t) + h^a \hat{q}(1 - \hat{\tau})$, where $(1 - \hat{\tau})$ is a smoothed indicator that depends on lower and upper sample quantiles denoted by $\hat{q}_a$ and $\hat{q}_b$. With q_a and q_b as the corresponding population quantiles, let $\mathcal{A}^* \equiv \{g : q_a^* < t < q_b^*\}$ be a fixed subset of the support for V that contains $\mathcal{A} \equiv \{t : q_a < t < q_b\}$. Define $\tau^*(t)$ as the indicator on $\mathcal{A}^*$, then letting

$$\Delta_b \equiv h^{-a} \sup_{\theta,t} \left| \left(\hat{f}_b(t) - E\hat{f}_b(t)\right) \right|$$

$$A_b \leq \Delta_b \left[h^a \sup_{\theta,t} \left| [\tau^*(t)/\hat{g}(t)] \right| + h^a \sup_{\theta,t} \left| [1 - \tau^*(t)]/\hat{\rho}(t) \right| \right]$$

On $\mathcal{A}^*$, $\inf \hat{g}(t) \xrightarrow{p} \underline{g} > 0$. On the complement, $\inf \hat{\tau} \xrightarrow{p} 0$. Therefore, $A_b \xrightarrow{p} 0$ if $\Delta_b \xrightarrow{p} 0$. From Lemma 4:

$$\Delta_b = O\left(h^{-a}\right) o_p\left(h^{-1-b} N^{-1/2+\delta}\right),\ \delta > 0.$$

Since $h = O\left(N^{-r}\right)$, $\Delta_b \xrightarrow{p} 0$ for $r < (1/2 - \delta)/(1 + a + b)$.

For A_u, $A_u \xrightarrow{p} 0$ if $\Delta_u \xrightarrow{p} 0$, where:

$$\Delta_u \equiv h^{-a} \sup_{\theta,t} \left| \left(\hat{f}_u(t) - E\hat{f}_u(t) \right) \right| \leq h^{-a} \sup_{\theta,t} \left| \hat{f}_u(t) \right| + h^{-a} \sup_{\theta,t} \left| E\hat{f}_u(t) \right|$$

With a similar argument holding for both terms, for the first term:

$$h^{-a} E \sup_{\theta,t} \left| \hat{f}_u(t) \right| \leq h^{-a-1} \frac{1}{N} \sum_j E\left[(1 - b_j) |Y_j| \right],$$

Employing the tail assumption on Y_j , it suffices to show convergence to zero for:

$$h^{-(1+a)} \int_{h^{-b}}^{\infty} y / \left(\left[1 + y^2 \right]^{(df+1)/2} \right) dy \leq h^{-(1+a)} \int_{h^{-b}}^{\infty} y / \left(y^{(df+1)} \right) dy.$$

With $df > 1$, the above bound is

$$O\left[h^{-(1+a)} h^{(df-1)b} \right],$$

which converges to zero for

$$b = \varepsilon + (a+1) / (df - 1), \varepsilon > 0.$$

Combining this restriction with that on r above (for $\Delta_b \xrightarrow{p} 0$) and letting $\lambda \equiv df/(df-1)$, the uniform convergence for term A follows with: $r < (1/2 - \delta)/[\lambda(1+\alpha) + \varepsilon]$.

B: Relative Bias

Let X_k be a continuous variable supported on $[a_k, b_k]$. For $c : 2a < c < 1$, write $\mathbf{1}(\mathcal{A})$ as an indicator on $\mathcal{A}$, and with X_k , $k = 1, ..., K_c$ as a continuous component of X, define:

$$S_N \equiv \{x : a_k + h^c < x_k < b_k - h^c,\ k = 1, ..., K_c\}$$

where the product is taken of the $k = 1, ..., K_c$ continuous X-variables. On S_N, $\sup_\theta B \xrightarrow{p} 0$ from Lemma 3. On the complement of S_N, it can be shown that B vanishes if the probability on this set vanishes sufficiently fast ($0 < 2\alpha < c$).

The proof for (2) can be based on a similar argument. Alternatively, we can exploit trimming to establish uniformity in an $o_p(1)$ neighborhood of θ_0. To outline the argument for one of the terms in (2), write:

$$\Delta \equiv \mathbf{1}\left(\hat{a} - X\left(\hat{\theta} - \theta \right) < X\theta < \hat{b} + X\left(\theta - \hat{\theta} \right) \right) \left| \nabla_\theta^k \left[\hat{f}(x\theta; \theta) - f(x\theta; \theta) \right] \right| / \hat{g}^*(x\theta; \theta)^a$$

Denote $|c|$ as the vector with i^{th} element $|c_i|$. Then, with $\hat{\delta} \equiv |X| \left|\hat{\theta} - \theta\right|$ and with τ as the indicator on $X\theta$ s.t. $\hat{a} - \hat{\delta} < X\theta < \hat{b} + \hat{\delta}$, Δ is bounded above by:

$$\begin{aligned} & \tau \left|\nabla_\theta^k \left[\hat{f}(x\theta;\theta) - f(x\theta;\theta)\right]\right| / \hat{g}^*(x\theta;\theta)^a \\ \leq & \frac{\tau}{\hat{g}^*(x\theta;\theta)^a}\left[\left|\nabla_\theta^k \left[\hat{f}(x\theta;\theta) - E\hat{f}(x\theta;\theta)\right]\right| + \left|\nabla_\theta^k \left[E\hat{f}(x\theta;\theta) - f(x\theta;\theta)\right]\right|\right] \end{aligned}$$

The proof for the first term is similar but simpler to that in (1) because $\hat{g}^*$ is uniformly close to g and in large samples τg is bounded away from 0. The argument for the second term follows from Lemma 3b with minor modifications.

The following lemma provides a result that is useful for the recentered test statistic.

Lemma 6. Let $\tau_k M_{ik} \equiv E(Y_i|X_{ik})$ and $\hat{M}_{ik} \equiv \hat{E}_I(Y_i|X_{ik})$. Writing the window $h = O(N^{-r})$, refer to r as a window parameter. Then, assume that this estimated expectation is regular (D3) with window parameter $r_I : 1/6 < r_I < 1/4$. Consider the estimated "outer" expectation $\hat{E}_o(\hat{M}|V)$ and assume that it is regular with outer window parameter $r_o < r_I$. Below, we subscript h according to the window parameter upon which it is based (e.g., $h_o = O(N^{-r_o})$)Then:

$$\Delta_E \equiv \frac{1}{N}\sum_i \left[\hat{E}_o\left(\hat{M}_{ik}|V_{ik}\right) - \hat{E}_o\left(M_{ik}|V_{ik}\right)\right]^2 = o_p\left(N^{-1/2}\right)$$

Proof. By definition, with $\delta_j \equiv \hat{M}_{jk} - M_{jk}$, $\Delta_E \leq \Delta_{E1} + \Delta_{E2}$, where:

$$\begin{aligned} \Delta_{E1} &\equiv \frac{1}{N}\sum_i \left[\sum_{j\neq i} \frac{1}{h_o^2 (N-1)^2} \delta_j^2 K_{ij}^2\right] \\ \Delta_{E2} &\equiv \frac{1}{N}\sum_i \left[\sum_s \sum_{r\neq s} \frac{1}{(N-1)^2} \frac{|\delta_r| K_{ir}}{h_o} \frac{|\delta_s| K_{is}}{h_o}\right] \end{aligned}$$

Note that

$$|a||b| \leq \max(a^2, b^2) \leq a^2 + b^2$$

Therefore, for Δ_{E2}, which converges in probability to 0 slower than Δ_{E1} :

$$0 < \Delta_{E2} < \frac{1}{N}\sum_i \left[\sum_s \sum_{r\neq s} \frac{1}{h_o^2 (N-1)^2} \left[\frac{\delta_r^2 K_{ir}^2}{h_o^2} + \frac{\delta_s^2 K_{is}^2}{h_o^2}\right]\right]$$

It suffices to show that $E(\Delta_{E2}) = o\left(N^{-1/2}\right)$. From above:

$$E(\Delta_{E2}) = O(1)\left[E\left(\frac{\delta_r^2 K_{ir}^2}{h_o^2}\right) + E\left(\frac{\delta_s^2 K_{is}^2}{h_o^2}\right)\right]$$

Proceeding with the first term (the analysis for the second is identical), write: $\delta_r = \delta_r[i] + \delta_r^*[i]$, where $\delta_r^*[i] = O(1/hN)$ is the component of δ_r that depends on i and $\delta_r[i]$ is the remaining component after the i^{th} term has been removed. It can be shown that

$$\begin{aligned} E\left(\frac{\delta_r^2 K_{ir}^2}{h_o^2}\right) &= E\left(\frac{\delta_r^2[i] K_{ir}^2}{h_o^2}\right) + o\left(N^{-1/2}\right) \\ &= \frac{1}{h_o} E\left[E\left(\delta_r^2[i] \mid X_r\right) E\left(\frac{1}{h_o} K_{ir}^2 \mid X_r\right)\right] + o\left(N^{-1/2}\right) \end{aligned}$$

The first inner expectation is uniformly $O[\max(h_I^4, 1/(Nh_I)]$ while the second inner expectation is uniformly $O(1)$. Therefore, with $h_e = O(N^{-r_e}), h = O(N^r)$, and $r_e < r$:

$$E\left(\frac{\delta_r^2 K_{ir}^2}{h_o^2}\right) = O\left[\max\left(h_I^3\right), 1/\left(Nh_I^2\right)\right] + o\left(N^{-1/2}\right) = o\left(N^{-1/2}\right),\ 1/6 < r < 1/4.$$

Both the gradient and the moment conditions for the test statistic can be written as the sum of two components, each of which depends on estimated weights. The next two subsections show that in each of these components the weights may be taken as known.

6.2.2 Estimated Weights: $[Y - E(Y|v)]\hat{w}$

One of the components of the test statistic and of the gradient for the estimator depends on a weighted distance between the dependent variable and its expectation conditioned on an index. The following lemmas simplify this component.

Lemma 7. Define:

$$\hat{w}_i \equiv \left\{ \nabla_\theta \hat{E}_a(Y_i|V_i) \quad \text{or} \quad \hat{E}(Y_i|X_{ki}) \right\},$$

where $\hat{E}_a$ is an adjusted expectation (D3) with window parameter $r : 1/8 < r < 1/4$. The expectation $\hat{E}$ is regular with window parameter $r_k = r$. With $S_i \equiv X_{ki}$ or the index, V_i, define $\tau(S_i)$ as the indicator on $a < S_i < b$. Assume $E\left|Y_j^2 | X_j\right| \leq \bar{\sigma}^2 = O(1)$. Then, with $u_i \equiv (Y_i - M_i)$:

$$D \equiv \sqrt{N}\langle \tau(V_i) u \tau(S_i)(\hat{w} - w)\rangle = o_p(1)$$

Proof. We provide the proof for $\hat{w}_i \equiv \nabla_\theta \hat{E}_a(Y_i|V_i)$, as the proof for the other weight is similar. Consider $\hat{w}_i^* \equiv \nabla_\theta \hat{E}(Y_i|V_i)$, where $\hat{E}$ is regular with window parameter r. Since:

$$\sqrt{N}\langle \tau(V_i) u(\hat{w} - \hat{w}^*)\rangle = o_p(1),$$

we need to establish convergence in probability to 0 for

$$D^* \equiv \sqrt{N}\langle \tau(V_i) u(\hat{w}^* - w)\rangle$$

Recalling from (D3) that for regular expectations: $\delta \equiv \hat{w}_i^* - w_i = \nabla_\theta \left(\hat{f}_i / \hat{g}_i \right) - \nabla_\theta \left(f_i / g_i, \right)$, this differential can be written as a sum of similar terms, one of which is given as

$$\nabla_\theta \hat{f}_i / \hat{g} - \nabla_\theta f_i / g_i = \left[g_i \left(\nabla_\theta \hat{f}_i - \nabla_\theta f \right) - \nabla_\theta f \left(\hat{g}_i - g_i \right) \right] / \hat{g}_i g_i$$

With similar arguments holding for the other terms, we analyze the first term. With $\Delta_i \equiv \nabla_\theta \hat{f}_i - \nabla_\theta f_i$, this term is given as:

$$\sqrt{N} \left\langle \tau \left(V_i \right) u \Delta / \hat{g} \right\rangle = D_1^* + o_p \left(1 \right), \; D_1^* \equiv \sqrt{N} \left\langle \tau \left(V_i \right) u \Delta / g \right\rangle .$$

Employing a mean-square convergence argument, $E \left[\left(D_1^* \right)^2 \right] = S + C$:

$$S \equiv E \left\langle u^2 \Delta^2 \right\rangle ; \; C \equiv \frac{1}{N} \sum_i \sum_{j \neq i} E \left(u_i u_j \Delta_i \Delta_j \right)$$

Taking an iterated expectation, S tends to zero. For C, write:

$$\Delta_i = \Delta_i \left[j \right] + \bar{\Delta}_i; \;\; \Delta_j = \Delta_j \left[i \right] + \bar{\Delta}_j,$$

where $\bar{\Delta}_i$ and $\bar{\Delta}_j$ do not depend on Y_i or on $Y_{j.}$ Then:

$$C = O(N) E \left(u_i \Delta_j \left[i \right] \right) E \left[u_j \Delta_i \left[j \right] \right] = O(N) O \left(\frac{1}{N h^2} \right)^2 = O \left(\frac{1}{N h^4} \right) = o \left(1 \right) .$$

With the exception of one weight component in the recentered moment conditions, the lemma above will be applied to simplify both the gradient for the estimator and the moment conditions. The complication, which is due to the recentering, is covered by Lemma 8 below.

Lemma 8. Referring to Lemma 6, let $M_{ik} \equiv E \left(Y_i | X_{ik} \right)$ and $\hat{M}_{ik} \equiv \tau_k \hat{E}_I \left(Y_i | X_{ik} \right)$. Let $\hat{E}_I$ and $\hat{E}_o$ be regular non-parametric expectations with respective windows r_I and r_o satisfying the restrictions in Lemma 6. Then:

$$\Delta \equiv \sqrt{N} \left\langle \hat{\tau} \left[Y - M \right] \left[\hat{E}_o (\hat{M} | V) - E (M_k | V) \right] \right\rangle \xrightarrow{p} 0$$

Proof. Let $\hat{w}_k \equiv \hat{E}_o (M_k | V)$, $w_k \equiv E \left(M_k | V \right)$, and $u \equiv [Y - M]$. Employing the same arguments as in the proof to Theorem 2 and Lemma 4.21 of Pakes and Pollard(1989), take the trimming function as known and write $\Delta = \Delta_1 + \Delta_2 + o_p(1)$,

$$\Delta_1 = \sqrt{N} \langle \tau u [\hat{w}_k - w_k] \rangle ; \; \Delta_2 = \sqrt{N} \left\langle \tau u [\hat{E}_o (\hat{M} | V) - \hat{E}_o (M_k | V)] \right\rangle$$

From Lemma 7, $\Delta_1 \xrightarrow{p} 0$. For the second term, the convergence rate in the expectation differential is not sufficient in itself to establish the desired result. Accordingly,

in what follows we show that Δ_2 simplifies to a term whose expected square converges to zero.

To simplify Δ_2, substitute from the definitions in the statement of the lemma to obtain:

$$\Delta_2 = \; N^{-1/2} \sum_{i=1}^{n} \frac{\tau_i u_i}{\hat{g}_i} \sum_{j \neq i}^{n} \frac{1}{h_o(N-1)} [\hat{M}_j - M_j] k_{ij}]$$

With Δ_2' defined by replacing $\hat{g}_i$ with g_i in Δ_2, it can be shown that $\Delta_2 = \Delta_2' + o_p(1)$.

By definition:

$$\Delta_2' = \; N^{-1/2} \sum_{i=1}^{n} \tau_i u_i \frac{1}{g_i} \sum_{j \neq i}^{n} \frac{1}{h_o(N-1)} [\frac{\hat{f}_{1j}}{\hat{g}_{1j}} - \frac{f_{1j}}{g_{1j}}] k_{ij}$$

It can also be shown that:

$$\Delta_2' = \; N^{-1/2} \sum_{i=1}^{n} \tau_i u_i \frac{1}{g_i} \sum_{j \neq i}^{n} \frac{1}{h_o(N-1)} [\frac{\hat{f}_{1j}}{\hat{g}_{1j}} - \frac{f_{1j}}{g_{1j}}] \left[\hat{g}_{1j}/g_{1j}\right] k_{ij} + o_p(1)$$

Writing Δ_2'' for the expression above:

$$\begin{aligned} \Delta_2'' \;\; &\equiv \;\; O\left(N^{-3/2} h_o^{-1}\right) \sum_{r=1}^{n} \frac{\tau_r u_r}{g_r} \sum_{j \neq r}^{n} [\hat{f}_{1j} - f_{1j} \hat{g}_{1j} M_j] k_{rj} \\ &= \;\; O\left(N^{-5/2} h_o^{-1} h_I^{-1}\right) \sum_{r=1}^{n} T_r, \;\; T_r \equiv \frac{\tau_r u_r}{g_r} \sum_{j \neq r}^{n} \sum_{l \neq j}^{n} \left[Y_l k_{lj}^1 - k_{lj}^1 M_j\right] k_{rj}. \end{aligned}$$

To complete the argument, we show that $E\left[\left(\Delta''\right)^2\right] = o(1)$. Squaring Δ_2'' and noting that $h_I^{-2} > h_o^{-2}$, the expectation of the cross-product terms is:

$$E(CP) = O(N^{-5}) O\left(h_I^{-4}\right) O(N^2) E(T_r T_s)$$

In T_r, for each j, there are $O(1)$ terms that depend on Y_r or on Y_s. Therefore, there are $O(N)$ such terms obtained by summing over j. Similarly, there are $O(N)$ such terms in T_s. Except for these $O(N^2)$ terms, all others vanish in expectation. Therefore:

$$E(CP) = O(N^{-5}) O\left(h_I^{-4}\right) O(N^2) O\left(N^2\right) = O\left(\frac{1}{N h^4}\right)$$

For $h_I = O(N^p)$, $p < 1/4$, the above expectation vanishes. The argument for the squared terms in Δ_2'' is similar.

6.2.3 Estimated Weights: $\left[\hat{E}\left(Y|V\right) - E\left(Y|V\right)\right]\hat{w}$,

This weighted component appears in both the test statistic and the gradient for the estimator. The lemmas below show that it is close in probability to a simplified term.

Lemma 9. With $h = O(N^{-r})$, $\frac{1}{8} < r < \frac{1}{4}$, then, with $\hat{w}$ as :

$$(a): \frac{\partial \hat{E}\left(Y|V\right)}{\partial \theta}; (b): \hat{E}\left(Y|X_k\right); \text{ or } (c): \hat{E}[\hat{\tau}_k \hat{E}\left(Y|X_k\right)|V],$$

$$\Delta \equiv \sqrt{N}\left[\left\langle \hat{\tau}_v \left(\hat{M} - M\right)\hat{w}\right\rangle - \left\langle \tau_v \left(\hat{M} - M\right) w\right\rangle\right] = o_p\left(1\right)$$

Proof. The arguments for (a), (b), and (c) are similar. For $\hat{w}$ in (c), write:

$$\Delta \equiv \sqrt{N}\left\langle \tau_v \left(\hat{M} - M\right)\left(\hat{w} - w\right)\right\rangle + \sqrt{N}\left\langle \tau_v \left(\hat{M} - M\right)\left(\hat{\tau}_v - \tau_v\right)\hat{w}\right\rangle$$

For the first term, the result follows from Lemmas 2 and 3. The argument for the second term is similar (see the section of the proof of Theorem 2 relating to indicators).

Lemma 10. (A Linear Characterization) Under the same window condition as in Lemma 9 and with $\hat{M}$ as the vector with i^{th} element $\hat{M}_i \equiv \hat{f}_i/\hat{g}_i$:

$$\sqrt{N}\left[\left\langle \left(\hat{M} - M\right) w\left(\hat{g}_v/g_v\right)\right\rangle - \left\langle \left(\hat{M} - M\right) w\right\rangle\right] = o_p\left(1\right)$$

Proof. The proof follows from Lemmas 2 and 3.