Econometric Reviews, 27(4–6):542–573, 2008

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474930801960444

Taylor & Francis
Taylor & Francis Group

A CLASS OF IMPROVED PARAMETRICALLY GUIDED NONPARAMETRIC REGRESSION ESTIMATORS

Carlos Martins-Filho,[1] Santosh Mishra,[1] and Aman Ullah[2]

[1]*Department of Economics, Oregon State University, Corvallis, Oregon, USA*
[2]*Department of Economics, University of California, Riverside, California, USA*

□ *In this article we define a class of estimators for a nonparametric regression model with the aim of reducing bias. The estimators in the class are obtained via a simple two-stage procedure. In the first stage, a potentially misspecified parametric model is estimated and in the second stage the parametric estimate is used to guide the derivation of a final semiparametric estimator. Mathematically, the proposed estimators can be thought as the minimization of a suitably defined Cressie–Read discrepancy that can be shown to produce conventional nonparametric estimators, such as the local polynomial estimator, as well as existing two-stage multiplicative estimators, such as that proposed by Glad (1998). We show that under fairly mild conditions the estimators in the proposed class are* $\sqrt{nh_n}$ *asymptotically normal and explore their finite sample (simulation) behavior.*

Keywords Asymptotic normality; Combined semiparametric estimation.

JEL Classification C14; C22.

1. INTRODUCTION

Nonparametric regression is a useful approach to tackle potential model misspecification. Although a vast and growing literature on the estimation of such models exists (Fan and Yao, 2003; Pagan and Ullah, 1999), much of the past literature has been devoted to the study of kernel based estimators. Prominent among these are Nadaraya–Watson (NW) and local linear (LL) estimators (Fan, 1992; Stone, 1977). Construction of these estimators depends on a bandwidth sequence h_n

Received November 15, 2005; Accepted December 11, 2006
Address correspondence to Carlos Martins-Filho, Department of Economics, Oregon State University, Ballard Hall 303, Corvallis, OR 97331-3612, USA; Fax: (541) 737-5917; E-mail: carlos.martins@orst.edu

such that $0 < h_n \to 0$ as the sample size $n \to \infty$. It is well known in the nonparametric literature that for fixed n, bandwidth size controls the tradeoff between pointwise bias and variance, and it is normally not possible to reduce bias without a simultaneous increase in variance or vice versa. Numerous attempts to bypass this tradeoff have emerged in both the nonparametric density and regression literatures, with particular interest in estimation procedures that preserve the magnitude of the variance while at the same time reducing pointwise bias. These attempts have included bias reduction via higher order kernels (Gasser et al., 1985; Müller, 1984), the specification of regression models that are a combination of parametric and nonparametric components (Fan and Ullah, 1999) and also "boosting" traditional nonparametric estimators (DiMarzio and Taylor, 2004). However, one of the most promising approaches, inspired by developments in the nonparametric density estimation literature (Hjort and Glad, 1995; Hjort and Jones, 1996; Jones et al., 1995; Naito, 2004), has been the parametrically guided nonparametric estimation procedure proposed by Glad (1998). Given $\{(y_i, x_i)\}_{i=1,2,\ldots}$ a sequence of independent and identically distributed pairs, with $E(y_i \mid x_i) = m(x_i)$, Glad considers the identity

$$m(x_i) \equiv m(x_i;\theta) r_m(x_i,\theta), \tag{1}$$

where $r_m(x_i,\theta) = \frac{m(x_i)}{m(x_i;\theta)}$, $m(x_i;\theta)$ for $\theta \in \Theta \subset \Re^p$ is a potentially misspecified parametric regression model. Since $E\left(\frac{y_i}{m(x_i;\theta)} \mid x_i\right) = r_m(x_i,\theta)$, Glad proposes an estimator $\hat{m}_G(x) = m(x;\hat{\theta}) r_m(x,\hat{\theta})$, where $r_m(x,\hat{\theta})$ is a nonparametric fit based on a regressand $\frac{y_i}{m(x_i;\hat{\theta})}$ with regressor x_i and $\hat{\theta}$ is a first stage parametric estimator based on the parametric model $E(y_i \mid x_i) = m(x_i;\theta)$. The intuition behind the procedure is that if the first stage parametric model is sufficiently "close" to $m(x_i)$, the multiplicative correction factor $r_m(x_i,\theta)$ will be easier to estimate nonparametrically leading to an improved $\hat{m}_G(x)$. In fact, Glad (1998) shows that when using a local polynomial estimator for the nonparametric fit, $\hat{m}_G(x)$ can have a smaller bias than traditional local polynomial estimators while maintaining the same variance.

The intuition supporting Glad's procedure can be used to define alternative parametrically guided estimators. Consider for example the identity

$$m(x_i) \equiv m(x_i;\theta) + r_a(x_i,\theta), \tag{2}$$

where $r_a(x_i,\theta) = m(x_i) - m(x_i;\theta)$. Since $E(y_i - m(x_i;\theta) \mid x_i) = r_a(x_i,\theta)$, an estimator $\hat{m}_A(x) = m(x;\hat{\theta}) + r_a(x,\hat{\theta})$ can be defined where $r_a(x,\hat{\theta})$ is a nonparametric fit based on a regressand $y_i - m(x_i;\hat{\theta})$ with regressor x_i, and

$\hat{\theta}$ is a first stage parametric estimator based on the possibly misspecified parametric model $E(y_i \mid x_i) = m(x_i; \theta)$. Here, rather than a multiplicative correction factor, $r_m(x_i, \hat{\theta})$, as in Glad, the potentially improved estimator is additively corrected (Rahman and Ullah, 2002).

The main contribution of this article is to show that Glad's multiplicatively corrected estimator $\hat{m}_G(x)$, the additively corrected estimator $\hat{m}_A(x)$, as well as the traditional NW and LL estimators belong to a vast class of parametrically indexed estimators. We show that all estimators in this class are asymptotically normal after proper normalization and that their asymptotic distributions differ only by their location. In other words, the estimators in this class have identical variance for their asymptotic distribution and differ only through the leading term in their bias. Regarding the previous literature, our asymptotic normality result is also useful in that known asymptotic normality results for NW and LL estimators appear as special cases, and asymptotic normality of the estimators proposed by Glad (1998) and Rahman and Ullah (2002) is obtained for the first time.[1]

The key insight in understanding how these estimators can be embedded in a single class is to realize that identities (1) and (2) are special cases of

$$m(x_i) \equiv m(x_i; \theta) + r_u(x_i, \theta) m(x_i; \theta)^{\alpha}, \tag{3}$$

where $r_u(x_i, \theta) = \frac{m(x_i) - m(x_i;\theta)}{m(x_i;\theta)^{\alpha}}$ and $\alpha \in \Re$. Note that (1) is obtained from (3) by taking $\alpha = 1$, and (2) is obtained by taking $\alpha = 0$. Since $E(\frac{y_i - m(x_i;\theta)}{m(x_i;\theta)^{\alpha}} \mid x_i) = r_u(x_i, \theta)$, an estimator $\hat{m}(x, \alpha) = m(x; \hat{\theta}) + r_u(x, \hat{\theta}) m(x; \hat{\theta})^{\alpha}$ can be defined where $r_u(x, \hat{\theta})$ is a nonparametric fit based on a regressand $\frac{y_i - m(x_i;\hat{\theta})}{m(x_i;\hat{\theta})^{\alpha}}$ with regressor x_i, and $\hat{\theta}$ is a first stage parametric estimator based on the possibly misspecified parametric model $E(y_i \mid x_i) = m(x_i; \theta)$.

To gain further insight into the nature of $\hat{m}(x, \alpha)$, we observe that it can be viewed either as: (1) the minimizer of a general loss function, or (2) the minimizer of a Cressie–Read power divergence statistic, subject to a suitably defined local moment condition. From the first point of view, our approach is similar to that of Naito (2004) which proposes a general loss function that embeds a number of parametrically guided nonparametric density estimators. To motivate this perspective, we give two examples.

Example 1. The NW estimator is defined as $\hat{m}_{NW}(x) \equiv \arg\min_c \frac{1}{n} \sum_{i=1}^{n} \left((y_i \quad c)\left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right)\right)^2 K_{h_n}(x_i - x)$, where $K_{h_n}(\cdot) = \frac{1}{h_n} K(\cdot / h_n)$. If an initial

[1]Glad (1998) established the order of the bias and variance for her estimator, but no result on its asymptotic distribution.

parametric regression estimator $m(x_i;\hat{\theta})$ is available, a transposed minimization can be defined in the residual space, i.e.,

$$\min_c \frac{1}{n}\sum_{i=1}^{n}\left(\begin{pmatrix} y_i - m(x_i;\hat{\theta}) & c - m(x;\hat{\theta})\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}\right)^2 K_{h_n}(x_i - x).$$

It is simple to show that this optimand is minimized by $\hat{m}_A(x)$, provided that the $r_a(x_i,\hat{\theta})$ is obtained via a NW estimator. Similarly, if $e' = (1\ 0)$ and $b' = (b_0\ b_1)$ we have that $\hat{m}_A(x) = \hat{b}_0$, where

$$\hat{b} \equiv \arg\min_{b_0,b_1} \frac{1}{n}\sum_{i=1}^{n}\left(\begin{pmatrix} y_i - m(x_i;\hat{\theta}) & b_0 + b_1(x_i - x) - m(x;\hat{\theta})\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}\right)^2 \times K_{h_n}(x_i - x) \tag{4}$$

provided that $r_a(x_i,\hat{\theta})$ is obtained via a LL estimator. In essence, the additively corrected estimator, $\hat{m}_A(x)$, can be viewed as the minimizer of an L_2 distance in a suitably transposed space of residuals.

Example 2. In the previous example, once the parametric model is chosen, minimization of the optimand in the transposed residual space occurs without accounting for the shape (variability) of $m(x;\theta)$ locally. Hence, we consider the minimizer of

$$\arg\min_c \frac{1}{n}\sum_{i=1}^{n}\left(\begin{pmatrix} y_i - m(x_i;\hat{\theta}) & c - m(x;\hat{\theta})\end{pmatrix}\begin{pmatrix}\frac{m(x;\hat{\theta})}{m(x_i;\hat{\theta})}\\-1\end{pmatrix}\right)^2 K_{h_n}(x_i - x), \tag{5}$$

where $\frac{m(x;\hat{\theta})}{m(x_i;\hat{\theta})}$ provides a measure of the local variability of $m(x;\hat{\theta})$. Again, it is simple to show that the minimizer of this optimand is $\hat{m}_G(x)$ provided that $r_m(x_i,\hat{\theta})$ is obtained via a NW estimator. Similarly, we have that $\hat{m}_G(x) = \hat{b}_0$, where

$$\hat{b} \equiv \arg\min_{b_0,b_1} \frac{1}{n}\sum_{i=1}^{n}\left(\begin{pmatrix} y_i - m(x_i;\hat{\theta}) & b_0 + b_1(x_i - x) - m(x;\hat{\theta})\end{pmatrix}\begin{pmatrix}\frac{m(x;\hat{\theta})}{m(x_i;\hat{\theta})}\\-1\end{pmatrix}\right)^2 \times K_{h_n}(x_i - x), \tag{6}$$

provided that $r_m(x_i,\hat{\theta})$ is obtained via a LL estimator.

We focus on a LL estimator and generalize the loss functions in Examples 1 and 2 by considering

$$L_n(b_0,b_1;x,\alpha,\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\begin{pmatrix} y_i - m(x_i;\hat{\theta}) & m(x;\hat{\theta}) - b_0 - b_1(x_i - x)\end{pmatrix}\begin{pmatrix}\hat{r}_i^{\alpha}\\1\end{pmatrix}\right)^2 \times K_{h_n}(x_i - x),$$

where $\alpha \in \Re$ and $\hat{r}_i = \frac{m(x;\hat{\theta})}{m(x_i;\hat{\theta})}$. Here, $\hat{r}_i$ captures the variability of the parametric function in the neighborhood of x and α determines how the variation in $\hat{r}_i$ contributes to the location of the residuals in the transposed space. The estimator $\hat{m}(x,\alpha)$ is given by $e'\hat{b}$, where $e' = (1 \;\; 0)$ and $\hat{b}' = (\hat{b}_0 \;\; \hat{b}_1)$ satisfies

$$\hat{b} \equiv \arg\min_{b_0,b_1} L_n(b_0, b_1; x, \alpha, \hat{\theta}), \quad \text{for a given choice of } \alpha \text{ and } \hat{\theta}. \tag{7}$$

To motivate $\hat{m}(x,\alpha)$ from the second point of view, we define the Cressie–Read discrepancy between two discrete distributions with common support $p = (p_1, \ldots, p_n)$ and $\pi = (\pi_1, \ldots, \pi_n)$ as

$$I_\lambda(p;\pi) = \frac{1}{\lambda(1+\lambda)} \sum_{i=1}^{n} p_i \left(\left(\frac{p_i}{\pi_i} \right)^{\lambda} - 1 \right),$$

for a given choice of λ.[2] Let $R_i(x) = (1 \;\; x_i - x)$, $Z_i(x) = (y_i - m(x_i;\theta))r_i^\alpha + m(x;\theta)$, where $r_i = \frac{m(x;\theta)}{m(x_i;\theta)}$, and suppose there exists a known function ϕ such that locally $E(\phi(Z_i(x), b_0, b_1) \mid x_i) = 0$ for a unique (b_0, b_1). Following Imbens et al. (1998), we seek

$$\begin{aligned} (\tilde{b}_0, \tilde{b}_1, \tilde{\pi}) &\equiv \arg\min_{b_0,b_1,\pi} I_\lambda(1/n, \ldots, 1/n; \pi) \\ &\text{subject to } \sum_{i=1}^{n} \phi(Z_i(x); b_0, b_1)\pi_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} \pi_i = 1. \end{aligned} \tag{8}$$

We choose $\lambda \to 0$ and following Lewbel (2007) define $\phi = (\phi_1, \phi_2)'$ as a vector valued function in $\Re^2$ with $\phi_1 = (Z_i(x) - b_0 - b_1(x_i - x))K_{h_n}(x_i - x)$ and $\phi_2 = (Z_i(x) - b_0 - b_1(x_i - x))(x_i - x)K_{h_n}(x_i - x)$. Then, the solution for the above minimization is attained by

$$\begin{aligned} (\tilde{b}_0, \tilde{b}_1, \tilde{\pi}) \equiv \arg\max_{b_0,b_1,\pi} \sum_{i=1}^{n} \ln(\pi_i) \qquad &\text{subject to } \sum_{i=1}^{n} \pi_i = 1 \\ \sum_{i=1}^{n} \pi_i (Z_i(x) - b_0 - b_1(x_i - x))K_{h_n}(x_i - x) = 0, \qquad &\text{and} \\ \sum_{i=1}^{n} \pi_i (Z_i(x) - b_0 - b_1(x_i - x))(x_i - x)K_{h_n}(x_i - x) = 0. \end{aligned} \tag{9}$$

[2]See Cressie and Read (1984) and Read and Cressie (1988).

By the Kuhn–Tucker Theorem, if $R'(x)K_\pi(x)R(x)$ is nonsingular, then we have

$$\tilde{b}_0 = e'(R'(x)K(x)R(x))^{-1}R'(x)K(x)Z(x) \qquad \text{and} \qquad \tilde{\pi}_i = 1/n \quad \text{for all } i,$$

with $R'(x) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ (x_1-x) & (x_2-x) & \cdots & (x_n-x) \end{pmatrix}$, $K(x) = \text{diag}\{K_{h_n}(x_i - x)\}_{i=1}^n$, $K_\pi(x) = \text{diag}\{\pi_i K_{h_n}(x_i - x)\}_{i=1}^n$, and $Z(x)$ an n-dimensional vector with ith element given by $Z_i(x)$.[3] Since $Z_i(x)$ depends on the unknown θ, a feasible version of the estimator is

$$\hat{b}_0 = e'(R'(x)K(x)R(x))^{-1}R'(x)K(x)\widehat{Z}(x),$$

where $\widehat{Z}_i = (y_i - m(x_i;\hat{\theta}))\hat{r}_i^{\alpha} + m(x;\hat{\theta})$. It is straightforward to verify that $\hat{b}_0 = \hat{m}(x,\alpha)$.

Regardless of how our estimator is motivated, intuitively we start with a regressand y_i and create a smoother modification of it, namely, $\widehat{Z}_i$. This modification itself helps in the reduction of bias because we deal with a potentially smoother version of y_i. It can be easily seen that when $m(x_i;\hat{\theta}) = c$ then $\widehat{Z}_i = y_i$, giving us an optimand that produces the traditional local linear estimator. When $\alpha = 1$, $\widehat{Z}_i = m(x;\hat{\theta})\frac{y_i}{m(x_i;\hat{\theta})}$ which gives us an optimand that produces Glad's estimator. Also, when $\alpha = 0$, $\widehat{Z}_i = (y_i - m(x_i;\hat{\theta})) + m(x;\hat{\theta})$ giving an optimand that produces the additively corrected estimator of Rahman and Ullah (2002).

We establish the asymptotic distribution of the proposed estimators in a two-step procedure. First, since we are dealing with a LL type estimator where Z_i replaces y_i, it is convenient to develop the asymptotic results with a nonstochastic $m(x_i;\theta_0)$, where θ_0 is interpreted as a quasi true parameter value. The infeasible estimator based on the quasi true value θ_0 is then shown to be asymptotically normal under suitable normalization. Second, we show the asymptotic equivalence of the infeasible estimator and its feasible counterpart, where the quasi true parameter θ_0 is estimated by pseudo maximum likelihood estimation (PMLE).

The structure of the article is as follows. This introduction is followed by the specification of the new class of estimators in Section 2. In Section 3 we provide results on the asymptotic behavior of the estimator in the class, and in Section 4 we provide a set of simulation results that shed light on the finite sample behavior of the estimator. The last section is a brief conclusion.

[3]If ϕ is such that $E(\phi(Z_i(x), b_0) \mid x_i) = 0$ for a unique b_0, with $\phi = (Z_i(x) - b_0)K_{h_n}(x_i - x)$, then the maximization in (9) gives $\tilde{\pi}_i = 1/n$ and $\tilde{b}_0 = (\sum_{i=1}^n K_{h_n}(x_i - x))^{-1}\sum_{i=1}^n K_{h_n}(x_i - x)Z_i(x)$ provided that $\sum_{i=1}^n K_{h_n}(x_i - x)\tilde{\pi}_i \neq 0$.

2. THE CLASS OF ESTIMATORS

We consider a sequence $\{(y_i, x_i)\}_{i=1}^n$ of independent two-dimensional random vectors with a common density, where y_i represents a regressand and x_i represents a regressor. We are primarily interested in the estimation of a regression model given by

$$y_i = m(x_i) + \varepsilon_i, \qquad \text{where } i = 1, \ldots, n, \tag{10}$$

$E(\varepsilon_i \mid x_i) = 0$ and $V(\varepsilon_i \mid x_i) = \sigma^2(x_i) < \infty$ for all x_i. The primary interest is on the estimation of the nonparametric regression function $m(\cdot)$, and to this end we propose a class of semiparametric regression estimators based on a two step estimation procedure. First, a parametric regression function $m(x_i; \theta)$ is stipulated and estimated via a parametric procedure that produces an estimator $m(x_i; \hat{\theta})$. The function $m(x; \hat{\theta})$ is assumed to belong to a class M of parametric functions that satisfies some smoothness conditions specified below, but is otherwise unrestricted. In the second step, the initial parametric estimate $m(x_i; \hat{\theta})$ is used to define the following optimand

$$L_n(b_0, b_1; x, \alpha, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left((e_i \quad m(x; \hat{\theta}) - b_0 - b_1(x_i - x)) \begin{pmatrix} \hat{r}_i^{\alpha} \\ 1 \end{pmatrix} \right)^2 K_{h_n}(x_i - x), \tag{11}$$

where $\alpha \in \Re$, $\hat{r}_i = \frac{m(x;\hat{\theta})}{m(x_i;\hat{\theta})}$, $e_i = y_i - m(x_i; \hat{\theta})$ and h_n is a nonstochastic bandwidth such that $0 < h_n \to 0$ as $n \to \infty$. The class of semiparametric estimator we propose is given by $\hat{m}(x, \alpha) = e' \hat{b}$, where $e' = (1 \;\; 0)$ and $\hat{b}' = (\hat{b}_0 \;\; \hat{b}_1)$ satisfies

$$\hat{b} = \arg\min_{b_0, b_1} L_n(b_0, b_1; x, \alpha, \hat{\theta}), \qquad \text{for a given choice of } \alpha \text{ and } \hat{\theta}. \tag{12}$$

We emphasize that the class of estimators $F \equiv \{\hat{m}(x, \alpha) : \alpha \in \Re \text{ and } m(x; \hat{\theta}) \in M\}$ depends on α, the stipulated parametric function $m(x; \theta)$, and also the estimator $\hat{\theta}$. Some well-known non-parametric estimators belong to the class F. For example, if $\alpha = 0$ and $m(x; \hat{\theta})$ is a parametric function of x belonging to some specified class M we have an additively corrected estimator; if $m(x; \hat{\theta}) = c$ for all x, $\hat{b}_0$ is the local linear estimator of Stone (1977) and Fan (1992);[4] and if $\alpha = 1$ and $m(x; \hat{\theta})$ is a parametric function of x belonging to some specified class M, we obtain the estimator proposed by Glad (1998).

[4]Mutatis mutandis local polynomial estimators of order $p \geq 2$ can also be obtained from the optimization in (12).

Using standard calculus and the algebra of local polynomial estimators (Ruppert and Wand, 1994) we obtain the following simple expression for the estimators in F

$$\hat{m}(x,\alpha) = e'(R')^{-1}R'(x)K(x)\widehat{Z}(x), \tag{13}$$

where $R'(x) = \begin{pmatrix} 1 & \cdots & 1 \\ (x_1-x) & \cdots & (x_n-x) \end{pmatrix}$, $K(x) = \text{diag}\{K_{h_n}(x_i-x)\}_{i=1}^n$ and $\widehat{Z}_i(x) = m(x;\hat{\theta}) + (y_i - m(x_i;\hat{\theta}))\hat{r}_i^{\alpha}$ is the ith element of the vector $\widehat{Z}(x)$. The expression is convenient in that it has the usual structure of local linear estimators with the exception of a modified regressand given by $\widehat{Z}(x)$. Hence, arguments typically used to establish the asymptotic properties of such estimators (Fan and Yao, 1998; Martins-Filho and Yao, 2007) can be used in the study of the asymptotic properties of $\hat{m}(x,\alpha)$. In what follows, it will be convenient to first consider the properties of an infeasible version of the estimators we propose, which is constructed by using a nonrandom parametric regression function $m(x;\theta)$ rather than $m(x,\hat{\theta})$. We label such estimator $\tilde{m}(x,\alpha)$ and first obtain the asymptotic properties of $\tilde{m}(x,\alpha)$. We then provide sufficient conditions for the asymptotic equivalence of $\tilde{m}(x,\alpha)$ and $\hat{m}(x,\alpha)$ under a suitable normalization.

3. ASYMPTOTIC PROPERTIES

3.1. The Estimator $\tilde{m}(x,\alpha)$

First, we give sufficient conditions for the $\sqrt{nh_n}$ asymptotic normality of $\tilde{m}(x,\alpha)$ and second we establish that $\sqrt{nh_n}(\hat{m}(x,\alpha) - \tilde{m}(x,\alpha)) = o_p(1)$ for all x and α. Throughout our developments, as well as in the statement of the regression model under consideration in (1), we have assumed for simplicity that there is only one regressor, i.e., $x_i \in \Re$. It should be transparent from the proofs below, that all results follow for the case where $x_i \in \Re^D$, D a finite positive integer, with appropriate adjustments on the relative speed of n and h_n^D. We start by providing a list of general assumptions and notation that will be selectively adopted in the lemma and theorems that follow. Throughout, C will represent a nonstochastic constant that may take different values in $\Re$, and the sequence of bandwidths h_n is such that, $nh_n^2 \to \infty$ as $n \to \infty$.

Assumption A1. 1. Let $g_X(x)$ be the common marginal density of x_i evaluated at x and assume that $g_X(x) < C$ for all x; 2. $g_X^{(d)}(x)$ is the dth order derivative of $g_X(x)$ evaluated at x and we assume that $|g_X^{(1)}(x)| < C$ for all x; 3. $|g_X(x) - g_X(x')| \leq C|x - x'|$ for all x, x'; 4. We denote the common joint density of (x_i, ε_i) evaluated at (x,ε) by $g(x,\varepsilon)$, the density of x_i

conditional on ε_i evaluated at x by $g_{x_i|\varepsilon_i}(x)$, and assume that $g_{x_i|\varepsilon_i}(x) < C$ for all x.

Assumption A2. 1. $K(x) : \Re \to \Re$ is a symmetric bounded function with compact support S_K such that 1. $\int |x|K(x)dx < \infty$; 2. $\int K(x)dx = 1$; 3. $\int xK(x)dx = 0$; 4. $\int x^2K(x)dx = \sigma_K^2$; 5. For all $x, x' \in S_K$ we have $|K(x) - K(x')| \leq C|x - x'|$; 6. $\int K^{2+\theta}(x)dx < \infty$ for some $\theta > 0$.

Assumption A3. 1. $|m^{(d)}(x)| < C$ for all x and $d = 1, 2$, where $m^{(d)}(x)$ is the dth order derivative of $m(x)$ evaluated at x.

Assumption A4. The parametric regression function $m(x;\theta)$ belongs to a class of parametrically indexed class M defined by the following characteristics: 1. $\theta \in \Theta$, Θ a compact subset of $\Re^q$; 2. $|m^{(d)}(x;\theta)| < C$ for all x, $\theta \in \Theta$ and $d = 1, 2$, where $m^{(d)}(x;\theta)$ is the dth order partial derivative of $m(x;\theta)$ with respect to its first argument evaluated at θ and x; 3. For all $x \in G$ an arbitrary compact subset of $\Re$, and $\theta \in \Theta$ there exist constants $0 < C_L \leq C_H < \infty$ such that $C_L < |m(x;\theta)| < C_H$; 4. $|\frac{dm(x;\theta)}{d\theta}| < C$ for all θ and $x \in G$, G a compact subset of $\Re$.

In what follows it will be convenient to write, $\tilde{m}(x,\alpha) - m(x) = \frac{1}{nh_n} \sum_{i=1}^n W_n\left(\frac{x_i - x}{h_n}, x\right)\widetilde{Z}_i^*$ where $\widetilde{Z}_i^* = \widetilde{Z}_i(x) - m(x) - m^{(1)}(x)(x_i - x)$, $W_n\left(\frac{x_i - x}{h_n}, x\right) = e' S_n^{-1}(x)\left(1 \ \frac{x_i - x}{h_n}\right)' K\left(\frac{x_i - x}{h_n}\right)$ with

$$S_n(x) = \frac{1}{nh_n} \begin{pmatrix} \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right) & \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\left(\frac{x_i - x}{h_n}\right) \\ \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\left(\frac{x_i - x}{h_n}\right) & \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\left(\frac{x_i - x}{h_n}\right)^2 \end{pmatrix} \equiv \begin{pmatrix} s_{n,0}(x) & s_{n,1}(x) \\ s_{n,1}(x) & s_{n,2}(x) \end{pmatrix},$$

and $\widetilde{Z}_i(x) = m(x;\theta) + (y_i - m(x_i,\theta))r_i^\alpha$ with $r_i = \frac{m(x;\theta)}{m(x_i,\theta)}$ is the ith component element of the vector $\widetilde{Z}(x)$. The following lemma is a special case of Theorem 1 in Martins-Filho and Yao (2007) for independent and identically distributed (IID) data. We provide a proof of Lemma 1 in the appendix to facilitate reading and understanding of our arguments as the proof for the non-IID case is substantially longer and more involved.

Lemma 1. *Assume* A1, A2, A3*, and let G be a compact subset of* $\Re$. *If* $s_{n,j}(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\left(\frac{x_i - x}{h_n}\right)^j$ *for* $j = 0, 1, 2$ *we have:*

(a) *If* $nh_n^3 \to \infty$, *then* $\sup_{x \in G} |s_{n,j}(x) - E(s_{n,j}(x))| = O_p\left(\left(\frac{ln(n)}{nh_n}\right)^{1/2}\right)$;
(b) *If* $\frac{nh_n^3}{ln(n)} \to \infty$, *then* $\sup_{x \in G} \frac{1}{h_n} |s_{n,j}(x) - E(s_{n,j}(x))| = o_p(1)$;

(c) *If* $\frac{nh_n^3}{ln(n)} \to \infty$, *then* $\tilde{m}(x,\alpha) - m(x) = \frac{1}{nh_n g_X(x)} \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^* + O_p(R_n(x))$, *where*

$$R_n(x) = \left|\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right| + \left|\frac{1}{n}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right|$$

with $\widetilde{Z}_i^* = \widetilde{Z}_i - m(x) - m^{(1)}(x)(x_i - x)$.

In the following theorem we establish the order in probability of the difference between $\tilde{m}(x,\alpha) - m(x)$ and $\frac{1}{nh_n g_X(x)} \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*$ uniformly in G. This result permits, under suitable normalization, the investigation of the asymptotic properties of $\tilde{m}(x,\alpha) - m(x)$ by restricting attention to $\frac{1}{nh_n g_X(x)} \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*$.

Theorem 1. *Assume* A1, A2, A3, *and* A4. *In addition assume that* $\frac{nh_n^3}{ln(n)} \to \infty$, *then for all* $x \in G$, G *a compact subset of* $\Re$ *we have,*

$$\sup_{x\in G}\left|\tilde{m}(x,\alpha) - m(x) - \frac{1}{nh_n g_X(x)}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right| = O_p\left(\left(\frac{h_n\, ln(n)}{n}\right)^{1/2}\right) + O_p(h_n^3).$$

The next theorem establishes the asymptotic normality of $\tilde{m}(x,\alpha) - m(x)$ under suitable normalization.

Theorem 2. *Assume* A1, A2, A3, *and* A4. *In addition assume that* $E(|\varepsilon_i|^{2+\delta}|x_i) < C$ *for some* $\delta > 0$, $\frac{nh_n^3}{ln(n)} \to \infty$ *and* $h_n^2\, ln(n) \to 0$, *then we have for all* $x \in G$ *a compact subset of* $\Re$

$$\sqrt{nh_n}(\tilde{m}(x,\alpha) - m(x) - B(x;\alpha,\theta)) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{g_X(x)}\int K^2(\psi)d\psi\right),$$

where $B(x;\alpha,\theta) = \frac{1}{2}h_n^2\sigma_K^2 B_c(x;\alpha,\theta) + o_p(h_n^2)$ *and*

$$B_c(x;\alpha,\theta) = \gamma^{(2)}(x) - B^{(2)}(x) = m^{(2)}(x) - (1-\alpha)m^{(2)}(x;\theta) - \alpha\left(\frac{2m^{(1)}(x)m^{(1)}(x;\theta) + m(x)m^{(2)}(x;\theta)}{m(x;\theta)}\right) + \frac{\alpha(\alpha+1)m(x)(m^{(1)}(x;\theta))^2}{m(x;\theta)^2} - \frac{\alpha(\alpha-1)(m^{(1)}(x;\theta))^2}{m(x;\theta)}.$$

It is easy to verify that the asymptotic bias for the local linear estimator of Stone (1977) can be obtained directly from our Theorem 1 by setting $\alpha = 0$ and $m(x;\theta) = c$. Furthermore, the results in Theorem 1 in Glad

(1998, p. 653) can also be obtained directly from our Theorem 1 by setting $\alpha = 1$ with $m(x;\theta) \in M$. Theorem 1 also reveals that the variance of the asymptotic distribution of the estimators in the class we propose do not depend on $m(x;\theta)$ or α. As such, their variance is equivalent to that of a one step estimator of $m(x)$ such as the local linear estimator of Stone (1977) or the two step estimator of Glad (1998). Asymptotically, the difference among the estimators in the class lies primarily on the bias term $B(x;\alpha,\theta)$, which clearly depends on α and $m(x;\theta)$, which ideally would be chosen simultaneously to minimize bias. However, it is instructive to consider their impact on bias separately.

The impact of $m(x;\theta)$: It is convenient to write $B_c(x;\alpha,\theta)$ as

$$B_c(x;\alpha,\theta) = m^{(2)}(x) - m^{(2)}(x;\theta)(1+\alpha(1-\rho_0(x;\theta))) + a(x;\theta)(\alpha(1+\rho_0(x;\theta)) - \alpha^2((1-\rho_0(x;\theta)) - 2\alpha\rho_1(x;\theta)),$$

where $\rho_0(x;\theta) = \frac{m(x)}{m(x;\theta)}$, $\rho_1(x;\theta) = \frac{m^{(1)}(x)}{m^{(1)}(x;\theta)}$, and $a(x;\theta) = \frac{(m^{(1)}(x;\theta))^2}{m(x;\theta)}$. We now make the following observations regarding the impact of $m(x;\theta)$ choice on the bias: (a) if the parametric guide $m(x;\theta)$ has kth order derivatives $(k = 0,1,2)$ evaluated at x that are equal to those of $m(x)$, implying that $\rho_0(x;\theta) = \rho_1(x;\theta) = 1$, then $B_c(x;\alpha,\theta) = 0$. In this case, $\tilde{m}(x,\alpha)$ has a leading bias term that is strictly smaller in absolute value than that of the LL (or NW) estimator for all α. Hence, in this case the choice of α has no impact on $B_c(x;\alpha,\theta)$; (b) if $m(x;\theta)$ and $m^{(1)}(x;\theta)$ are "close" to $m(x)$ and $m^{(1)}(x)$, in that $\rho_0(x;\theta) = 1+\epsilon$, and $\rho_1(x;\theta) = 1+\frac{\epsilon}{2}$ for $\epsilon \in B(0;\delta)$ a small neighborhood $-\delta$ of zero, then $B_c(x;\alpha,\theta) \simeq m^{(2)}(x) - m^{(2)}(x;\theta)$. Given that $m^{(2)}(x;\theta) \neq 0$, sufficient conditions for bias reduction of $\tilde{m}(x,\alpha)$ relative to the LL estimator are given by: (i) $m^{(2)}(x)$ and $m^{(2)}(x;\theta)$ have the same sign; (ii) $|m^{(2)}(x)| > |m^{(2)}(x;\theta)|$. As in (a), $B_c(x;\alpha,\theta)$ does not depend on α; (c) if $m(x;\theta)$ and $m^{(1)}(x;\theta)$ are not "close" to $m(x)$ and $m^{(1)}(x)$ in the manner described in (b), then α plays a crucial role in obtaining bias reduction. This observation stresses the importance of considering the broader class of estimators we propose, since bias reduction can be attained (or improved) relative to the estimators currently available. We illustrate this point with two simple examples.

Example 1. Suppose $m(x) = 1 + x + 3x^2$, $m(x;\theta) = x\theta$ and assume that $x_i \sim U[0.6,1]$ and independent, and $y_i \mid x_i = x \sim N(m(x),1)$ for $i = 1,2,\ldots,n$. We consider the estimation of $m(x)$ guided by $m(x;\theta)$ with $\alpha = 2$ and compare it to $\alpha = 0$ (additively corrected and LL estimators), and $\alpha = 1$ (Glad's estimator).[5] The gains of considering $\alpha = 2$ are significant as $\frac{\int_{0.6}^{1}(B_c(x;1,\theta_0))^2 dx}{\int_{0.6}^{1}(B_c(x;2,\theta_0))^2 dx} = 4.4$, which measures the gains relative

[5]Since the parametric guide is linear, the LL and additively corrected estimators coincide.

to the Glad estimator, and $\frac{\int_{0.6}^{1}(B_c(x;0,\theta_0))^2dx}{\int_{0.6}^{1}(B_c(x;2,\theta_0))^2dx} = 13.2$ which measures the gains relative to the additively corrected and LL estimators.[6] In this example, Glad's estimator has smaller bias than the LL estimator, but by considering other estimators in our proposed class, bias can be significantly reduced.

Example 2. Suppose $m(x) = (1+x)e^{0.4x}$, $m(x;\theta) = x\theta$ and assume that $x_i \sim U[0.4, 1]$ and independent, and $y_i \mid x_i = x \sim N(m(x), 1)$ for $i = 1, 2, \ldots, n$. We consider the estimation of $m(x)$ guided by $m(x;\theta)$ with $\alpha = -0.25$ and compare it to $\alpha = 0$ (additively corrected and LL estimators), and $\alpha = 1$ (Glad's estimator). Here, $\frac{\int_{0.4}^{1}(B_c(x;0,\theta_0))^2dx}{\int_{0.4}^{1}(B_c(x;-0.25,\theta_0))^2dx} = 2.71$, and $\frac{\int_{0.4}^{1}(B_c(x;1,\theta_0))^2dx}{\int_{0.4}^{1}(B_c(x;-0.25,\theta_0))^2dx} = 44.4$. In this example, Glad's estimator does not reduce bias relative to the LL or additively corrected estimators, however, once again by considering our broader class of estimators, we are able to reduce bias significantly relative to the LL estimator.

The impact of α : Since the bias of the estimators in the class we consider generally depend on α, a natural question that arises is whether or not an optimal estimator can be defined (or chosen) based on α for given $m(x;\theta)$ and bandwidth h_n. A commonly used criteria for estimator selection is mean integrated square error (MISE), hence we define $MISE(\alpha) = E(\int(\tilde{m}(x,\alpha) - m(x,\alpha))^2dx)$ for a specified parametric guide $m(x;\theta)$. Given that the asymptotic variance of $\tilde{m}(x,\alpha)$ is not a function of α, minimization of $MISE(\alpha)$ is equivalent to minimization of

$$\int B_c^2(x;\alpha,\theta)dx. \tag{14}$$

Ignoring the terms in $B_c(x;\alpha,\theta)$ with order smaller than h_n^2, we obtain after some standard algebra the following equation that must be solved to obtain the value of α that minimizes (14),

$$A(x;\theta)\alpha^3 + B(x;\theta)\alpha^2 + C(x;\theta)\alpha + D(x;\theta) = 0, \tag{15}$$

where

$$A(x;\theta) = 4\int Q_1(x;\theta)^2dx, \qquad B(x;\theta) = 2\int Q_1(x;\theta)Q_3(x;\theta)dx,$$
$$C(x;\theta) = \int(4Q_2(x;\theta)Q_1(x;\theta) + 2Q_3(x;\theta)^2)dx,$$
$$D(x;\theta) = 2\int Q_2(x;\theta)Q_3(x;\theta)dx,$$

[6]Here θ_0 is calculated by minimizing the Kullback–Leibler discrepancy or maximizing the likelihood function. For $\alpha = 1, 0$ the bias term doesn't involve θ_0.

with

$$Q_1(x;\theta) = \left(\frac{m(x)}{m(x;\theta)^2} - \frac{1}{m(x;\theta)}\right)(m^{(1)}(x;\theta))^2,$$

$$Q_2(x;\theta) = m^{(2)}(x) - m^{(2)}(x;\theta) \qquad \text{and}$$

$$Q_3(x;\theta) = m^{(2)}(x;\theta) - \frac{2}{m(x;\theta)}m^{(1)}(x)m^{(1)}(x;\theta) + \frac{1}{m(x;\theta)}\left(m^{(1)}(x;\theta)\right)^2 + \frac{m(x)}{m(x;\theta)^2}\left(m^{(1)}(x;\theta)\right)^2 - \frac{m(x)}{m(x;\theta)^2}m^{(2)}(x;\theta).$$

First, we observe that as a polynomial of order 3 in α, Eq. (15) may have multiple roots depending on $A(x;\theta)$, $B(x;\theta)$, $C(x;\theta)$, and $D(x;\theta)$. Second, these terms involve integrals of functions of $m(x)$, $m(x;\theta)$ as well as their first and second derivatives, all of which are in practice unknown. Hence, to render Eq. (15) operational, the unknown functions, $m(x)$, $m^{(1)}(x)$, $m^{(2)}(x)$, $m(x;\theta)$, $m^{(1)}(x;\theta)$, and $m^{(2)}(x;\theta)$ must be replaced by suitable estimates. Given the first step parametric estimators, it is straightforward to obtain $m(x;\hat{\theta})$, $m^{(1)}(x;\hat{\theta})$, and $m^{(2)}(x;\hat{\theta})$. The remaining unknown functions, $m(x)$, $m^{(1)}(x)$, and $m^{(2)}(x)$, can be estimated by a traditional local polynomial estimator of order 3. Solving an estimated version of Eq. (15) produces a data-driven, and consequently stochastic, α that gives the researcher a sample driven guidance to the choice of α, or equivalently, the preferred loss function. The difficulty here is that a data-driven stochastic α redefines the structure of our proposed estimator rendering potentially invalid the asymptotic results of Theorem 1.

The difficulties outlined in the previous paragraph are similar to those faced by Naito (2004) in the context of density estimation and are conceptually no different from those involved in the choice of h_n, the bandwidth (Ruppert et al., 1995). In general, the expression for an optimal bandwidth that minimize MISE depends on the unknown second derivative of $m(x)$ and the $g_X(x)$, which need to be estimated based on the available data. The resulting data driven bandwidth based on such estimates is stochastic and the derived asymptotic properties of estimators based on such bandwidths are not generally available.

3.2. The Estimator $\hat{m}(x, \alpha)$

We now consider the case where the parametric guide results from a first stage estimation procedure, i.e., we have $m(x;\hat{\theta})$. Theorem 2 shows that under fairly mild conditions there is no impact on the asymptotic distribution obtained in Theorem 1 when we consider a stochastic parametric guide $m(x;\hat{\theta})$. Clearly, the parametric guide used in the first step of the estimation is almost surely an incorrect specification for the

regression model. Hence, we assume that the first stage estimator is a pseudo maximum likelihood estimator and that θ_0 is the pseudoparameter value that minimizes the Kullback–Liebler distance between the assumed parametric joint density of (y_i, x_i) and its true joint density. Hence, if $h(y, x; \theta) = g_X(x)h_\theta(y \mid x)$ is the assumed parametric joint density of (y_i, x_i) and $h(y, x) = g_X(x)h(y \mid x)$ is the true joint density,

$$\theta_0 \equiv \arg\min_{\theta\in\Theta} E\left(\ln\left(\frac{h(y \mid x)}{h_\theta(y \mid x)}\right)\right) \equiv \arg\min_{\theta\in\Theta} \int\int \ln\left(\frac{h(y \mid x)}{h_\theta(y \mid x)}\right) h(y, x)\, dy\, dx.$$

We now make the following additional assumption that assures that the pseudomaximum likelihood estimator $\hat{\theta}$ satisfies $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$ (White, 1982).

Assumption A5 . 1. $E(\ln(h(y, x)))$ exists and $|\ln(h(y, x; \theta))| \le \mu(y, x)$ for all $\theta \in \Theta$, where $\mathrm{E}(\mu(y, x))$ exists and $E\left(\ln\left(\frac{h(y \mid x)}{h_\theta(y \mid x)}\right)\right)$ has a unique minimum at θ_0; 2. $\frac{\partial \ln(h(y,x;\theta))}{\partial \theta_i}$ for $i = 1, \ldots, q$ are continuously differentiable functions of θ; 3. $\left|\frac{\partial^2 \ln(h(y,x;\theta))}{\partial \theta_i \partial \theta_j}\right| \le m_1(y, x)$ and $\left|\frac{\partial \ln(h(y,x;\theta))}{\partial \theta_i}\frac{\partial \ln(h(y,x;\theta))}{\partial \theta_j}\right| \le m_2(y, x)$ for all $\theta \in \Theta$, $i, j = 1, \ldots, q$ where $E(m_1(y, x)), E(m_2(y, x))$ exist; 4. θ_0 is in the interior of Θ, $n^{-1}\sum_{t=1}^{n} \frac{\partial \ln(h(y_t,x_t;\theta_0))}{\partial \theta_i}\frac{\partial \ln(h(y_t,x_t;\theta_0))}{\partial \theta_j}$ is nonsingular and $n^{-1}\sum_{t=1}^{n} \frac{\partial^2 \ln(h(y_t,x_t;\theta_0))}{\partial \theta_i \partial \theta_j}$ has constant rank in a neighborhood of θ_0.

Theorem 2 below is the main result of the article. It establishes that under A5 there is no asymptotic loss in estimating θ_0. As such, under suitable normalization, the infeasible and the feasible versions of the estimators in the class we propose are asymptotically equivalent.

Theorem 3. *Assume* A1, A2, A3, A4, *and* A5. *In addition assume that* $E(|\varepsilon_i|^{2+\delta} | x_i) < C$ *for some* $\delta > 0$, $\frac{nh_n^3}{ln(n)} \to \infty$ *and* $h_n^2\, ln(n) \to 0$, *then for all* $x \in G$ *a compact subset of* $\Re$, *we have*

$$\sqrt{nh_n}(\hat{m}(x, \alpha) - m(x) - B(x; \alpha, \theta_0)) \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{g_X(x)} \int K^2(\psi) d\psi\right),$$

where $B(x; \alpha, \theta_0) = \frac{1}{2} h_n^2 \sigma_K^2 B_c(x; \alpha, \theta_0) + o_p(h_n^2)$.

It is worth mentioning that relative to traditional nonparametric regression estimators, such as NW and LL estimators, $\hat{m}(x, \alpha)$ is more expensive from a computational perspective, given the need to obtain a first stage parametric guide $m(x; \hat{\theta})$. When the parametric guide is linear in θ, the additional computational cost is negligible, but the cost can increase rapidly for nonlinear parametric guides.

Given the asymptotic equivalence of $\tilde{m}(x,\alpha)$ and $\hat{m}(x,\alpha)$, all comments made following Theorem 1 regarding the impact of $m(x;\theta)$ and α on the magnitude of the bias term apply to $\hat{m}(x,\alpha)$. In particular, since in practice it is not possible to evaluate the distance between $m(x;\theta)$ and $m(x)$, bias reduction could be attained by exploring different estimators in our proposed class through a suitable choice of α. As illustrated in Examples 1 and 2 following Theorem 1, linear parametric guides in combination with a suitable choice of α can be effective in producing significant bias reduction. It is important to recognize that this improvement is only possible by considering a class of estimators indexed by α, and may not result if attention is limited to Glad's estimator or the additively corrected estimator of Rahman and Ullah (2002) (see Examples 1 and 2). Hence, as a practical guide for implementation of $\hat{m}(x,\alpha)$, we suggest that a linear parametric guide be chosen followed by the algorithm we propose on p. 552 to select α.

4. SIMULATION

In this section we provide some experimental evidence on the finite sample behavior of our proposed estimator. We also compare its performance to that of the estimator proposed by Glad (1998), and the local linear estimator of Stone (1977) and Fan (1992). We consider the same two DGPs studied by Glad. The DGPs and parametric guides considered are given in Table 1.

$\{x_i, \varepsilon_i\}$ are identically and independently distributed with x_i drawn from a uniform distribution and ε_i drawn from a normal distribution with 0 mean and standard deviation given in Table 1. The sample size considered are $n = \{50, 100, 200, 400\}$ and the number of replications $M = 500$. We consider 51 values of the parameter α varying from -5 to 5 with steps of size 0.1.[7] An optimal bandwidth that minimizes the mean integrated square error (MISE) is used in the simulation. Given the local

TABLE 1 DGPs and guides

	$m(x)$	DGP1 $= 2 + \sin 2\pi x$	DGP2 $= 2 + x - 2x^2 + 3x^5$
TrueDGP	σ	0.50	0.70
Parametric guides	$m(x;\hat{\theta})$	$P_1^1 = \hat{\theta}_0 + \hat{\theta}_1 \sin 2\pi x$	$P_1^2 = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 x^2 + \hat{\theta}_3 x^5$
		$P_2^1 = \hat{\theta}_0 + \hat{\theta}_1 x$	$P_2^2 = \hat{\theta}_0 + \hat{\theta}_1 x$
		$P_3^1 = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 x^2 + \hat{\theta}_3 x^3$	$P_3^2 = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 e^x$

[7]The choice of α is guided by the intention of including both positive and negative values of α, as well as taking into account the special cases where $\alpha = 0$ and $\alpha = 1$ that correspond to the additively corrected estimator and that of Glad (1998).

linear structure of our estimator it is straightforward to obtain the optimal bandwidth that minimizes the asymptotic approximation for mean integrated squared error (Ruppert et al., 1995). It is given by

$$\hat{h}_n = \arg\min_{h_n} \frac{1}{4}\sigma_K^2 h_n^4 \int B_c(x;\alpha,\theta)^2\,dx + \frac{1}{nh_n}\int K^2(\psi)\,d\psi \int \frac{\sigma^2(x)}{g_X(x)}\,dx,$$

where $\hat{h}_n = n^{-1/5}\left(\frac{\int \sigma^2(x) g_X(x)^{-1}\,dx \int K(\psi)^2\,d\psi}{\sigma_K^4 \int B_c(x;\alpha,\theta)^2\,dx}\right)^{1/5}$. In our simulations, the unknown components of the optimal bandwidth expression—$m(x)$, $m^{(1)}(x)$, $m^{(2)}(x)$—are known, but in practice they must be estimated.[8] It is important to note that here the optimal bandwidth is a function of α. In all cases, we use the standard Gaussian kernel.

Tables 2 and 3 provide the simulation results for DGP1 and DGP2, respectively. In Table 2 the true DGP is given as $2 + \sin 2\pi x$, $P_1^1 = \hat{\theta}_0 + \hat{\theta}_1 \sin 2\pi x$, $P_2^1 = \hat{\theta}_0 + \hat{\theta}_1 x$, $P_3^1 = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 x^2 + \hat{\theta}_3 x^3$ and the standard deviation of the error is 0.50. The rows associated with Glad, Add, and LL represent the results corresponding to the multiplicative corrected estimator of Glad, the additively corrected estimator, and the local linear estimator. The row best is associated with the value of α that produces the estimator in the class with smallest MSE, given the chosen parametric

TABLE 2 Bias (B), Variance (V), Mean Square Error (MSE), and Eff for DGP1

	Guide	P_1^1				P_2^1				P_3^1			
Model	n	B^2	V	MSE	Eff	B^2	V	MSE	Eff	B^2	Var	MSE	Eff
Best	50	0.61	327.23	327.84	1.00	0.64	328.23	328.27	1.00	2.16	329.23	331.39	1.00
	100	0.44	168.44	168.88	1.00	0.50	170.44	170.95	1.00	0.45	168.44	168.89	1.00
	200	0.31	85.25	85.56	1.00	0.31	85.65	85.97	1.00	0.31	85.25	85.57	1.00
	400	0.03	39.32	39.35	1.00	0.04	38.32	38.36	1.00	0.03	39.32	39.35	1.00
Glad	50	5.96	337.23	343.88	0.95	51.70	337.23	388.93	0.84	5.95	337.23	343.18	0.96
	100	4.37	169.44	173.20	0.97	42.63	169.44	212.07	0.80	4.40	169.44	173.84	0.97
	200	2.71	86.05	88.75	0.96	26.36	86.05	112.42	0.76	2.72	86.05	88.79	0.96
	400	1.27	40.32	41.59	0.95	12.98	40.32	53.30	0.71	1.29	40.32	41.51	0.95
Add	50	2.14	338.00	340.42	0.96	59.51	397.23	456.74	0.72	1.69	337.25	338.94	0.97
	100	1.16	171.44	172.10	0.97	44.02	198.44	242.46	0.70	1.15	171.44	172.59	0.97
	200	0.61	90.25	90.86	0.94	27.22	88.25	115.47	0.74	0.61	91.25	91.86	0.93
	400	0.46	43.32	43.78	0.90	12.95	45.32	58.27	0.65	0.46	47.33	47.79	0.82
LL	50	59.51	397.23	456.74	0.71	59.51	397.23	456.74	0.72	59.51	397.23	456.74	0.72
	100	44.02	198.44	242.46	0.69	44.02	198.44	242.46	0.70	44.02	198.44	242.46	0.69
	200	27.22	88.25	115.47	0.74	27.22	88.25	115.47	0.74	27.22	88.25	115.47	0.74
	400	12.95	45.32	58.27	0.67	12.95	45.32	58.27	0.65	12.95	45.32	58.27	0.67

Note: All entries for bias squared, variance, and mean square error are multiplied by 10^4.

[8]Suitable estimators are given in the comments following Theorem 2.

TABLE 3 Bias (B), Variance (V), Mean Square Error (MSE) And Eff for DGP2

	Guide	P_1^2				P_2^2				P_3^2			
Model	n	B^2	V	MSE	Eff	B^2	V	MSE	Eff	B^2	V	MSE	Eff
Best	50	0.91	602.13	603.04	1.00	0.92	608.82	609.74	1.00	1.01	606.16	607.18	1.00
	100	0.84	298.73	299.57	1.00	0.85	303.43	304.28	1.00	0.80	305.63	306.43	1.00
	200	0.38	152.51	152.90	1.00	0.39	157.38	157.77	1.00	0.39	152.65	153.05	1.00
	400	0.08	76.99	77.08	1.00	0.08	83.84	83.93	1.00	0.08	80.54	80.62	1.00
Glad	50	3.05	606.21	609.26	0.99	33.87	609.53	643.40	0.94	3.96	607.97	611.93	0.99
	100	2.71	299.39	302.10	0.98	28.48	306.56	335.04	0.90	2.92	308.64	311.56	0.98
	200	1.32	153.64	154.96	0.98	15.25	160.02	175.27	0.90	1.59	152.54	154.13	0.99
	400	0.43	85.43	85.86	0.90	8.43	85.77	94.20	0.89	0.88	80.39	81.27	0.98
Add	50	3.23	608.44	611.67	0.98	36.64	611.02	647.67	0.94	3.02	607.20	610.22	0.99
	100	2.61	303.10	305.71	0.98	29.32	307.77	337.09	0.90	2.53	310.76	313.29	0.97
	200	1.40	161.41	162.81	0.93	15.98	154.10	170.08	0.92	0.89	159.86	160.75	0.95
	400	0.70	77.53	78.23	0.98	8.82	83.78	92.61	0.90	0.33	88.18	88.51	0.90
LL	50	36.64	611.02	647.67	0.93	36.64	611.02	647.67	0.94	36.64	611.02	647.67	0.93
	100	29.32	307.77	337.09	0.88	29.32	307.77	337.09	0.90	29.32	307.77	337.09	0.90
	200	15.98	154.10	170.08	0.89	15.98	154.10	170.08	0.92	15.98	154.10	170.08	0.91
	400	8.82	83.78	92.61	0.83	8.82	83.78	92.61	0.90	8.82	83.78	92.61	0.92

Note: All entries for bias squared, variance and mean square error are multiplied by 10^4.

guide. We define $\text{Eff}^j = \frac{MSE^{best}}{MSE^j} \leq 1$, and consequently desirable estimators must have high Eff.

In Table 3 the true DGP is given as $2 + x - 2x^2 + 3x^5$, $P_1^2 = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 x^2 + \hat{\theta}_3 x^5$, $P_2^2 = \hat{\theta}_0 + \hat{\theta}_1 x$, $P_3^2 = \hat{\theta}_0 + \hat{\theta}_1 x + \hat{\theta}_2 e^x$. The standard deviation of the error is 0.70. All other entries in Table 3 correspond to those in Table 2.

To evaluate the performance of our class of estimators, we estimate the model using all 51 values of α and report the results for that α which yields the lowest sample MISE.[9] We use r and j to denote the index of replication and parametric guide, respectively. Let

$$B_i^j = \frac{1}{M}\sum_{r=1}^{M}[\hat{m}_j^r(x_i, \alpha, \hat{\theta}) - m(x_i)],$$

$$S_i^j = \frac{1}{M}\sum_{r=1}^{M}\left[\hat{m}_j(x_i, \alpha, \hat{\theta}) - \frac{1}{M}\sum_{r=1}^{M}\hat{m}_j^r(x_i, \alpha, \hat{\theta})\right]^2,$$

$B^j = [B_1^j, \ldots, B_n^j]'$, and $S^j = [S_1^j, \ldots, S_n^j]'$, where $\hat{m}_j^r(x_i, \alpha, \hat{\theta})$ is estimated conditional mean for the rth replication and the jth parametric guide. $m(x_i)$ is the true nonparametric function. Let $B^{2^j} = n^{-1}\sum_{i=1}^{n}(B_i^j)^2$, $V^j = n^{-1}\sum_{i=1}^{n} S_i^j$, and $MSE_i^j = (B_i^j)^2 + S_i^j$ be the squared bias, variance and

[9]As we do not prove the strict convexity of MISE with respect to α there may be several α that minimize MISE.

mean square error of estimates, respectively. Thus MSE for model j is given by $MSE^j = n^{-1} \sum_{i=1}^{n} MSE_i^j$. While comparing two estimators, the one with higher Eff^j is preferable.

We find that for both DGPs when the parametric guide is correct, i.e., coincides with the true regression, there is substantial bias reduction for all sample sizes and all combined estimators vis a vis the local linear estimator.[10] If the linear guide is a poor approximation for the true regression, we find that both Glad and the additively corrected estimator provide negligible improvement over the local linear estimator. This conclusion coincides with the results obtained in the simulations of Glad (1998). It can be theoretically shown that, when the parametric guide is linear the additively corrected estimator and the local linear estimator have the same leading term for the bias. Hence, for both DGPs we find that the Eff statistic is the same for the additively corrected and local linear estimators in the case of a linear parametric guide. As the variances of the estimators in the class do not depend on α and first stage estimation, it is expected that variances across parametric guides and α should be of similar magnitude, which is observed in our simulation results. For DGP1 we find that Glad's estimator is closest to the best model when the parametric guide is not linear. However, for the best model there is significant improvement in terms of bias reduction. For DGP2 we find that the estimators' performance depends on the parametric guide. Both estimators outperform the local linear estimator when the parametric guide is not linear. Also, it is observed that in all cases the best model does not coincide with either the additively corrected or Glad's estimators.

Finally, we note that the optimal α obtained via the grid search described above and used in the simulations is very close to the α obtained by solving Eq. (15). This suggests that the asymptotic approximation for the bias in Theorem 2 seems to be fairly reasonable for sample sizes of relative small size, i.e., $n = 400$. For example, we find that for DGP1 with $n = 400$, the α obtained from the grid search method are -1.7, and -1.8 for parametric guides P_2^1 and P_3^1, respectively. The corresponding numbers obtained from solving Eq. (11) are -1.75 and -1.76 respectively.[11] For DGP 2 we derive a similar conclusion.[12]

[10]Under the unrealistic assumption of a correctly specified parametric DGP, a suitable parametric estimator (possibly unbiased and efficient in an appropriately defined class) can be chosen, and the bias-variance tradeoff intrinsic to all nonparametric estimators considered herein can be bypassed.

[11]When the parametric guide is equal to $m(x)$ the left-hand side of Eq. (15) is identically zero for all x.

[12]In fact, optimal values of α do not vary significantly with n for $n \geq 100$ in our simulations.

5. CONCLUSION

This article proposes a class of nonparametric regression estimators that improve the bias of traditional kernel based nonparametric estimators without an increase in variance. This class of estimators is associated with the minimization of a new loss function, which depends on a parameter α, that includes as special cases well-known estimators such as the local linear (polynomial), Nadaraya–Watson, Glad (1998), and an additively corrected estimator. The estimators in the class can be obtained in a two stage procedure. In the first stage, a parametric estimation reduces the variability of the regressand, and in the second stage a local linear (polynomial) estimator is fitted to the modified regressand that incorporates the impact of the curvature of the first stage parametric fit in the neighborhood of x.

We obtain reduced bias due to smaller variability of the modified regressand. Inclusion of α allows for a larger scope for bias reduction compared to the existing combined estimators. The variance of the estimators in the class does not change asymptotically, although our simulations reveal that in finite samples variance is also reduced relative to that of the local linear estimator.

Bias and variance of the estimators in the class are derived and asymptotic normality is established. As the second stage modified regressand includes the parametric fit, first asymptotic normality is established for a nonstochastic parametric guide. Subsequently, it is shown that when the first stage estimator is obtained via pseudo maximum likelihood estimation, the final estimator inherits the asymptotic properties of the estimator obtained with a nonstochastic guide.

We perform a small Monte Carlo study to evaluate the performance of the new estimator relative to that of the existing alternatives. The indexing parameter α is allowed to vary over a range negative and positive values. Our simulations provide following conclusions: (1) when the parametric guide coincides with the true regression, all combined estimators outperform the local linear estimator; (2) even when the parametric guide is highly misspecified there exists an estimator in the proposed class that provides significant bias reduction vis a vis the local linear estimator. This is significant, since currently available estimators that attempt to reduce bias with parametric guides, do not significantly reduce bias relative to the local linear estimator when the guide is severely misspecified.

6. APPENDIX

Proof of Lemma 1. (a) We prove the case where $j = 0$. Similar arguments can be used for $j = 1, 2$ given A2. Let $B(x_0, r) = \{x \in \Re : |x - x_0| < r\}$ for $r \in \Re^+$. G compact implies that there exists $x_0 \in G$ such

that $G \subseteq B(x_0, r)$. Therefore for all $x, x' \in G$, $|x - x'| < 2r$. Let $h_n > 0$ be a sequence such that $h_n \to 0$ as $n \to \infty$ where $n \in \{1, 2, 3, \dots\}$. For any n, by the Heine–Borel theorem there exists a finite collection of sets $\{B(x_k, (\frac{n}{h_n^3})^{-1/2})\}_{k=1}^{l_n}$ such that $G \subset \bigcup_{k=1}^{l_n} B(x_k, (\frac{n}{h_n^3})^{-1/2})$ for $x_k \in G$ with $l_n < (\frac{n}{h_n^3})^{1/2} r$. For $x \in B(x_k, (\frac{n}{h_n^3})^{-1/2})$,

$$|s_{n,0}(x) - s_{n,0}(x_k)| \le (nh_n)^{-1} \sum_{t=1}^{n} C|h_n^{-1}(x_k - x)| < C(nh_n)^{-1/2} \qquad \text{and}$$

$$|E(s_{n,0}(x_k)) - E(s_{n,0}(x))| < C(nh_n)^{-1/2}.$$

Hence, $|s_{n,0}(x) - E(s_{n,0}(x))| \le |s_{n,0}(x_k) - E(s_{n,0}(x_k))| + 2C(nh_n)^{-1/2}$ and

$$\sup_{x \in G} |s_{n,0}(x) - E(s_{n,0}(x))| \le \max_{1 \le k \le l_n} |s_{n,0}(x_k) - E(s_{n,0}(x_k))| + 2C(nh_n)^{-1/2}.$$

Since $2\left(\frac{nh_n}{\ln(n)}\right)^{1/2} C(nh_n)^{-1/2} \to 0$, then to prove (a) it suffices to show that there exists a constant $\Delta > 0$ such that for all $\epsilon > 0$ there exists N such that for all $n > N$, $P\left(\left(\frac{nh_n}{\ln(n)}\right)^{1/2} \max_{1 \le k \le l_n} |s_{n,0}(x) - E(s_{n,0}(x))| \ge \Delta\right) \le \epsilon$. Let $\varepsilon_n = \left(\frac{\ln(n)}{nh_n}\right)^{1/2} \Delta$. Then, for every n,

$$P\left(\max_{1 \le k \le l_n} |s_{n,0}(x_k) - E(s_{n,0}(x_k))| \ge \varepsilon_n\right) \le \sum_{k=1}^{l_n} P(|s_{n,0}(x_k) - E(s_{n,0}(x_k))| \ge \varepsilon_n).$$

But $|s_{n,0}(x_k) - E(s_{n,0}(x_k))| = |\frac{1}{n} \sum_{i=1}^{n} W_{in}|$ where $W_{in} = \frac{1}{h_n} K(\frac{x_i - x_k}{h_n}) - \frac{1}{h_n} E(K(\frac{x_i - x_k}{h_n}))$ with $E(W_{in}) = 0$ and $|W_{in}| \le \frac{C}{h_n}$. Since $\{W_{in}\}_{i=1}^{n}$ is an independent sequence, by Bernstein's inequality

$$P(|s_{n,0}(x_k) - E(s_{n,0}(x_k))| \ge \varepsilon_n) < 2 \exp\left(\frac{-nh_n \varepsilon_n^2}{2h_n \bar{\sigma}^2 + \frac{2C\varepsilon_n}{3}}\right)$$

where $\bar{\sigma}^2 = n^{-1} \sum_{i=1}^{n} V(W_{in}) = h_n^{-2} E(K^2(\frac{x_i - x_k}{h_n})) - (h_n^{-1} E(K(\frac{x_i - x_k}{h_n})))^2$. Under assumptions A1 and A2, we have that $h_n \bar{\sigma}^2 \to B_{\bar{\sigma}^2}$ by Lebesgue's dominated convergence theorem for some constant $B_{\bar{\sigma}^2}$. Let $c_n = 2h_n \bar{\sigma}^2 + \frac{2}{3} C\varepsilon_n$. Then, $\frac{-nh_n \varepsilon_n^2}{2h_n \bar{\sigma}^2 + \frac{2C\varepsilon_n}{3}} = \frac{-\Delta^2 \ln(n)}{c_n}$. Hence, for any $\epsilon > 0$ there exists N such that for all $n > N$,

$$P\left(\max_{1 \le k \le l_n} |s_{n,0}(x_k) - E(s_{n,0}(x_k))| \ge \varepsilon_n\right)$$

$$< 2l_n n^{-\Delta^2/c_n} < 2\left(\frac{n}{h_n^2}\right)^{1/2} r n^{-\Delta^2/c_n} < 2(nh_n^2)^{-1/2} r < \epsilon$$

since $c_n \to 2B_{\bar{\sigma}^2}$ and therefore there exists $\Delta^2 > 2B_{\bar{\sigma}^2}$.

(b) The result follows directly from part (a) and the assumption that $\frac{nh_n^3}{\ln(n)} \to \infty$.

(c) Let $S(x) = \begin{pmatrix} g_X(x) & 0 \\ 0 & g_X(x)\sigma_K^2 \end{pmatrix}$ and $A_n(x) \equiv \tilde{m}(x,\alpha) - m(x) - \frac{1}{nh_n g_X(x)} \sum_{i=1}^n K(\frac{x_i - x}{h_n})\widetilde{Z}_i^*$, then

$$
\begin{aligned}
|A_n| &= \frac{1}{nh_n}\left|\sum_{i=1}^n \left(W_n\left(\frac{x_i - x}{h_n}, x\right) - \frac{1}{g_X(x)}K\left(\frac{x_i - x}{h_n}\right)\right)\widetilde{Z}_i^*\right| \\
&= \frac{1}{nh_n}\left|(1,0)(S_n^{-1}(x) - S^{-1}(x))\begin{pmatrix} \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^* \\ \sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\frac{x_i - x}{h_n}\widetilde{Z}_i^* \end{pmatrix}\right| \\
&\le \frac{1}{h_n}((1,0)(S_n^{-1}(x) - S^{-1}(x))^2(1,0)')^{1/2} \\
&\quad \times \frac{1}{n}\left(\left|\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right| + \left|\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\frac{x_i - x}{h_n}\widetilde{Z}_i^*\right|\right).
\end{aligned}
$$

By part (b) $B_n(x) \equiv \frac{1}{h_n}((1,0)(S_n^{-1}(x) - S^{-1}(x))^2(1,0)')^{1/2} = O_p(1)$ uniformly in G. Hence, if we put $R_{n,1}(x) \equiv n^{-1}(|\sum_{i=1}^n K(\frac{x_i - x}{h_n})\widetilde{Z}_i^*| + |\sum_{i=1}^n K(\frac{x_i - x}{h_n})\frac{x_i - x}{h_n}\widetilde{Z}_i^*|)$ the proof is complete.

Proof of Theorem 1. Given that $g_X(x) < C$ for all x from A1 and part (c) of Lemma 1, we have

$$
\begin{aligned}
&\left|\hat{m}(x,\alpha) - m(x) - \frac{1}{nh_n g_X(x)}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right| \\
&\le Ch_n B_n(x)\left(\left|\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right| + \left|\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right|\right) \\
&= Ch_n B_n(x)(|c_1(x)| + |c_2(x)|)
\end{aligned}
$$

Since $B_n(x) = O_p(1)$ uniformly in G from part (b) of Lemma 1, it suffices to investigate the order in probability of $|c_1(x)|$ and $|c_2(x)|$. Here, we establish the order of $c_1(x)$ noting that the proof for $c_2(x)$ follows a similar argument given assumption A2. We write $c_1(x) = I_{1n} - I_{2n} - I_{3n} + I_{4n}$, where

$$
\begin{aligned}
I_{1n}(x) &= \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right) \\
&\quad \times \left(-B^{(1)}(x)(x_i - x) - \frac{1}{2}B^{(2)}(x)(x_i - x)^2 - o((x_i - x)^2)\right),
\end{aligned}
$$

$$I_{2n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) \times \left(-\gamma^{(1)}(x)(x_i - x) - \frac{1}{2}\gamma^{(2)}(x)(x_i - x)^2 - o((x_i - x)^2)\right),$$

$$I_{3n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) m^{(1)}(x)(x_i - x),$$

$$I_{4n}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) r_i^{\alpha}\varepsilon_i,$$

where $B(v) = m(v,\theta)(\frac{m(x;\theta)}{m(v,\theta)})^{\alpha}$, $\gamma(v) = m(v)(\frac{m(x;\theta)}{m(v,\theta)})^{\alpha}$, and $B^{(d)}(v)$ and $\gamma^{(d)}(v)$ are derivatives of order d evaluated at v, whose existence follows from Assumption A4. Now,

$$I_{1n}(x) = -B^{(1)}(x)h_n s_{n,1}(x) - \frac{1}{2}B^{(2)}(x)h_n^2 s_{n,2}(x) - o(1)h_n^2 s_{n,2}(x),$$

and since from A4 $|B^{(1)}(x)|, |B^{(2)}(x)| < C$ for all $x \in G$, we have

$$\begin{aligned}
\sup_{x\in G}|I_{1n}(x)| &\le \mathrm{C}h_n \sup_{x\in G}|s_{n,1}(x)| + \mathrm{C}h_n^2 \sup_{x\in G}|s_{n,2}(x)| + o(1)h_n^2 \sup_{x\in G}|s_{n,2}(x)| \\
&\le \mathrm{C}h_n \sup_{x\in G}|s_{n,1}(x)| + \mathrm{C}h_n^2 \sup_{x\in G}|s_{n,2}(x) - E(s_{n,2}(x))| \\
&\quad + \mathrm{C}h_n^2 \sup_{x\in G}|E(s_{n,2}(x))| + o(1)h_n^2 \sup_{x\in G}|s_{n,2}(x) - E(s_{n,2}(x))| \\
&\quad + o(1)h_n^2 \sup_{x\in G}|E(s_{n,2}(x))| \\
&\le h_n O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right) + h_n^2 O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right) + \mathrm{C}h_n^2,
\end{aligned}$$

where the last inequality follows from part (a) of Lemma 1 and the fact that $\sup_{x\in G}|E(s_{n,2}(x))| = O(1)$. Similarly, given that $|\gamma^{(1)}(x)|, |\gamma^{(2)}(x)| < C$ for all $x \in G$ we have that

$$\sup_{x\in G}|I_{2n}(x)| \le h_n O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right) + h_n^2 O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right) + \mathrm{C}h_n^2. \tag{16}$$

$I_{3n}(x) = m^{(1)}(x)h_n s_{n,1}(x)$, and consequently, from part (a) of Lemma 1 and the fact that $|m^{(1)}(x)| < C$ for all $x \in G$ by A4,

$$\sup_{x\in G}|I_{3n}(x)| \le \mathrm{C}h_n O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right). \tag{17}$$

Now, we consider

$$I_{4n}(x) = m(x;\theta)^{\alpha}\left(\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right)\frac{\varepsilon_i}{m(x_i,\theta)^{\alpha}}\right)$$

$$= m(x;\theta)^{\alpha} q(x) \qquad \text{where } q(x) \equiv \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right)\frac{\varepsilon_i}{m(x_i,\theta)^{\alpha}}$$

$$|I_{4n}(x)| \leq C|q(x)| \qquad \text{by A4.}$$

Now consider an open covering $\{B(x_k, (\frac{n}{h_n^2})^{-1/2})\}_{k=1}^{l_n}$ such that $G \subset \bigcup_{k=1}^{l_n} B\left(x_k, \left(\frac{n}{h_n^2}\right)^{-1/2}\right)$ for $x_k \in G$ with $l_n < \left(\frac{n}{h_n^2}\right)^{1/2} r$ as described in the proof of Lemma 1. Then, we can write

$$|q(x)| \leq |\, q(x) - q(x_k)| + |q(x_k)|.$$

Now, observe that

$$|q(x) - q(x_k)| = \left|\frac{1}{nh_n}\sum_{i=1}^{n}\left(K\left(\frac{x_i - x}{h_n}\right) - K\left(\frac{x_i - x_k}{h_n}\right)\right)\frac{\varepsilon_i}{m(x_i,\theta)^{\alpha}}\right|$$

$$\leq \frac{1}{nh_n}\sum_{i=1}^{n} C\left|\frac{x_k - x}{h_n}\right|\frac{|\varepsilon_i|}{|m(x_i,\theta)^{\alpha}|} \qquad \text{by A2}$$

$$\leq C(nh_n^2)^{-1/2}\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i| \qquad \text{by A4.}$$

Also, by the fact that the conditional variance of ε_i is bounded for all x, we have that $E(|q(x) - E(q(x_k))|) \leq C(nh_n^2)^{-1/2}$, hence

$$\sup_{x\in G}|q(x)| \leq \max_{1\leq k\leq l_n}|q(x_k)| + C(nh_n^2)^{-1/2}\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|$$

$$\leq \max_{1\leq k\leq l_n}|q(x_k)| + C(nh_n^2)^{-1/2}.$$

The last inequality follows since by Kolmogorov's law of large numbers $\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i| = O_p(1)$. Hence, we now focus on $\max_{1\leq k\leq l_n}|q(x_k)|$. First, put $f(x_i,\varepsilon_i) = \frac{\varepsilon_i}{m(x_i,\theta)^{\alpha}}$ and let

$$q^{B}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) f(x_i,\varepsilon_i) I(|f(x_i,\varepsilon_i)| \leq B_n),$$

where $B_1 \le B_2 \le \cdots$ such that $\sum_{i=1}^{\infty} B_i^{-s} < \infty$ for some $s > 0$ and $I(\cdot)$ is the indicator function. Consider

$$\sup_{x \in G} |q(x) - q^B(x) - E(q(x) - q^B(x))| \le T_{1n} + T_{2n}$$

where

$$\begin{aligned} T_{1n} &= \sup_{x \in G} |q(x) - q^B(x)| \\ &= \sup_{x \in G} \left| \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n} \right) f(x_i, \varepsilon_i) I(|f(x_i, \varepsilon_i)| > B_n) \right|, \\ T_{2n} &= \sup_{x \in G} |E(q(x) - q^B(x))|. \end{aligned}$$

By the Borel–Cantelli Lemma, for all $\epsilon > 0$ and for all m such that $m' < m < n$ we have that $P(|f(x_m, \varepsilon_m)| \le B_n) > 1 - \epsilon$, and by Chebyshev's inequality and the increasing nature of the B_i sequence, for $n > N \in \Re$ we have $P(|f(x_i, \varepsilon_i|) < B_n) > 1 - \epsilon$. Hence, for $n > \max\{N, m\}$ we have that for all $i \le n$ $P(|f(x_i, \varepsilon_i)| < B_n) > 1 - \epsilon$, and therefore $I(|f(x_i, \varepsilon_i)| > B_n) = 0$ with probability 1, which gives $T_{1n} = o_{as}(1)$. Now,

$$\begin{aligned} |E(q(x) - q^B(x))| &\le \frac{1}{nh_n} \sum_{i=1}^{n} \int \int_{|f(x_i, \varepsilon_i)| > B_n} K\left(\frac{x_i - x}{h_n} \right) |f(x_i, \varepsilon_i)| g(x_i, \varepsilon_i) dx_i d\varepsilon_i \\ &\le \frac{1}{C_L^{\alpha} h_n} \int \int_{|\varepsilon_i| > B_n C_L} K\left(\frac{x_i - x}{h_n} \right) |\varepsilon_i| g(x_i, \varepsilon_i) dx_i d\varepsilon_i \\ &\le C \int K(\psi_i) \int_{|\varepsilon_i| > B_n C_L} |\varepsilon_i| g(x + h_n \psi_i, \varepsilon_i) d\varepsilon_i d\psi_i \\ &\le C \sup_{x \in G} \int_{|\varepsilon_i| > B_n C_L} |\varepsilon_i| g(x, \varepsilon_i) d\varepsilon_i. \end{aligned}$$

By Hölder's inequality, for $s > 1$,

$$\begin{aligned} &\int_{|\varepsilon_i| > B_n C_L} |\varepsilon_i| g(x, \varepsilon_i) d\varepsilon_i \\ &\quad \le \left(\int |\varepsilon_i|^s g(x, \varepsilon_i) d\varepsilon_i \right)^{1/s} \left(\int I(|\varepsilon_i| > B_n C_L) g(x, \varepsilon_i) d\varepsilon_i \right)^{1 - 1/s}. \end{aligned}$$

The first integral after the inequality is uniformly bounded, and since $g_{x_i \mid \varepsilon_i}(x) < C$, we have by Chebyshev's Inequality

$$\left(\int I(|\varepsilon_i| > B_n C_L) g(x, \varepsilon_i) d\varepsilon_i \right)^{1-1/s} \leq C(P(|\varepsilon_i| > B_n C_L))^{1-1/s} \leq C B_n^{1-s},$$

hence $T_{2n} = O(B_n^{1-s})$ and $\sup_{x \in G} |q(x) - q^B(x) - E(q(x) - q^B(x))| = O_{as}(B_n^{1-s})$. Now,

$$P\left(\max_{1 \leq k \leq l_n} |q^B(x_k) - E(q^B(x_k))| \geq \epsilon_n \right) \leq \sum_{i=1}^{l_n} P(|q^B(x_k) - E(q^B(x_k))| \geq \epsilon_n)$$

and let $q^B(x_k) - E(q^B(x_k)) = n^{-1} \sum_{i=1}^n Z_{in}$, where

$$\begin{aligned} Z_{in} = {} & \frac{1}{h_n} K\left(\frac{x_i - x}{h_n} \right) f(x_i, \varepsilon_i) I(|f(x_i, \varepsilon_i)| \leq B_n) \\ & - E\left(\frac{1}{h_n} K\left(\frac{x_i - x}{h_n} \right) f(x_i, \varepsilon_i) I(|f(x_i, \varepsilon_i| \leq B_n) \right). \end{aligned}$$

Since $f(x_i, \varepsilon_i) I(|f(x_i, \varepsilon_i)| \leq B_n) \leq B_n$ we have that $|Z_{in}| \leq C B_n / h_n$, $E(Z_{in}) = 0$, and since $\{Z_{in}\}_{i=1}^n$ is an independent sequence, by Bernstein's inequality

$$P(|q^B(x_k) - E(q^B(x_k))| \geq \epsilon_n) < 2 \exp\left(\frac{-n h_n \epsilon_n^2}{h_n \left(2\bar{\sigma}^2 + \frac{2 C B_n \epsilon_n}{3 h_n} \right)} \right),$$

where

$$\begin{aligned} \bar{\sigma}^2 = n^{-1} \sum_{i=1}^n V(Z_i) = {} & \frac{1}{h_n^2} E\left(K^2\left(\frac{x_i - x}{h_n} \right) f(x_i, \varepsilon_i)^2 I(|f(x_i, \varepsilon_i)| \leq B_n) \right) \\ & - \left(\frac{1}{h_n} E\left(K\left(\frac{x_i - x}{h_n} \right) f(x_i, \varepsilon_i) I(|f(x_i, \varepsilon_i)| \leq B_n) \right) \right)^2 \\ \leq {} & \frac{1}{h_n^2} E\left(K^2\left(\frac{x_i - x}{h_n} \right) \frac{1}{m(x_i, \theta)^{2\alpha}} \right) E(\varepsilon_i^2 \mid x_i) \\ & + \left(\frac{1}{h_n} \int K\left(\frac{x_i - x}{h_n} \right) \frac{1}{|m(x_i, \theta)^{\alpha}|} g_X(x_i) \right. \\ & \left. \times \int_{\left| \frac{\varepsilon_i}{m(x_i, \theta)^{\alpha}} \right| \leq B_n} |\varepsilon_i| g_{\varepsilon_i | x_i}(\varepsilon_i) d\varepsilon_i dx_i \right)^2 \\ \leq {} & h_n^{-1} C \int K^2(\psi) g_X(x + h_n \psi) d\psi \\ & + C^2 \left(\int K(\psi) g_X(x + h_n \psi) d\psi \right)^2 \leq C / h_n, \end{aligned}$$

where the bound follows form Assumption A1 and the fact that $E(\varepsilon_i^2 \mid x_i) = \sigma^2(x) < \infty$ for all x. Given that $\epsilon_n = \left(\frac{\ln(n)}{nh_n}\right)^{1/2}\Delta$, we have

$$P\Big(\max_{1\le k\le l_n}|q^B(x_k) - E(q^B(x_k))| \ge \epsilon_n\Big) < 2\frac{n^{1/2}}{h_n}n^{-\frac{\Delta^2}{c_n}},$$

where $c_n = 2h_n\bar{\sigma}^2 + \frac{2}{3}CB_n\epsilon_n$. Now, $h_n\bar{\sigma}^2 = O(1)$ and $c_n = o(1)$ provided $B_n\epsilon_n \to 0$ as $n \to \infty$. Hence, for Δ^2 sufficiently large

$$P\Big(\max_{1\le k\le l_n}|q^B(x_k) - E(q^B(x_k))| \ge \epsilon_n\Big) < \frac{n^{1-\frac{\Delta^2}{c_n}}}{(nh_n^2)^{1/2}} < \epsilon.$$

Hence,

$$\begin{aligned}\sup_{x\in G}|q(x)| &\le \sup_{x\in G}|q^B(x) - E(q^B(x))| + \sup_{x\in G}|q(x) - q^B(x) - E(q(x) - q^B(x))| \\ &= \sup_{x\in G}|q^B(x) - E(q^B(x))| + O_{as}(B_n^{1-s}) \\ &\le \max_{1\le k\le l_n}|q^B(x_k) - E(q^B(x_k))| + C(nh_n^2)^{-1/2} + O_{as}(B_n^{1-s}).\end{aligned}$$

By choosing $B_n = n^{\frac{1}{s}+\delta}$, $\delta > 0$, and $s > 2$ we have $O(B_n^{1-s}) = o(n^{-1/2})$. Then,

$$\begin{aligned}\sup_{x\in G}|q(x)| &= O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right) \qquad \text{and therefore} \\ \sup_{x\in G}|I_{4n}(x)| &= O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right).\end{aligned} \tag{18}$$

Combining the results for $I_{1n}(x)$, $I_{2n}(x)$, $I_{3n}(x)$, and $I_{4n}(x)$, we have $h_n \sup_{x\in G}|c_1(x)| = O_p\big(\big(\frac{h_n\ln(n)}{n}\big)^{1/2}\big) + O_p(h_n^3)$. Hence, given that $B_n(x) = O_p(1)$, we have that

$$\begin{aligned}&\sup_{x\in G}\left|\tilde{m}(x,\alpha) - m(x) - \frac{1}{nh_n g_X(x)}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right| \\ &\quad = O_p\left(\left(\frac{h_n\ln(n)}{n}\right)^{1/2}\right) + O_p(h_n^3).\end{aligned}$$

Proof of Theorem 2. From Theorem 1, and given that $h_n^2\ln(n) \to 0$, we have that

$$\sqrt{nh_n}\left(\tilde{m}(x,\alpha) - m(x) - \frac{1}{nh_n g_X(x)}\sum_{i=1}^n K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right) = o_p(1),$$

thus we focus on the asymptotic behavior of

$$\sqrt{nh_n}\left(\frac{1}{nh_n g_X(x)}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right)\widetilde{Z}_i^*\right)$$
$$= \sqrt{nh_n}\frac{1}{g_X(x)}(I_{1n}(x) - I_{2n}(x) - I_{3n}(x) + I_{4n}(x)),$$

where $I_{in}(x)$ for $i = 1, \ldots, 4$ are as defined in the proof of Theorem 1. We first examine $\sqrt{nh_n}\frac{1}{g_X(x)}I_{4n}(x) = \sum_{i=1}^{n} Z_{in}$, where $Z_{in} = \frac{K(\frac{x_i - x}{h_n})r_i^{\alpha}\varepsilon_i}{\sqrt{nh_n}g_X(x)}$. Note that $E(Z_{in}) = 0$, $V(Z_{in}) = \frac{\sigma^2}{nh_n g_X(x)^2}E(K^2(\frac{x_i - x}{h_n})r_i^{2\alpha})$, and define

$$s_n^2 = \sum_{i=1}^{n} V(Z_{in}) = \frac{\sigma^2}{g_X(x)^2 h_n}E\left(K^2\left(\frac{x_i - x}{h_n}\right)r_i^{2\alpha}\right).$$

By Lebesgue's dominated convergence theorem and assumption A4, we have that $s_n^2 \to \frac{\sigma^2}{g_X(x)}\int K^2(\psi)d\psi$.

By Liapounov's central limit theorem $\sum_{i=1}^{n}\frac{Z_{in}}{s_n} \xrightarrow{d} N(0,1)$ provided that $\sum_{i=1}^{n} E(|\frac{Z_{in}}{s_n}|^{2+\delta}) = o(1)$ for some $\delta > 0$. To verify this, note that

$$\begin{aligned}
\sum_{i=1}^{n} E\left(\left|\frac{Z_{in}}{s_n}\right|^{2+\delta}\right) &= (s_n^2)^{-1-\delta/2}\frac{g_X(x)^{-2-\delta}}{(nh_n)^{\delta/2}}\frac{1}{h_n}E\left(\left|K\left(\frac{x_i - x}{h_n}\right)\varepsilon_i r_i^{\alpha}\right|^{2+\delta}\right) \\
&= (s_n^2)^{-1-\delta/2}\frac{g_X(x)^{-2-\delta}}{(nh_n)^{\delta/2}}\frac{1}{h_n} \\
&\quad \times E\left(K^{2+\delta}\left(\frac{x_i - x}{h_n}\right)E(|\varepsilon_i|^{2+\delta}|x_i)|r_i^{\alpha}|^{2+\delta}\right) \\
&\leq (s_n^2)^{-1-\delta/2}\frac{g_X(x)^{-2-\delta}}{(nh_n)^{\delta/2}}\frac{C}{h_n}E\left(K^{2+\delta}\left(\frac{x_i - x}{h_n}\right)\right) \\
&\to Cg_X(x)\int K^{2+\delta}(\psi)d\psi,
\end{aligned}$$

given $E(|\varepsilon_i|^{2+\delta}|x_i) < C$ and assumptions A1 and A2. Now, since $\frac{C}{h_n}E(K^{2+\delta}(\frac{x_i - x}{h_n})) \to Cg_X(x)\int K^{2+\delta}(\psi)d\psi$, $s_n^2 \to \frac{\sigma^2}{g_X(x)}\int K^2(\psi)d\psi$ and $nh_n \to \infty$ we have $\sum_{i=1}^{n} E(|\frac{Z_{in}}{s_n}|^{2+\delta}) = o(1)$.

From Theorem 1 we have that,

$$\frac{-1}{nh_n g_X(x)}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right)B^{(1)}(x)(x_i - x) \leq Ch_n O_p\left(\left(\frac{nh_n}{\ln(n)}\right)^{-1/2}\right)$$

and therefore $\frac{-1}{\sqrt{nh_n}g_X(x)}\sum_{i=1}^{n}K(\frac{x_i-x}{h_n})B^{(1)}(x)(x_i-x) \le (h_n^2\ln(n))^{1/2}O_p(1) = o(1)$. Similarly,

$$\frac{-1}{\sqrt{nh_n}g_X(x)}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h_n}\right)\gamma^{(1)}(x)(x_i-x) \le (h_n^2\ln(n))^{1/2}O_p(1) = o(1)$$

and $\frac{1}{\sqrt{nh_n}g_X(x)}\sum_{i=1}^{n}K(\frac{x_i-x}{h_n})m^{(1)}(x)(x_i-x) \le (h_n^2\ln(n))^{1/2}O_p(1) = o(1)$. Now, let

$$v_n(x) = \frac{-B^{(2)(x)}h_n^2}{2nh_ng_X(x)}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)^2,$$

then $\frac{E(v_n(x))}{h_n^2} \to \frac{-B^{(2)}(x)}{2}\sigma_K^2 g_X(x)$ and

$$V\left(\frac{v_n(x)}{h_n^2}\right) = \left(\frac{B^{(2)}(x)}{2}\right)^2\frac{1}{nh_n^2g_X^2(x)} \times\left(h_n\int\psi^4K^2(\psi)g_X(x+h_n\psi)d\psi - h_n^2\sigma_K^4\right)\to 0.$$

Hence, by Chebyshev's inequality $\frac{v_n(x)}{h_n^2} - \frac{-B^{(2)}(x)}{2}\sigma_K^2 g_X(x) = o_p(1)$. Following the same arguments,

$$\frac{1}{h_n^2}\frac{-\gamma^{(2)(x)}h_n^2}{2nh_ng_X(x)}\sum_{i=1}^{n}K\left(\frac{x_i-x}{h_n}\right)\left(\frac{x_i-x}{h_n}\right)^2 - \frac{-\gamma^{(2)}(x)}{2}\sigma_K^2g_X(x) = o_p(1).$$

Hence,

$$\sqrt{nh_n}(\tilde{m}(x,\alpha) - m(x) - B(x;\alpha,\theta)) \xrightarrow{d} N\left(0, \frac{\sigma^2}{g_X(x)}\int K^2(\psi)d\psi\right),$$

where $B(x;\alpha,\theta) = \frac{1}{2}h_n^2\sigma_K^2(\gamma^{(2)}(x) - B^{(2)}(x)) + o_p(h_n^2)$. Simple manipulations give,

$$\begin{aligned}B_c(x;\alpha,\theta) = \gamma^{(2)}(x) - B^{(2)}(x) &= m^{(2)}(x) - (1-\alpha)m^{(2)}(x;\theta)\\ &\quad - \alpha\left(\frac{2m^{(1)}(x)m^{(1)}(x;\theta) + m(x)m^{(2)}(x;\theta)}{m(x;\theta)}\right)\\ &\quad + \frac{\alpha(\alpha+1)m(x)(m^{(1)}(x;\theta))^2}{m(x;\theta)^2} + \frac{\alpha(\alpha-1)m^{(1)}(x;\theta)}{m(x;\theta)}\end{aligned}$$

and therefore we can write $B(x;\alpha,\theta) = \frac{1}{2}h_n^2\sigma_K^2B_c(x;\alpha,\theta) + o_p(h_n^2)$.

Proof of Theorem 3. We prove the theorem by establishing that

$$\sqrt{nh_n}(\tilde{m}(x,\alpha)-\hat{m}(x,\alpha))=e'S_n^{-1}(x)\begin{pmatrix}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^n K\left(\frac{x_i-x}{g_n}\right)q_i\\ \frac{1}{\sqrt{nh_n}}\sum_{i=1}^n K\left(\frac{x_i-x}{g_n}\right)\left(\frac{x_i-x}{h_n}\right)q_i\end{pmatrix}=o_p(1),$$

where $q_i=\widetilde{Z}_i-\widehat{Z}_i$. Since $S_n^{-1}(x)=O_p(1)$ and K has compact support, it suffices to show that $\alpha_n=\frac{1}{\sqrt{nh_n}}\sum_{i=1}^n K(\frac{X_i-x}{h_n})q_i=o_p(1)$. We write

$$\begin{aligned}\alpha_n=&\sqrt{nh_n}\Bigg(\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i-x}{h_n}\right)(m(x;\theta_0)-m(x;\hat{\theta}))\\&+\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i-x}{h_n}\right)(m(x_i)-m(x_i,\theta_0))(r_i^\alpha-\hat{r}_i^\alpha)\\&+\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i-x}{h_n}\right)(r_i^\alpha-\hat{r}_i^\alpha)\varepsilon_i\\&+\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i-x}{h_n}\right)(m(x_i;\hat{\theta})-m(x_i,\theta_0))\hat{r}_i^\alpha\Bigg)\\=&\sqrt{nh_n}(Q_{1n}(x)+Q_{2n}(x)+Q_{3n}(x)+Q_{4n}(x))\end{aligned}$$

and treat each $\sqrt{nh_n}Q_{jn}$ for $j=1,\ldots,4$ separately:

$$\begin{aligned}\sqrt{nh_n}Q_{1n}(x)=\sqrt{nh_n}&((s_{n,0}(x)-g_X(x))(m(x;\theta_0)-m(x;\hat{\theta}))\\&+g_X(x)(m(x;\theta_0)-m(x;\hat{\theta})))\end{aligned}$$

Now, note that by Taylor's theorem

$$|m(x;\hat{\theta})^\alpha-m(x;\theta_0)^\alpha|\le|\alpha\|(m(x;\theta_m))^{\alpha-1}|\left|\frac{\partial m(x;\theta_m)}{\partial\theta}\right||\hat{\theta}-\theta_0|,$$

$$\theta_m=\lambda\hat{\theta}+(1-\lambda)\theta_0,\qquad\text{where }\lambda\in(0,1),$$

and given $0<C_L\le|m(x;\theta)|\le C_H<\infty$ for all $\theta\in\Theta$, $x\in G$, and $|\frac{\partial m(x;\theta)}{\partial\theta}|<C$, we have that $\sup_{x\in G}|m(x;\hat{\theta})^\alpha-m(x;\theta_0)^\alpha|\le C|\hat{\theta}-\theta_0|=n^{-1/2}O_p(1)$ by Theorem 3.2 in White (1982) for every α, which gives $m(x;\theta_0)-m(x;\hat{\theta})=O_p(n^{-1/2})$. Together with the fact that $|g_X(x)|<C$ and since from Lemma 1 $s_{n,0}(x)-g_X(x)=O_p(h_n)$ we have that $\sqrt{nh_n}Q_{1n}(x)=h_n^{3/2}O_p(1)+$

$h_n^{1/2} O_p(1) = o_p(1)$. Now, note that

$$|Q_{2n}(x)| \le \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) |m(x_i) - m(x_i, \theta_0)||r_i^\alpha - \hat{r}_i^\alpha|$$

$$\le C \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) |r_i^\alpha - \hat{r}_i^\alpha| \le C|s_{n,0}(x)| \sup_{x \in G} |r_i^\alpha - \hat{r}_i^\alpha|$$

and,

$$|r_i^\alpha - \hat{r}_i^\alpha|$$

$$\le \frac{|m(x_i; \hat{\theta})^\alpha - m(x_i, \theta_0)^\alpha \| m(x; \theta_0)^\alpha| + |m(x; \theta_0)^\alpha - m(x; \hat{\theta})^\alpha \| m(x_i, \theta_0)^\alpha|}{|m(x_i, \theta_0)^\alpha||m(x_i; \hat{\theta})^\alpha|}.$$

Since, as established above, $\sup_{x \in G} |m(x; \hat{\theta})^\alpha - m(x; \theta_0)^\alpha| = n^{-1/2} O_p(1)$ we have $\sup_{x \in G} |r_i^\alpha - \hat{r}_i^\alpha| = n^{-1/2} O_p(1)$ and consequently $\sqrt{nh_n} Q_{2n}(x) = h_n^{3/2} O_p(1) = o_p(1)$, since $s_{n,0}(x) = O_p(h_n)$. Now

$$|Q_{3n}(x)| \le \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) |r_i^\alpha - \hat{r}_i^\alpha||\varepsilon_i|$$

$$\le \sup_{x \in G} |r_i^\alpha - \hat{r}_i^\alpha| \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) |\varepsilon_i|$$

$$\le n^{-1/2} O_p(1) \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) |\varepsilon_i|$$

$$= n^{-1/2} O_p(1) \frac{1}{h_n} E\left(K\left(\frac{x_i - x}{h_n}\right) |\varepsilon_i|\right)$$

$$= n^{-1/2} O_p(1) \frac{1}{h_n} E\left(K\left(\frac{x_i - x}{h_n}\right)\right) E(|\varepsilon_i \| x_i)$$

$$\le n^{-1/2} O_p(1) \qquad \text{given that } E(|\varepsilon_i \| x_i) < C,$$

which gives $\sqrt{nh_n} Q_{3n}(x) = o_p(1)$. Similarly,

$$|Q_{4n}(x)| \le \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) |m(x_i, \hat{\theta}) - m(x_i, \theta_0) \| \hat{r}_i^\alpha|$$

$$\le \sup_{x_i \in G} |m(x_i, \hat{\theta}) - m(x_i, \theta_0)| \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_n}\right) |\hat{r}_i^\alpha|$$

$$\le n^{-1/2} O_p(1) s_{n,0}(x),$$

where the last inequality follows from the fact that $\hat{r}_i^{\alpha} = O_p(1)$ and $\sup_{x_i \in G} |m(x_i, \hat{\theta}) - m(x_i, \theta_0)| = O_p(n^{-1/2})$. Finally, since $s_{n,0}(x)$ uniformly converges to $g_X(x)$ by Lemma 1, we have $\sqrt{nh_n} Q_{4n}(x) = o_p(1)$, which concludes the proof.

ACKNOWLEDGMENTS

We thank participants of the Second Conference on Information and Entropy Econometrics, Amos Golan, Essie Maasoumi, Peter Phillips, Jeff Racine, and an anonymous referee for helpful comments. The authors retain responsibility for any remaining errors.

REFERENCES

DiMarzio, M., Taylor, C. C. (2004). Boosting kernel density estimates: A bias reduction technique? *Biometrika* 91:226–233.

Cressie, N., Read, T. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B* 46:440–464.

Fan, J. (1992). Design adaptive nonparametric regression. *Journal of the American Statistical Association* 87:998–1004.

Fan, J., Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85:645–660.

Fan, Y., Ullah, A. (1999). Asymptotic normality of a combined regression estimator. *Journal of Multivariate Analysis* 71:191–240.

Fan, J., Yao, Q. (2003). *Nonlinear Time Series*. New York: Springer Verlag.

Gasser, T., Müller, H.-G., Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society B* 47:238–252.

Glad, I. (1998). Parametrically guided non-parametric regression. *Scandinavian Journal of Statistics* 25:649–668.

Hjort, N., Glad, I. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics* 23:882–904.

Hjort, N., Jones, M. C. (1996). Locally parametric nonparamteric density estimation. *The Annals of Statistics* 24:1619–1647.

Imbens, G., Spady, R. H., Johnson, P. (1998). Information theoretic approaches to inferences in moment condition models. *Econometrica* 66:333–357.

Jones, M. C., Linton, O., Nielsen, J. P. (1995). A simple and effective bias reduction method for density and regression estimation. *Biometrika* 82:327–338.

Lewbel, A. (2007). A local generalized method of moments estimator. *Economics Letters* 94:124–128.

Martins-Filho, C., Yao, F. (2007). Nonparametric regression estimation with general parametric error covariance. Working paper, Department of Economics, Oregon State University. http://oregonstate.edu/~martinsc/martins-filho-yao-v3-07.pdf.

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

Naito, K. (2004). Semiparametric density estimation by local L_2 fitting. *The Annals of Statistics* 32:1162–1191.

Pagan, A., Ullah, A. (1999). *Nonparametric Econometrics*. New York: Cambridge University Press.

Rahman, M., Ullah, A. (2002). Improved combined parametric and nonparametric regression: Estimation and hypothesis testing. In: Ullah, A., Wan, A., Chaturvedi, A., eds. *Handbook of Applied Econometrics and Statistical Inference*. New York: Marcel Dekker.

Read, T., Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. New York: Springer Verlag.

Ruppert, D., Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics* 22:1346–1370.

Ruppert, D., Sheather, S. J., Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90:1257–1270.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics* 5:595–620.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50:1–25.