Nonparametric Statistics
Vol. 19, No. 1, January 2007, 23–62

Finite sample performance of kernel-based regression methods for non-parametric additive models under common bandwidth selection criterion

CARLOS MARTINS-FILHO*† and KE YANG‡

†Department of Economics, Oregon State University, Ballard Hall 303, Corvallis, OR 97331-3612 USA
‡College of Business, Alfred University, Olin Building 411, Alfred, NY 14802-1205 USA

(Received 17 April 2006; revised 12 December 2006)

In this paper, we investigate the finite sample performance of four kernel-based estimators that are currently available for additive non-parametric regression models – the classic backfitting estimator (CBE), the smooth backfitting estimator, the marginal integration estimator, and two versions of a two-stage estimator of which the first is proposed by Kim, Linton and Hengartner (1999) and the second is proposed in this paper. The bandwidths are selected for each estimator by minimizing their respective asymptotic approximation of the mean average squared errors. In our simulations, we are particularly concerned with the performance of these estimators under this unified data-driven bandwidth selection method, since in this case both the asymptotic and the finite sample properties of all estimators are currently unavailable. The comparison is based on the estimators' average squared error. Our Monte Carlo results seem to suggest that the CBE is the best performing kernel-based procedure.

Keywords: Additive non-parametric regression; Local linear estimation; Backfitting estimation; Smooth backfitting; Marginal integration

AMS Subject Classifications: 62G05; 62G08; 62G20

1. Introduction

Given a random vector (Y, X), $Y \in \Re$ and $X \in \Re^d$, the conditional expectation $E(Y|X = x) = m(x)$, where $x' = (x_1, \ldots, x_d)$ can be estimated non-parametrically under certain regularity conditions. Stone [1] showed that the best rate obtainable in the estimation of $m(x)$ is $n^{s/(2s+d)}$, where s is the degree of smoothness of the function m. The fact that the optimal rate depends inversely on d is known as the curse of dimensionality in non-parametric regression estimation. However, as shown by Stone [2], if $m(x)$ has an additive structure, i.e.

$$E(Y|X = x) = \alpha + \sum_{\delta=1}^{d} m_\delta(x_\delta) \tag{1}$$

*Corresponding author. Email: carlos.martins@oregonstate.edu; Fax: +1 541 737 5917

Nonparametric Statistics
ISSN 1048-5252 print/ISSN 1029-0311 online
http://www.tandf.co.uk/journals
DOI: 10.1080/10485250701297933

with $E(m_\delta(\cdot)) = 0$, each of the component functions $m_\delta(\cdot)$ can be estimated at an optimal rate $n^{s/(2s+1)}$, which does not depend on d. This circumvention of the curse of dimensionality, as well as the ease of interpreting the impacts of different regressors on the regressand has led to the popularity of additive non-parametric regression models in both theoretical and applied literatures.[†]

Four estimators have emerged as viable alternatives for the regression model in equation (1): the Classic backfitting estimator (CBE), proposed by Buja *et al.* [5]; the Marginal integration estimator (MIE), proposed by Newey [6], Tjøstheim and Auestad [7] and Linton and Nielsen [8]; a two-stage estimator (2SE), proposed by Linton [9] and Kim *et al.* [10]; and the smooth backfitting estimator (SBE), recently proposed by Mammen *et al.* [11]. All these estimators share, among other things, the use of kernel-based non-parametric estimation methods, such as Nadaraya-Watson or local polynomial fitting in intermediate stages,[‡] but they differ in how the additive structure constraint is utilized to produce final estimators of the component functions.

The CBE has been the most studied of these procedures. Using local polynomial as the intermediate smoother, CBE converges to the true regression function at an optimal rate of $n^{s/(2s+1)}$ (see [14] for the bivariate model and [15] for the multivariate model), but it is not oracle efficient, i.e. the estimator of each component function does not have the same asymptotic bias as when all other components are known. Compared with CBE, the MIE is computationally more expensive, but it reaches the oracle efficiency bounds (see [8] for $d = 2$ and [16] for $d > 2$). In addition, MIE is more robust against model mis-specification, according to a simulation study in [17]. However, the MIE becomes less efficient as the correlation among regressors increases, due to the fact that it needs to estimate the model at many out-of-sample points. The 2SE proposed by Kim *et al.* [10] reduces asymptotic variance by combining the MIE with a one-step backfitting. They also suggest the use of an internalized Nadaraya-Watson smoother in the MIE to avoid estimating the model at out-of-sample points. The 2SE is more efficient than MIE when an oversmoothing bandwidth is applied to the second stage estimation. More recently, Mammen *et al.* [11] proposed a smooth backfitting procedure that is motivated by the projection interpretation of kernel estimators, suggested by Mammen *et al.* [18]. Its implementation relies on iterative calculation of a system of first order equations from a suitably defined distance minimization criterion. The SBE does not have the drawbacks of CBE, MIE or 2SEs. It reaches both the optimal convergence rate and the oracle efficiency bound. In addition, the asymptotic expressions of SBE for one component function do not rely on other components that completely circumvents the problem caused by the correlation among regressors. A simulation study in [19] shows that SBE is computationally quite efficient even for a high dimensional model, e.g. $d = 100$.

For empirical researchers, how these different procedures perform in finite samples is of essential interest. First, the slower convergence rate of non-parametric estimators compared with parametric estimators suggests that their finite sample properties may be quite different from what is suggested by the asymptotic theory. Second, unfortunately, all asymptotic properties obtained for these estimators rely on bandwidths being non-stochastic. In practice, however, bandwidths are chosen by data driven methods, such as cross validation, and various plug-in methods (see for example [20] and [21]). Therefore, a carefully designed Monte Carlo simulation based on data driven bandwidth selection methods would be valuable to reveal the relative estimation accuracy of these procedures in various scenarios.

[†]See, *inter alia*, [3] and [4].

[‡]Alternative non-parametric smoothing methods, e.g. spline or wavelet method, could potentially be used, but such methods have not received the attention given to kernel-based methods. See [12] and [13].

There is a small number of simulation studies dealing with additive non-parametric regression (see [17] and [19]). The current literature generally makes comparisons based on bandwidth selection methods that favor one of the competing estimators. A variety of bandwidth selection methods have been proposed for different estimators. These include the direct plug-in method proposed by Opsomer and Ruppert [21] for the CBE with local polynomial smoothing; rule of thumb bandwidths suggested by Linton and Nielsen [8] for the MIE and by Kim *et al.* [10] for the 2SE; cross-validation methods proposed by Nielsen and Sperlich [19] and penalized least square methods and plug-in methods proposed by Mammen and Park [22] for the SBE. Here, to accurately assess the relative performance of the estimators, a unified plug-in method is proposed for bandwidth selection in the simulation, which is inspired by the DPI method of Opsomer and Ruppert [21] and involves a common criterion function for bandwidth selection. To the best of our knowledge, this is the first simulation study for all four of the available estimators with a unified bandwidth selection method. We are also particularly interested in the impact of different degrees of regressor dependency on the estimation of m_δ. Robustness against model mis-specification is not an objective of our simulation, i.e. in all experiments we conducted here, the underlying models are always assumed to be additive.[†] Ultimately, our objective is to provide applied researchers with information that allows one for a more accurate comparison of these competing estimation alternatives in a finite sample setting.

Besides this introduction the paper has five more sections. Section 2 describes in a unified notation the estimators under study and their properties. Section 3 provides asymptotic conditional bias and variance for the SBE, MIE and 2SE estimators, a plug-in formula to select bandwidths and a description of how the bandwidth selection method is implemented. Section 4 presents the data generation processes used in the simulation and section 5 discusses the results and makes some recommendations. Section 6 provides a brief conclusion with some directives for future research.

2. Estimators under study

For computational convenience, notation and exposition, a bivariate model is used in this paper, but the conclusions extend to higher dimensions. Let (Y, X, Z) be a random vector with joint density $f(y, x, z)$ such that $E(Y|X = x, Z = z) = m(x, z) = \alpha + m_1(x) + m_2(z)$, with $E(m_1(X)) = E(m_2(Z)) = 0$ and $V(Y|X = x, Z = z) = \sigma^2$. Here α and σ^2 are unknown parameters and $m_1(\cdot)$ and $m_2(\cdot)$ are real valued functions with regularity properties that will be made explicit later in this section. Suppose a random sample of size n, denoted by $\{y_i, x_i, z_i\}_{i=1}^n$ is available. Our primary interest is on the estimation of $m(x, z) = \alpha + m_1(x) + m_2(z)$. Let $\vec{y} = (y_1, \ldots, y_n)'$, and define similarly $\vec{x}$ and $\vec{z}$. In addition, let $\vec{m}_1(\vec{x}) = (m_1(x_1), \ldots, m_1(x_n))'$, and similarly define $\vec{m}_2(\vec{z})$.

Since a local linear smoother will be used in defining the estimators under study, we first introduce some notations. Let $K_{h_j}(\cdot) = K(\cdot/h_j)/h_j$, where $K(\cdot)$ is an univariate kernel function and h_j, $j = 1, 2$, are the bandwidths used for the estimation of m_1 and m_2, respectively. The local linear smoothing matrix with respect to $\vec{x}$ and $\vec{z}$ are defined as

$$S_1 = \begin{pmatrix} s_1(x_1) \\ \vdots \\ s_1(x_n) \end{pmatrix}, \quad \text{and} \quad S_2 = \begin{pmatrix} s_2(z_1) \\ \vdots \\ s_2(z_n) \end{pmatrix}, \tag{2}$$

[†]See [17] and [23] for simulation studies that address model mis-specifications.

where $s_1(x), s_2(z) : \Re \to \Re^n$ are

$$s_1(x) = e(R_X(x)'W_X(x)R_X(x))^{-1}R_X(x)'W_X(x),$$
$$s_2(z) = e(R_Z(z)'W_Z(z)R_Z(z))^{-1}R_Z(z)'W_Z(z), \tag{3}$$

where $e = (1, 0)$, $W_X(x) = \text{diag}\{K_{h_1}(x_i - x)\}_{i=1}^n$, $W_Z(z) = \text{diag}\{K_{h_2}(z_i - z)\}_{i=1}^n$, $R_X(x) = (\vec{1}_n, \vec{x} - \vec{1}_n x)$, $R_Z(z) = (\vec{1}_n, \vec{z} - \vec{1}_n z)$ and $\vec{1}_n$ is a one vector of size n.

Given a bivariate non-parametric estimator $\hat{m}(x, z)$ for $m(x, z)$, one can, in general, define estimators for $m_1(\cdot)$, $m_2(\cdot)$ and α as solutions for the following minimization problem:

$$\begin{aligned} \text{minimize} \quad & \iint \{\hat{m}(x, z) - m_1(x) - m_2(z) - \alpha\}^2 dP(x, z) \\ \text{subject to} \quad & m_1 \in \mathcal{H}_1, m_2 \in \mathcal{H}_2, \quad \text{and} \quad \alpha \in \mathbb{R}, \end{aligned} \tag{4}$$

where $P(\cdot, \cdot)$ is a joint measure, while $\mathcal{H}_1$ and $\mathcal{H}_2$ are function classes members of which satisfy the identification conditions such as $\iint m_1(x) dP(x, z) = 0$ and $\iint m_2(z) dP(x, z) = 0$. Note that given $E(m_1(X)) = E(m_2(Z)) = 0$, a suitable estimator for α is $\bar{y} = 1/n \sum_{i=1}^n y_i$. For the rest of our discussion we will simply assume α is known since $\bar{y}$ converges to α with rate $\sqrt{n}$.

2.1 *Classic backfitting estimator* (*CBE*)

In the minimization problem defined in equation (4) if we take the measure $P(x, z)$ to be the joint probability measure of X and Z, i.e. $dP(x, z) = f_{XZ}(x, z)dxdz$, with $f_{XZ}(x, z)$ being the joint density of X and Z, the solution to the minimization problem should satisfy the following equations:

$$m_1(x) = \int \hat{m}(x, z) \frac{f_{XZ}(x, z)}{f_X(x)} dz - \int m_2(z) \frac{f_{XZ}(x, z)}{f_X(x)} dz - \bar{y} \tag{5}$$

$$m_2(z) = \int \hat{m}(x, z) \frac{f_{XZ}(x, z)}{f_Z(z)} dx - \int m_1(x) \frac{f_{XZ}(x, z)}{f_Z(z)} dx - \bar{y} \tag{6}$$

where $f_X(\cdot)$ and $f_Z(\cdot)$ are marginal densities of X and Z, respectively. Replacing the conditional expectations appearing in equations (5) and (6) with appropriate local linear projections, the CBE can be expressed as the solution for

$$\begin{pmatrix} I_n & S_1^* \\ S_2^* & I_n \end{pmatrix} \begin{pmatrix} \vec{m}_1^{\text{CB}}(\vec{x}) \\ \vec{m}_2^{\text{CB}}(\vec{z}) \end{pmatrix} = \begin{pmatrix} S_1^* \\ S_2^* \end{pmatrix} \vec{y}, \tag{7}$$

with $S_d^* = (I_n - (1/n)\vec{1}_n\vec{1}_n')S_d \equiv D_n S_d$ for $d = 1, 2$, where I_n is an identity matrix.

2.2 *Marginal integration estimator* (*MIE*)

In the minimization problem defined in equation (4), if we let $dP(x, z) = f_X(x) f_Z(z)dxdz$ instead, the solutions to the minimization problem satisfy

$$m_1(x) = \int \hat{m}(x, z) f_Z(z) dz - \bar{y} \tag{8}$$

$$m_2(z) = \int \hat{m}(x, z) f_X(x) dx - \bar{y}. \tag{9}$$

The MIE is essentially an empirical version of equations (8) and (9), with $f_Z(z)$ and $f_X(x)$ replaced by empirical frequencies. More precisely, the MIE is defined by first obtaining an estimator $\hat{m}(x, z)$, in this case a bivariate local linear estimator, defined as

$$\hat{m}(x, z; h_1, h_2) = e_2(X(x, z)'W(x, z)X(x, z))^{-1}X(x, z)'W(x, z)\vec{y}, \tag{10}$$

where $e_2 = (1, 0, 0)$, $X(x, z) = (\vec{1}_n, \vec{x} - \vec{1}_n x, \vec{z} - \vec{1}_n z)$ and $W(x, z) = \text{diag}\{K_{h_1}(x_i - x) \times K_{h_2}(z_i - z)\}_{i=1}^n$. Then,

$$m_1^{\text{MI}}(x) = \frac{1}{n}\sum_{i=1}^{n}\hat{m}(x, z_i) - \bar{y}, \quad m_2^{\text{MI}}(z) = \frac{1}{n}\sum_{i=1}^{n}\hat{m}(x_i, z) - \bar{y}. \tag{11}$$

2.3 *Two-stage estimators (2SE1 & 2SE2)*

The 2SE is an effort to improve upon MIE by accounting for the dependency between X and Z in solving equations (5) and (6). This is accomplished by estimating $\int(m_2(z)f_{XZ}(x, z)/f_X(x))dz \equiv E(m_2(z)|X = x)$ by $s_1(x)\vec{m}_2^{\text{MI}}(\vec{z})$, $\int(m_1(x)f_{XZ}(x, z)/f_Z(z))dx \equiv E(m_1(x)|Z = z)$ by $s_2(z)\vec{m}_1^{\text{MI}}(\vec{x})$, where $\vec{m}_1^{\text{MI}}(\vec{x}) = (m_1^{\text{MI}}(x_1), \ldots, m_1^{\text{MI}}(x_n))$, $\vec{m}_2^{\text{MI}}(\vec{z}) = (m_2^{\text{MI}}(z_1), \ldots, m_2^{\text{MI}}(z_n))$. In addition, $\int(\hat{m}(x, z)f_{XZ}(x, z)/f_X(x))dz \equiv E(\hat{m}(x, z)|X = x)$ and $\int(\hat{m}(x, z)f_{XZ}(x, z)/f_Z(z))dx \equiv E(\hat{m}(x, z)|Z = z)$ are estimated, respectively, by $s_1(x)\vec{y}$ and $s_2(z)\vec{y}$. Kim *et al.* [10] consider the case where m_1^{MI} and m_2^{MI} are based on a bivariate internalized Nadaraya-Watson estimate for $m(x, z)$[†] and define the 2SE1 as,

$$\begin{aligned}\vec{m}_1^{2S1}(x) &= s_1(x)(\vec{y} - \vec{m}_2^{\text{MI}}(\vec{z}) - \vec{1}_n\bar{y}) = s_1(x)(\vec{y} - \vec{\gamma}_2^P(\vec{z})) \quad \text{and} \\ \vec{m}_2^{2S1}(z) &= s_2(z)(\vec{y} - \vec{m}_1^{\text{MI}}(\vec{x}) - \vec{1}_n\bar{y}) = s_2(z)(\vec{y} - \vec{\gamma}_1^P(\vec{x})),\end{aligned} \tag{12}$$

where $\vec{\gamma}_1^P(\vec{x}) = (\gamma_1^P(x_1), \ldots, \gamma_1^P(x_n))'$ and $\vec{\gamma}_2^P(\vec{z})$ are similarly defined with

$$\gamma_1^P(x_i) = \frac{1}{n}\sum_{j=1}^{n}K_{g_1}(x_j - x_i)\frac{\hat{f}_Z(z_j)}{\hat{f}_{XZ}(x_j, z_j)}y_j, \quad \gamma_2^P(z_i) = \frac{1}{n}\sum_{j=1}^{n}K_{g_2}(z_j - z_i)\frac{\hat{f}_X(x_j)}{\hat{f}_{XZ}(x_j, z_j)}y_j \tag{13}$$

and $\hat{f}_X(x_i)$, $\hat{f}_Z(z_i)$ and $\hat{f}_{XZ}(x_i, z_i)$ are kernel density estimates with bandwidth g_1 and g_2 associated with X and Z, respectively.

Since the internalized Nadaraya-Watson smoother does not produce an equivalent kernel vector that sums to one, the 2SE1 may not be accurate even in the simplest case, where $\vec{y}$ is a constant vector. To achieve better finite sample performance, we propose an alternative two-stage estimation procedure, 2SE2 as follows:

- First, pilot estimators for $m_1(x_i)$ and $m_2(z_i)$, $i = 1, \ldots, n$ are obtained by

$$m_1^P(x_i) = \frac{1}{n}\sum_{j=1}^{n}K_{g_1}(x_j - x_i)\frac{\hat{f}_Z(z_j)}{\hat{f}_{XZ}(x_j, z_j)}(y_j - \bar{y}) \tag{14}$$

$$m_2^P(z_i) = \frac{1}{n}\sum_{j=1}^{n}K_{g_2}(z_j - z_i)\frac{\hat{f}_X(x_j)}{\hat{f}_{XZ}(x_j, z_j)}(y_j - \bar{y}); \tag{15}$$

[†]See [24] and [10] for details.

- Second, the final 2SE2 is obtained with a one step backfitting procedure,

$$\vec{m}_1^{2S2}(x) = s_1(x)(\vec{y} - \vec{1}_n\bar{y} - \vec{m}_2^P(\vec{z})) \quad \text{and} \quad \vec{m}_2^{2S2}(z) = s_2(z)(\vec{y} - \vec{1}_n\bar{y} - \vec{m}_1^P(\vec{x})), \quad (16)$$

where $\vec{m}_1^P(\vec{x}) = (m_1^P(x_1), m_1^P(x_2), \ldots, m_1^P(x_n))'$ and $\vec{m}_2^P(\vec{z})$ are similarly defined.

We expect that 2SE2 will outperform 2SE1 in general, and particularly so when α is of relatively large scale.

2.4 *Smooth backfitting estimator* (*SBE*)

The local linear SBE is motivated by the following minimization problem

$$\begin{aligned}\text{minimize} \quad & \iint \sum_{i=1}^{n} \{Y_i - \alpha - m_1(x) - m_2(z) - m_1^{(1)}(x)(x_i - x) \\ & - m_2^{(1)}(z)(z_i - z)\}^2 K_{h_1}(x_i - x) K_{h_2}(z_i - z) dx dz, \end{aligned} \quad (17)$$

subject to the identification conditions

$$\int \sum_{i=1}^{n} m_1(x) K_{h_1}(x_i - x) dx = \int \sum_{i=1}^{n} m_2(z) K_{h_2}(z_i - z) dz = 0. \quad (18)$$

Note that the minimization is with respect to α, $m_1(x)$ and $m_2(z)$ and their first derivatives $m_1^{(1)}(x)$ and $m_2^{(1)}(z)$. Again, α can simply be estimated by $\bar{y}$, so the first order conditions of the above minimization with respect to $m_1(x)$ and $m_1^{(1)}(x)$ are given by

$$\begin{pmatrix} m_1^{SB}(x) \\ m_1^{(1),SB}(x) \end{pmatrix} = \begin{pmatrix} \tilde{m}_1(x) \\ \tilde{m}^{(1)}(x) \end{pmatrix} - \hat{M}_X(x)^{-1} \int \hat{S}_{XZ}(x, z) \begin{pmatrix} m_2^{SB}(z) \\ m_2^{(1),SB}(z) \end{pmatrix} dz, \quad (19)$$

where $\begin{pmatrix} \tilde{m}_1(x) \\ \tilde{m}^{(1)}(x) \end{pmatrix}$ is a local linear projection of $(\vec{y} - \vec{1}_n\bar{y})$ onto the subset of $\Re^n$ where $\vec{x}$ takes values and

$$\hat{M}_X(x) = \begin{pmatrix} \hat{f}_X(x) & \hat{f}_X^X(x) \\ \hat{f}_X^X(x) & \hat{f}_X^{XX}(x) \end{pmatrix}, \quad \hat{S}_{XZ}(x, z) = \begin{pmatrix} \hat{f}_{XZ}(x, z) & \hat{f}_{XZ}^Z(x, z) \\ \hat{f}_{XZ}^X(x, z) & \hat{f}_{XZ}^{XZ}(x, z) \end{pmatrix}$$

with

$$\hat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x), \quad \hat{f}_{XZ}(x, z) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x) K_{h_2}(z_i - z),$$

$$\hat{f}_X^X(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x)(x_i - x),$$

$$\hat{f}_{XZ}^Z(x, z) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x) K_{h_2}(z_i - z)(z_i - z),$$

$$\hat{f}_X^{XZ}(x, z) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x)(x_i - x)(z_i - z) \quad \text{and}$$

$$\hat{f}_{XZ}^{XZ} = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x_i - x) K_{h_2}(z_i - z)(x_i - x)(z_i - z).$$

Similar first order conditions as in equation (19) can be defined for $m_2(z)$ and $m_2^{(1)}(z)$. With starting values set to $m_j^{\mathrm{SB}}(\cdot)$, $m_j^{(1),\mathrm{SB}}(\cdot)$, $\tilde{m}_j(\cdot)$, $\tilde{m}_j^{(1)}(\cdot)$ for $j = 1, 2$, the smooth backfitting estimator is obtained by iterative calculation of equation (19) and its analogue with respect to Z, until $m_j^{\mathrm{SB}}(\cdot)$, $j = 1, 2$ converge under a suitably chosen criterion. In implementing the algorithm, the integral in the updating equation (19) can be approximated with a weighted average of the integrand evaluated over a grid in the support of Z (or X).

3. Asymptotic approximations and bandwidth selection

The plug-in bandwidth selection methods, which we consider for all estimators, depend on obtaining suitable asymptotic approximations for the conditional mean average squared errors (MASE). By definition, for a generic estimator $\hat{m}(x, z)$ of $m(x, z)$, we have

$$\begin{aligned}\mathrm{MASE}(\hat{m}|\vec{x}, \vec{z}) &= \frac{1}{n}\sum_{i=1}^{n}(E(\hat{m}(x_i, z_i) - m(x_i, z_i)|\vec{x}, \vec{z}))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(E(\hat{m}(x_i, z_i)|\vec{x}, \vec{z}) - m(x_i, z_i))^2 + \frac{1}{n}\sum_{i=1}^{n} V(\hat{m}(x_i, z_i)|\vec{x}, \vec{z}). \quad (20)\end{aligned}$$

Since conditional on the regressors MASE can be written as the averaged squared conditional bias and averaged conditional variance of the the estimator, we need expressions for the bias and variance in order to obtain data dependent expressions for h_1 and h_2 that minimize an asymptotic approximation for MASE (AMASE). To this end we make the following general assumptions that are necessary to obtain the conditional bias and variance of the estimators under study:

ASSUMPTION 1 *The kernel $K(\cdot)$ is such that K: $[-1, 1] \to [0, B_K)$ for some finite $B_K > 0$, $K(\psi) = K(-\psi)$ for $\psi \in \Re$, $\mu_1 = \int \psi K(\psi) d\psi = 0$, $\mu_2 = \int \psi^2 K(\psi) d\psi < \infty$ and there exists a constant c such that $|K(u) - K(v)| \leq c|u - v|$ for all $u, v \in \Re$. In addition, $\int K^2(\psi) d\psi$ exists and we write $R_K = \int K^2(\psi) d\psi$.*

ASSUMPTION 2 *The second derivative of the functions $m_1(x)$, $m_2(z)$, $f_X(x)$, $f_Z(z)$ and $f_{XZ}(x, z)$, denoted by $m_1^{(2)}(x)$, $m_2^{(2)}(z)$, $f_X^{(2)}(x)$, $f_Z^{(2)}(z)$ and $\partial^2 f_{XZ}(x, z)/\partial_d \partial_d$, respectively, all exist and are continuous over their compact supports given by S_X, S_Z and $S_X \times S_Z$. We assume further that there exist generic constants $0 < b_f < B_f$ that are, respectively, lower and upper bounds on f_X, f_Z and f_{XZ}.*

ASSUMPTION 3 *There exist non-stochastic bandwidths g_1, h_1 and g_2, h_2 associated with regression directions $m_1(\cdot)$ and $m_2(\cdot)$, respectively. These bandwidths are such that $g_1, h_1, g_2, h_2 \to 0$, $nh_1h_2, ng_1g_2 \to \infty$ as $n \to \infty$, and that $g_d \sim h_d$ (same order) for $d = 1, 2$.*

3.1 *Classic backfitting estimator (CBE)*

Opsomer and Ruppert (1997) show that when $(nh_1)/(\log n)$, $(nh_2)/(\log n) \to \infty$ it is possible to obtain asymptotic approximations for the conditional bias and variance of $m_1^{\mathrm{CB}}(x)$ and $m_2^{\mathrm{CB}}(z)$. These asymptotic approximations are most commonly used in obtaining the estimator's mean average squared error (AMASE). Since AMASE is highly non-linear on the

bandwidths, the minimization of AMASE can only be accomplished by a numerical procedure. However, in the special case of independent regressors, it is possible to obtain an analytical solution for the optimal bandwidths. Expressions for the optimal h_1 and h_2, in the sense that they minimize the AMASE, for CBE are given by:

$$h_1^{\mathrm{CB}} = \left(\frac{\sigma^2 R_K}{n\mu_2^2\theta_{11}} n^{-1} \sum_{i=1}^{n} f_X(x_i)^{-1}\right)^{1/5} \quad \text{and} \quad h_2^{\mathrm{CB}} = \left(\frac{\sigma^2 R_K}{n\mu_2^2\theta_{22}} n^{-1} \sum_{i=1}^{n} f_Z(z_i)^{-1}\right)^{1/5}, \tag{21}$$

where $\theta_{11} = n^{-1}\sum_{i=1}^{n}(m_1^{(2)}(x_i) - E(m_1^{(2)}(x_i)))^2$ and $\theta_{22} = n^{-1}\sum_{i=1}^{n}(m_2^{(2)}(z_i) - E(m_2^{(2)}(z_i)))^2$. A few points are worth noting regarding the practical use of these expressions: (a) θ_{11}, θ_{22}, f_X and f_Z are unknown, rendering h_1^{CB} and h_2^{CB} inadequate for producing feasible CBE. In practice, the unknown quantities θ_{11}, θ_{22}, f_X and f_Z must be estimated to render the expressions in equation (21) useful; and (b) their relatively simple analytical form derives from assuming independence of the regressors. A simulation study in [21] suggests that these bandwidths are quite robust in increasing correlation between regressors. Therefore, we adopt this method in implementing bandwidth selection in our simulations.

3.2 *Smooth backfitting estimator (SBE)*

The following theorem is a trivial extension of the results in [19] and [22] to give an approximation for conditional bias, variance and MASE of SBE.

THEOREM 1 *Under Assumptions* 1–3 *and if* $nh_1h_2^2, nh_2h_1^2 \to \infty$, *for* $(x, z) \in S_X \times S_Z$, *the conditional bias and variance of local linear SBE can be written as*:

$$E(m_1^{\mathrm{SB}}(x) - m_1(x)|\vec{x}) = \frac{1}{2}\mu_2 h_1^2 (m_1^{(2)}(x) - E(m_1^{(2)}(X))) + o_p(h_1^2) \tag{22}$$

$$V(m_1^{\mathrm{SB}}(x)|\vec{x}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + o_p((nh_1)^{-1}). \tag{23}$$

Mutatis mutandis, *similar expressions for* $m_2^{\mathrm{SB}}(z)$ *are obtained. The conditional MASE of local linear SBE for* $m(x, z)$ *is given by*,

$$\begin{aligned}\mathrm{MASE} = {} & \frac{1}{4}\mu_2^2 h_1^4 \theta_{11} + \frac{1}{4}\mu_2^2 h_2^4 \theta_{22} + \frac{1}{4}\mu_2^2 h_1^2 h_2^2 \theta_{12} + \frac{1}{nh_1}\sigma^2 R_K \sum_{i=1}^{n} f_X(x_i)^{-1} \\ & + \frac{1}{nh_2}\sigma^2 R_K \sum_{i=1}^{n} f_Z(z_i)^{-1} + o_p(h_1^4 + h_2^4) + o_p((nh_1)^{-1} + (nh_2)^{-1}),\end{aligned} \tag{24}$$

where θ_{11}, θ_{22} *are the same as defined in equation* (21) *and* $\theta_{12} = \sum_{i=1}^{n}(m_1^{(2)}(x_i) - E(m_1^{(2)}(x_i)))(m_2^{(2)}(z_i) - E(m_2^{(2)}(z_i)))$

The plug-in estimators for bandwidths that minimize the AMASE, denoted by h_1^{SB} and h_2^{SB}, can be obtained from the following procedures:

1. Fit the model with local linear SBE with a preliminary bandwidth, denote the estimates $\hat{m}_1^0(\vec{x})$, $\hat{m}_2^0(\vec{z})$. Use the residuals to calculate $\hat{\sigma}^2$;
2. Project $\hat{m}_1^0(\vec{x})$ onto the subset of $\Re^n$, where $\vec{x}$ takes values using a local cubic kernel smoother to obtain estimates for $m_1^{(2)}(x_i)$, denoted by $\check{m}_1^{(2)}(x_i)$, similarly get $\check{m}_2^{(2)}(z_i)$ for

all $i = 1, \ldots, n$. Estimate θ_{11}, θ_{12} and θ_{22} by averaging over sample points. Denote the estimates by $\hat{\theta}_{11}, \hat{\theta}_{12}$ and $\hat{\theta}_{22}$; and

3. Plug $\hat{\theta}_{11}, \hat{\theta}_{12}, \hat{\theta}_{22}$ and $\hat{\sigma}^2$ into AMASE and find $h_1^{\mathrm{SB}} > 0$, $h_2^{\mathrm{SB}} > 0$ that minimize the AMASE.

This procedure is a revised version of the plug-in method in [22], where an iterative SBE fitting is used for the optimal bandwidth searching. Our procedure is computationally simpler as it requires only one SBE fitting, which should be the most time consuming part in the search procedure. The relative performance of these two alternative procedures for bandwidth selection has not been investigated. Note that in general, no analytical expressions for h_1^{SB} and h_2^{SB} are available. In the special case where X is independent from Z, the term θ_{12} is of order $O_p(n^{-1})$ and, therefore, can be ignored in AMASE. In this case, the h_1^{SB} and h_2^{SB} can be written as

$$h_1^{\mathrm{SB}} = \left(\frac{\sigma^2 R_K n^{-1} \sum_{i=1}^n f_X(x_i)^{-1}}{n \mu_2^2 \theta_{11}} \right)^{1/5} \quad \text{and} \quad h_2^{\mathrm{SB}} = \left(\frac{\sigma^2 R_K n^{-1} \sum_{i=1}^n f_Z(z_i)^{-1}}{n \mu_2^2 \theta_{22}} \right)^{1/5}. \tag{25}$$

These expressions are identical to h_1^{CB} and h_2^{CB} and the plug-in bandwidth for univariate local linear regression of Ruppert *et al.* [25]. The only difference here is that the unknown quantities are estimated using SBE. Based on the good performance of this bandwidth for CBE in the presence of dependence among regressors, we conjecture that it should work reasonably well for SBE.

3.3 *Marginal integration estimator (MIE)*

Linton and Nielsen [8] show that when $nh_1h_2^2, nh_2h_1^2 \to \infty$ then $\sqrt{nh_j}(m_j^{\mathrm{MI}}(\cdot) - E(m_j^{\mathrm{MI}}(\cdot)))$, for $j = 1, 2$, are asymptotically normal. However, the AMASE for the MIE, even under regression independence, does not produce closed analytical expressions for optimal bandwidths similar to those for CBE and SBE. The AMASE for the MIE and the optimal bandwidths that minimize AMASE are presented in the following theorem, the proof of which is straightforward, compared with the results in [8], and is omitted.

THEOREM 2 *Let $(x, z) \in S_X \times S_Z$ and assume that X and Z are independent. Assume that Assumptions* 1–3 *are holding and that $nh_1h_2^2, nh_2h_1^2 \to \infty$.*

(i) *The conditional bias and variance of $m_1^{MI}(x)$ for $x \in S_X$ are given by,*

$$E(m_1^{MI}(x) - m_1(x)|\vec{x}, \vec{z}) = \frac{1}{2} h_1^2 \mu_2 m_1^{(2)}(x) + \frac{1}{2} h_2^2 \mu_2 E(m_2^{(2)}(Z)) + o_p(h_1^2 + h_2^2) \tag{26}$$

and

$$V(m_1^{MI}(x)|\vec{x}, \vec{z}) = \frac{1}{nh_1} \sigma^2 R_K f_X(x)^{-1} + o_p((nh_1)^{-1}). \tag{27}$$

Mutatis mutandis identical expressions for $m_2^{MI}(z)$ are obtained. For $m^{MI}(x, z) = \bar{y} + m_1^{MI}(x) + m_2^{MI}(z)$ we have,

$$E(m^{\mathrm{MI}}(x, z) - m(x, z)|\vec{x}, \vec{z}) = \frac{1}{2} h_1^2 \mu_2 \left(m_1^{(2)}(x) + E(m_1^{(2)}(X)) \right) + \frac{1}{2} h_2^2 \mu_2 \left(m_2^{(2)}(z) + E(m_2^{(2)}(Z)) \right) + o_p(h_1^2 + h_2^2)$$

and

$$V(m^{\mathrm{MI}}(x,z)|\vec{x},\vec{z}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + \frac{1}{nh_2}\sigma^2 R_K f_Z(z)^{-1} + o_p((nh_1)^{-1} + (nh_2)^{-1}). \tag{28}$$

(ii) *The conditional MASE for the MIE is given by*,

$$\mathrm{MASE} = \frac{1}{4}h_1^4\mu_2^2\psi_{11} + \frac{1}{2}h_1^2h_2^2\mu_2^2\psi_{12} + \frac{1}{4}h_2^4\mu_2^2\psi_{22} + \frac{\sigma^2 R_K}{n}\left(\frac{1}{nh_1}\sum_{i=1}^{n} f_X(x_i)^{-1} + \frac{1}{nh_2}\sum_{i=1}^{n} f_Z(z_i)^{-1}\right) + o_p\left(h_1^4 + h_2^4 + (nh_1)^{-1} + (nh_2)^{-1}\right),$$

where $\psi_{d\delta} = 1/n\sum_{i=1}^{n}(m_d^{(2)}(x_i) + E(m_d^{(2)}(x_i)))(m_\delta^{(2)}(z_i) + E(m_\delta^{(2)}(z_i)))$ *for* $d, \delta = 1, 2$.

(iii) *The bandwidths that minimize the conditional AMASE, disregarding the term* $o_p(\cdot)$, *denoted by* $h_1^{\mathrm{MI}}, h_2^{\mathrm{MI}}$, *must satisfy*,

$$(h_1^{\mathrm{MI}})^5\mu_2^2\psi_{11} + (h_1^{\mathrm{MI}})^3(h_2^{\mathrm{MI}})^2\mu_2^2\psi_{12} = \frac{\sigma^2 R_K}{n}\left(\frac{1}{n}\sum_{i=1}^{n} f_X(x_i)^{-1}\right) \tag{29}$$

$$(h_2^{\mathrm{MI}})^5\mu_2^2\psi_{22} + (h_2^{\mathrm{MI}})^3(h_1^{\mathrm{MI}})^2\mu_2^2\psi_{12} = \frac{\sigma^2 R_K}{n}\left(\frac{1}{n}\sum_{i=1}^{n} f_Z(z_i)^{-1}\right). \tag{30}$$

As in the case of CBE, these optimal bandwidths depend on unknown quantities in equation (29) and (30) that have to be estimated to render them useful. Specifically, it is necessary to estimate $\psi_{d\delta}$, f_X and f_Z. Hence, the stochastic nature of the estimates of h_1^{MI} and h_2^{MI} and their dependence on the regressand produce the same non-linearities and difficulties that were alluded to when discussing CBE.

3.4 *Two stage estimator (2SE1 & 2SE2)*

In this section we obtain the conditional MASE for the 2SEs. The next two theorems provide a simplified version of the conditional bias, variance and MASE for 2SE1 and 2SE2, respectively. The more general results and their proofs are given in Theorem 1 and 2 in the Appendix. The proofs depend on Lemma 1 that establishes uniform convergence of certain bounded functions of X and Z. These results are then used to construct conditional MASE and to obtain optimal bandwidths for the two stage estimators. As in the case of CBE, SBE and MIE estimation, certain requirements on the speed of convergence to zero of the bandwidths are necessary.

THEOREM 3 *Suppose that Assumptions* 1–3 *hold*, $ng_1^3(ln(g_1))^{-1} \to \infty$ *and* $n(g_1g_2)^3(ln(g_1g_2))^{-1} \to \infty$. *Put* $\gamma_1(x) = \alpha + m_1(x)$ *and* $\gamma_2(z) = \alpha + m_2(z)$. *If* X *and* Z *are independent, and under the assumption that the bandwidths used in the first stage* $-g_1, g_2-$*are identical to those* $-h_1, h_2-$*used in the second stage of the estimation, we have that*:

(i) *The conditional bias for* $m_1^{2S1}(x)$ *is given by*,

$$E(m_1^{2S1}(x) - m_1(x)|\vec{x},\vec{z}) = \frac{1}{2}h_1^2\mu_2 m_1^{(2)}(x) - \frac{1}{2}h_2^2\mu_2 E(m_2^{(2)}(Z)) + \frac{1}{2}h_2^2\mu_2\int f_Z^{(2)}(v)\gamma_2(v)dv + o_p(h_1^2+h_2^2)$$

and the conditional variance is given by

$$V(m_1^{2S1}(x)|\vec{x},\vec{z}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + o_p((nh_1)^{-1}).$$

Mutatis mutandis, similar expressions for $m_2^{2S1}(z)$ *are obtained*;

(ii) *The conditional bias and variance for* $m^{2S1}(x,z)$ *are given by*,

$$E(m^{2S1}(x,z) - m(x,z)|\vec{x},\vec{z}) = \frac{1}{2}h_1^2\mu_2\left(m_1^{(2)}(x) - E(m_1^{(2)}(X)) + \int f_X^{(2)}(v)\gamma_1(v)dv\right) + \frac{1}{2}h_2^2\mu_2\left(m_2^{(2)}(z) - E(m_2^{(2)}(Z)) + \int f_z^{(2)}(v)\gamma_2(v)dv\right) + o_p(h_1^2+h_2^2)$$

and

$$V(m^{2S1}(x,z)|\vec{x},\vec{z}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + \frac{1}{nh_2}\sigma^2 R_K f_Z(z)^{-1} + o_p((nh_1)^{-1} + (nh_2)^{-1});$$

(iii) *The conditional MASE for the 2SE1 is given by*

$$\text{MASE} = \frac{1}{4}h_1^4\mu_2^2\phi_{11} + \frac{1}{4}h_2^4\mu_2^2\phi_{22} + \frac{1}{2}h_1^2h_2^2\mu_2\phi_{12} + \sigma^2 R_K n^{-1}\left(\frac{1}{nh_1}\sum_{i=1}^n f_X(x_i)^{-1} + \frac{1}{nh_2}\sum_{i=1}^n f_Z(z_i)^{-1}\right) + o_p\left(h_1^4 + h_2^4 + (nh_1)^{-1} + (nh_2)^{-1}\right),$$

where

$$\phi_{11} = \frac{1}{n}\sum_{i=1}^n\left(m_1^{(2)}(x_i) - E(m_1^{(2)}(X)) + \int f_X^{(2)}(v)\gamma_1(v)dv\right)^2$$

$$\phi_{22} = \frac{1}{n}\sum_{i=1}^n\left(m_2^{(2)}(z_i) - E(m_1^{(2)}(X)) + \int f_X^{(2)}(v)\gamma_2(v)dv\right)^2$$

$$\phi_{12} = \frac{1}{n}\sum_{i=1}^n\left(m_1^{(2)}(x_i) - E(m_1^{(2)}(X)) + \int f_X^{(2)}(v)\gamma_1(v)dv\right) \times\left(m_2^{(2)}(z_i) - E(m_2^{(2)}(Z)) + \int f_Z^{(2)}(v)\gamma_2(v)dv\right).$$

THEOREM 4 *Suppose that Assumptions* 1–3 *hold, that* $ng_1^3(ln(g_1))^{-1}\to\infty$ *and* $n(g_1g_2)^{2p+1}(ln(g_1g_2))^{-1}\to\infty$. *If X and Z are independent, and under the assumption that*

the bandwidths used in the first stage $-g_1, g_2-$ *are identical to those* $-h_1, h_2-$ *used in the second stage of the estimation, we have that*:

(i) *The conditional bias for* $m_1^{2S2}(x)$ *is given by*,

$$E(m_1^{2S2}(x) - m_1(x)|\vec{x}, \vec{z}) = \frac{1}{2}h_1^2\mu_2 m_1^{(2)}(x) - \frac{1}{2}h_2^2\mu_2 E(m_2^{(2)}(Z)) + \frac{1}{2}h_2^2\mu_2 \int f_Z^{(2)}(v)m_2(v)dv + o_p(h_1^2 + h_2^2)$$

and the conditional variance is given by

$$V(m_1^{2S2}(x)|\vec{x}, \vec{z}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + o_p((nh_1)^{-1}).$$

Mutatis mutandis, similar expressions for $m_2^{2S2}(z)$ *are obtained*;

(ii) *The conditional bias and variance for* $m^{2S2}(x, z)$ *are given by*,

$$\begin{aligned} &E(m^{2S2}(x, z) - m(x, z)|\vec{x}, \vec{z}) \\ &= \frac{1}{2}h_1^2\mu_2 \left(m_1^{(2)}(x) - E(m_1^{(2)}(X)) + \int f_X^{(2)}(v)m_1(v)dv\right) \\ &\quad + \frac{1}{2}h_2^2\mu_2 \left(m_2^{(2)}(z) - E(m_2^{(2)}(Z)) + \int f_Z^{(2)}(v)m_2(v)dv\right) \\ &\quad + o_p(h_1^2 + h_2^2) \end{aligned}$$

and

$$V(m^{2S2}(x, z)|\vec{x}, \vec{z}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + \frac{1}{nh_2}\sigma^2 R_K f_Z(z)^{-1} + o_p\left((nh_1)^{-1} + (nh_2)^{-1}\right);$$

(iii) *The conditional* MASE *for* 2SE2 *is given by*

$$\begin{aligned} \text{MASE} &= \frac{1}{4}h_1^4\mu_2^2\chi_{11} + \frac{1}{4}h_2^4\mu_2^2\chi_{22} + \frac{1}{2}h_1^2h_2^2\mu_2\chi_{12} \\ &\quad + \sigma^2 R_K n^{-1}\left(\frac{1}{nh_1}\sum_{i=1}^n f_X(x_i)^{-1} + \frac{1}{nh_2}\sum_{i=1}^n \frac{1}{f_Z}(z_i)^{-1}\right) \\ &\quad + o_p\left(h_1^4 + h_2^4 + (nh_1)^{-1} + (nh_2)^{-1}\right), \end{aligned}$$

where

$$\chi_{11} = \frac{1}{n}\sum_{i=1}^n \left(m_1^{(2)}(x_i) - E(m_1^{(2)}(X)) + \int f_X^{(2)}(v)m_x(v)dv\right)^2$$

$$\chi_{22} = \frac{1}{n}\sum_{i=1}^n \left(m_2^{(2)}(z_i) - E(m_2^{(2)}(Z)) + \int f_Z^{(2)}(v)m_2(v)dv\right)^2$$

$$\chi_{12} = \frac{1}{n}\sum_{i=1}^n \left(m_1^{(2)}(x_i) - E(m_1^{(2)}(X)) + \int f_X^{(2)}(v)m_1(v)dv\right) \times \left(m_2^{(2)}(z_i) - E(m_2^{(2)}(Z)) + \int f_Z^{(2)}(v)m_2(v)dv\right).$$

A number of remarks are in order regarding Theorems 3 and 4.

1. Although the conditional bias of all estimators under study are of similar order, the 2SE conditional bias in direction m_d ($d = 1, 2$) under independence of X and Z have two extra terms of order $O(h_\delta^2)$ for $\delta \neq d$, if compared with the bias of the univariate local linear estimator, i.e. $(1/2)h_d^2\mu_2 m_d^{(2)}(x)$. The impact of these terms on the conditional bias of the estimators is unclear, since their sign and magnitude depends on the data generating process. Likewise, it is not possible to ascertain the relative magnitude of these terms and those of similar order which appear in the conditional bias expression for CBE, SBE and MIE. In the case where X and Z are not independent (Theorems 1 and 2 in the Appendix), comparisons are made even more difficult by the presence of an additional term of order $O(h_\delta^2)$. Kim *et al.* [10] are able to eliminate these extra terms with undersmoothing in the first stage estimation, i.e. letting g_1, g_2 degenerate at a faster speed relative to h_1, h_2 (see Theorems 1 and 2 in the Appendix). Note that this oracle property of the estimation procedure can be obtained in the context of backfitting by choosing bandwidths that oversmooth at the last step of the backfitting algorithm.
2. When X and Z are independent, both CBE and SBE with local linear smoother produce conditional bias and variance are given by

$$E(m_1^{\mathrm{CB}}(x) - m_1(x)|\vec{x}, \vec{z}) = \frac{1}{2}h_1^2\mu_2(m_1^{(2)}(x) - E(m_1^{(2)}(X))) + o_p(h_1^2 + h_2^2) \tag{31}$$

and

$$V(m_1^{\mathrm{CB}}(x)|\vec{x}, \vec{z}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + o_p((nh_1)^{-1}). \tag{32}$$

Hence, for both m_1^{CB} and m_1^{SB} the biases depend only on the curvature of m_1, weighted by the density. On the other hand, the biases of the m_1^{2S1} and m_1^{2S2}, as well as that of the m_1^{MI} depend on the curvature of the other component function, even when X and Z are independent. As pointed out by Opsomer and Ruppert [14, p. 198], it seems natural to expect estimators for an additive model, where the regressors are independent to have asymptotic bias for one of the component functions to be independent of the other. Whether this theoretical advantage of CBE and SBE translates into better estimation accuracy in finite sample is a question we want to answer with our simulations.
3. The 2SEs have conditional variances that are of the same order and identical (of order $O((nh_d)^{-1})$) to that of CBE, SBE and MIE and a univariate local linear estimator.

Given the AMASE results from Theorems 3 and 4 the optimal bandwidths that minimize the conditional AMASE for 2SE1 and 2SE2 must satisfy the following two sets of equations:

$$(h_1^{\mathrm{2S1}})^5\mu_2^2\phi_{11} + (h_2^{\mathrm{2S1}})^2(h_1^{\mathrm{2S1}})^3\mu_2\phi_{12} = \sigma^2 R_K\left(\frac{1}{n}\sum_{i=1}^{n} f_X(x_i)^{-1}\right) \tag{33}$$

$$(h_2^{\mathrm{2S1}})^5\mu_2^2\phi_{22} + (h_1^{\mathrm{2S1}})^2(h_2^{\mathrm{2S1}})^3\mu_2\phi_{12} = \sigma^2 R_K\left(\frac{1}{n}\sum_{i=1}^{n} f_Z(z_i)^{-1}\right) \tag{34}$$

and

$$(h_1^{\mathrm{2S2}})^5\mu_2^2\chi_{11} + (h_2^{\mathrm{2S2}})^2(h_1^{\mathrm{2S2}})^3\mu_2\chi_{12} = \sigma^2 R_K\left(\frac{1}{n}\sum_{i=1}^{n} f_X(x_i)^{-1}\right) \tag{35}$$

$$(h_2^{\mathrm{2S2}})^5\mu_2^2\chi_{22} + (h_1^{\mathrm{2S2}})^2(h_2^{\mathrm{2S1}})^3\mu_2\chi_{12} = \sigma^2 R_K\left(\frac{1}{n}\sum_{i=1}^{n} f_Z(z_i)^{-1}\right). \tag{36}$$

3.5 *Data driven bandwidth selection*

The choice of data driven bandwidth for the Monte Carlo experiments was based on two considerations. First, we want to have a bandwidth selection rule that interferes minimally with the performance of the estimators. By this, we mean a bandwidth estimator that transfers minimal noise from the estimation of f_X, f_Z, $\theta_{d\delta}$, $\psi_{d,\delta}$, $\phi_{d,\delta}$ and $\chi_{d\delta}$ for $d, \delta = 1, 2$, $\int f_X^{(2)}(v)\gamma_1(v)dv$, $\int f_Z^{(2)}(v)\gamma_2(v)dv$, $\int f_X^{(2)}(v)m_1(v)dv$ and $\int f_Z^{(2)}(v)m_2(v)dv$ to the estimation of m_1 and m_2. This provides an ideal setting to compare the performance of the estimators, as any differences can be attributed to the structure of the estimators themselves and not to the estimation of the unknowns in the expressions for the optimal bandwidths. Second, we want to compare the performance of the estimators when using bandwidth selection rules proposed in the previous section and those already proposed in the literature.

3.5.1 True bandwidths. Elimination of the noise that is generated by the estimation of the parameters in the expression for optimal bandwidths – equation (21) for CBE, equation (25) for SBE, equations (29) and (30) for MIE, equations (33) and (34) for 2SE1 and equations (35) and (36) for 2SE2 – can be accomplished in a Monte Carlo study setting since the true values of these unknowns can be obtained directly from the specification of the DGP. Hence, the first set of bandwidths that we use are based on complete information about the normally unknown functionals that appear on the specification of the optimal bandwidths.[†] In this case the only difficulty involves the evaluation of the integrals that define the expectations that appears in $\psi_{d,\delta}$, $\phi_{d,\delta}$ and $\chi_{d\delta}$ for $d, \delta = 1, 2$ and $\int f_X^{(2)}(v)\gamma_1(v)dv$, $\int f_Z^{(2)}(v)\gamma_2(v)dv$, $\int f_X^{(2)}(v)m_1(v)dv$ and $\int f_Z^{(2)}(v)m_2(v)dv$. These expectations can be difficult to compute, depending on the nature of m_d. In our study, all integrals were calculated numerically using the Gauss-Legendre quadrature method.

3.5.2 Estimated bandwidths. The estimated bandwidths for the CBE were obtained using the procedure proposed by Opsomer and Ruppert [21] to estimate θ_{11}, θ_{22} and σ^2. We assumed that f_X and f_Z are uniform densities over a compact support and the terms $n^{-1}\sum_{i=1}^n f_X(x_i)^{-1}$ and $n^{-1}\sum_{i=1}^n f_Z(z_i)^{-1}$ are estimated by $\max_i(x_i) - \min_i(x_i)$ and $\max_i(z_i) - \min_i(z_i)$, respectively, where $\max_i(x_i)$ and $\min_i(x_i)$ are the maximum and minimum sample values in $\vec{x}$.

Since the SBE share the same analytical solutions of optimal bandwidth with the CBE, the same bandwidths are used for SBE as those for CBE.

Two different estimated bandwidths are considered for MIE. The first were proposed by Linton and Nielsen [8] and take the form,

$$\ddot{h}_1 = \left(\frac{\ddot{\sigma}^2 R_K(\max_i(x_i) - \min_i(x_i))}{n\mu_2^2(\hat{\beta}_1 + \hat{\beta}_2)^2}\right)^{1/5} \text{ and } \ddot{h}_2 = \left(\frac{\ddot{\sigma}^2 R_K(\max_i(z_i) - \min_i(z_i))}{n\mu_2^2(\hat{\beta}_1 + \hat{\beta}_2)^2}\right)^{1/5},$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are OLS estimates of the parameters associated with $x_i^2/2$ and $z_i^2/2$ of a regression of y_i on a constant, $x_i^2/2$, $z_i^2/2$, x_i, z_i and $x_i z_i$. $\ddot{\sigma}^2$ is the typical estimate for the variance in a classical linear regression model. The second procedure involves the numerical solution of equations (29) and (30).[‡] Once again, we assumed that f_X and f_Z are uniform densities over a compact support and estimated their inverses by $\max_i(x_i) - \min_i(x_i)$ and

[†]Note that the true optimal bandwidths are different across samples since MASE is evaluated at sample points.

[‡]Numerical solutions for the non-linear systems defined by equations (29) and (30), (33) and (34) as well as (35) and (36) are obtained using a quasi-Newton method (step-by-step line search) with an analytical Jacobian. See [26].

$\max_i(z_i) - \min_i(z_i)$. $\psi_{d\delta}$ were estimated using the same procedure for the estimation of $\theta_{d\delta}$ with the necessary sign changes inside the summations.

We also consider two different estimated bandwidths for 2SE1. The first is the simple rule of thumb proposed in Kim *et al.* [10] in which h_1 and h_2 are selected as follows,

$$h_1^K = n^{-1/5}\frac{1}{2}\hat{\sigma}_X \text{ and } h_2^K = n^{-1/5}\frac{1}{2}\hat{\sigma}_Z,$$

where $\hat{\sigma}_X = \sqrt{1/n\sum_{i=1}^n (x_i - \bar{x})^2}$ and $\hat{\sigma}_Z = \sqrt{1/n\sum_{i=1}^n (z_i - \bar{z})^2}$. These estimated bandwidths produce an estimator that we label $2SE1_K$ in the tables describing the simulation results. The second bandwidth selection procedure we consider for the 2SE1 is based on the numerical solution of equations (33) and (34). To this end the unknown quantities $\phi_{d\delta}$ must be estimated together with f_X, f_Z and σ^2. The estimation of $\phi_{d\delta}$ depends on the estimation of two parts $-m_d^{(2)}(x_i) - E(m_d^{(2)})$ and $\int f_X^{(2)}(v)\gamma_1(v)dv$ (or $\int f_Z^{(2)}(v)\gamma_2(v)dv$). The first term is estimated as in the case of CBE, the second term can be interpreted as $E\gamma_1(v)\frac{f_X^{(2)}(v)}{f_X(v)}$, which is estimated by $n^{-1}\sum_{i=1}^n \hat{\gamma}_1(x_i)\frac{\hat{f}_X^{(2)}(x_i)}{\hat{f}_X(x_i)}$, where $\hat{\gamma}_1$ comes from a preliminary CBE and $\hat{f}$ is estimated by a kernel-density estimator with a Silverman's rule-of-thumb bandwidth. σ^2 is estimated as in the case for CBE.

Finally, the estimated bandwidths used to produce the 2SE2 are the result of the numerical solution for equations (35) and (36). As in the case for 2SE1, the unknowns that appear in the above mentioned equations, i.e. $\chi_{d\delta}$ must be estimated together with f_X, f_Z and σ^2. We follow the same estimation procedure described above for 2SE1 with the exception that $\hat{\gamma}_d$ is substituted by $\hat{m}_d$.

4. The data generating process (DGP)

The data used in this study is generated from a fully specified bivariate additive model. First, the independent variables $\{x_i\}_{i=1}^n$ and $\{z_i\}_{i=1}^n$ are generated from a bivariate normal distribution with joint density given by

$$\begin{pmatrix} x_i \\ z_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1/9 & c/9 \\ c/9 & 1/9 \end{pmatrix}\right),$$

where $c = 0, 0.25, 0.75$, gives the desired correlation between the random variables. We allow for different correlation values because one of our objectives is to evaluate how regressor dependency impacts the performance of the estimators. One of the assumptions required to obtain expressions for the conditional mean and variance of the estimators under study is that f_{XZ} have compact support. To satisfy this assumption we discard every generated data point that is outside [0, 1] and resample until all generated pseudorandom numbers are within this interval. The regression error ϵ_i is generated independently as a standard normal, and the regressands $\{y_i\}_{i=1}^n$ are obtained in accordance with three models:

$$\text{Model 1}: \quad y_i = m_1(x_i) + m_2(z_i) + \epsilon_i \tag{37}$$

$$\text{Model 2}: \quad y_i = m_1(x_i) + m_3(z_i) + \epsilon_i \tag{38}$$

$$\text{Model 3}: \quad y_i = m_2(x_i) + m_3(z_i) + \epsilon_i, \tag{39}$$

where $m_1(x) = 1 - 6x + 36x^2 - 53x^3 + 22x^5$, $m_2(x) = \sin(5\pi x)$ and $m_3(x) = \exp(3x)$. The fact that these functions have very different curvatures makes the use of a common

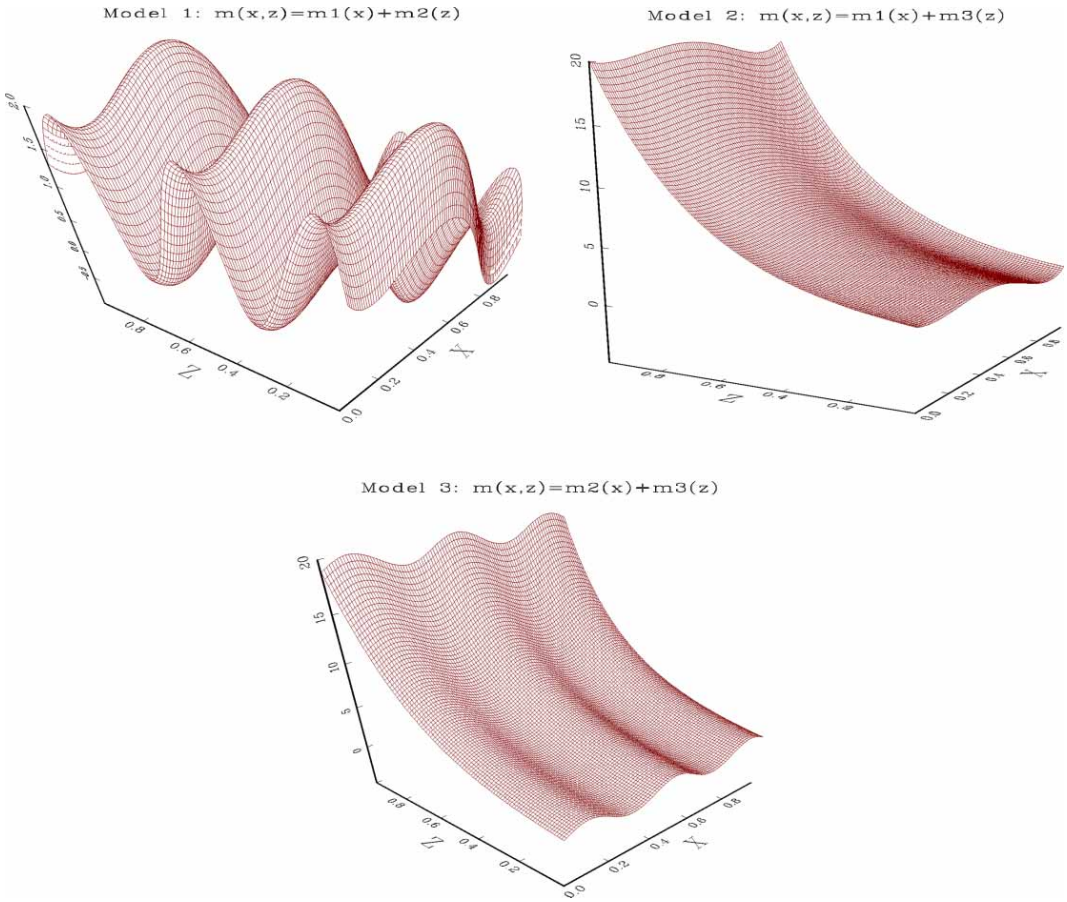

Figure 1. 3D Plot of true models.

bandwidth inadequate. Figure 1 provides graphs of the three models over the relevant range of X and Z.

We generate samples of size $n = 200, 350, 500$ and for all sample sizes we generated 500 replications. Samples of relatively small size are used for two reasons. First, the small sample sizes reduce the computational burden in the Monte Carlo. Second, we wanted to evaluate the estimators performance under fairly undesirable conditions.

5. Estimation results

A Gaussian kernel is used to construct the estimators. Computer codes for the estimation were written in GAUSS 5.0 and estimation was done on a PC running on a 3.1 Ghz Intel® Pentium® IV processor. Table 1 provides the computational time (in seconds) for all estimators considered for an experiment using model 1.† The columns listed under h^{CB}, h^{SB}, h^{MI} and h^{2S} represent the elapsed time to calculate the estimated bandwidths described in section 3.5.2, and the columns under $\hat{m}(x, z)$ represent the elapsed time to calculate the estimators once bandwidths are available. The time to compute the bandwidths for MIE, 2SE1 and 2SE2 is larger than that necessary to obtain bandwidths for CBE and SBE. This comes as no surprise

†There is a small variation in computing time for different models, but none of the conclusions described in the text are changed.

Table 1. Computation time (seconds) by estimator.

	CBE		SBE		MIE		2SE1/2SE2	
c	h^{CB}	$\hat{m}(x,z)$	h^{SB}	$\hat{m}(x,z)$	h^{MI}	$\hat{m}(x,z)$	h^{2S}	$\hat{m}(x,z)$
				$n = 200$				
0	2.641	0.953	2.641	3.547	2.750	68.687	2.922	0.672
0.25	2.656	0.953	2.656	3.687	2.735	68.500	2.734	0.672
0.75	2.640	0.953	2.640	3.547	2.703	68.310	2.719	0.656
				$n = 350$				
0	17.562	7.938	17.562	11.453	17.735	871.094	18.469	5.890
0.25	17.515	6.718	17.515	11.313	17.687	868.344	18.687	4.813
0.75	17.562	9.125	17.562	11.433	17.703	872.781	18.344	4.781
				$n = 500$				
0	52.375	19.734	52.375	25.984	52.969	3630.735	54.391	14.188
0.25	52.454	19.625	52.454	26.016	52.781	3616.375	54.078	14.110
0.75	52.515	19.625	52.515	25.859	52.844	3641.562	54.078	14.078

as the former require the numerical solution of a non-linear set of equations, whereas the latter are based on a closed form expression. However, the extra computational burden is very moderate, and in no case greater than 1.5 seconds.

Computational time does vary significantly across estimators. MIE is, by far, the most demanding with regards to computing time of all estimators under study, due to the fact that it evaluates the model at n^2 points, whereas the others require evaluation only at n points. Since MIE underperformed compared with all other estimators in a preliminary full set of simulations, particularly in models where the correlation among independent variables are high ($c = 0.75$), we did not include MIE in the reported tables. Once the bandwidths are selected, the 2SE1 and 2SE2 are faster to implement than all other estimators. Although both CBE and SBE are based on iterative procedures, in our simulation, convergence occurs in just a few steps, even in the case where X and Z are highly correlated. SBE takes more time to compute than the CBE due to the extra integral term in updating equation (19). Finally, we observe the expected significant increase in computational time for all estimators, as the sample size n increases.

The analysis of the experimental results focuses on the average squared error (ASE) of the estimators, their average bias (AB), average variance and on the estimation of the bandwidths across all replications. Let $\bar{y}^r$, $\hat{m}_1^r(x_i)$ and $\hat{m}_2^r(z_i)$ represent estimates for replication $r = 1, \ldots, 500$ based on CBE, SBE, 2SE1 or 2SE2 and define the ASE^r and the AB^r for $\hat{m}_1^r(x_i)$ in the r^{th} replication as,

$$\mathrm{ASE}^r(\hat{m}_1) = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}_1^r(x_i) - (m_1(x_i) - E(m_1(X)))^2,$$

$$\mathrm{AB}^r(\hat{m}_1) = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}_1^r(x_i) - (m_1(x_i) - E(m_1(X)))$$

and similarly for $\hat{m}_2^r(z_i)$.[†] For $\hat{m}^r(x_i, z_i) = \bar{y}^r + \hat{m}_1^r(x_i) + \hat{m}_2^r(z_i)$, we put

$$\mathrm{ASE}^r(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}^r(x_i, z_i))^2, \text{ and } \mathrm{AB}^r(\hat{m}_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}^r(x_i, z_i)).$$

[†]We note that the functions m_d used in the DGP do not satisfy $E(m_d) = 0$ for $d = 1, 2, 3$. Therefore, the estimators considered in the study are estimating $m_d - E(m_d)$. As such, the definition for $\mathrm{ASE}^r(\hat{m}_d)$ and $\mathrm{AB}^r(\hat{m}_d)$ incorporates the constants $E(m_d)$.

Table 2. Average squared error using true bandwidth.

	$n = 200$				$n = 350$				$n = 500$			
c	CBE	SBE	2SE1	2SE2	CBE	SBE	2SE1	2SE2	CBE	SBE	2SE1	2SE2
Model 1												
0	0.084	0.088	0.120	0.089	0.054	0.055	0.071	0.056	0.039	0.041	0.051	0.041
0.25	0.090	0.094	0.125	0.095	0.056	0.059	0.075	0.059	0.042	0.045	0.056	0.045
0.75	0.083	0.095	0.150	0.106	0.051	0.063	0.098	0.068	0.040	0.051	0.076	0.055
Model 2												
0	0.061	0.200	2.550	0.142	0.039	0.143	1.441	0.090	0.028	0.104	0.933	0.055
0.25	0.055	0.251	2.749	0.165	0.036	0.201	1.561	0.093	0.028	0.172	1.094	0.073
0.75	0.062	1.007	8.877	0.783	0.038	0.930	6.289	0.580	0.028	0.905	5.037	0.463
Model 3												
0	0.079	0.314	4.671	0.235	0.051	0.212	2.721	0.142	0.039	0.170	1.809	0.101
0.25	0.086	0.366	5.095	0.310	0.055	0.264	2.812	0.161	0.040	0.230	2.045	0.128
0.75	0.079	1.060	11.111	0.933	0.052	0.985	8.330	0.751	0.040	0.955	6.925	0.648

The numbers reported in the tables are average squared errors and bias across all replications. Since some preliminary finite sample experimental evidence on the performance of these estimators are already available ([21,17,19]), we are primarily interested in the relative performance of the estimators.

Tables 2 and 3 provide ASE across experiments using true and estimated bandwidths, respectively, for all estimators, for different sample sizes and for various correlation levels. Some general regularities are promptly identified. As expected, increases in sample size reduce ASE for all estimators and across all correlation levels with true and estimated bandwidths.

The effects of increased correlation on the ASE of the estimators are quite different. For the classic backfitting estimator ASE is similar across correlation levels for each sample size, but they do differ across models. In some cases the results even show mild decrease in ASE as correlation increases. These regularities are true when true and estimated bandwidths are used. Results are quite different for SBE, MIE and 2SEs. All estimators seem to be impacted by increased correlation, with ASE increasing as c grows. This is true when true or estimated bandwidths are used. It is apparent, however, that ASE is not significantly affected by mild correlation among the regressors. The increase is significant, however, when the correlation moves from low levels 0.25 to 0.75. For SBE and 2SEs, the impact of increases in c on their ASE do vary across models. In model 1, the increases in c only mildly increases the ASE of SBE, 2SE1 and 2SE2. However, the same increase in c causes much more trouble for SBE and 2SEs. This seems to suggest that it is the combination of correlation and unbalanced scale across component functions that causes the increases in ASE for SBE and 2SEs. Increases in sample size do not seem to reduce the disparity in ASE across models.[†]

One should also observe that, as expected, across all experiments and estimators the reported ASEs increase from table 2 to table3, confirming that in finite samples the noise introduced by estimated bandwidths impacts the performance of the estimators. Also as expected, increased sample size dampens this impact.

The most noticeable result from tables 2 and 3 is that, as measured by ASE, the CBE is superior to all estimators across all correlation levels, sample sizes and models. The second

[†]In the preliminary simulation, the MIE seems to be the most sensitive of the estimators with respect to increases in c. This coincides with the fact that its asymptotic variance increase significantly with c. Intuitively, this loss of accuracy is caused by the fact that the MIE needs to estimate the function at many out-of-sample points. When the correlation is high, the values of the function at those points are very hard to capture due to their distance from the observed values of the function.

Table 3. Average squared error using estimated bandwidth.

	$n = 200$					$n = 350$					$n = 500$				
c	CBE	SBE	2SE1	$2SE1_K$	2SE2	CBE	SBE	2SE1	$2SE1_K$	2SE2	CBE	SBE	2SE1	$2SE1_K$	2SE2
Model 1															
0	0.097	0.100	0.139	0.150	0.102	0.063	0.065	0.080	0.088	0.066	0.048	0.049	0.059	0.063	0.049
0.25	0.099	0.102	0.147	0.156	0.104	0.065	0.067	0.085	0.094	0.067	0.047	0.050	0.061	0.069	0.050
0.75	0.095	0.106	0.174	0.195	0.119	0.059	0.070	0.106	0.124	0.075	0.045	0.057	0.080	0.095	0.060
Model 2															
0	0.073	0.251	4.572	7.035	0.180	0.045	0.170	2.012	4.05	0.105	0.031	0.121	1.199	2.720	0.062
0.25	0.069	0.311	5.040	7.284	0.216	0.042	0.230	2.339	4.32	0.112	0.031	0.192	1.458	2.999	0.085
0.75	0.077	1.043	12.970	15.677	0.915	0.044	0.951	8.131	11.099	0.648	0.032	0.922	6.143	8.950	0.505
Model 3															
0	0.092	0.329	6.335	6.247	0.250	0.061	0.211	2.905	3.673	0.142	0.046	0.168	1.808	2.465	0.101
0.25	0.101	0.408	6.694	6.819	0.328	0.062	0.263	2.967	3.853	0.158	0.047	0.228	2.033	2.772	0.124
0.75	0.097	1.114	14.472	13.800	1.025	0.063	1.009	9.731	10.274	0.778	0.049	0.967	7.683	8.507	0.656

Table 4. Average bias using true bandwidth.

	$n = 200$				$n = 350$				$n = 500$			
c	CBE	SBE	2SE1	2SE2	CBE	SBE	2SE1	2SE2	CBE	SBE	2SE1	2SE2
Model 1												
0	0.010	0.010	0.164	−0.008	0.005	0.006	0.119	−0.010	0.005	0.006	0.100	−0.006
0.25	−0.000	−0.000	0.168	−0.012	0.003	0.004	0.126	−0.006	0.004	0.005	0.108	−0.001
0.75	−0.024	−0.025	0.192	−0.085	−0.016	−0.016	0.163	−0.064	−0.012	−0.012	0.146	−0.055
Model 2												
0	−0.028	−0.029	1.510	0.149	−0.007	−0.008	1.149	0.151	−0.010	−0.011	0.923	0.109
0.25	0.018	0.018	1.601	0.196	0.020	0.020	1.222	0.172	0.030	0.031	1.030	0.168
0.75	0.044	0.045	2.824	0.396	0.052	0.053	2.394	0.360	0.060	0.060	2.160	0.328
Model 3												
0	−0.022	−0.023	2.082	0.240	−0.014	−0.014	1.595	0.201	−0.006	−0.007	1.302	0.173
0.25	−0.020	0.047	2.160	0.277	−0.011	−0.012	1.609	0.204	0.007	0.007	1.391	0.219
0.75	0.152	0.152	3.302	0.603	0.160	0.161	2.893	0.591	0.136	0.137	2.633	0.563

best is the 2SE2 that we propose, followed in order by SBE, and 2SE1 estimation. The SBE is fairly accurate in model 1 but surprisingly loses accuracy in models 2 and 3. We suspect the the reason is that approximation of the integral is less accurate in models 2 and 3, since the range of $m_2(\cdot)$ is much wider in the latter two cases. An increase of the number of grid points on which the integral is approximated should be able to improve the overall performance of SBE, but the computation time will certainly increase. From table 3, we observe that our proposed bandwidth estimation procedure for 2SE1 outperforms the bandwidth selection procedure proposed by Kim *et al.* [10] (2SE1) across all experiments.

Tables 4 and 5 provide average bias for all estimators across experiments using true and estimated bandwidths, respectively. As in the case of ASE, some general regularities can be noticed. With almost no exception (these involve 2SE2) the CBE and SBE show smallest biases across estimators followed by 2SE2, when bandwidths are estimated. As in the case of ASE, our proposed bandwidth selection methods reduce the bias of the 2SE1.

Tables 4 and 5 also reveal that the average bias increases with c across all experiments and estimators. Again, this is particularly noticeable when $c = 0.75$. The impact of sample size on bias when true bandwidths are used is different across estimators. For CBE and 2SE1 no discernible pattern is observed, but for 2SE2 the bias falls with sample size. When bandwidths are estimated the bias falls for all estimators and models as n increases, except for CBE. Combining the results from tables 2 and 3 with tables 4 and 5, we can conclude that the variance of the estimators decreases with sample size across all experiments for all estimators.[†]

Tables 6 and 7 provide ASEs for the estimation of m_d for $d = 1, 2$ for all correlation levels and sample sizes using true and estimated bandwidths. The general regularities observed for ASE in tables 2 and 3 seem to apply in each regression direction. In addition, these results suggest that the ASE per direction is impacted significantly by the curvature of the functions being estimated and that the curvature of one function impacts the ASE properties of the other regression direction for all estimators.

Tables 8 and 9 provide the average true and estimated bandwidths across experiments for different sample sizes, correlations and models. Tables 8 and 9 reveal that true and estimated bandwidths for all estimators are quite insensitive to correlation levels. They do, however, noticeably change across models. Although expected gains from increased sample size do not appear dramatic for the sample sizes considered in this experiment, our

[†]Note that for any estimator considered the variance for the r^{th} replication can be obtained by $\text{ASE}^r - \text{AB}^r$.

Table 5. Average bias using estimated bandwidth.

	$n = 200$					$n = 350$					$n = 500$				
c	CBE	SBE	2SE1	$2SE1_K$	2SE2	CBE	SBE	2SE1	$2SE1_K$	2SE2	CBE	SBE	2SE1	$2SE1_K$	2SE2
Model 1															
0	0.010	0.010	0.176	0.213	−0.003	0.005	0.006	0.118	0.155	−0.007	0.005	0.006	0.097	0.128	−0.004
0.25	−0.000	−0.000	0.192	0.215	−0.007	0.003	0.004	0.126	0.161	−0.004	0.004	0.005	0.105	0.138	0.000
0.75	−0.024	−0.025	0.197	0.247	−0.084	−0.016	−0.016	0.162	0.209	−0.061	−0.012	−0.012	0.140	0.185	−0.053
Model 2															
0	−0.028	−0.029	1.976	2.573	0.150	−0.007	−0.008	1.329	1.968	0.155	−0.010	−0.011	1.022	1.612	0.108
0.25	0.018	0.018	2.136	2.652	0.192	0.020	0.020	1.469	2.055	0.170	0.030	0.031	1.174	1.720	0.166
0.75	0.044	0.045	3.411	3.823	0.348	0.052	0.053	2.724	3.230	0.333	0.060	0.060	2.383	2.912	0.300
Model 3															
0	−0.022	−0.023	2.377	2.427	0.221	−0.014	−0.014	1.613	1.866	0.184	−0.006	−0.007	1.270	1.532	0.161
0.25	−0.020	−0.021	2.424	2.524	0.264	−0.011	−0.012	1.617	1.900	0.184	0.007	0.007	1.359	1.630	0.201
0.75	0.152	0.152	3.743	3.698	0.556	0.160	0.161	3.106	3.222	0.564	0.136	0.137	2.760	2.926	0.538

Table 6. ASE regression direction using true bandwidth.

	CBE		SBE		2SE1		2SE2	
c	m_1	m_2	m_1	m_2	m_1	m_2	m_1	m_2
Model 1								
				$n = 200$				
0	0.030	0.056	0.029	0.059	0.038	0.066	0.032	0.060
0.25	0.034	0.060	0.035	0.061	0.043	0.069	0.037	0.063
0.75	0.033	0.060	0.037	0.072	0.051	0.087	0.046	0.078
				$n = 350$				
0	0.018	0.036	0.019	0.038	0.023	0.041	0.019	0.038
0.25	0.021	0.036	0.0224	0.038	0.026	0.042	0.022	0.038
0.75	0.019	0.037	0.247	0.051	0.033	0.058	0.029	0.051
				$n = 500$				
0	0.012	0.027	0.013	0.028	0.016	0.03	0.013	0.028
0.25	0.015	0.028	0.016	0.030	0.018	0.03	0.015	0.029
0.75	0.017	0.029	0.020	0.043	0.028	0.04	0.025	0.040
	CBE		SBE		2SE1		2SE2	
c	m_1	m_3	m_1	m_3	m_1	m_3	m_1	m_3
Model 2								
				$n = 200$				
0	0.032	0.107	0.180	0.116	0.650	0.725	0.073	0.142
0.25	0.027	0.103	0.244	0.111	0.700	0.867	0.090	0.150
0.75	0.036	0.114	1.470	0.302	2.984	2.668	0.915	0.486
				$n = 350$				
0	0.020	0.071	0.126	0.0	0.363	0.436	0.040	0.092
0.25	0.018	0.063	0.193	0.068	0.393	0.499	0.048	0.087
0.75	0.021	0.066	1.332	0.276	2.106	1.972	0.665	0.360
				$n = 500$				
0	0.014	0.052	0.092	0.054	0.231	0.285	0.025	0.062
0.25	0.013	0.042	0.166	0.046	0.272	0.364	0.036	0.063
0.75	0.016	0.048	1.291	0.254	1.670	1.621	0.534	0.300
	CBE		SBE		2SE1		2SE2	
c	m_2	m_3	m_2	m_3	m_2	m_3	m_2	m_3
Model 3								
				$n = 200$				
0	0.052	0.134	0.309	0.149	1.205	1.300	0.144	0.191
0.25	0.055	0.121	0.361	0.132	1.334	1.377	0.191	0.188
0.75	0.057	0.141	1.416	0.302	3.683	3.466	1.082	0.510
				$n = 350$				
0	0.033	0.071	0.203	0.080	0.692	0.758	0.077	0.103
0.25	0.034	0.084	0.255	0.091	0.711	0.786	0.091	0.120
0.75	0.036	0.091	1.321	0.241	2.785	2.683	0.842	0.424
				$n = 500$				
0	0.025	0.055	0.1613	0.060	0.456	0.524	0.054	0.079
0.25	0.026	0.056	0.222	0.061	0.515	0.600	0.066	0.089
0.75	0.027	0.062	1.279	0.204	2.326	2.213	0.702	0.352

proposed bandwidth estimation procedure produces bandwidths that are much closer to the true bandwidths than those produced by the procedures suggested by Linton and Nielsen [8] and Kim *et al.* [10] for MIE and 2SE1 estimation, respectively. In addition, the true bandwidths are identical (up to two decimal points) for all estimators, across all models and experiments. All estimated bandwidths for models 1 and 2 undersmooth if compared with the true bandwidths reported in table 6. For model 3 bandwidths oversmooth if compared

Table 7. ASE by regression direction using estimated bandwidth.

	CBE		SBE		2SE1		$2SE1_K$		2SE2	
c	m_1	m_2	m_1	m_2	m_1	m_2	m_1	m_2	m_1	m_2
Model 1										
					$n = 200$					
0	0.036	0.064	0.035	0.066	0.046	0.075	0.053	0.073	0.038	0.068
0.25	0.039	0.064	0.040	0.065	0.052	0.076	0.057	0.077	0.042	0.067
0.75	0.039	0.068	0.041	0.079	0.056	0.095	0.064	0.098	0.051	0.086
					$n = 350$					
0	0.020	0.044	0.021	0.045	0.025	0.049	0.031	0.045	0.021	0.046
0.25	0.024	0.043	0.025	0.044	0.029	0.049	0.034	0.047	0.025	0.044
0.75	0.021	0.043	0.026	0.056	0.034	0.064	0.041	0.067	0.031	0.056
					$n = 500$					
0	0.014	0.034	0.013	0.028	0.018	0.037	0.022	0.033	0.015	0.035
0.25	0.016	0.032	0.016	0.030	0.020	0.036	0.024	0.036	0.017	0.033
0.75	0.018	0.033	0.022	0.043	0.028	0.050	0.034	0.053	0.026	0.044
	CBE		SBE		2SE1		$2SE1_K$		2SE2	
c	m_1	m_3	m_1	m_3	m_1	m_3	m_1	m_3	m_1	m_3
Model 2										
					$n = 200$					
0	0.036	0.115	0.231	0.131	1.174	1.179	1.813	1.805	0.102	0.156
0.25	0.033	0.112	0.304	0.130	1.300	1.435	1.883	2.017	0.133	0.166
0.75	0.041	0.123	1.517	0.405	4.225	3.655	5.033	4.387	1.210	0.527
					$n = 350$					
0	0.022	0.075	0.021	0.045	0.51	0.571	1.038	1.080	0.050	0.096
0.25	0.020	0.067	0.025	0.044	0.59	0.689	1.109	1.203	0.063	0.093
0.75	0.022	0.070	0.026	0.056	2.66	2.440	3.557	3.217	0.808	0.392
					$n = 500$					
0	0.016	0.054	0.108	0.057	0.301	0.349	0.689	0.723	0.030	0.064
0.25	0.015	0.044	0.185	0.051	0.368	0.451	0.763	0.859	0.045	0.066
0.75	0.016	0.050	1.317	0.276	2.001	1.901	2.861	2.647	0.628	0.327
	CBE		SBE		2SE1		$2SE1_K$		2SE2	
c	m_2	m_3	m_2	m_3	m_2	m_3	m_2	m_3	m_2	m_3
Model 3										
					$n = 200$					
0	0.058	0.143	0.316	0.161	1.644	1.687	1.614	1.678	0.159	0.195
0.25	0.065	0.129	0.401	0.152	1.755	1.735	1.782	1.789	0.210	0.192
0.75	0.063	0.152	1.461	0.342	4.691	4.272	4.477	4.128	1.249	0.484
					$n = 350$					
0	0.039	0.074	0.198	0.083	0.749	0.785	0.941	0.983	0.079	0.103
0.25	0.040	0.086	0.250	0.095	0.760	0.822	0.983	1.035	0.093	0.120
0.75	0.041	0.096	1.339	0.264	3.203	3.016	3.367	3.175	0.907	0.413
					$n = 500$					
0	0.029	0.057	0.156	0.062	0.462	0.515	0.628	0.681	0.055	0.077
0.25	0.031	0.058	0.218	0.063	0.519	0.586	0.706	0.777	0.068	0.087
0.75	0.033	0.066	1.288	0.223	2.553	2.385	2.800	2.612	0.742	0.343

with the true bandwidths reported in table 6. How much under or oversmoothing occurs depends largely on the degree of curvature of the m_d that compose the models. When there is more curvature, as in the case of models 1 and 3 the degree of under and oversmoothing seems to increase, indicating that increased curvature makes for more difficult bandwidth estimation.

Table 8. Average true bandwidths.

	CBE/SBE		2SE1		2SE2	
c	h_1	h_2	h_1	h_2	h_1	h_2
Model 1						
			$n = 200$			
0	0.062	0.036	0.061	0.036	0.062	0.036
0.25	0.062	0.036	0.062	0.036	0.062	0.036
0.75	0.063	0.036	0.063	0.036	0.063	0.036
			$n = 350$			
0	0.055	0.032	0.055	0.032	0.055	0.032
0.25	0.055	0.032	0.055	0.032	0.055	0.032
0.75	0.056	0.032	0.056	0.032	0.056	0.032
			$n = 500$			
0	0.051	0.030	0.051	0.030	0.051	0.030
0.25	0.051	0.030	0.051	0.030	0.051	0.030
0.75	0.052	0.030	0.052	0.030	0.052	0.030
Model 2						
			$n = 200$			
0	0.061	0.066	0.055	0.059	0.061	0.066
0.25	0.062	0.067	0.055	0.059	0.062	0.066
0.75	0.063	0.067	0.056	0.059	0.063	0.067
			$n = 350$			
0	0.055	0.059	0.049	0.052	0.055	0.059
0.25	0.055	0.059	0.049	0.052	0.055	0.059
0.75	0.056	0.060	0.050	0.053	0.056	0.060
			$n = 500$			
0	0.051	0.055	0.046	0.049	0.051	0.055
0.25	0.051	0.055	0.046	0.049	0.051	0.055
0.75	0.052	0.056	0.046	0.049	0.052	0.056
Model 3						
			$n = 200$			
0	0.036	0.066	0.035	0.061	0.036	0.066
0.25	0.036	0.066	0.035	0.062	0.036	0.066
0.75	0.036	0.068	0.035	0.062	0.036	0.067
			$n = 350$			
0	0.032	0.059	0.032	0.055	0.032	0.059
0.25	0.032	0.059	0.032	0.055	0.032	0.059
0.75	0.032	0.060	0.031	0.056	0.032	0.060
			$n = 500$			
0	0.030	0.055	0.029	0.051	0.030	0.055
0.25	0.030	0.055	0.029	0.051	0.030	0.055
0.75	0.030	0.056	0.029	0.052	0.030	0.056

6. Conclusions

Additive non-parametric regression models have gained increased popularity by their ease of interpretation and the fact that these models allow for the circumvention of the curse of dimensionality. Classic backfitting, smooth backfitting, marginal integration and two stage estimators have recently emerged as viable alternatives for the estimation of additive non-parametric regression models. Little is known about the finite and asymptotic properties of all estimators when bandwidths are selected by data driven procedures. Applied researchers are not only uninformed about the estimators' properties but are also unaware of their relative performance. In this paper, we provided experimental evidence on the finite sample properties of these estimators and on their relative performances. We also propose a modification of the two-stage estimator first introduced by Kim *et al.* [10] that outperforms the original two-stage estimator.

Table 9. Average estimated bandwidths.

	CBE/SBE		2SE1		$2SE1_K$		2SE2	
c	h_1	h_2	h_1	h_2	h_1	h_2	h_1	h_2
Model 1								
				$n = 200$				
0	0.054	0.042	0.054	0.042	0.042	0.042	0.054	0.042
0.25	0.054	0.040	0.054	0.040	0.042	0.042	0.054	0.040
0.75	0.057	0.043	0.057	0.043	0.041	0.041	0.057	0.043
				$n = 350$				
0	0.051	0.039	0.051	0.039	0.038	0.038	0.051	0.0394
0.25	0.050	0.038	0.050	0.038	0.038	0.038	0.050	0.0388
0.75	0.052	0.038	0.052	0.038	0.036	0.036	0.052	0.0389
				$n = 500$				
0	0.047	0.037	0.047	0.037	0.035	0.035	0.047	0.037
0.25	0.048	0.035	0.048	0.035	0.035	0.035	0.048	0.035
0.75	0.049	0.036	0.049	0.036	0.034	0.034	0.049	0.036
Model 2								
				$n = 200$				
0	0.054	0.048	0.054	0.048	0.042	0.042	0.054	0.048
0.25	0.054	0.046	0.054	0.046	0.042	0.042	0.054	0.046
0.75	0.055	0.041	0.054	0.041	0.041	0.041	0.055	0.041
				$n = 350$				
0	0.050	0.047	0.050	0.047	0.038	0.038	0.050	0.0476
0.25	0.049	0.045	0.049	0.045	0.038	0.038	0.049	0.0450
0.75	0.051	0.040	0.051	0.040	0.036	0.036	0.051	0.0401
				$n = 500$				
0	0.047	0.045	0.047	0.045	0.035	0.035	0.047	0.045
0.25	0.047	0.043	0.047	0.043	0.035	0.035	0.047	0.043
0.75	0.049	0.038	0.049	0.038	0.034	0.034	0.049	0.038
Model 3								
				$n = 200$				
0	0.041	0.047	0.041	0.047	0.042	0.042	0.041	0.047
0.25	0.043	0.048	0.043	0.047	0.042	0.042	0.043	0.048
0.75	0.043	0.040	0.043	0.040	0.041	0.041	0.043	0.040
				$n = 350$				
0	0.039	0.047	0.039	0.047	0.038	0.038	0.039	0.047
0.25	0.039	0.046	0.039	0.046	0.038	0.038	0.039	0.046
0.75	0.039	0.040	0.039	0.040	0.036	0.036	0.039	0.040
				$n = 500$				
0	0.036	0.045	0.036	0.045	0.035	0.035	0.036	0.045
0.25	0.036	0.045	0.036	0.045	0.035	0.035	0.036	0.045
0.75	0.036	0.039	0.036	0.039	0.034	0.034	0.036	0.039

Although the theoretic results suggest that both smooth backfitting and two-stage estimators could reach the oracle efficiency bound, our experiments suggest that in the finite sample the classic backfitting estimator seems to emerge as the best estimator among those currently available in the literature. This superiority is based on an evaluation of the estimators' ASE under estimated and true bandwidths. Separate evidence on their bias is also provided to support this conclusion. Although Monte Carlo studies suffer from the problem of specificity, we believe that the results here are strong enough to recommend the use of classic backfitting estimation.

Acknowledgement

We thank two anonymous referees for helpful comments. The authors retain responsibility for any remaining errors.

References

[1] Stone, C., 1980, Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, **8**, 1348–1360.
[2] Stone, C., 1985, Additive regression and other nonparametric models. *Annals of Statistics*, **6**, 689–705.
[3] Hastie, T.J. and Tibshirani, R.J., 1990, Generalized additive models (New York: Chapman and Hall).
[4] Pagan, A. and Ullah, A., 1999, *Nonparametric econometrics* (Cambridge: Cambridge University Press).
[5] Buja, A., Hastie, T.J. and Tibshirani, R.J., 1989, Linear smoothers and additive models. *Annals of Statistics*, **17**, 453–555.
[6] Newey, W., 1994, Kernel estimation of partial means. *Econometric Theory*, **10**, 233–253.
[7] Tjøstheim, D. and Auestad, B., 1994, Nonparametric identification of nonlinear time series projections. *Journal of the American Statistical Association*, **89**, 1398–1409.
[8] Linton, O. and Nielsen, J.P., 1995, A Kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–100.
[9] Linton, O., 1997, Efficient estimation of additive nonparametric regression models. *Biometrika*, **84**, 469–474.
[10] Kim, W., Linton, O. and Hengartner, N., 1999, A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8**, 278–297.
[11] Mammen, E., Linton, O. and Nielsen, J., 1999, The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, **27**, 1443–1490.
[12] Wahba, G., 1990, *Spline models for observational data* (Philadelphia, PA: SIAM).
[13] Horowitz, J.L. and Mammen, E., 2005, Oracle-efficient nonparametric estimation of an additive model with an unknown link function, working paper, Department of Economics, Northwestern University.
[14] Opsomer, J. and Ruppert, D., 1997, Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, **25**, 186–211.
[15] Opsomer, J., 2000, Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**, 166–179.
[16] Linton, O. and Härdle, W., 1996, Estimation of additive regression models with known links. *Biometrika*, **83**, 529–540.
[17] Sperlich, S., Linton, O. and Härdle, W., 1999, Integration and backfitting methods in additive models - finite sample properties and comparison. *TEST*, **8**, 1–39.
[18] Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P., 2001, A general projection framework for constrained smoothing. *Statistical Science*, **16**(3), 232–248
[19] Nielsen, J.P. and Sperlich, S., 2005, Smooth backfitting in practice. *Journal of Royal Statistical Society*, Series B, **67**, 43–61.
[20] Silverman, B.W., 1986, *Density estimation for statistics and data analysis* (New York: Chapman and Hall).
[21] Opsomer, J. and Ruppert, D., 1998, A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, **93**, 605–619.
[22] Mammen, E. and Park, B.U., 2005, Bandwidth selection for smooth backfitting in additive models. *The Annals of Statistics*, **33**, 1260–1294
[23] Dette, H., Lieres und Wilkau, C.V. and Sperlich, S., 2005, A comparison of different nonparametric methods for inference on additive models. *Journal of Nonparametric Statistics*, **17**(1), 57–81.
[24] Jones, M.C., Davies, S.J. and Park, B.U., 1994, Versions of kernel type regression estimators. *Journal of the American Statistical Association*, **89**, 825–832.
[25] Ruppert, D., Sheather, S.J. and Wand, M.P., 1995, An effective bandwidth selector for least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.
[26] Dennis, Jr., J.E. and Schnabel, R., 1983, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Englewood Cliffs, NJ: Prentice-Hall).
[27] Fan, J., 1992, Design adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Appendix: Theorems and Proofs

LEMMA 1 *Assume Assumptions* 1–3 *hold and suppose that* $\phi(x, z) : \Re^2 \to \Re$ *is a continuous function,* G_1 *a compact subset of* $\Re$, *and* $|\phi(x, z)| < B_\phi < \infty$. *Let*

$$s_j(x) = (ng_1)^{-1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{g_1}\right)\left(\frac{x_i - x}{g_1}\right)^j \phi(x_i, z_i) \text{ with } j = 0, 1, 2.$$

a) *If* $ng_1^{2p+1}(\ln(g_1))^{-1} \to \infty$ *for* $p > 0$, *then* $\sup_{x \in G_1} |s_j(x) - E(s_j(x))| = o_p(g_1^p)$.

b) *Let G_2 be a compact subset of $\Re^2$ and*

$$\hat{s}(x, z) = (ng_1g_2)^{-1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{g_1}\right)\left(\frac{z_i - z}{g_2}\right)^j \phi(x_i, z_i).$$

If $n(g_1g_2)^{2p+1}(\ln(g_1g_2))^{-1} \to \infty$ for $p > 0$, then $\sup_{(x,z)\in G_2} |\hat{s}(x, z) - E(\hat{s}(x, z))| = o_p((g_1g_2)^p)$.

Proof a) We prove the case where $j = 0$. Similar arguments can be used for $j = 1, 2$. Let $B(x_0, r) = \{x \in \Re : |x - x_0| < r\}$ for $r \in \Re^+$. G_1 compact implies that there exists $x_0 \in G_1$ such that $G_1 \subseteq B(x_0, r)$. Therefore, for all $x, x' \in G_1$ $|x - x'| < 2r$. Let $g_1 > 0$ be a sequence such that $g_1 \to 0$ as $n \to \infty$ where $n \in \{1, 2, 3 \ldots\}$. For any n, by the Heine–Borel theorem there exists a finite collection of sets $\{B(x_k, g_1^a)\}_{k=1}^{l_n}$ such that $G_1 \subset \cup_{k=1}^{l_n} B(x_k, g_1^a)$ for $x_k \in G_1$ with $l_n < g_1^{-a} r$ for $a \in (0, \infty)$. By assumption $|s_0(x) - s_0(x_k)| \leq (ng_1)^{-1} \sum_{i=1}^{n} c|g_1^{-1}(x_k - x)|B_\phi < B_\phi c g_1^{a-2}$ for $x \in B(x_k, g_1^a)$. Similarly, $|E(s_0(x_k)) - E(s_0(x))| < B_\phi c g_1^{a-2}$ for $x \in B(x_k, g_1^a)$. Hence, $|s_0(x) - E(s_0(x))| \leq |s_0(x_k) - E(s_0(x_k))| + 2B_\phi c g_1^{a-2}$ for $x \in B(x_k, g_1^a)$ and

$$\sup_{x\in G_1} |s_0(x) - E(s_0(x))| \leq \max_{1\leq k\leq l_n} |s_0(x_k) - E(s_0(x_k))| + 2B_\phi c g_1^{a-2}.$$

To show that $\lim_{n\to\infty} P(\sup_{x\in G_1} |s_0(x) - E(s_0(x))| \geq g_1^p \epsilon) = 0$ for $p > 0$, we need $g_1^{a-p-2} \to 0$ as $n \to \infty$ and $\lim_{n\to\infty} P(\max_{1\leq k\leq l_n} |s_0(x) - E(s_0(x))| \geq g_1^p \epsilon) = 0$. But

$$P(\max_{1\leq k\leq l_n} |s_0(x_k) - E(s_0(x_k))| \geq g_1^p \epsilon) \leq \sum_{k=1}^{l_n} P(|s_0(x_k) - E(s_0(x_k))| \geq g_1^p \epsilon).$$

Put $W_{\text{in}} = g_1^{-2} E(K^2\left(\frac{x_i - x_k}{g_1}\right)\phi^2(x_i, z_i)) - (g_1^{-1} E(K\left(\frac{x_i - x_k}{g_1}\right)\phi(x_i, z_i)))^2$ and use Bernstein's inequality to obtain

$$P(|s_0(x_k) - E(s_0(x_k))| \geq g_1^p \epsilon) < 2 \exp\left(\frac{-n g_1^{2p} \epsilon^2}{2(\bar{\sigma}^2 + B_W g_1^p \epsilon/3)}\right),$$

where $\bar{\sigma}^2 = n^{-1} \sum_{i=1}^{n} V(W_{in})$. Under Assumptions 1–3 and the fact that $\phi(x, z)$ and $f_{XZ}(x, z)$ are continuous we have that $g_1\bar{\sigma}^2 = O(1)$. Hence, for the desired result, the right-hand side of the inequality must approach zero as $n \to \infty$. It suffices to show that $(ng_1^{2p}\epsilon^2)/(2\bar{\sigma}^2 + 2/3B_W g_1^p \epsilon) + a \ln(g_1) \to \infty$, which given that $g_1\bar{\sigma}^2 = O(1)$ will result if $(ng_1^{2p+1})/(\ln(g_1)) \to \infty$.

b) Let $\theta = (x, z)'$, a typical element in $\Re^2$. Let $B(\theta_0, r) = \{\theta \in \Re^2 : ||\theta - \theta_0|| < r\}$ for $r \in \Re^+$. G_2 compact implies that there exists $\theta_0 \in G_2$ such that $G_2 \subseteq B(\theta_0, r)$. Therefore, for all $\theta, \theta' \in G_2$ $||\theta - \theta'|| < 2r$. Let $g_1, g_2 > 0$ be a sequence such that $g_1, g_2 \to 0$ as $n \to \infty$, where $n \in \{1, 2, 3 \ldots\}$. For any n, by the Heine–Borel theorem there exists a finite collection of sets $\{B(\theta_k, r_n)\}_{k=1}^{l_n}$ such that $G_2 \subset \cup_{k=1}^{l_n} B(\theta_k, r_n)$ for $\theta_k \in G_2$ with $l_n < r_n^{-1} r^2$, $r_n = (g_1g_2)^a$ for $a \in (0, \infty)$. For $\theta \in B(\theta_k, r_n)$, $|\hat{s}(\theta) - \hat{s}(\theta_k)| \leq B_\phi c(g_1 + g_2)(g_1g_2)^{a-2}$.

Similarly, $|E(\hat{s}(\theta_k)) - E(\hat{s}(\theta))| < B_\phi c(g_1+g_2)(g_1g_2)^{a-2}$. Hence,

$$\sup_{\theta\in G_2} |\hat{s}(\theta) - E(\hat{s}(\theta))| \le \max_{1\le k\le l_n} |\hat{s}(\theta_k) - E(s(\theta_k))| + 2B_\phi c(g_1+g_2)(g_1g_2)^{a-2}.$$

To show that $\lim_{n\to\infty} P(\sup_{\theta\in G_2} |\hat{s}(\theta) - E(\hat{s}(\theta))| \ge (g_1g_2)^p\epsilon) = 0$ for $p > 0$ it suffices to have $(g_1g_2)^{a-p-2} = O(1)$ and $\lim_{n\to\infty} P(\max_{1\le k\le l_n} |\hat{s}(\theta) - E(\hat{s}(\theta))| \ge (g_1g_2)^p\epsilon) = 0$. But

$$P(\max_{1\le k\le l_n} |\hat{s}(\theta_k) - E(\hat{s}(\theta_k))| \ge (g_1g_2)^p\epsilon) \le \sum_{k=1}^{l_n} P(|\hat{s}(\theta_k) - E(\hat{s}(\theta_k))| \ge (g_1g_2)^p\epsilon).$$

Put $W_{in} = \frac{1}{g_1g_2}K_1\left(\frac{x_i-x}{g_1}\right)K\left(\frac{z_i-z}{g_2}\right) - E\left(\frac{1}{g_1g_2}K\left(\frac{x_i-x}{g_1}\right)K\left(\frac{z_i-z}{g_2}\right)\right)$ and using Bernstein's inequality, we have

$$P(|\hat{s}(\theta_k) - E(\hat{s}(\theta_k))| \ge (g_1g_2)^p\epsilon) < 2\exp\left(\frac{-n(g_1g_2)^{2p+1}\epsilon^2}{2g_1g_2\bar{\sigma}^2 + 2B_W(g_1g_2)^p\epsilon/3}\right),$$

where $\bar{\sigma}^2 = 1/n\sum_{i=1}^n V(W_{in})$.

Hence, for the desired result the right-hand side of the inequality must approach zero as $n\to\infty$. For this it suffices to have $(n(g_1g_2)^{2p+1})/(\ln(g_1g_2)) \to \infty$. ■

THEOREM 1 *Suppose that Assumptions* 1–3 *hold*, $ng_1^3(\ln(g_1))^{-1} \to \infty$ *and* $n(g_1g_2)^3(\ln(g_1g_2))^{-1} \to \infty$. *Put* $\gamma_1(x) = \alpha + m_1(x)$ *and* $\gamma_2(z) = \alpha + m_2(z)$. *Then, the conditional bias of* $m_1^{2S1}(x)$ *for* $x \in S_X$ *is given by*,

$$\begin{aligned} E(m_1^{2S1}(x) - m_1(x)|\vec{x},\vec{z}) &= \frac{h_1^2}{2}\mu_2 m_1^{(2)}(x) - \frac{1}{2}g_2^2\mu_2 E(m_2^{(2)}(Z)|\vec{x}) \\ &- \frac{1}{2}g_2^2\mu_2 E\left(\int f_X^{(2)}(v)m(v,z_i)dv|\vec{x}\right) \\ &+ \frac{1}{2}g_2^2\mu_2 E\left(\int m(v,Z)f_X(v)f_{XZ}^{-1}(v,Z)\sum_{d=1}^{2}\frac{\partial^2 f_{XZ}(v,Z)}{\partial_d\partial_d}dv|\vec{x}\right) \\ &+ o_p(h_1^2) + o_p(g_2^2) \end{aligned}$$

and

$$V(m_1^{2S1}(x)|\vec{x},\vec{z}) = \frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1} + o_p((nh_1)^{-1}). \tag{A1}$$

Mutatis mutandis similar expressions are obtained for m_2. *The conditional bias and variance of* $m^{2S1}(x,z)$ *are*

$$\begin{aligned} E(m^{2S1}(x,z) - m(x,z)|\vec{x},\vec{z}) &= \frac{h_1^2}{2}\mu_2 m_1^{(2)}(x) - \frac{1}{2}g_2^2\mu_2 E(m_2^{(2)}(Z)|\vec{x}) \\ &- \frac{1}{2}g_2^2\mu_2 E\left(\int f_X^{(2)}(v)m(v,Z)dv|\vec{x}\right) \end{aligned}$$

$$+\frac{1}{2}g_2^2\mu_2 E\left(\int m(v,Z)f_X(v)f_{XZ}^{-1}(v,Z)\sum_{d=1}^{2}\frac{\partial^2 f_{XZ}(v,Z)}{\partial_d\partial_d}dv|\vec{x}\right)$$
$$+\frac{h_2^2}{2}\mu_2 m_2^{(2)}(z)-\frac{1}{2}g_1^2\mu_2 E(m_1^{(2)}(X)|\vec{z})$$
$$-\frac{1}{2}g_1^2\mu_2 E\left(\int f_Z^{(2)}(v)m(x_i,v)dv|\vec{z}\right)$$
$$+\frac{1}{2}g_1^2\mu_2 E\left(\int m(X,v)f_Z(v)f_{XZ}^{-1}(X,v)\sum_{d=1}^{2}\frac{\partial^2 f_{XZ}(X,v)}{\partial_d\partial_d}dv|\vec{z}\right)$$
$$+o_p(h_1^2)+o_p(g_2^2)+o_p(h_2^2)+o_p(g_1^2)$$

and

$$V(m^{2S1}(x,z)|\vec{x},\vec{z})=\frac{1}{nh_1}\sigma^2 R_K f_X(x)^{-1}+\frac{1}{nh_2}\sigma^2 R_K f_Z(z)^{-1}+o_p((nh_1)^{-1}+(nh_2)^{-1}). \tag{A2}$$

Proof Let $\epsilon'=(\epsilon_1,\ldots,\epsilon_n)$, where $\epsilon_i=y_i-\alpha-m_1(x_i)-m_2(z_i)$ and $\gamma_2^P(\vec{z})$ be as defined in equation (13). By construction,

$$\begin{aligned}m_1^{2S1}(x)&=s_1(x)(\vec{y}-\vec{\gamma}_2^P(\vec{z}))\\&=s_1(x)(1_n\alpha+\vec{m}_1(\vec{x})+\vec{m}_2(\vec{z})+\epsilon-\vec{\gamma}_2^P(\vec{z}))\\&=s_1(x)(\vec{m}_1(\vec{x})+\epsilon)+s_1(x)(\vec{\gamma}_2(\vec{z})-\vec{\gamma}_2^P(\vec{z})),\end{aligned} \tag{A3}$$

where $\vec{\gamma}_2(\vec{z})'=(\gamma_2(z_1),\ldots,\gamma_2(z_n))$. Under our assumptions and using the results of Fan [27] for local linear estimation,

$$E(s_1(x)(\vec{m}_1(\vec{x})+\epsilon)|\vec{x},\vec{z})=s_1(x)\vec{m}_1(\vec{x})=m_1(x)+\frac{h_1^2}{2}\mu_2 m_1^{(2)}(x)+o_p(h_1^2). \tag{A4}$$

We now look at the second term in equation (A3). Note that the i^{th} element of $-(\vec{\gamma}_2(\vec{z})-\vec{\gamma}_2^P(\vec{z}))$ is

$$\begin{aligned}\gamma_2^P(z_i)-\gamma_2(z_i)&=\frac{1}{n}\sum_{k=1}^{n}\frac{1}{g_2}K\left(\frac{z_k-z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}y_k-\gamma_2(z_i)\\&=\hat{L}_{1n}(z_i)+\hat{L}_{2n}(z_i)+\hat{L}_{3n}(z_i),\end{aligned}$$

where

$$\hat{L}_{1n}(z_i)=\frac{1}{n}\sum_{k=1}^{n}\frac{1}{g_2}K\left(\frac{z_k-z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}\epsilon_k$$
$$\hat{L}_{2n}(z_i)=\frac{1}{n}\sum_{k=1}^{n}\frac{1}{g_2}K\left(\frac{z_k-z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}(m_2(z_k)-m_2(z_i))$$
$$\hat{L}_{3n}(z_i)=\frac{1}{n}\sum_{k=1}^{n}\left\{\frac{1}{g_2}K\left(\frac{z_k-z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}m(x_k,z_i)-\gamma_2(z_i)\right\}.$$

If $\hat{L}_1' = (\hat{L}_{1n}(z_1), \ldots, \hat{L}_{1n}(z_n))$, $\hat{L}_2' = (\hat{L}_{2n}(z_1), \ldots, \hat{L}_{2n}(z_n))$ and $\hat{L}_3' = (\hat{L}_{3n}(z_1), \ldots, \hat{L}_{3n}(z_n))$, then the last term in equation (A3) can be written as

$$s_1(x)(\vec{\gamma}_2(\vec{z}) - \vec{\gamma}_2^P(\vec{z})) = -s_1(x)(\hat{L}_1 + \hat{L}_2 + \hat{L}_3) \tag{A5}$$

and $E(s_1(x)(\vec{\gamma}_2(\vec{z}) - \vec{\gamma}_2^P(\vec{z}))|\vec{x}, \vec{z}) = -s_1(x)(E(\hat{L}_1|\vec{x}, \vec{z}) + E(\hat{L}_2|\vec{x}, \vec{z}) + E(\hat{L}_3|\vec{x}, \vec{z}))$. By assumption $E(\hat{L}_1|\vec{x}, \vec{z}) = 0$, we now treat $\hat{L}_2$ and $\hat{L}_3$ separately. In what follows we define $\bar{f}_X(x) = E(\hat{f}_X(x)) = g_1^{-1} \int K\left(\frac{v-x}{g_1}\right) f_X(v)dv$ and $\bar{f}_{XZ}(x, z) = E(\hat{f}_{XZ}(x, z)) = (g_1 g_2)^{-1} \iint K\left(\frac{v-x}{g_1}\right) K\left(\frac{u-z}{g_2}\right) f_{XZ}(u, v)dudv$, and

$$L_{2n}(z_i) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} (m_2(z_k) - m_2(z_i)).$$

Given that there exists $0 < B_{dm_2}$ such that $|m_2^{(1)}(z)| < B_{dm_2}$ for all $z \in S_Z$ compact, we have that by using the Mean Value Theorem

$$|\hat{L}_{2n}(z_i) - L_{2n}(z_i)| \leq B_{dm_2} B_Z \sup_{(x_k, z_k) \in S_X \times S_Z} \left| \frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k, z_k)} - \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} \right| \hat{f}_Z(z_i) \tag{A6}$$

for a bound B_Z on $|z_k - z_i|$. Hence, it follows from Assumption 2 that there exists $0 < B_1, B_{f_Z}$, B_{f_X} such that $|f_Z(z)| < B_{f_Z}$ for all $z \in S_Z$, $|f_X(x)| < B_{f_X}$ for all $x \in S_X$ and $|f_{XZ}(x, z)|^{-1} < B_1$, for all $(x, z) \in S_X \times S_Z$. Therefore, we have

$$\begin{aligned}
&\sup_{z_i \in S_Z} |\hat{L}_{2n}(z_i) - L_{2n}(z_i)| \\
&\quad \leq B_{dm_2} B_Z \sup_{(x_k, z_k) \in S_X \times S_Z} \left| \frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k, z_k)} - \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} \right| \times (\sup_{z_i \in S_Z} |\hat{f}_Z(z_i) - \bar{f}_Z(z_i)| + B_{f_Z}) \\
&\quad \leq \left(\sup_{z_i \in S_Z} |\hat{f}_Z(z_i) - \bar{f}_Z(z_i)| + B_{f_Z} \right) B_{dm_2} B_Z \left(\sup_{x_k \in S_X} |\hat{f}_X(x_k) - \bar{f}_X(x_k)| \right. \\
&\qquad \times \sup_{(x_k, z_k) \in S_X \times S_Z} |\hat{f}_{XZ}^{-1}(x_k, z_k) - \bar{f}_{XZ}^{-1}(x_k, z_k)| + B_1 \sup_{x_k \in S_X} |\hat{f}_X(x_k) - \bar{f}_X(x_k)| \\
&\qquad \left. + B_{f_X} \sup_{(x_k, z_k) \in S_X \times S_Z} |\hat{f}_{XZ}^{-1}(x_k, z_k) - \bar{f}_{XZ}^{-1}(x_k, z_k)| \right).
\end{aligned} \tag{A7}$$

Let $a^n(z_i) = (\mu_2/2) g_2^2 m_2^{(2)}(z_i)$ and note that by the triangle inequality we have

$$\begin{aligned}
\sup_{z_i \in S_Z} |\hat{L}_{2n}(z_i) - a^n(z_i)| \leq & \sup_{z_i \in S_Z} |\hat{L}_{2n}(z_i) - L_{2n}(z_i)| + \sup_{z_i \in S_Z} |E(L_{2n}(z_i)) - a^n(z_i)| \\
& + \sup_{z_i \in S_Z} |L_{2n}(z_i) - E(L_{2n}(z_i))|.
\end{aligned} \tag{A8}$$

Since $E(L_{2n}(z_i)) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} (m_2(z_k) m_2(z_i)) f_{XZ}(x_k, z_k) dx_k dz_k$ and $m_2(z_k) - m_2(z_i) = m_2^{(1)}(z_i)(z_k - z_i) + (1/2) m_2^{(2)}(z_i)(z_k - z_i)^2 + (z_k - z_i)^2 o(1)$ we have that

$$E(L_{2n}(z_i)) = g_2 m_2^{(1)}(z_i) F_{1n}(z_i) + \frac{1}{2} g_2^2 m_2^{(2)}(z_i) F_{2n}(z_i) + g_2^2 F_{2n}(z_i) o(1),$$

where

$$F_{1n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} \frac{z_k - z_i}{g_2} f_{XZ}(x_k, z_k) dx_k dz_k$$

and

$$F_{2n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} \frac{(z_k - z_i)^2}{g_2^2} f_{XZ}(x_k, z_k) dx_k dz_k.$$

Let $F_{11n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \frac{z_k - z_i}{g_2} \left(\frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} - \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)}\right) (\bar{f}_{XZ}(x_k, z_k) - f_{XZ}(x_k, z_k)) dx_k dz_k$ and $F_{12n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \frac{z_k - z_i}{g_2} \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)} (\bar{f}_{XZ}(x_k, z_k) - f_{XZ}(x_k, z_k)) dx_k dz_k$, then given that $\int \psi K(\psi) d\psi = 0$ we can write $F_{1n}(z_i) = -F_{11n}(z_i) - F_{12n}(z_i)$. We now write

$$F_{12n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)} \times \left(\frac{1}{2}\mu_2 \sum_{i=1}^{2} \frac{\partial^2 f_{XZ}}{\partial_i \partial_i}(x_k, z_k) g_i^2 + o(g_1^2 + g_2^2)\right) dx_k dz_k.$$

Given that $\int \psi K(\psi) d\psi = 0$, and by Lebesgue's dominated convergence theorem, we have $F_{12n}(z_i) = o(g_1^2 + g_2^2)$ uniformly in S_Z. Since $F_{11n}(z_i)$ is clearly of smaller order, and $g_1 \sim g_2$, we conclude that $F_{1n}(z_i) = o(g_2^2)$ uniformly in S_Z.

Let $F_{21n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \left(\frac{z_k - z_i}{g_2}\right)^2 \left(\frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} - \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)}\right) (\bar{f}_{XZ}(x_k, z_k) - f_{XZ}(x_k, z_k)) dx_k dz_k$ and $F_{22n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \left(\frac{z_k - z_i}{g_2}\right)^2 \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)} (\bar{f}_{XZ}(x_k, z_k) - f_{XZ}(x_k, z_k)) dx_k dz_k$, then given that $\int \psi^2 K(\psi) d\psi = \mu_2$ we can write $F_{2n}(z_i) = \mu_2 - F_{21n}(z_i) - F_{22n}(z_i)$. We now write

$$F_{22n}(z_i) = g_2^{-1} \iint K\left(\frac{z_k - z_i}{g_2}\right) \left(\frac{z_k - z_i}{g_2}\right)^2 \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)} \times \left(\frac{1}{2}\mu_2 \sum_{i=1}^{2} \frac{\partial^2 f_{XZ}}{\partial_i \partial_i}(x_k, z_k) g_i^2 + o(g_1^2 + g_2^2)\right) dx_k dz_k.$$

Given that $\int \psi^2 K(\psi) d\psi = \mu_2$, and by Lebesgue's dominated convergence theorem, we have $F_{22n}(z_i) = O(g_1^2 + g_2^2) + o(g_1^2 + g_2^2)$ uniformly in S_Z. Since $F_{21n}(z_i)$ is clearly of smaller order, and $g_1 \sim g_2$, we conclude that $F_{2n}(z_i) = \mu_2 + O(g_2^2) + o(g_2^2)$ uniformly in S_Z. Combining the results for F_{1n} and F_{2n} we conclude that

$$E(L_{2n}(z_i)) - \frac{1}{2} g_2^2 \mu_2 m_2^{(1)}(z_i) \equiv E(L_{2n}(z_i)) - a^n(z_i) = o(g_2^2) \text{ uniformly in } S_Z. \quad \text{(A9)}$$

Finally, by Lemma 1 $\sup_{z_i \in S_Z} |L_{2n}(z_i) - E(L_{2n}(z_i))| = o_p(1)$, hence, combining equations (A7), (A8) and (A9) we have $\hat{L}_{2n}(z_i) = (\mu_2/2) g_2^2 m_2^{(2)}(z_i) + o_p(g_2^2)$ uniformly in S_Z.

Let $L_{3n}(z_i) = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}(x_k, z_k)} m(x_k, z_i) - \gamma_2(z_i)\right)$, then if there exists $0 < B_m$ such that $m(x, z) < B_m$ for all $(x, z) \in S_X \times S_Z$ we have

$$|\hat{L}_{3n}(z_i) - L_{3n}(z_i)| \leq B_m \sup_{(x_k, z_k) \in S_X \times S_Z} \left| \frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k, z_k)} - \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} \right| \sup_{z_i \in S_Z} \hat{f}_Z(z_i),$$

which is similar in structure to inequality equation (A6). Hence, using the same arguments we have that $\hat{L}_{3n}(z_i) = L_{3n}(z_i) + o_p(g_2^2)$ uniformly in S_Z. Let

$$A_1^n(z_i) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} \quad \text{and}$$

$$A_2^n(z_i) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} m_1(x_k).$$

Now note that,

$$\begin{aligned} L_{3n}(z_i) &= \alpha A_1^n(z_i) + A_2^n(z_i) + m_2(z_i) A_1^n(z_i) - \gamma_2(z_i) \\ &= \gamma_2(z_i)(A_1^n(z_i) - 1) + A_2^n(z_i) \end{aligned}$$

and consequently $E(L_3^n(z_i)) = \gamma_2(z_i)(E(A_1^n(z_i)) - 1) + E(A_2^n(z_i))$. We look at each expectation separately.

$$\begin{aligned} E(A_1^n(z_i)) &= \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} f_{XZ}(x_k, z_k) dx_k dz_k \\ &= \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \bar{f}_X(x_k) dx_k dz_k \\ &\quad - \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} (\bar{f}_{XZ}(x_k, z_k) - f_{XZ}(x_k, z_k)) dx_k dz_k \\ &= \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) (\bar{f}_X(x_k) - f_X(x_k)) dx_k dz_k \\ &\quad + \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) f_X(x_k) dx_k dz_k \\ &\quad - \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right) \frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} (\bar{f}_{XZ}(x_k, z_k) - f_{XZ}(x_k, z_k)) dx_k dz_k \end{aligned}$$

$$
\begin{aligned}
&= 1 + \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right)(\bar{f}_X(x_k) - f_X(x_k))dx_k dz_k \\
&\quad - \left(\iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right)\left(\frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} - \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)}\right)(\bar{f}_{XZ}(x_k, z_k)\right. \\
&\quad \left. - f_{XZ}(x_k, z_k))dx_k dz_k + \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right)\frac{f_X(x_k)}{f_{XZ}(x_k, z_k)}(\bar{f}_{XZ}(x_k, z_k)\right. \\
&\quad \left. - f_{XZ}(x_k, z_k))dx_k dz_k\right) \\
&= 1 + C_1^n(z_i) - (C_2^n(z_i) + C_3^n(z_i)).
\end{aligned}
$$

Also, using the fact that $E(m_1(X)) = 0$, we can similarly write

$$
\begin{aligned}
E(A_2^n(z_i)) &= \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right)(\bar{f}_X(x_k) - f_X(x_k))m_1(x_k)dx_k dz_k \\
&\quad - \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right)\frac{f_X(x_k)}{f_{XZ}(x_k, z_k)}(\bar{f}_{XZ}(x_k, z_k) - f_{XZ}(x_k, z_k))m_1(x_k)dx_k dz_k \\
&\quad - \iint \frac{1}{g_2} K\left(\frac{z_k - z_i}{g_2}\right)\left(\frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)} - \frac{f_X(x_k)}{f_{XZ}(x_k, z_k)}\right)m_1(x_k)(\bar{f}_{XZ}(x_k, z_k) \\
&\quad - f_{XZ}(x_k, z_k))dx_k dz_k \\
&= D_1^n(z_i) + D_2^n(z_i) + D_3^n(z_i).
\end{aligned}
$$

By Taylor's Theorem, for all $(x, z) \in S_X \times S_Z$ and $\delta > 0$

$$
-g_1^2\frac{1}{2}f_X^{(2)}\mu_2 - \frac{1}{2}\mu_2 g_1^2\delta < \bar{f}_X(x) - f_X(x) < g_1^2\frac{1}{2}f_X^{(2)}(x)\mu_2 + \frac{1}{2}\mu_2 g_1^2\delta \text{ and}
$$

$$
-\frac{\mu_2}{2}\sum_{i=1}^{2}\frac{\partial^2 f_{XZ}(x, z)}{\partial_i \partial_i}g_i^2 - \frac{\mu_2}{2}\sum_{i=1}^{2}g_i^2\delta < \bar{f}_{XZ}(x, z) - f_{XZ}(x, z) < \frac{\mu_2}{2}\sum_{i=1}^{2}\frac{\partial^2 f_{XZ}(x, z)}{\partial_i \partial_i}g_i^2 + \frac{\mu_2}{2}\sum_{i=1}^{2}g_i^2\delta.
$$

Therefore, given Assumption 2 and provided that S_X is bounded

$$
\begin{aligned}
\frac{C_1^n(z_i)}{g_2^2} &= \frac{\mu_2}{2}\int f_X^{(2)}(v)dv + o(1) \text{ uniformly in } S_Z, \text{ and} \\
\frac{C_3^n(z_i)}{g_2^2} &= \frac{\mu_2}{2}\int f_X(v)\frac{\partial^2 f_{XZ}(v, z_i)}{\partial_1 \partial_1}\frac{1}{f_{XZ}(v, z_i)}dv \\
&\quad + \frac{\mu_2}{2}\int f_X(v)\frac{\partial^2 f_{XZ}(v, z_i)}{\partial_2 \partial_2}\frac{1}{f_{XZ}(v, z_i)}dv + o(1)
\end{aligned}
$$

uniformly in S_Z. We ignore $C_2^n(z_i)$ as it is of order smaller than that of $C_1^n(z_i)$ and $C_3^n(z_i)$. Also,

$$
\begin{aligned}
\frac{D_1^n(z_i)}{g_2^2} &= \frac{\mu_2}{2}\int f_X^{(2)}(v)m_1(v)dv + o(1) \text{ and} \\
\frac{D_2^n(z_i)}{g_2^2} &= \frac{\mu_2}{2}\int f_X(v)m_1(v)\frac{\partial^2 f_{XZ}(v, z_i)}{\partial_1 \partial_1}\frac{1}{f_{XZ}(v, z_i)}dv \\
&\quad + \frac{\mu_2}{2}\int f_X(v)m_1(v)\frac{\partial^2 f_{XZ}(v, z_i)}{\partial_2 \partial_2}\frac{1}{f_{XZ}(v, z_i)}dv + o(1)
\end{aligned}
$$

uniformly in S_Z. As above, we ignore $D_3^n(z_i)$ as it is of order smaller than $D_1^n(z_i)$ and $D_2^n(z_i)$. Now, note that

$$\sup_{z_i \in S_Z} |L_{3n}(z_i) - E(L_{3n}(z_i))| \leq B_m \sup_{z_i \in S_Z} |A_1^n(z_i) - E(A_1^n(z_i))| + \sup_{z_i \in S_Z} |A_2^n(z_i) - E(A_2^n(z_i))|.$$

By Lemma 1

$$\frac{1}{g_2^2} \sup_{z_i \in S_Z} |L_{3n}(z_i) - E(L_{3n}(z_i))| = o_p(1) \tag{A10}$$

given that $\frac{\bar{f}_X(x_k)}{\bar{f}_{XZ}(x_k, z_k)}$ is bounded. Let $\tau_n(z_i) = \gamma_2(z_i)(T_{1n}(z_i) - 1) + T_{2n}(z_i)$, where

$$\begin{aligned}
T_{1n}(z_i) &= 1 + \frac{\mu_2}{2} g_2^2 \int f_X^{(2)}(v) dv - \frac{\mu_2}{2} g_2^2 \int f_X(v) \frac{\partial^2 f_{XZ}(v, z_i)}{\partial_1 \partial_1} \frac{1}{f_{XZ}(v, z_i)} dv \\
&\quad - \frac{\mu_2}{2} g_2^2 \int f_X(v) \frac{\partial^2 f_{XZ}(v, z_i)}{\partial_2 \partial_2} \frac{1}{f_{XZ}(v, z_i)} dv \\
T_{2n}(z_i) &= \frac{\mu_2}{2} g_2^2 \int m_1(v) f_X^{(2)}(v) dv - \frac{\mu_2}{2} g_2^2 \int f_X(v) m_1(v) \frac{\partial^2 f_{XZ}(v, z_i)}{\partial_1 \partial_1} \frac{1}{f_{XZ}(v, z_i)} dv \\
&\quad - \frac{\mu_2}{2} g_2^2 \int f_X(v) m_1(v) \frac{\partial^2 f_{XZ}(v, z_i)}{\partial_2 \partial_2} \frac{1}{f_{XZ}(v, z_i)} dv.
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{1}{g_2^2} \sup_{z_i \in S_Z} |E(L_{3n}(z_i)) - \tau_n(z_i)| &\leq \frac{1}{g_2^2} \sup_{z_i \in S_Z} |E(A_1^n(z_i)) - T_{1n}(z_i)| \\
&\quad + \frac{1}{g_2^2} \sup_{z_i \in S_Z} |E(A_2^n(z_i)) - T_{2n}(z_i)|.
\end{aligned} \tag{A11}$$

Hence,

$$\begin{aligned}
\frac{1}{g_2^2} \sup_{z_i \in S_Z} |\hat{L}_{3n}(z_i) - \tau_n(z_i)| &\leq \frac{1}{g_2^2} \sup_{z_i \in S_Z} |\hat{L}_{3n}(z_i) - L_{3n}(z_i)| + \frac{1}{g_2^2} \sup_{z_i \in S_Z} |L_{3n}(z_i) - E(L_{3n}(z_i))| \\
&\quad + \frac{1}{g_2^2} \sup_{z_i \in S_Z} |E(L_{3n}(z_i)) - \tau_n(z_i)|
\end{aligned} \tag{A12}$$

and combining equations (A10),(A11) and (A12), we obtain

$$\begin{aligned}
\hat{L}_{3n}(z_i) &= \frac{\mu_2}{2} g_2^2 \int f_X^{(2)}(v) m(v, z_i) dv - \frac{\mu_2}{2} g_2^2 \left(\int m(v, z_i) f_X(v) \frac{\partial^2 f_{XZ}(v, z_i)}{\partial_1 \partial_1} f_{XZ}^{-1}(v, z_i) dv \right. \\
&\quad \left. + \int m(v, z_i) f_X(v) \frac{\partial^2 f_{XZ}(v, z_i)}{\partial_2 \partial_2} f_{XZ}^{-1}(v, z_i) dv \right) + o_p(g_2^2)
\end{aligned}$$

uniformly in S_Z. Hence, combining the approximations for $\hat{L}_{2n}(z_i)$ and $\hat{L}_{3n}(z_i)$, we have

$$s_1(x)\hat{L}_2 = \frac{1}{2} g_2^2 \mu_2 E(m_2^{(2)}(z_i)|\vec{x}) + o_p(h_1^2) + o_p(g_2^2) \quad \text{and}$$

$$s_1(x)\hat{L}_3 = \frac{1}{2}g_2^2\mu_2 E\left(\int f_X^{(2)}(v)m(v,z_i)dv|\vec{x}\right) - \frac{1}{2}g_2^2\mu_2 E\left(\int m(v,z_i)f_X(v)f_{XZ}^{-1}(v,z_i)\sum_{d=1}^{2}\frac{\partial^2 f_{XZ}(v,z_i)}{\partial_d\partial_d}dv|\vec{x}\right) + o_p(h_1^2) + o_p(g_2^2),$$

which completes the proof of part a) of the theorem.
b) Let $[a_{ij}]_{i=1,j=1}^{m,p}$ denote an $m \times p$ matrix with typical element a_{ij}. We write $\vec{y} - \vec{\gamma}_2^P(\vec{z}) = \left(I - \frac{1}{ng_2}B_n\right)\vec{y}$ where

$$B_n = \left[K\left(\frac{z_j - z_i}{g_2}\right)\frac{\hat{f}_X(x_j)}{\hat{f}_{XZ}(x_j,z_j)}\right]_{i=1,j=1}^{n,n}.$$

Hence, $E(m_1^{2S1}(x)|\vec{x},\vec{z}) = s_1(x)\left(I - \frac{1}{ng_2}B_n\right)\vec{m}(\vec{x},\vec{z})$ and

$$nh_1V(m_1^{2S1}(x)|\vec{x},\vec{z}) = nh_1\sigma^2 s_1(x)\left(I - \frac{1}{ng_2}B_n\right)\left(I - \frac{1}{ng_2}B_n\right)' s_1(x)' \tag{A13}$$

$$= \sigma^2\left(nh_1 s_1(x)s_1(x)' - \frac{nh_1}{ng_2}s_1(x)B_n's_1(x)' - \frac{nh_1}{ng_2}B_n s_1(x)' + \frac{nh_1}{n^2g_2^2}s_1(x)BB's_1(x)'\right)$$

$$= \sigma^2\left(V_{1n}(x) + V_{2n}(x) + V_{3n}(x) + V_{4n}(x)\right). \tag{A14}$$

From [27] we have $V_{1n}(x) \xrightarrow{p} \frac{1}{f_X(x)}\int K^2(v)dv$. Now,

$$V_{2n}(x) = e\left(\frac{R_X'(x)W_X(x)R_X(x)}{nh_1}\right)^{-1}\frac{R_X'(x)W_X(x)B_nW_X(x)R_X(x)}{n^2g_2h_1} \times \left(\frac{R_X'(x)W_X(x)R_X(x)}{nh_1}\right)^{-1}e',$$

since $(R_X'(x)W_X(x)R_X(x)/nh_1)^{-1}$ converges in probability to a finite matrix we focus on

$$\frac{R_X'(x)W_X(x)B_nW_X(x)R_X(x)}{n^2g_2h_1} \equiv \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22}\end{pmatrix},$$

where

$$m_{11} = h_1 \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right) \frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \frac{1}{nh_1 g_2} \sum_{j=1}^{n} K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right),$$

$$m_{12} = h_1^2 \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right) \frac{x_i - x}{h_1} \frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \frac{1}{nh_1 g_2} \sum_{j=1}^{n} K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right),$$

$$m_{21} = h_1^2 \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right) \frac{x_i - x}{h_1} \frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \frac{1}{nh_1 g_2} \sum_{j=1}^{n} \times K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right) \frac{x_j - x}{h_1},$$

$$m_{22} = h_1^3 \frac{1}{nh_1} \sum_{i=1}^{n} \frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} K\left(\frac{x_i - x}{h_1}\right) \frac{x_i - x}{h_1} \frac{1}{nh_1 g_2} \sum_{j=1}^{n} \times K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right) \frac{x_j - x}{h_1}.$$

We now show that $m_{ij} = o_p(1)$ for all i, j. First,

$$\begin{aligned} m_{11} &= h_1 \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right) (\hat{f}_{XZ}(x, z_i) - f_{XZ}(x, z_i)) \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} - \frac{f_X(x_i)}{f_{XZ}(x_i, z_i)}\right) \\ &\quad + h_1 \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right) (\hat{f}_{XZ}(x, z_i) - f_{XZ}(x, z_i)) \frac{f_X(x_i)}{f_{XZ}(x_i, z_i)} \\ &\quad + h_1 \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right) f_{XZ}(x, z_i) \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} - \frac{f_X(x_i)}{f_{XZ}(x_i, z_i)}\right) \\ &\quad + h_1 \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right) f_{XZ}(x, z_i) \frac{f_X(x_i)}{f_{XZ}(x_i, z_i)} \\ &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 \end{aligned}$$

and $|m_{11}| \leq |\alpha_1| + |\alpha_2| + |\alpha_3| + |\alpha_4|$. Now we note that

$$|\alpha_1| \leq h_1 \sup_{z_i \in S_Z} |\hat{f}_{XZ}(x, z_i) - f_{XZ}(x, z_i)| \sup_{x_i, z_i \in S_X \times S_Z} \left| \frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} - \frac{f_X(x_i)}{f_{XZ}(x_i, z_i)} \right| \times \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right),$$

$$|\alpha_2| \leq h_1 B_{f_X} B_{f_{X,Z}}^{-1} \sup_{z_i \in S_Z} |\hat{f}_{XZ}(x, z_i) - f_{XZ}(x, z_i)| \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right),$$

$$|\alpha_3| \leq h_1 B_{f_{X,Z}} \sup_{x_i,z_i \in S_X \times S_Z} \left| \frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i,z_i)} - \frac{f_X(x_i)}{f_{XZ}(x_i,z_i)} \right| \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right),$$

$$|\alpha_4| \leq h_1 B_{f_X} \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_1}\right).$$

Hence, by Lemma 1 all expressions to the left of the inequalities converge in probability to zero, and consequently $m_{11} \xrightarrow{p} 0$. Now, note that m_{12} (m_{21} is identical in structure) is identical to m_{11} except for the presence in the summand of $(x_i - x)/h_1$, but since $K(\cdot) = 0$ outside of its compact support and S_X is compact, we have by Lemma 1 that $m_{12} \xrightarrow{p} 0$. The same argument is also applied to show that $m_{22} \xrightarrow{p} 0$ and, therefore, V_{2n}, $V_{3n} \xrightarrow{p} 0$.

$$V_{4n} = e\left(\frac{R_X'(x)W_X(x)R_X(x)}{nh_1}\right)^{-1} \frac{R_X'(x)W_X(x)B_nB_n'W_X(x)R_X(x)}{n^3 g_2^2 h_1} \times \left(\frac{R_X'(x)W_X(x)R_X(x)}{nh_1}\right)^{-1} e'$$

and as in the case of V_{2n}, we focus on showing that the matrix $(R_X'(x)W_X(x)B_nB_n'W_X(x)R_X(x))/(n^3 g_2^2 h_1)$ converges in probability to zero. Note that,

$$\frac{R_X'(x)W_X(x)B_nB_n'W_X(x)R_X(x)}{n^3 g_2^2 h_1} \equiv \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix},$$

where

$$u_{11} = g_2 \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i,z_i)}\right)^2 \frac{1}{ng_2^2} \sum_{l=1}^{n} K\left(\frac{x_l - x}{h_1}\right) K_1\left(\frac{z_i - z_l}{g_2}\right) \times \frac{1}{nh_1 g_2} \sum_{j=1}^{n} K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right),$$

$$u_{12} = h_1 g_2 \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i,z_i)}\right)^2 \frac{1}{ng_2^2} \sum_{l=1}^{n} K\left(\frac{x_l - x}{h_1}\right) K_1\left(\frac{z_i - z_l}{g_2}\right) \times \frac{1}{nh_1 g_2} \sum_{j=1}^{n} K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right) \frac{x_j - x}{h_1},$$

$$u_{21} = h_1 g_2 \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i,z_i)}\right)^2 \frac{1}{ng_2^2} \sum_{l=1}^{n} K\left(\frac{x_l - x}{h_1}\right) \frac{x_l - x}{h_1} K\left(\frac{z_i - z_l}{g_2}\right) \times \frac{1}{nh_1 g_2} \sum_{j=1}^{n} K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right),$$

$$u_{22} = h_1^2 g_2 \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i,z_i)}\right)^2 \frac{1}{ng_2^2} \sum_{l=1}^{n} K\left(\frac{x_l - x}{h_1}\right) \frac{x_l - x}{h_1} K\left(\frac{z_i - z_l}{g_2}\right) \times \frac{1}{nh_1 g_2} \sum_{j=1}^{n} K\left(\frac{z_i - z_j}{g_2}\right) K\left(\frac{x_j - x}{h_1}\right) \frac{x_j - x}{h_1}.$$

We will argue that $u_{ij} \xrightarrow{p} 0$ for all i, j. First, we observe that

$$
\begin{aligned}
|u_{11}| &= g_2 \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \right)^2 \left| \hat{f}^2_{XZ}(x, z_i) - f^2_{XZ}(x, z_i) \right| \\
&\quad + g_2 \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \right)^2 f^2_{XZ}(x, z_i) \\
&\leq g_2 \left(\sup_{z_i \in S_Z} \left| \hat{f}^2_{XZ}(x, z_i) - f^2_{XZ}(x, z_i) \right| + B^2_{f_{XZ}} \right) \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \right)^2 \\
&\leq g_2 \left(\sup_{z_i \in S_Z} \left| \hat{f}^2_{XZ}(x, z_i) - f^2_{XZ}(x, z_i) \right| + B^2_{f_{XZ}} \right) \left(\frac{1}{n} \sum_{i=1}^{n} \left| \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \right)^2 \right. \right. \\
&\quad \left. \left. - \left(\frac{f_X(x_i)}{f_{XZ}(x_i, z_i)} \right)^2 \right| + B \right) \\
&\leq g_2 \left(\sup_{z_i \in S_Z} \left| \hat{f}^2_{XZ}(x, z_i) - f^2_{XZ}(x, z_i) \right| + B^2_{f_{XZ}} \right) \left(\sup_{(x_i, z_i) \in S_X \times S_Z} \left| \left(\frac{\hat{f}_X(x_i)}{\hat{f}_{XZ}(x_i, z_i)} \right)^2 \right. \right. \\
&\quad \left. \left. - \left(\frac{f_X(x_i)}{f_{XZ}(x_i, z_i)} \right)^2 \right| + B \right)
\end{aligned}
$$

and, since by Lemma 1 $\hat{f}_{XZ}(x, z) - f_{XZ}(x, z) = o_p(1)$ and $\hat{f}_X(x) - f_X(x) = o_p(1)$ uniformly, $u_{11} = o_p(1)$. We also note that u_{21}, u_{12} and u_{22} differ from u_{11} only in that $(x_j - x)/h_1$ and $(x_l - x)/h_1$ appear in the summands. Again, since $K(\cdot) = 0$ outside of its compact support and S_X compact, we have by Lemma 1 that $u_{21}, u_{12}, u_{22} = o_p(1)$ and consequently $V_{4n} = o_p(1)$. ■

THEOREM 2 *Suppose that Assumptions* 1–3 *hold, that* $ng_1^3(\ln(g_1))^{-1} \to \infty$ *and* $n(g_1 g_2)^{2p+1}(\ln(g_1 g_2))^{-1} \to \infty$ *and let* $\mu(x, z) = m_1(x) + m_2(z)$. *Then, the conditional bias of* $m_1^{2S2}(x_i)$ *is given by*,

$$
\begin{aligned}
E(m_1^{2S2}(x) - m_1(x)|\vec{x}, \vec{z}) &= \frac{h_1^2}{2} \mu_2 m_1^{(2)}(x) - \frac{1}{2} g_2^2 \mu_2 E(m_2^{(2)}(Z)|\vec{x}) \\
&\quad - \frac{1}{2} g_2^2 \mu_2 E\left(\int f_X^{(2)}(v) \mu(v, Z) dv | \vec{x} \right) \\
&\quad + \frac{1}{2} g_2^2 \mu_2 E\left(\int \mu(v, Z) f_X(v) f_{XZ}^{-1}(v, Z) \sum_{d=1}^{2} \frac{\partial^2 f_{XZ}(v, Z)}{\partial_d \partial_d} dv | \vec{x} \right) \\
&\quad + o_p(h_1^2) + o_p(g_2^2)
\end{aligned}
$$

and

$$
V(m_1^{2S2}(x)|\vec{x}, \vec{z}) = \frac{1}{nh_1} \sigma^2 R_K f_X^{-1}(x) + o_p((nh_1)^{-1}). \tag{A15}
$$

Mutatis mutandis, similar expressions are obtained for m_2. The conditional bias and variance of $m^{2S2}(x_i, z_i)$ are

$$\begin{aligned}
E(m^{2S2}(x,z) - m(x,z)|\vec{x},\vec{z}) &= \frac{h_1^2}{2}\mu_2 m_1^{(2)}(x) - \frac{1}{2}g_2^2\mu_2 E(m_2^{(2)}(Z)|\vec{x}) \\
&- \frac{1}{2}g_2^2\mu_2 E\left(\int f_X^{(2)}(v)\mu(v,Z)dv|\vec{x}\right) \\
&+ \frac{1}{2}g_2^2\mu_2 E\left(\int \mu(v,Z)f_X(v)f_{XZ}^{-1}(v,Z)\sum_{d=1}^{2}\frac{\partial^2 f_{XZ}(v,Z)}{\partial_d\partial_d}dv|\vec{x}\right) \\
&+ \frac{h_2^2}{2}\mu_2 m_2^{(2)}(z) - \frac{1}{2}g_1^2\mu_2 E(m_1^{(2)}(x_i)|\vec{z}) - \frac{1}{2}g_1^2\mu_2 E\left(\int f_Z^{(2)}(v)\mu(X,v)dv|\vec{z}\right) \\
&+ \frac{1}{2}g_1^2\mu_2 E\left(\int \mu(X,v)f_Z(v)f_{XZ}^{-1}(X,v)\sum_{d=1}^{2}\frac{\partial^2 f_{XZ}(X,v)}{\partial_d\partial_d}dv|\vec{z}\right) \\
&+ o_p(h_1^2) + o_p(g_2^2) + o_p(h_2^2) + o_p(g_1^2)
\end{aligned}$$

and

$$\begin{aligned}
V(m^{2S2}(x,z)|\vec{x},\vec{z}) &= \frac{1}{nh_1}\sigma^2 R_K f_X^{-1}(x) + \frac{1}{nh_2}\sigma^2 R_K f_Z^{-1}(z) \\
&+ o_p\left((nh_1)^{-1}\right) + o_p\left((nh_2)^{-1}\right). \qquad \text{(A16)}
\end{aligned}$$

Proof Let $\epsilon' = (\epsilon_1, \cdots, \epsilon_n)$ where $\epsilon_i = y_i - \alpha - m_1(x_i) - m_2(z_i)$. By construction,

$$\begin{aligned}
m_1^{2S2}(x) &= s_1(x)(\vec{y} - \vec{1}_n\bar{y} - \vec{m}_2^P(\vec{z})) \\
&= s_1(x)\vec{1}_n(\alpha - \bar{y}) + s_1(x)(\vec{m}_1(\vec{x}) + \epsilon) + s_1(x)(\vec{m}_2(\vec{z}) - \vec{m}_2^P(\vec{z})). \qquad \text{(A17)}
\end{aligned}$$

Note that for the first term it is easy to show that $E(\alpha - \bar{y}|\vec{x},\vec{z}) = O_p(n^{-1/2})$, the second term is identical to the first term that appeared in equation (A3) in the proof of Theorem 1. Now we look at the i^{th} element of $-(\vec{m}_2(\vec{z}) - \vec{m}_2^P(\vec{z}))$, which is

$$\begin{aligned}
m_2^P(z_i) - m_2(z_i) &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{g_2}K\left(\frac{z_k - z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}(y_k - \bar{y}) - m_2(z_i) \\
&= \hat{L}_{0n}(z_i) + \hat{L}_{1n}(z_i) + \hat{L}_{2n}(z_i) + \hat{L}_{3n}(z_i),
\end{aligned}$$

where

$$\begin{aligned}
\hat{L}_{0n}(z_i) &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{g_2}K\left(\frac{z_k - z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}(\alpha - \bar{y}) = A_1^n(z_i)(\alpha - \bar{y}) \\
\hat{L}_{1n}(z_i) &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{g_2}K\left(\frac{z_k - z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}\epsilon_k \\
\hat{L}_{2n}(z_i) &= \frac{1}{n}\sum_{k=1}^{n}\frac{1}{g_2}K\left(\frac{z_k - z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k,z_k)}(m_2(z_k) - m_2(z_i))
\end{aligned}$$

$$\hat{L}_{3n}(z_i) = \frac{1}{n}\sum_{k=1}^{n}\left\{\frac{1}{g_2}K\left(\frac{z_k - z_i}{g_2}\right)\frac{\hat{f}_X(x_k)}{\hat{f}_{XZ}(x_k, z_k)}\mu(x_k, z_i) - m_2(z_i)\right\}$$
$$= A_{2n}(z_i) + m_2(z_i)(A_{1n}(z_i) - 1),$$

where $A_{1n}(z_i)$, $A_{2n}(z_i)$, $\hat{L}_{1n}(z_i)$ and $\hat{L}_{2n}(z_i)$ are as defined in the proof of Theorem 1. Hence, using the convergence results of Theorem 1 we obtain the desired expression for the conditional bias of $m_1^{2S2}(x)$. For the conditional variance we note that, for $\bar{\epsilon} = n^{-1}\vec{1}_n'\epsilon$

$$m_1^{2S2}(x) - E(m_1^{2S2}(x)|\vec{x}, \vec{z}) = s_1(x)\left(I - \frac{1}{ng_2}B_n\right)(\epsilon - \bar{\epsilon})$$

and consequently,

$$V(m_1^{2S2}(x)|\vec{x}, \vec{z}) = s_1(x)\left(I - \frac{1}{ng_2}B_n\right)E\left((\epsilon - \bar{\epsilon})(\epsilon - \bar{\epsilon})'|\vec{x}, \vec{z}\right)\left(I - \frac{1}{ng_2}B_n\right)' s_1(x)'$$
$$= \sigma^2 s_1(x)\left(I - \frac{1}{ng_2}B_n\right)\left(I - \frac{1}{ng_2}B_n\right)' s_1(x)'$$
$$- \sigma^2 s_1(x)\left(I - \frac{1}{ng_2}B_n\right)\frac{\vec{1}_n\vec{1}_n'}{n}\left(I - \frac{1}{ng_2}B_n\right)' s_1(x)' \equiv V_1 - V_2$$

$$nh_1 V(m_1^{2S2}(x)|\vec{x}, \vec{z}) = nh_1(V_1 - V_2).$$

The first term in the conditional variance expression is identical to equation (A10) in the proof of Theorem 1, and the second term can easily be shown to be $o_p(1)$. ■