Additive nonparametric regression estimation via *backfitting* and marginal integration under common bandwidth selection criterion: Small sample performance [*]

by

Fernando A. Boeira Sabino da Silva[*]

Abstract

Additive nonparametric regression estimation via backfitting and marginal integration under common bandwidth selection criterion: Small sample performance.

In this paper, we conducted a Monte Carlo investigation to reveal some characteristics of finite sample distributions of the backfitting (B) and Marginal Integration (MI) estimators for an additive bivariate regression. We are particularly interested in

[*]I would like to thank Professor Carlos Martins-Filho for the great contributions throughout this work and Professor Pranab K. Sen for the support. Also, I thank Edward Carlstein, Haipeng Shen, Marcelo Cortes Neri, Cristiano Fernandes and Suman Sen for their comments and suggestions. The author holds responsibility for any remaining errors.

[*]Supported by UNC at Chapel Hill and UFRGS.

providing some evidence on how the different methods for the selection of bandwidth, such as the plug-in method, influence the finite sample properties of the MI and B estimators. We are particularly concerned with the performance of these estimators when bandwidth selection is done based in data driven methods, since in this case the aymptotics properties of these estimators are currently unavailable. The impact of ignoring the dependency between regressors is also investigated. Finally, differently from what occurs at the present time, when the B and MI estimators are used *ad-hoc*, our objective is to provide information that allows for a more accurate comparison of these two competing alternatives in a finite sample setting.

1 Introduction

The estimation of additive nonparametric regressions has been recently discussed in several studies. The hypothesis of additivity is of practical and theoretical interest. From a practical viewpoint, this supposition facilitates interpretation and reduces the computational demand for an unrestricted nonparametric regression. Theoretically speaking, it guarantees rates of convergence for nonparametric estimators that are reasonably quick and independent from the dimensionality problem identified by Friedman & Stuetzle (1981).[1] In addition, with this hypothesis, there is no need to assume some kind of hardly justifiable metric when the variables are measured in different units or are highly correlated (Buja, Hastie & Tibshirani 1989). Currently, there are four viable estimators for an additive nonparametric model - the Backfitting estimator (B-estimator), the Marginal Integration estimator (MI-estimator), a two stage estimator (2S-estimator) and the method called Smooth Backfitting.[2] The B-estimator is based on Friedman & Stuetzle (1981); however, it became popular

[1] Let (X, Y) be a random vector with joint density f, $X \in \mathbb{R}^d$, $Y \in \mathbb{R}$, d is a finite positive integer. Our goal is to estimate $E(Y|X = x) = m(x)$. Stone (1985) has shown if an additive constraint is imposed in $m(x)$, i.e., $E(Y|X = x) = \alpha + \sum_{i=1}^{d} m_i(x_i)$ with $E(m_i(x_i)) = 0$, each of regressions $m_i(\cdot)$ can be estimated at their optimal rate $n^{s/(2s+1)}$ where s is the degree of smoothers of m (which does not depende on d).

[2] The estimators differ in how the additivity constraint is used to produce final estimators of m_i.

through the studies carried out by Hastie & Tibshirani (1986, 1990). Its properties were studied in Buja, Hastie & Tibshirani (1989) and Opsomer & Ruppert (1997). At present, little is known about the statistical properties of the B-estimator. In general, it is still not possible to construct asymptotically valid confidence intervals for the estimated regression, even when the bandwidth $h_n \to 0$ at a desired rate. The knowledge about the B-estimator properties is even scarcer, when h_n is chosen by minimizing the criterion functions most widely used in the literature. Consequently, in practice, little is known about the asymptotic properties and in finite samples of the B-estimator. The MI-estimator was introduced in the seminal articles written by Linton & Nielsen (1995) and Linton & Härdle (1996). One of the most attractive properties of the MI-estimator is that it can be shown to be asymptotically normal when the regressor specific bandwidth h_n converges to zero at a preset rate. Nevertheless, its asymptotic distribution is still unknown when h_n is chosen by data driven methods currently available in the literature, such as cross validation and several plug-in methods, including those proposed by Silverman (1986) and Opsomer & Ruppert (1998). The difficulty in establishing the asymptotic normality in this setting is two-fold. Firstly, data driven h_n are stochastic sequences that may interact detrimentally with regressors and the regressand, which creates an additional difficulty in establishing the asymptotic normality of the MI-estimator. Secondly, data driven h_n are chosen by minimizing a criterion function (loss or risk). For the most widely used

criterion functions, the resulting optimal sequence of h_n, do not converge to zero at the rate that is necessary to obtain asymptotic normality. Just like the B-estimator, little is known, in practice, about both asymptotic and finite sample distributional properties of the MI-estimator. The 2S-estimator (Kim *et al* 1999) is also $n^{s/(2s+1)}$ asymptotically normal under some conditions on the rate of convergence of the bandwidths, but like the MI estimator the 2S-estimator rely on nonstochastic bandwidth. Thus, the same comments made about MI estimator apply to 2S-estimator. Smooth Backfitting was proposed by Mammen, Linton & Nielsen (1999) and Nielsen & Sperlich (2005). This method outperforms the method analyzed in Opsomer and Ruppert and the asymptotic properties are also known under weak conditions.[3]

To make currently available (asymptotic) distributional results useful we have to adapt them to the case in which h_n is a data dependent stochastic sequence. An alternative is to provide experimental evidence of the performance of the estimators based on several methods for the selection of bandwidth h_n by means of a Monte Carlo investigation. Therefore, in this paper, we will conduct a Monte Carlo investigation in order to show some characteristics of the distributions in finite samples of B and MI-estimators for an additive bivariate regression. We are particularly interested in providing some evidence of how the different methods for the selection

[3]In this paper we just compare two estimators: Backfitting (B-estimator) and Marginal Integration (MI-estimator).

of bandwidth h_n, such as plug-in methods, impact the finite sample properties of these estimators. Also, we attempt to offer some evidence of the behavior of different estimators of h_n relatively to the optimal sequence of h_n that minimizes a chosen loss function. The impact of ignoring the dependency between regressors in the estimation of the bandwidth is also investigated. This is common practice and should impact estimators' performance. Finally, differently from what occurs currently when the B and MI-estimators are used *ad-hoc*, the aim is to provide users with information that allows for a more accurate selection of which estimator should be used in a finite sample setting. Besides this introduction the paper has five more sections. Section 2 describes the specification of the model and the two estimators under analysis in a unified format. Section 3 describes the methods for the selection of bandwidth h_n under study. Section 4 presents the data-generating process to be used in the Monte Carlo investigation. Section 5 discusses the results of the analysis. Section 6 provides a brief conclusion.

2 Specification of the Model and the Estimators under Analysis

The statistical model considered here is that of a bivariate additive nonparametric regression adjusted by a local linear smoother. It is assumed that $\left\{(y_t, x_t, z_t)'\right\}_{t=1}^{n}$ form a sequence of realizations of a random vector IID (Y, X, Z) with $E(Y|\ X = x, Z = z) = m_1(x) + m_2(z)$, $V(Y\ |X = x, Z = z) = \sigma^2$ and $E(m_1(X)) = E(m_2(Z)) = 0$. $m_1(\cdot)$ and $m_2(\cdot)$ are real valued functions with some regularity conditions (see Buja, Hastie & Tibshirani 1989), including a suitably chosen degree of differentiability. It is convenient for our purposes to define the following vectors: $Y = (Y_1, ..., Y_n)'$, $X = (X_1, ..., X_n)'$, $Z = (Z_1, ..., Z_n)'$, $\overrightarrow{m}_1(X) = (m_1(X_1), ..., m_1(X_n))'$, $\overrightarrow{m}_2(Z) = (m_2(Z_1), ..., m_2(Z_n))'$, $e_t^k = (0, ..., 1, ..., 0)'$ is a vector of length k, where number one appears in the t^{th} position of the vector, and for any constant c, $\overrightarrow{c}_n = (c, ..., c)'$ is a vector of length n. We denote the marginal densities of X and Z by $f_X(x)$ and $f_Z(z)$ and the joint maginal density of (X, Z) by $f_{XZ}(x, z)$. Also, we denote by $K_d : \mathbb{R}^d \to \mathbb{R}$ a d-variate symmetric kernel function with $d = 1, 2$ and by h_{1n} and h_{2n} the bandwidths associated with the estimation of m_1 and m_2, respectively. By using the previously introduced notation, define two estimating weight functions as:

$$\mathbf{s}_1(x) : \mathbb{R} \to \mathbb{R}^n : \mathbf{s}_1(x) = e_1^{2'} \left(\mathbf{R}_X(x)'\mathbf{V}_X(x)\mathbf{R}_X(x)\right)^{-1} \mathbf{R}_X(x)'\mathbf{V}_X(x)$$

and

$$\mathbf{s}_2(z) : \mathbb{R} \to \mathbb{R}^n : \mathbf{s}_2(z) = e_1^{2'} \left(\mathbf{R}_Z(z)'\mathbf{V}_Z(z)\mathbf{R}_Z(z)\right)^{-1} \mathbf{R}_Z(z)'\mathbf{V}_Z(z) \tag{1}$$

where $V_X(x) = diag\left\{K_1\left(\frac{X_t - x}{h_{1n}}\right)\right\}_{t=1}^{n}$, $R_X(x) = \left(\overrightarrow{1_n}, \overline{x} - \overrightarrow{1_n}x\right)$ and similarly for Z.

Let $\mathbf{S}_1$ and $\mathbf{S}_2$ represent the matrices whose rows are the smoothers at $\mathbf{X}$ and $\mathbf{Z}$:

$$\mathbf{S}_1 = \begin{pmatrix} \mathbf{s}_1(x_1) \\ . \\ . \\ . \\ \mathbf{s}_1(x_n) \end{pmatrix} \text{ and } \mathbf{S}_2 = \begin{pmatrix} \mathbf{s}_2(z_1) \\ . \\ . \\ . \\ \mathbf{s}_2(z_n) \end{pmatrix}$$

Define the vector of the values estimated at points $X_1, ..., X_n$ by $\widehat{\mathbf{m}} = \widehat{\mathbf{m}}_1 + \widehat{\mathbf{m}}_2$, where $\widehat{\mathbf{m}}_1$ and $\widehat{\mathbf{m}}_2$ are the solutions to the following system of estimating equations:

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{m}}_1 \\ \widehat{\mathbf{m}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \end{bmatrix} \mathbf{Y} \tag{2}$$

where $\mathbf{I}_n$ is an identity matrix of dimension n and $\mathbf{S}_d^* = (\mathbf{I}_n - \mathbf{1}\mathbf{1}'/n)\mathbf{S}_d$, $d = 1, 2$.[4] In practice, the system is solved by using the backfitting algorithm, however, in

[4]The adjustment of the smoothers is necessary to guarantee the uniqueness of the solutions (if they exist), see Hastie & Tibshirani(1990).

the bivariate case, when the local linear estimator is used, the backfitting algorithm converges to an explicit solution to $\overrightarrow{\mathbf{m}}_1(\mathbf{X})$ and $\overrightarrow{\mathbf{m}}_2(\mathbf{Z})$ given by

$$\overrightarrow{\mathbf{m}}_1^b(\mathbf{X}) = \left(\mathbf{I}_n - \left(\mathbf{I}_n - \mathbf{S}_1^*\mathbf{S}_2^*\right)^{-1}\left(\mathbf{I}_n - \mathbf{S}_1^*\right)\right)\mathbf{Y}$$

and

$$\overrightarrow{\mathbf{m}}_2^b(\mathbf{Z}) = \left(\mathbf{I}_n - \left(\mathbf{I}_n - \mathbf{S}_2^*\mathbf{S}_1^*\right)^{-1}\left(\mathbf{I}_n - \mathbf{S}_2^*\right)\right)\mathbf{Y} \tag{3}$$

if the inverses exist. The existence of these estimators and their stochastic properties are still, in general, unknown; however, by using the local linear estimator, Opsomer & Ruppert(1997,1998) derived a series of results (for large samples), which is shown below. In our case, there is a solution if:

A1: The kernel K is bounded, continuous, has compact support and its first derivate has a finite number of sign changes over its support. In addition, $\mu_j(K) \equiv \int u^j K(u)du = 0$ for all odd j and $\mu_2(K) \neq 0$.

A2: The densities $f(x,z)$, $f_X(x)$ and $f_Z(z)$ are bounded, continuous and have compact support, and their first derivates have a finite number of sign changes over their supports. Also, $f_X(x) > 0$ and $f_Z(z) > 0$ for all $(x,z) \in \text{supp}(f)$ and

$$\sup\left|\frac{f(x,z)}{f_X(x)f_Z(z)} - 1\right| < 1.$$

A3: When $n \to \infty$, h_{1n}, $h_{2n} \to 0$ and $nh_{1n}/\log(n)$, $nh_{2n}/\log(n) \to \infty$.

A4: The second derivates of m_1 and m_2 exist and are bounded and continuous.

The MI estimator for a bivariate regression function (for an additive model) was proposed in Linton & Nielsen (1995) and it is based on the fact that for any function Q such that $\int dQ(z) = 1$, $\int m(x,z)dQ(z) = m_1(x) + c_1$ where $c_1 = \int m_2(z)dQ(z)$ and similarly for $\int m_2(z)dQ(z)$ and similarly for $\int m(x,z)dQ(x)$. Assuming (y_i, x_i, z_i) are independent and identically distributed, $E(\epsilon_i|x_{i,}z_i) = 0$ and $Var(\epsilon_i|x_{i,}z_i) = \sigma^2$, and $(x_{i,}z_i)$ has joint density $f(x,z)$ and marginals $f_X(x)$ and $f_Z(z)$. The idea of the MI estimator using a bivariate local linear estimator at $X = x$, $Z = z$ is to find an estimator for $\widehat{\mathbf{m}}(x,z;h_{1n},h_{2n}) = e_1^{3'}\left(\mathbf{X}(x,z)'\mathbf{W}\left(x,z\right)\mathbf{X}\left(x,z\right)\right)^{-1}\mathbf{X}(x,z)'\mathbf{W}\left(x,z\right)\mathbf{Y}$, where $\mathbf{X}(x,z) = \left(\overrightarrow{1}_n, X - \overrightarrow{x}, Z - \overrightarrow{z}\right)$ and

$$\mathbf{W}(x,z) = diag\left\{\frac{1}{h_{1n}h_{2n}}K_2\left(\frac{1}{h_{1n}}\left(X_t - x\right), \frac{1}{h_{2n}}\left(Z_t - z\right)\right)\right\}_{t=1}^{n}. \tag{4}$$

using $Q(\cdot)$ to be the empirical distribution function. We then define the matrix

$$\widehat{\mathbf{m}}(\mathbf{X},\mathbf{Z}) = \begin{pmatrix} \widehat{m}(x_1,z_1) & \widehat{m}(x_1,z_2) & . \ . \ . & \widehat{m}(x_1,z_n) \\ \widehat{m}(x_2,z_1) & \widehat{m}(x_2,z_2) & . \ . \ . & \widehat{m}(x_2,z_n) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \widehat{m}(x_n,z_1) & \widehat{m}(x_n,z_2) & . \ . \ . & \widehat{m}(x_n,z_n) \end{pmatrix}.$$

The MI estimator for $\overrightarrow{\mathbf{m}}_1^{mi}(\mathbf{X})$ and $\overrightarrow{\mathbf{m}}_2^{mi}(\mathbf{Z})$, using the identity function as linking function and without considering an intercept (see Linton & Nielsen,1995 and Linton

& Hardle, 1996), is respectively given by $\overrightarrow{\mathbf{m}}_1^{mi}(\mathbf{X}) = \frac{1}{n}\widehat{\mathbf{m}}(\mathbf{X},\mathbf{Z})\overrightarrow{1}_n$, and $\overrightarrow{\mathbf{m}}_2^{mi}(\mathbf{Z}) = \frac{1}{n}\widehat{\mathbf{m}}(\mathbf{X},\mathbf{Z})'\overrightarrow{1}_n$. The weighting functions Q_1 and Q_2 (see Linton & Nielsen, 1995) used for the estimation were the empirical distribution functions $F_{xn}(x)$ and $F_{zn}(z)$ that converge in distribution to $F_X(x)$ and $F_Z(z)$ respectively. The approximations provided in Linton & Nielsen(1995, p.95) are still valid when the empirical functions are written in lieu of Q. Particularly, when x and z are independent, the empirical functions will be the optimal weighting functions in the sense that they minimize the variances of the asymptotic approximations.

The definitions provided above consider h_{1n} and h_{2n} as known nonstochastic sequences that converge to zero at a specified rate. For the B-estimator, Opsomer & Ruppert(1997) show that when, $n \to \infty$, h_{1n}, $h_{2n} \to 0$ and $\frac{nh_{1n}}{\log n}$, $\frac{nh_{2n}}{\log n} \to \infty$ it is possible to obtain an asymptotic approximation to the conditional bias and conditional variance of $\overrightarrow{m}_1^b(X_i)$ and $\overrightarrow{m}_2^b(Z_i)$, where $\overrightarrow{m}_1^b(X_i)$ and $\overrightarrow{m}_2^b(Z_i)$ are the ith elements of $\overrightarrow{\mathbf{m}}_1^b(\mathbf{X})$ and $\overrightarrow{\mathbf{m}}_2^b(\mathbf{Z})$, respectively.[5] For the MI-estimator, Linton & Nielsen (1995) show that when h_{1n}, $h_{2n} \to 0$ and $nh_{1n}h_{2n}^2$, $nh_{2n}h_{1n}^2 \to \infty$, then $\sqrt{nh_{1n}}\left(\overrightarrow{m}_1^{mi}(X_i) - E\left(\overrightarrow{m}_1^{mi}(X_i)\right)\right)$ and $\sqrt{nh_{2n}}\left(\overrightarrow{m}_2^{mi}(Z_i) - E\left(\overrightarrow{m}_2^{mi}(Z_i)\right)\right)$ are asymptotically normal, where $\overrightarrow{m}_1^{mi}(X_i)$ and $\overrightarrow{m}_2^{mi}(Z_i)$ are the ith elements of $\overrightarrow{\mathbf{m}}_1^{mi}(\mathbf{X})$ and $\overrightarrow{\mathbf{m}}_2^{mi}(\mathbf{Z})$, respectively.[6]

[5] The approximation is valid under another three suppositions (see Opsomer & Ruppert,1997).

[6] It is possible to show that the data-driven bandwidth selection methods currently used in the

3 Methods for Data Driven Bandwidth Selection

One of the most important steps in estimating the nonparametric regression models is the selection of smoothing parameters or bandwidths h_n. In essence, once the smoother is selected, the selection of the smoothing parameters is tantamount to the selection of the smooth itself (see Martins-Filho & Bin, 1999 and Silva, 2001). In this paper, two methods for the automatic selection of the bandwidth h_n[7] are considered. These two methods are variants of *plug in* methods[8], that use analytical optimization.

An appropriate error criterion (see Ruppert & Wand, 1994 and Ruppert, Sheather & Wand, 1995) is the weighted conditional $MISE$ given by (in the case of $\mathbf{X}$)

literature, including cross validation and several *plug-in* methods, do not produce sequences $\{h_{1n}\}$ and $\{h_{2n}\}$ that converge to zero at the desired rates. For proofs, see Martins-Filho (2001).

[7] The focus is on h_n fixed within the support used.

[8] An alternative would be the use of cross validation. However, Jones, Marron & Sheather (1996) comment that plug-in methods are better than cross validation methods, in simulation studies and asymptotically. Park, Byeong, Marron (1990), Simonoff (1996) and Opsomer and Ruppert (1998) have shown cross-validation methods possess several undesirable properties. The plug-in methods demand less computational time, do not show undersmoothing of the cross validation method, and the rate of convergence $\left(\widehat{h}_n - h_n\right) \to 0$ when $\widehat{h}_n$ is chosen by plug-in methods is quicker than the rate of convergence of $\left(\widetilde{h}_n - h_n\right) \to 0$ when $\widetilde{h}_n$ is obtained by cross validation.

$$MISE\left(\widehat{m}_p\left(\cdot;h_n\right)|X_1,...,X_n\right) = E\int\left[\left\{\widehat{m}_p(x;h_n)-m(x)\right\}^2 \mid X_1,..,X_n\right] f_X(x)dx. \tag{5}$$

where $f_X(x)$ represents the density of X with support $[a,b]$. Also, assume that the errors are homoskedastic with variance σ^2. For p odd Ruppert & Wand (1994) show that

$$\begin{aligned} MISE\left(\widehat{m}_p\left(\cdot;h_n\right)|X_1,...,X_n\right) &= \left[\frac{h_n^{p+1}\mu_{p+1}\left(K_{(p)}\right)}{(p+1)!}\right]^2\int m^{(p+1)}(x)^2 f_X(x)dx + \\ &\quad \frac{R(K_{(p)})\sigma^2(b-a)}{nh_n} + o_p\left[h_n^{2p+2}+(nh_n)^{-1}\right]. \end{aligned} \tag{6}$$

where $\mu_j(K) = \int u^j K(u)du$, $K_{(p)}(u) = \left\{|M_p(u)|\,/\,|N_p|\right\}K(u)$, N_p is a matrix $(p+1)\times(p+1)$ whose (i,j)th is equal to $\mu_{i+j-2}(K)$, $M_p(u)$ is the same as N_p but with the first column replaced by $(1,u,u^2,...,u^p)'$ and $R(K_{(p)}) \equiv \mu_0\left(K_{(p)}^2\right)$. The minimizer of (6) is asymptotically

$$\widetilde{h}_n = \left[\frac{(p+1)\left(p!\right)^2 R\left(K_{(p)}\right)\sigma^2(b-a)}{2n\mu_{p+1}\left(K_{(p)}\right)^2\int m^{(p+1)}(u)^2 f_X(u)du}\right]^{1|(2p+3)} \tag{7}$$

if $\int m^{(p+1)}(u)^2 f_X(u)du$ is different from zero. A convenient error criterion, which uses only the fitted values at the observation points, is the conditional $MASE$, dis-

cussed by Hardle, Hall & Marron(1988). In the univariate case, the $MASE$ of m can be written as[9]

$$MASE\left(\widehat{m}_p\left(\cdot;h_n\right)|X_1,...,X_n\right)=\frac{1}{n}\sum_{i=1}^{n}E\left\{\left(\left\{\widehat{m}_p(x;h_n)-m(x)\right\}^2\mid X_1,..,X_n\right)\right\}. \quad (8)$$

The basic principle of plug-in methods is the direct estimation of the estimates of σ^2 and of the functionals that appear in the expressions describing the values of the smoothing parameters h_n, after the criterion to be used for the nonparametric estimation has been minimized.

The plug-in method proposed in Linton & Nielsen(1995) is based upon the following rule of thumb (ROT):

$$h_{inROT}=\left\{\frac{\widetilde{\sigma}^2R(K_{(1)})(b_i-a_i)}{\mu_2(K_{(1)})^2\left(\widehat{\theta}_1+\widehat{\theta}_2\right)^2}\right\}^{1/5}n^{-1/5}, \quad (9)$$

where $i=1,2$, b_i and a_i denote the sample maximum and minimum of the regressor of interest, $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are the coefficients of $x^2/2$ and $z^2/2$ obtained from an ordinary least-squares regression of y on a constant, x, z, $x^2/2$, $z^2/2$ and xz, and $\widetilde{\sigma}^2$ is obtained from the residuals of this regression. This rule is asymptotically optimal in terms of the AMISE criterion (see equation 7), when $p=1$, x and z are independent and

[9]Note that (8) is a discrete approximation to (5).

the bivariate regression model $m(x,z)$ is a quadratic function. $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are merely approximations to the second derivate that will appear in (7) when $p=1$.

Another plug-in method used was proposed in Opsomer & Ruppert (1998). The aim, in this case, is to choose $h_{1n}, h_{2n} \in \mathbb{R}$ such that

$$MASE(h_{1n}, h_{2n} \mid \mathbf{X}, \mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} E\left(\widehat{m}(X_i, Z_i) - m(X_i, Z_i) \mid \mathbf{X}, \mathbf{Z}\right)^2. \tag{10}$$

From the corollary 4.2 by Opsomer & Ruppert (1997), the asymptotic approximation to the conditional $MASE$ given in (10)above, when the additive model is fitted by local linear regression, denoted by $AMASE$, is given by:

$$\begin{aligned} AMASE(h_{1n}, h_{2n} \mid \mathbf{X}, \mathbf{Z}) = {} & \frac{\mu_2(K_{(1)})^2}{4}\left(h_{1n}^4\theta_{11} + h_{1n}^2 h_{2n}^2\theta_{12} + h_{2n}^4\theta_{22}\right) + \\ & \sigma^2 R(K_{(1)})\left(\frac{b_x - a_x}{nh_{1n}} + \frac{b_z - a_z}{nh_{2n}}\right) \end{aligned} \tag{11}$$

where

$$\begin{aligned} \theta_{11} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{t}_i' D^2\mathbf{m}_1 + \mathbf{v}_i' E\left(m_1^{(2)}(X_i) \mid \mathbf{Z}\right)\right)^2, \\ \theta_{22} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}_i' D^2\mathbf{m}_2 + \mathbf{t}_i' E\left(m_2^{(2)}(Z_i) \mid \mathbf{X}\right)\right)^2, \\ \theta_{12} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{t}_i' D^2\mathbf{m}_1 + \mathbf{v}_i' E\left(m_1^{(2)}(X_i) \mid \mathbf{Z}\right)\right)\left(\mathbf{v}_i' D^2\mathbf{m}_2 + \mathbf{t}_i' E\left(m_2^{(2)}(Z_i) \mid \mathbf{X}\right)\right)^2 \end{aligned}$$

and $\mathbf{t}_i'$ and $\mathbf{v}_j$ represent the ith row and the jth column of $(\mathbf{I} - \mathbf{T}_{12}^*)^{-1}$, provided the inverse matrix exists and $[\mathbf{T}_{12}^*]_{ij} = \frac{1}{n}\frac{f_{XZ}(X_i, Z_j)}{f_X(X_i) f_Z(Z_j)} - \frac{1}{n}$.

By denoting the values of the bandwindths that minimize $AMASE$ by $h_{1nAMASE}$ and $h_{2nAMASE}$ and under the assumption of independence between $\mathbf{X}$ and $\mathbf{Z}$, it is possible to write

$$h_{1nAMASE} = \left(\frac{R(K_{(1)})\sigma^2 (b_1 - a_1)}{n\mu_2(K_{(1)})^2 \theta_{11}} \right)^{1|5}$$

and

$$h_{2nAMASE} = \left(\frac{R(K_{(1)})\sigma^2 (b_2 - a_2)}{n\mu_2(K_{(1)})^2 \theta_{22}} \right)^{1|5}. \tag{12}$$

The estimation strategy used consists in obtaining the estimates for σ^2 and θ_{ii}, $i = 1, 2$ and directly substitute them in (12). The plug-in rule (PI) used was: σ^2 was estimated by $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(y_t - \widehat{m}_1^b (X_i) - \widehat{m}_2^b (Z_i) \right)^2$ where $\widehat{m}_1^b (X_i)$ and $\widehat{m}_2^b (Z_i)$ are the solutions to the backfitting algorithm given in (3) and $\widehat{\theta}_{11}$ and $\widehat{\theta}_{22}$ were estimated by equation (9) proposed by Opsomer & Ruppert(1998), that is, $\widehat{\theta}_{11} = \frac{1}{n} Tr \mathbf{V}_1^{(2)*} \mathbf{Y}\mathbf{Y}^T \mathbf{V}_1^{(2)*\prime}$ and $\widehat{\theta}_{22} = \frac{1}{n} Tr \mathbf{V}_2^{(2)*} \mathbf{Y}\mathbf{Y}' \mathbf{V}_2^{(2)*\prime}$ where

$$\begin{aligned} \mathbf{V}_1^{(2)} &= \mathbf{S}_1^{(2)} \left(\mathbf{I}_n - S_2^* S_1^* \right)^{-1} \left(\mathbf{I}_n - \mathbf{S}_2^* \right), \mathbf{V}_2^{(2)} = \mathbf{S}_2^{(2)} \left(\mathbf{I}_n - \mathbf{S}_1^* \mathbf{S}_2^* \right)^{-1} \left(\mathbf{I}_n - \mathbf{S}_1^* \right), \\ \mathbf{V}_1^{(2)*} &= \left(\mathbf{I} - \mathbf{1}\mathbf{1}'/n \right) \mathbf{V}_1^{(2)}, \mathbf{V}_2^{(2)*} = \left(\mathbf{I} - \mathbf{1}\mathbf{1}'/n \right) \mathbf{V}_2^{(2)} \end{aligned} \tag{13}$$

and $\mathbf{S}_1^{(2)}$ and $\mathbf{S}_2^{(2)}$ represent the matrices whose rows can be written as $\left(\mathbf{s}_{1,x}^{(2)} \right)' = 2! e_3^{4\prime} \left(\mathbf{R}_X(x)' \mathbf{V}_X(x) \mathbf{R}_X(x) \right)^{-1} \mathbf{R}_X(x)' \mathbf{V}_X(x)$ and $\left(\mathbf{s}_{2,z}^{(2)} \right)' = 2! e_3^{4\prime} \left(\mathbf{R}_Z(z)' \mathbf{V}_Z(z) \mathbf{R}_Z(z) \right)^{-1} \mathbf{R}_Z(z)' \mathbf{V}_Z(z)$. The rule of thumb (ROT) described above was used to estimate the matrices $\mathbf{V}_X(x)$ and $\mathbf{V}_Z(z)$ which appear in $\left(\mathbf{s}_{1,x}^{(2)} \right)'$ and $\left(\mathbf{s}_{2,z}^{(2)} \right)'$. The direct plug-in rule (DPI) used

is described in page 612 (Opsomer and Ruppert, 1998). We obtain the estimates for σ^2 and θ_{ii}, $i = 1, 2$ using second-order approximations.[10]

[10]The PI rule used estimates σ^2 and θ_{ii}, $i = 1, 2$ using first-order approximations.

4 The Data Generating Process

The data used in the study were generated by a bivariate additive nonparametric regression model fitted by local linear regression, with varying correlation to evaluate the robustness to lack of independence between regressors. It is assumed that $\left\{(y_t, x_t, z_t)'\right\}_{t=1}^{n}$ form a sequence of realizations of a $\Re^3$- valued random vector $(\mathbf{Y}, \mathbf{X}, \mathbf{Z})$ and $\{\epsilon_t\}_{t=1}^{n}$ is a sequence of realizations of a random variable with distribution $N(0,1)$. The model used here can be described by

$$Y_t = m_1(X_t) + m_2(Z_t) + \epsilon_t \tag{14}$$

where $m_1(x_t) = -6x_t + 36x_t^2 - 53x_t^3 + 22x_t^5$, $m_2(z_t) = sin(z_t)$,[11] $X_t = S_t$, $Z_t = 5\pi W_t$, with $\{W_t, S_t\}_{t=1}^{n}$ generated by a joint density function with the desired correlation, given by $\begin{pmatrix} W_t \\ S_t \end{pmatrix}$ $N\left(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1/9 & c/9 \\ c/9 & 1/9 \end{pmatrix}\right)$, where $c = 0$ (independence), .25 ("low" correlation), .75 ("high" correlation).

The existence of a solution to the backfitting algorithm is generally unknown, but in the case in which local linear estimators are used, Opsomer & Ruppert (1997, 1998) derived a series of sufficient conditions that guarantee the existence of a single solution in the bivariate case (see conditions A1 to A4 described in section 2).

[11] We chose functions that have very different curvatures.

Because of A2 we rejected all observations for which one of the regressors exceeded $\pm 1.5\sigma$ of the mean (or equivalently outside the interval [0,1]), and in this case, we replaced them by new observations that fell within these limits. We considered samples of 100, 150 and 200 observations[12], each of which was replicated 1600, 1200 and 800 times, respectively.

In this study, a Gaussian kernel[13] was used. Some important results within this context are given next. For the Gaussian kernel, we obtain: $\mu_1(K_{(1)}) = 0$, $\mu_2(K_{(1)}) = 1$ and $R(K_{(1)}) = (2\sqrt{\pi})^{-1}$.

[12] We chose small samples for two reasons. First, the small sample sizes reduce the computational burden in a Monte Carlo setting. Second, we want to evaluate the estimators under very undesirable conditions.

[13] A choice as an Epanechnikov kernel or any kernel with compact support would be desirable since it would satisfy assmption A2 (compact support) that is necessary to guarantee some of the theoretical results regarding these estimators. However, the MI estimator is often not defined in this case due to the singularity of the matrix $W(x, z)$. Linton and Nielsen (1995) and Linton and Härdle (1996) used a Gaussian kernel in their applications even tough they have an explicit assumption on kernel support compactness.

5 Results

A simulation study was carried out to evaluate and compare the performance of the B and MI-estimators in finite samples for a bivariate additive regression. Such study is necessarily restrictive because there are many possibilities regarding the selection of the regression function, the density of regressors, the correlation between them, the error density, the sample size, the type of polynominal regression, the kernel function, the chosen bandwidth, the type of squared error criterion function, among other factors.

By looking at figure 2 by Opsomer & Ruppert(1997, p.191) we can note that the correlation 0.75 is outside the bounds set by assumption A2 of the referred article (p.190), when one normal bivariate distribution is used. Apparently, this does not affect the convergence. This supports the idea that correlation within these bounds, although sufficient, is not a necessary condition for the convergence of backfitting estimators. The kernel function used also does not satisfy condition A1. Likewise, this does not seem to affect the application of the results derived by Opsomer & Ruppert(1997).

The primary aim of the article is to compare the performance of B and MI-estimators in finite samples. For this purpose, we computed the average squared error $ASE = \frac{1}{n}\sum_{t=1}^{n}\left(\widehat{m}_1(X_t) + \widehat{m}_2(Z_t) - m_1(X_t) - m_2(Z_t)\right)^2$ in the simulation studies.

After that, we calculated the mean of replications in order to estimate the $MASE$. By comparing the values presented in Tables 1 and 2 we observed that B-estimators had a better performance than MI-estimators.

Table 1. $MASE$ estimates using backfitting with bandwidths PI and the true $AMASE$

	h_{PI}			h_{AMASE}		
	[14]$n = 100$	$n = 150$	$n = 200$	$n = 100$	$n = 150$	$n = 200$
$\rho = 0$	0.606	0.474	0.396	0.324	0.288	0.262
$\rho = 0.25$	0.600	0.476	0.404	0.323	0.287	0.265
$\rho = 0.75$	0.592	0.470	0.400	0.321	0.282	0.258

[14]The bandwidth h_{2nPI} had overflow problems in the simulation study. The data-generating process was repeated three times when $\rho = 0$ and once when $\rho = 0.25$ and $\rho = 0.75$.

Table 2. $MASE$ estimates using Marginal Integration with ROT bandwidths and the true $AMASE$

	h_{ROT}			h_{AMASE}		
	$n = 100$	$n = 150$	$n = 200$	$n = 100$	$n = 150$	$n = 200$
$\rho = 0$	2.189	0.842	0.629	0.583	0.466	0.365
$\rho = 0.25$	[15]30.530	0.948	1.950	0.679	0.468	0.386
$\rho = 0.75$	11.551	5.186	7.095	3.041	5.979	2.879

By analyzing Tables 1 and 2, it is possible to observe a series of important facts.[16] Firstly, note that the denial of the independence hypothesis between regressors does not affect the estimation made with the backfitting algorithm, no matter if the correlation between regressors is low ($\rho = 0.25$) or high ($\rho = 0.75$). This does not occur when the Marginal Integration is used. In this case, the impact of ignoring dependency remarkably influences the results obtained.

Also, note that the bandwidths used in this Monte Carlo investigation are chosen so as to minimize $MASE$. Thus, the comparison between estimators should be made using the $MASE$ criterion. However, if the median of the replications is used

[15] Frequently the MI estimator is not defined. The problem emerges due to the singularity of the $W(x, z)$ matrix. When the bandwidths were numerically too small we found large $ASE's$ for those replications. It's reasonable to expect that for larger samples this problem would disappear.

[16] As a general result we can notice that increases in sample sizes reduce MASE for all estimators.

to compare the estimators the results show visible differences. The results obtained were somehow expected. Opsomer & Ruppert(1997, p.198) comment that there is an interesting difference between both estimators when $\mathbf{X}$ and $\mathbf{Z}$ are independent. In this case, it is natural to expect that the asymptotic bias of estimators of an additive model for estimating one of the component functions does not depend on the behavior of the other function. Opsomer & Ruppert(1997) show that the B estimator has such property, whereas the MI estimator does not. Except if the bias effects of the component functions happen to offset each other, this will likely result in an increased bias relative to the backfitting estimator. The comparison between asymptotic variances is more straightforward due to the similar format of the expressions for both estimators. In this case, it is possible to show that the asymptotic variance of B-estimators is always smaller than that of MI-estimators, unless $\mathbf{X}$ and $\mathbf{Z}$ are independent.

The comparison between both estimators is clearer when the true bandwidths $h_{1nAMASE}$ and $h_{2nAMASE}$ are used. In a simulation study like this nothing is unknown in (12), that is, there will be no "noise" inherent to the estimation process when the two estimators are compared. In this case, there noticeably exist strong signs of the superiority of B-estimators.

In an attempt to clarify the superiority of B-estimators, the $MASE$of these estimators was calculated using the bandwidths h_{inROT}, $i = 1, 2$, directly. These bandwidths

were constructed in a format that is appropriate for the estimation via Marginal Integration. We suspect that even when using an appropriate rule for the estimation of MI-estimators, the performance of B-estimators would still be superior, which could be confirmed here. Nevertheless, something amazing occurred, as can be observed when we compare Tables 3 and 1 . Apparently, the estimation of the second derivative made in Opsomer & Ruppert(1998) deteriorates the performance of B-estimators in finite samples instead of improving it. Albeit unexpected, the result is interesting, since little is known about the properties of this estimator in finite samples. Table 4 shows the results using the DPI bandwidth proposed in Opsomer and Ruppert (1998). Among all estimated bandwidths used in this study the DPI bandwidth emerges as the best alternative. This superiority is based on an evaluation of the estimators' ASE.

Table 3. $MASE$ estimates using backfitting with ROT bandwidth

	h_{nROT}		
	$n = 100$	$n = 150$	$n = 200$
$\rho = 0$	0.436	0.376	0.328
$\rho = 0.25$	0.437	0.371	0.338
$\rho = 0.75$	0.434	0.367	0.325

Table 4. $MASE$ estimates using backfitting with DPI bandwidths and the true $AMASE$[17]

	h_{ROT}		
	$n = 100$	$n = 150$	$n = 200$
$\rho = 0$	0.344	0.306	0.274
$\rho = 0.25$	0.349	0.298	0.279
$\rho = 0.75$	0.346	0.298	0.266

Figures 1 to 3 show the densities [18] of $\log(h_{inAMASE}) - \log(h_{inPI})$, $\log(h_{inAMASE}) - \log(h_{inROT})$ and $\log(h_{inAMASE}) - \log(h_{inDPI})$, $i = 1, 2$ for the levels of correlation used and for the samples sized 100, 150 and 200, each of which was replicated 800, 600 and 400 times, respectively. As can be observed, the densities for the different levels of correlation are quite close. Seemingly the level of correlation between the covariates has little effect on the estimated bandwidths, which justifies the use of independence assumption in the computation of h_{nPI} , h_{nROT} and h_{nDPI}. Estimators h_{1nPI} and h_{1nDPI} display a very small bias (undersmoothing) in the estimation of m_1 (low-degree polynomial) whereas estimator h_{1nROT} has a stronger bias, causing an oversmoothing in the estimation of m_1. In this case, both estimators have a similar variability. Estimators h_{1nPI} and h_{1nROT} have a marked bias in the estimation of m_2

[17] The samples were replicated 800, 600 and 400 times, respectively.

[18] Estimated by Sheather-Jones (1991) bandwidth.

(undersmoothing)[19], however, the bias of estimators h_{2nPI} is stronger and also those densities display more variability. Estimator h_{nDPI} shows a small bias (oversmoothing) in the estimation of m_2 Estimators h_{nROT} and h_{nDPI} have a similar variability in the estimation of m_1 and m_2[20]. Overall the estimator h_{nDPI} has better relative performance than the other bandwidth estimators used.

[19] m_2 is a sine function (therefore less subject to first-order approximations than m_1). 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 the degree of undersmoothing and oversmoothing seems to increase.

[20] Note that the performance of both estimators improves as the sample size increases.

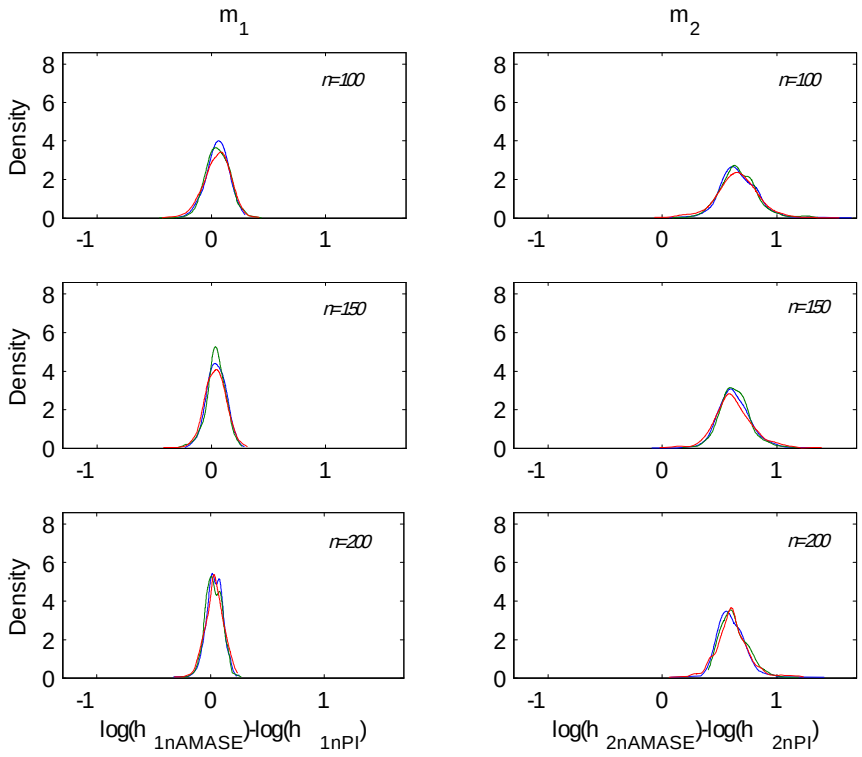

Figure 1: Density of the PI estimators for the three levels of correlation

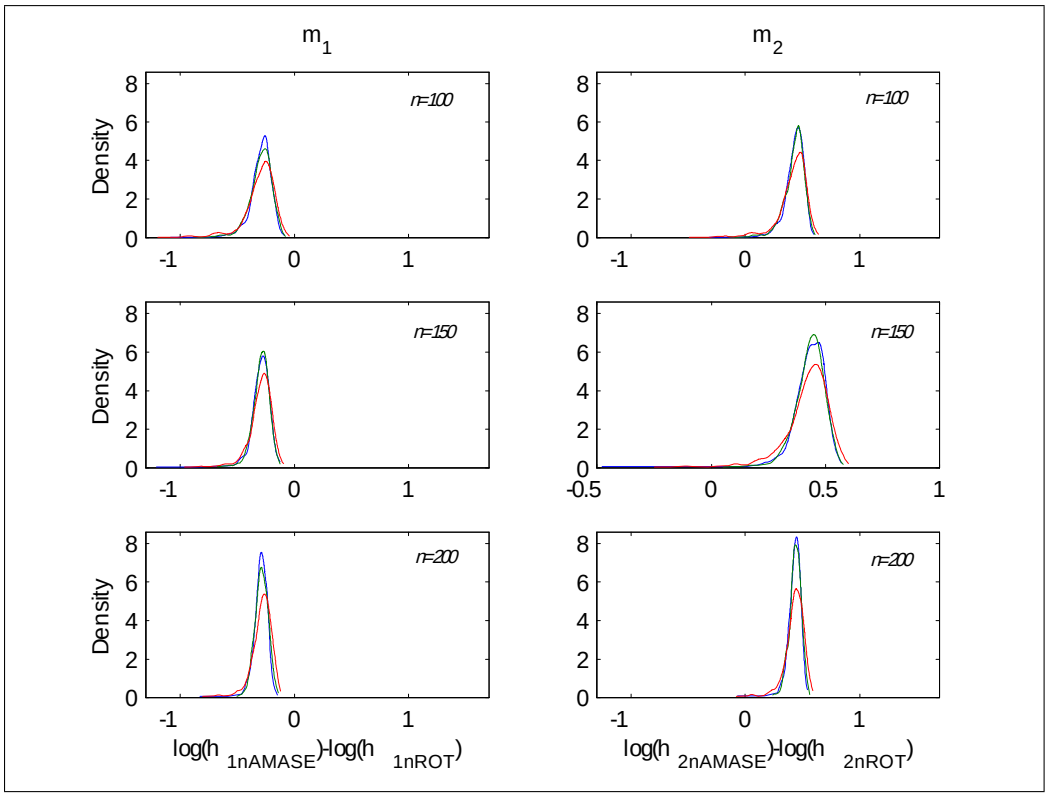

Figure 2: Density of the ROT estimators for the three levels of correlation

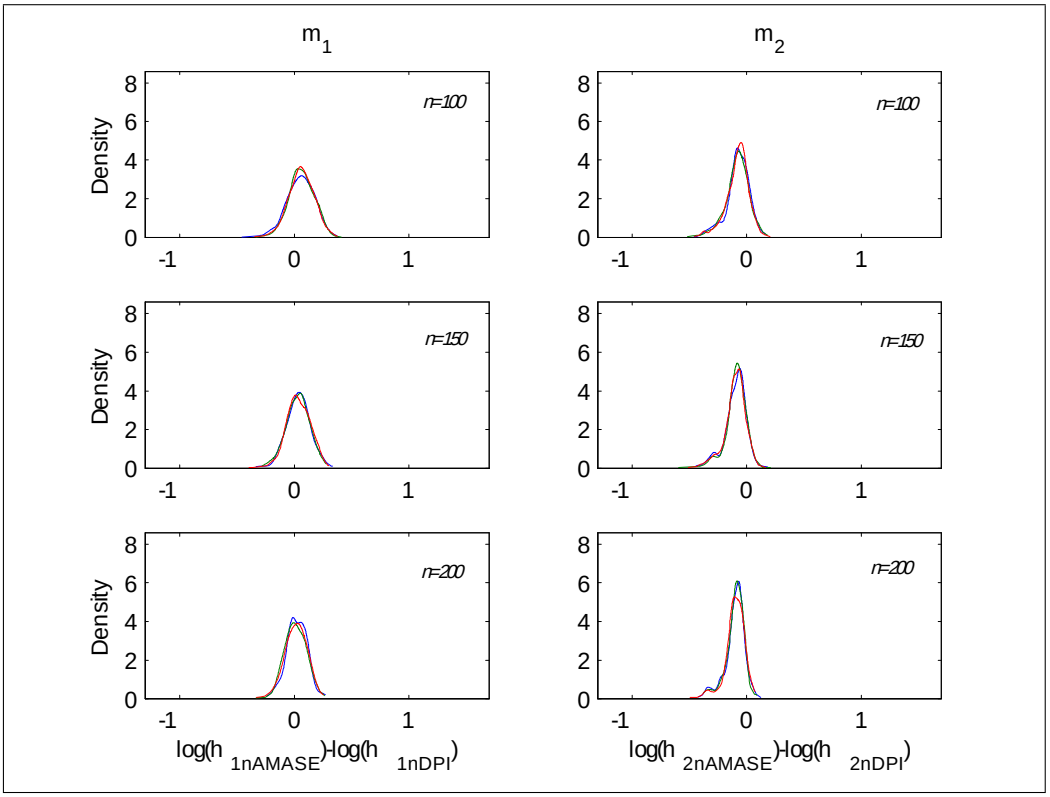

Figure 3: Density of the DPI estimators for the three levels of correlation

6 Conclusions

The current literature proposes, basically, four methods for the estimation of an additive nonparametric regression. In this paper we compared estimator B and MI. The comparison made by means of a Monte Carlo investigation suggests that the B-estimator has a superior performance to the MI- estimator. This superiority is based on the evaluation of the estimators'ASE under true and estimated bandwidths. Although the simulation study presented here has a reduced scope, this is confirmed in a more comprehensive study, see Martins-Filho(2001).[21]

The estimator proposed by Linton & Nielsen is based on an excellent idea, but it involves the product of the bandwidths. In the bivariate case, if the estimates for the two bandwidths are undersmoothed or oversmoothed the effect will be magnified. In addition, as mentioned in Silva(2001, p.16), the estimation via Marginal Integration is computationally more demanding [22], which is inconvenient to the users. In fact, the MI-estimator presents problems associated with unrestricted multivariate regressions,

[21] Aside from the bandwidths used in this article, it is also used cross validation.

[22] The difference between the simulation studies was remarkable. For samples sized 100 a replication with Marginal Integration lasted on average 14.5s. By using *Backfitting* with the bandwidth PI it lasted on average 0.605s, 0.685 with bandwidth DPI and 0.357s when the bandwidth ROT was used. A Pentium 4, 2.8 Ghz, 480MB of RAM was used. The programs were created in Gauss version 3.0 and are available from the author upon request.

which is undesirable.

The main objective of the article was to compare two alternative estimation procedures for the estimation of an additive nonparametric regression. The main findings are summarized below.

1. The lack of the independence assumption between the regressors does not affect the estimation made via the backfitting algorithm. This does not happen when the Marginal Integration is used.

2. An interesting difference between the B estimator and the MI estimator occurs when the $\mathbf{X}$ and $\mathbf{Z}$ regressors are independent. In this case, it is expected that the asymptotic bias of estimators of an additive model for estimating one of the component functions does not depend on the behavior of the other component. The B-estimator has such property, while the MI-estimator does not. For this reason, in general, the MI-estimator will present a stronger bias in relation to the B-estimator.

3. The asymptotic variance of B-estimators is always smaller than that of the MI-estimators, unless the regressors are independent.

4. In general, the MI-estimator needs to compute a higher number of operations than the B-estimator in order to estimate the additive components (see Kim, Linton and Hengartner, 1999), that is, the computational demand of the MI-estimator is greater than that of the B-estimator.

5. In the bivariate case, the MI-estimator involves the products of two bandwidths.

If the estimates of the bandwidths are undersmoothed or oversmoothed the effect will be magnified. This works similarly to the curse of dimensionality.

References

[1] Buja, A., Hastie, T. and Tibshirani, R.(1989). "Linear Smoothers and Additive Models(with discussion)". *Annals of Statistics,* **17,** 453-555.

[2] Friedman, J.H. and Stuetzle, W.(1981). "Projection Pursuit Regression", *Journal of the American Statistical Association,* **76**, 817-823.

[3] Härdle, W., Hall, P. and Marron, J. S.(1988), "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?", *Journal of the American Statistical Association,* **83**, 86-95.

[4] Hastie, T. J. and Tibshirani, R. J.(1986). "Generalized Additive Models". *Statistical Science*, **1**, N°3, 297-318.

[5] Hastie, T. J. and Tibshirani, R. J. *Generalized Additive Models*. Chapman and Hall, Washington, DC, 1990.

[6] Jones, M.C., Marron, J.S. and Sheater, S.J.(1996), "A Brief Survey of Bandwidth Selection for Density Estimation". *Journal of the American Statistical Association,* **91**, 401-407.

[7] Kim, W., Linton, O.B. & Hengartner, N.W.(1999). "A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap

Confidence Intervals". *Journal of Computational and Graphical Statistics,* **8**, 2, 279-297.

[8] Linton, O. and Nielsen, J.P.(1995). "A Kernel Method of Estimating Structured Nonparametric Regression based on Marginal Integration". *Biometrika*, **82**, 1, p.93-100.

[9] Linton, O. and Härdle, W.(1996). "Estimation of additive regression models with known links". *Biometrika*, **83**, 3, p.529-540.

[10] Mammen, E., Linton, O. & Nielsen, J. (1999). "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions". *Annals of Statistics*, **27**, 5, p. 1443-1490.

[11] Martins-Filho, C.B.(2001). "Additive Nonparametric Regression Estimator via *Backfitting* and Marginal Integration: Finite Sample Performance". *Working paper.* OSU, Department of Economics.

[12] Martins-Filho, C.B. & Bin, O.(1999). "Estimation of hedonic price functions via additive nonparametric regression". *Working paper.* OSU, Department of Economics.

[13] Nielsen, J.P. & Sperlich, S. (2005). "Smooth backfitting in practice". *J. R. Statist. Soc. B*, **67**, 1, p.43-61.

[14] Opsomer, J.D. and Ruppert, D.(1997). "Fitting a Bivariate Additive Model by Local Polynomial Regression", *The Annals of Statistics.*, **25**, 1, 186-211.

[15] Opsomer, J.D. and Ruppert, D.(1998). "A Fully Automated Bandwidth Selection Method for Fitting Additive Models", *Journal of the American Statistical Association,* **93**, 442, 605-619.

[16] Park, B. U. & Marron, J. S. (1990). "Comparison of data-driven bandwidth selectors", *Journal of the American Statistical Association*, **85**, 66-72.

[17] Sheather, S. J. and Wand, M. P.(1995), "An Effective Bandwidth Selector for Local Least Squares Regression", *Journal of the American Statistical Association*, **25**, 186-211.

[18] Ruppert, D. and Wand, M. P.(1994), "A Multivariate Locally Weighted Least Squares Regression", *Annals of Statistics,* **22**, 1346-1370.

[19] Sheater, S. J & Jones, M. C.(1991). "A reliable data-based bandwidth selection method for kernel density estimation". *Journal of the Royal. Statistical Society,* Series B*,* **53**, 3, 683-690.

[20] Silva, F. A. B. S. S.(2001). "Estimação de Regressões Aditivas via *Backfitting* e Integração Marginal: Performance em Amostras Finitas", Dissertação de Mestrado, Fundação Getúlio Vargas/EPGE, Rio de Janeiro, Brasil.

[21] Silverman, B.W. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[22] Simonoff, J. *Smoothing Methods in Statistics*, Springer, New York, 1996.

[23] Stone, C. J. (1985). "Additive regression and other nonparametric models", *Annals of Statistics,* **13,** 1, 689-705.