DiSES Working Papers

Università degli Studi di Salerno
Dipartimento di Scienze Economiche e Statistiche
>>>**www.dises.unisa.it**

Bias-corrected inference for multivariate nonparametric regression: model selection and oracle property

Francesco Giordano
Maria Lucia Parrella

Working Paper No. 3.232

ISSN: 1971-3029

Dipartimento di Scienze Economiche e Statistiche
Università Degli Studi di Salerno
Via Ponte Don Melillo – 84084; Fisciano (SA) – Italy

Tel	+39-089-96.21.55
Fax	+39-089-96.20.49
E-mail	dises@unisa.it
Web	www.dises.unisa.it

BIAS-CORRECTED INFERENCE FOR MULTIVARIATE NONPARAMETRIC REGRESSION: MODEL SELECTION AND ORACLE PROPERTY

Francesco Giordano[*] Maria Lucia Parrella[†]

Abstract. The local polynomial estimator is particularly affected by the *curse of dimensionality*. So, the potentialities of such a tool become ineffective for large dimensional applications. Motivated by this, we propose a new estimation procedure based on the local linear estimator and a *nonlinearity sparseness condition*, which focuses on the number of covariates for which the gradient is not constant. Our procedure, called BID for *Bias-Inflation-Deflation*, is automatic and easily applicable to models with many covariates without any additive assumption to the model. It simultaneously gives a consistent estimation of *a*) the optimal bandwidth matrix, *b*) the multivariate regression function and *c*) the multivariate, bias-corrected, confidence bands. Moreover, it automatically identify the relevant covariates and it separates the nonlinear from the linear effects. We do not need *pilot bandwidths*. Some theoretical properties of the method are discussed in the paper. In particular, we show the nonparametric oracle property. For linear models, the BID automatically reaches the optimal rate $O_p(n^{-1/2})$, equivalent to the parametric case. A simulation study shows a good performance of the BID procedure, compared with its direct competitor.

Keywords: multivariate nonparametric regression, multivariate bandwidth selection, multivariate confidence bands.

AMS 2010 classifications: 62G08, 62G05, 62G15, 62G10, 62H99.

JEL classifications: C14, C15, C18, C88.

Acknowledgements: The authors thank professor S.N. Lahiri for helpful comments on this work. This work has been partially supported by the PRIN research grant “La previsione economica e finanziaria: il ruolo dell’informazione e la capacità di modellare il cambiamento” - prot. 2010J3LZEN005.

The authors gratefully acknowledge support from the University of Salerno grant program “Sistema di calcolo ad alte prestazioni per l’analisi economica, finanziaria e statistica (High Performance Computing - HPC)” - prot. ASSA098434, 2009".

1. Introduction

Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a set of $\mathbb{R}^{d+1}$-valued random vectors, where the Y_i are the dependent variables and the $\mathbf{X}_i$ are the $\mathbb{R}^d$-valued covariates of the following model

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i. \tag{1}$$

[*] DISES, Via Ponte Don Melillo, 84084, Fisciano (SA), Italy, giordano@unisa.it

[†] DISES, Via Ponte Don Melillo, 84084, Fisciano (SA), Italy, mparrell@unisa.it

The function $m(\mathbf{X}_i) = E(Y_i|\mathbf{X}_i) : \mathbb{R}^d \to \mathbb{R}$ is the multivariate conditional mean function. The errors ε_i are assumed to be *i.i.d.* and independent of $\mathbf{X}_i$. We use the notation $\mathbf{X}_i = (X_i(1), \ldots, X_i(d))$ to refer to the covariates and $\mathbf{x} = (x_1, \ldots, x_d)$ to denote the target point at which we want to estimate m. We indicate with $f_X(\mathbf{x})$ the density function of the covariate vector, having support $supp(f_X) \subseteq \mathbb{R}^d$ and assumed to be positive. Besides, $f_\varepsilon(\cdot)$ is the density function of the errors, assumed to be $N(0, \sigma_\varepsilon^2)$.

Our goal is to estimate the function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ at a point $\mathbf{x} \in supp(f_X)$, supposing that the parametric form of the function m is completely unknown without imposing *any additive* assumption. We assume that the number of covariates d is high but only some covariates are relevant. The analysis of this framework raises the problem of the *curse of dimensionality*, which usually concerns nonparametric estimators, and consequently the problem of *variable selection*, which is necessary to pursue dimension reduction.

In the last years, many papers have studied this nonparametric framework. A good review is given in Comminges and Dalalyan (2012). For variable selection, we mention the penalty based methods for semiparametric models of Li and Liang (2008) and Dai and Ma (2012); the neural network based method of La Rocca and Perna (2005); the empirical based method of Variyath et al. (2010). Some other methods contextually perform variable selection and estimate the multivariate regression function consistently. See, for example, the COSSO of Lin and Zhang (2006), the ACOSSO of Storlie et al. (2011), the LAND of Zhang et al. (2011) and the RODEO of Lafferty and Wasserman (2008). All these methods are appealing for their approaches, but some typical drawbacks are: the difficulty to analyze theoretically the properties of the estimators; the computational burden; the difficulty to implement the procedures, which generally depend crucially on some regularization parameters, quite difficult to set; the necessity of considering stringent assumptions on the functional space (for example, imposing an additive model).

The aim of this paper is to propose a nonparametric multivariate regression method which mediates among the following priorities: the need of being automatic, the need of scaling to high dimension and the need of adapting to large classes of functions. In order to pursue this, we work around the local linear estimator and its properties. Our work has been inspired by the RODEO method of Lafferty and Wasserman (2008). As a consequence, some of the theoretical results presented in Lafferty and Wasserman (2008) have represented the building blocks of our research. In particular, we borrow the idea of using an iterative procedure in order to "adjust" the multivariate estimation one dimension at a time. Anyway, the BID procedure substantially works differently from the RODEO, since they have different targets. In the RODEO procedure, a technique is proposed in order to check the relevance of the covariate, which is iteratively repeated along each relevant dimension and along a grid of decreasing bandwidths, in order to find the correct *order* of the bandwidth matrix. In this way, it performs (nonlinear) variable selection, bandwidth selection and multivariate function estimation. However, it also leaves some unresolved issues. First of all, it does not identify the relevant *linear covariates*. Moreover, it does not estimate the optimal bandwidth matrix, so its final function estimation is not the optimal one. Finally, it can be applied only with uniform covariates whereas our method can be extended to non-uniform designs. In addition, we improve the rate of convergence of the final estimator, as in Comminges and Dalalyan (2012) and Bertin and Lecué (2008)).

In particular, the contributions of this paper are described in the following.

- we propose a plug-in method for the estimation of the optimal bandwidth matrix which is completely automatic and easily applicable to models with many covariates. It is based on the assumption that each covariate may have a different bandwidth value. Note that the bandwidths have a central role in the proposed BID procedure, since they are used to make variable selection and model selection as well;

- our method has the *nonparametric oracle property*, as defined in Storlie et al. (2011). In particular, it selects the correct subset of predictors with probability tending to one, and estimates the non-zero parameters as efficiently as if the set of relevant covariates were known in advance. Moreover, it automatically separates the *linearities* from the *nonlinearities*. We show that the rate of convergence of the final estimator is not sensitive to the number of relevant *linear covariates* involved in the model (*i.e.*, those for which the partial derivative is constant with respect to the same covariate), even when the model is not additive. As a consequence, the effective dimension of the model can increase, without incurring in the *curse of dimensionality*, as long as the number of *nonlinear covariates* (*i.e.*, those whose partial derivative is not constant with respect to the same covariate) is fixed;

- our procedure includes a consistent bias-corrected estimator for the multivariate regression function, and also for its multivariate confidence bands. These can be used, for example, to make model selection.

- the proposed method does not need any *regularization* parameter, contrary to *LASSO* based techniques, and it does not use any *additive* assumption.

In the next section we introduce the notation. The BID algorithm is presented in section 3. In section 4, we propose a method for the estimation of the optimal bandwidth matrix, while section 5 presents the estimators of the functionals for the derivation of the bias-corrected multivariate confidence bands. Section 6 contains the theoretical results. In section 7 we show a way to remove the uniformity assumption for the design matrix. A simulation study concludes the paper. The assumptions and the proofs are collected in the appendix.

2. Fundamentals of the Local Linear estimators

The BID smoothing procedure is based on the use of the local linear estimator (LLE). The last is a nonparametric tool whose properties have been deeply studied. See Ruppert and Wand (1994), among others. It corresponds to perform a locally weighted least squares fit of a linear function, equal to

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ Y_i - \beta_0(\mathbf{x}) - \boldsymbol{\beta}_1^T(\mathbf{x})(\mathbf{X}_i - \mathbf{x}) \right\}^2 K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}) \tag{2}$$

where the function $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$ gives the local weights and $K(\mathbf{u})$ is the Kernel function, a d-variate probability density function. The $d \times d$ matrix $\mathbf{H}$ represents the smoothing parameter, called the *bandwidth matrix*. It controls the variance of the Kernel function and regulates the amount of local averaging on each dimension, and so the local smoothness of the estimated regression function. Denote with $\boldsymbol{\beta}(\mathbf{x}) = (\beta_0(\mathbf{x}), \boldsymbol{\beta}_1^T(\mathbf{x}))^T$ the vector of coefficients to estimate at point $\mathbf{x}$. Using the matrix notation, the solution of the minimization problem in (2) can be written in closed form

$$\hat{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{H}) = (\boldsymbol{\Gamma}^T \, \mathbf{W} \, \boldsymbol{\Gamma})^{-1} \, \boldsymbol{\Gamma}^T \, \mathbf{W} \, \boldsymbol{\Upsilon}, \tag{3}$$

where $\hat{\boldsymbol{\beta}}(\mathbf{x};\mathbf{H}) = (\hat{\beta}_0(\mathbf{x};\mathbf{H}), \hat{\boldsymbol{\beta}}_1^T(\mathbf{x};\mathbf{H}))^T$ is the estimator of the vector $\boldsymbol{\beta}(\mathbf{x})$ and

$$\boldsymbol{\Upsilon} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\Gamma} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T \\ \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T \end{pmatrix}, \mathbf{W} = \begin{pmatrix} K_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{x}) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & K_{\mathbf{H}}(\mathbf{X}_n - \mathbf{x}) \end{pmatrix}.$$

Let $\mathbb{D}_m(\mathbf{x})$ denote the gradient of $m(\mathbf{x})$. Note from (2) that $\hat{\boldsymbol{\beta}}(\mathbf{x};\mathbf{H})$ gives an estimation of the function $m(\mathbf{x})$ and its gradient. In particular,

$$\hat{\boldsymbol{\beta}}(\mathbf{x};\mathbf{H}) = \begin{pmatrix} \hat{\beta}_0(\mathbf{x};\mathbf{H}) \\ \hat{\boldsymbol{\beta}}_1(\mathbf{x};\mathbf{H}) \end{pmatrix} \equiv \begin{pmatrix} \hat{m}(\mathbf{x};\mathbf{H}) \\ \hat{\mathbb{D}}_m(\mathbf{x};\mathbf{H}) \end{pmatrix}. \tag{4}$$

Despite its conceptual and computational simplicity, the practical implementation of the LLE is not trivial in the multivariate case.

One of the difficulties of the LLE is given by the selection of the smoothing matrix $\mathbf{H}$, which crucially affects the properties of the local polynomial estimator. An optimal bandwidth $\mathbf{H}^{opt}$ exists and can be obtained taking account of the bias-variance trade-off. In order to simplify the analysis, often $\mathbf{H}$ is taken to be of simple form, such as $\mathbf{H} = h\mathbf{I}_d$ or $\mathbf{H} = diag(h_1, \dots, h_d)$, where $\mathbf{I}_d$ is the identity matrix, but even in such cases the estimation of the optimal $\mathbf{H}$ is difficult, because it is computationally cumbersome and because it involves the estimation of some unknown functionals of the process. As a consequence, few papers deal with this topic in the multivariate context (among which Ruppert (1997) and Yang and Tschernig (1999)). One of the contributions of this paper is to propose a novel method for the estimation of the multivariate optimal bandwidth which can be efficiently implemented with many covariates. It is described in sections 4 and 5.

A second problem with the LLE is its bias. In particular, supposing that $\mathbf{x}$ is an interior point and $\mathbf{H} = diag(h_1, \dots, h_d)$, we know from Theorem 2.1 in Ruppert and Wand (1994) that the main terms in the asymptotic expansion of the bias and variance are

$$Abias\{\hat{m}(\mathbf{x};\mathbf{H})|\mathbf{X}_1, \dots, \mathbf{X}_n\} = \frac{1}{2}\mu_2 \sum_{j=1}^{d} \frac{\partial^2 m(\mathbf{x})}{\partial x_j \partial x_j} h_j^2 \tag{5}$$

$$Avar\{\hat{m}(\mathbf{x};\mathbf{H})|\mathbf{X}_1, \dots, \mathbf{X}_n\} = \frac{\rho_0 \sigma_\varepsilon^2}{n f_X(\mathbf{x}) \prod_{j=1}^{d} h_j}, \tag{6}$$

where μ_2 and ρ_0 are moments of the Kernel function, defined as

$$\mu_r = \int u_1^r K(\mathbf{u}) d\mathbf{u}, \qquad \rho_r = \int u_1^r K^2(\mathbf{u}) d\mathbf{u}, \qquad r = 0, 1, 2, \dots.$$

Note from (5) that the bias is influenced by the partial derivatives of m with respect to all the regressors. So, for a finite n, there is a bias component which makes the tests and the confidence intervals based on the LLE not centered around the true value of the function $m(\mathbf{x})$, even when the bandwidth matrix is the optimal one. As a consequence, some bias correction should be considered in order to calibrate the nonparametric inference based on the LLE, but this is difficult to obtain. There are few papers which consider some kind of bias reduction of the multivariate LLE, among which Lin and Lin (2008) and Choi et al. (2000). An interesting contribution of the BID procedure is that it produces a bias corrected estimation of the multivariate function $m(\mathbf{x})$ and its multivariate confidence bands. We explain how in section 5.

```
1. Define the *bias-inflating* bandwidth $\mathbf{H}_U = diag(h_U, \dots, h_U)$, where $h_U$ is a relatively high value (for example, 0.9).

2. Initialise the sets of covariates $C = \{1, 2, \dots, d\}$ and $A = \varnothing$.

3. For each covariate $X(j), j \in C$, do:
   a) using $\mathbf{H}_U$, compute the statistic $Z_j$ and the threshold $\lambda_j$, by (7) and (8)
   b) if $|Z_j| > \lambda_j$ then (*nonlinear covariate*):
        – define the *bias-deflating* bandwidth matrix $\mathbf{H}_L = diag(h_U, \dots, h_L, \dots, h_U)$, which is equal to $\mathbf{H}_U$ except for position $(j, j)$, where $h_L$ is a relatively small value (for example $h_L = h_U d/n$)
        – using $\mathbf{H}_U$ and $\mathbf{H}_L$, estimate the marginal bias $\hat{b}_j(\mathbf{x}, K)$ and the optimal bandwidth $\tilde{h}_j$, as shown in sections 4 and 5
      else
        – set $\tilde{h}_j = h_U$, $\hat{b}_j(\mathbf{x}, K) = 0$ and move the covariate $j$ from $C$ to $A$

4. For each covariate $X(j), j \in A$, do:
   a) using $\tilde{\mathbf{H}} = diag(\tilde{h}_1, \dots, \tilde{h}_d)$, compute the statistic $N_j$ and the threshold $\omega_j$, by equations (9) and (10)
   b) if $|N_j| < \omega_j$ then (*irrelevant covariate*) remove the covariate $j$ from $A$

5. Output:
   a) the final estimated optimal bandwidth $\tilde{\mathbf{H}} = diag(\tilde{h}_1, \dots, \tilde{h}_d)$
   b) the bias corrected estimate $\hat{m}(\mathbf{x}; \tilde{\mathbf{H}}) - \sum_{j=1}^{d} \hat{b}_j(\mathbf{x}, K)\tilde{h}_j^2$
   c) the sets of *nonlinear covariates* $C$ and *linear covariates* $A$.
```

Table 1: The basic BID smoothing algorithm

3. The BID method

The main idea of our procedure is to "explore" the multivariate regression function, searching for relevant covariates. These are divided into: a) the set of *nonlinear covariates* and b) the set of *linear covariates*. The covariates are defined *linear/nonlinear*, depending on the marginal relation between the dependent variable and such covariates, measured by a partial derivative which is constant/nonconstant with respect to the covariate itself.

The box in table 1 reports the steps of the basic algorithm used to analyse the case when all the covariates follow a Uniform distribution, assuming the hypotheses (A1)-(A6) reported in the appendix. In section 7 we extend the applicability of the procedure to those setups where the covariates are not uniformly distributed.

The name BID is an acronym for *Bias-Inflation-Deflation*. The reason for this name to the procedure is the following. The basic engine of the procedure is a double estimation for each dimension, which is included in step 3b and described in detail in sections 4 and 5. In the first estimation, of bias-inflation, we fix all the bandwidths equal to a large value h_U, such that we *oversmooth* along all the directions (note that this is equivalent to estimating locally an hyperplane). In the second estimation, of bias-deflation, we consider a second estimation with a small bandwidth $h_L << h_U$, such that we *undersmooth*. The comparison between the two estimations allows to determine the right degree of "peaks" and "valleys" for the *nonlinear directions*, whereas the *linear directions* remain oversmoothed.

Before describing the procedure in detail, some preliminary considerations are necessary. First of all, it is known that the LLE are usually analyzed under the assumption that $||\mathbf{H}|| \to 0$ when $n \to \infty$ (so the bandwidths of all the covariates must tend to zero for $n \to \infty$). This is required in order to control the bias of $\hat{m}(\mathbf{x}; \mathbf{H})$, so that it can be asymptotically zero. Anyway, we show in Lemma 2 that the bias of $\hat{m}(\mathbf{x}; \mathbf{H})$ does not depend on the bandwidths of the *linear covariates*, because for such covariates the second partial derivative is zero (so the sum in (5) is actually to be taken for $j \in C$). So, in order to gain efficiency, the BID lets the bandwidths of the linear covariates to remain large, while only the bandwidths of the *nonlinear covariates* tend to zero for $n \to \infty$ (see Theorem 1).

The BID procedure performs variable selection through steps 3b and 4b. In particular, step 3b concerns the identification of the *nonlinear covariates*, by means of the derivative expectation statistic Z_j proposed in Lafferty and Wasserman (2008). It is equal to

$$Z_j = \frac{\partial \hat{m}(\mathbf{x}; \mathbf{H})}{\partial h_j} = \mathbf{e}_1^T \mathbf{B} \mathbf{L}_j (\mathbf{I} - \mathbf{\Gamma} \mathbf{B}) \, \mathbf{\Upsilon}, \tag{7}$$

where $\mathbf{L}_j = diag\left(\frac{\partial \log K((X_{1j}-x_j)/h_j)}{\partial h_j}, \ldots, \frac{\partial \log K((X_{nj}-x_j)/h_j)}{\partial h_j}\right)$, $\mathbf{e}_1$ is the unit vector with a one in the first position and $\mathbf{B} = \left(\mathbf{\Gamma}^T \mathbf{W} \mathbf{\Gamma}\right)^{-1} \mathbf{\Gamma}^T \mathbf{W}$. The statistic in (7) reflects the sensitivity of the estimator $\hat{m}(\mathbf{x}; \mathbf{H})$ to the bandwidth of $X(j)$, so it is expected to take non-zero values for the *nonlinear covariates* and null value otherwise. Using the results shown in Lafferty and Wasserman (2008), the threshold can be fixed to

$$\lambda_j = \sqrt{\hat{\sigma}_\varepsilon^2 \mathbf{e}_1^T \mathbf{G}_j \mathbf{G}_j^T \mathbf{e}_1 2 \log n}, \quad j = 1, \ldots, d, \tag{8}$$

where $\mathbf{G}_j = \mathbf{B}\mathbf{L}_j(\mathbf{I} - \mathbf{\Gamma}\mathbf{B})$ and $\hat{\sigma}_\varepsilon^2$ is some consistent estimator of σ_ε^2. After step 3b, the set $A = \overline{C}$ contains both the *linear* and irrelevant covariates. In order to separate them, step 4b performs a threshold condition on the partial derivative coefficients, basing on

$$N_j \equiv \widehat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \tilde{\mathbf{H}}) = \mathbf{e}_{j+1} \tilde{\mathbf{B}} \, \mathbf{\Upsilon}, \quad j \in \overline{C}, \tag{9}$$

where $\tilde{\mathbf{B}}$ is the same matrix as $\mathbf{B}$ replacing the bandwidth matrix $\mathbf{H}$ with the estimated one, $\tilde{\mathbf{H}}$.

Such a statistic is expected to be approximately equal to zero for irrelevant covariates. The normal asymptotic distribution of the local polynomial estimator, shown in Lu (1996), can be used to derive the threshold

$$\omega_j = \sqrt{\hat{\sigma}_\varepsilon^2 \mathbf{e}_{j+1}^T \tilde{\mathbf{B}} \tilde{\mathbf{B}}^T \mathbf{e}_{j+1} 2 \log n}, \tag{10}$$

The statistics in (9) and its distribution derive from well-established results. In particular, the threshold (10) is based on the tail bounds for Normal distribution. One can show that $P\left(|N_j| > \omega_j\right) \to 0$ when $n \to \infty$ if the covariate j is irrelevant. But note that such tests are performed using the estimated bandwidth $\tilde{\mathbf{H}}$, which does not satisfy the classic assumption $||\mathbf{H}|| \to 0$. A theoretical justification of our proposal is thus required and it is given in Lemma 2 (see the appendix). In particular, we show that the bandwidths h_j, when j is an *irrelevant* covariate, do not influence the bias of the estimator $\widehat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \tilde{\mathbf{H}})$, under the assumptions (A1)-(A6). Therefore, such bandwidths can be fixed large, to gain efficiency. This is what is done through the estimated matrix $\tilde{\mathbf{H}}$ (see sections 4 and 5). In the same way, for j a *linear* covariate, the bias of the estimator $\widehat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \tilde{\mathbf{H}})$, under the assumptions (A1)-(A6), does not depend on h_j. So we can also fix large the bandwidths for linear covariates (see Lemma 2 and Corallary 1 in the appendix).

4. The optimal bandwidth matrix

In this section we propose a methodology to estimate the multivariate optimal bandwidth, when it is assumed to be of type $\mathbf{H} = diag(h_1, \dots, h_d)$. Here we assume, for simplicity, that the *nonlinear covariates* are the first k regressors $X(1), \dots, X(k)$.

When considering a given n, the optimal multivariate bandwidth $\mathbf{H}^{opt}$ must be chosen taking account of the bias-variance trade-off. It is defined as

$$\mathbf{H}^{opt} = \arg\min_{\mathbf{H}} \left[Abias^2\{\hat{m}(\mathbf{x}; \mathbf{H})\} + Avar\{\hat{m}(\mathbf{x}; \mathbf{H})\} \right],$$

where, for simplicity, we omit the conditioning on $\mathbf{X}_1, \dots, \mathbf{X}_n$ from the notation. It is known that a simple solution is available if we assume that $\mathbf{H} = h\mathbf{I}_d$. It is given by

$$h^{opt} = \left\{ \frac{d\rho_0 \sigma_\varepsilon^2}{n f_X(\mathbf{x}) \left[\mu_2 \sum_{j=1}^d \frac{\partial m(\mathbf{x})}{\partial x_j \partial x_j} \right]^2} \right\}^{1/(d+4)}. \tag{11}$$

Clearly, the assumption of a common bandwidth for all the dimensions is unsatisfactory, because some of the covariates are assumed to be irrelevant but also because we can observe different curvatures of the function $m(\mathbf{x})$ along the d directions. On the other side, the assumption of a diagonal matrix $\mathbf{H}$ with different bandwidths for the covariates is more realistic, but difficult to deal with when considering its estimation.

The bandwidth selection method proposed here is based on the idea of "marginalizing" the $AMSE$. Let us reformulate (5) and (6) as functions of the bandwidth h_j, conditioned to the other bandwidths $i \neq j = 1, \dots, d$. Denote with $\mathbf{H}_{(j)} = (h_1, \dots, h_{j-1}, h_{j+1}, \dots, h_d)$ the vector of the "given" bandwidths, for $j = 1, \dots, d$. For the asymptotic bias we have

$$\begin{aligned} Abias\{h_j | \mathbf{H}_{(j)}\} &= \frac{1}{2}\mu_2 \sum_{i \neq j}^{d} \frac{\partial^2 m(\mathbf{x})}{\partial x_i \partial x_i} h_i^2 + \frac{1}{2}\mu_2 \frac{\partial^2 m(\mathbf{x})}{\partial x_j \partial x_j} h_j^2 && (12) \\ &= a(\mathbf{x}, K, \mathbf{H}_{(j)}) + b_j(\mathbf{x}, K) h_j^2 && (13) \end{aligned}$$

where $a(\mathbf{x}, K, \mathbf{H}_{(j)}) = \sum_{i \neq j}^{d} b_i(\mathbf{x}, K) h_i^2$ represents the bias cumulated on the axes $i \neq j$. For the variance we have

$$Avar\{h_j | \mathbf{H}_{(j)}\} = \frac{\rho_0 \sigma_\varepsilon^2}{f(\mathbf{x}) \prod_{i \neq j}^{d} h_i} \frac{1}{nh_j} = c(\mathbf{x}, K, \mathbf{H}_{(j)}) \frac{1}{nh_j}, \tag{14}$$

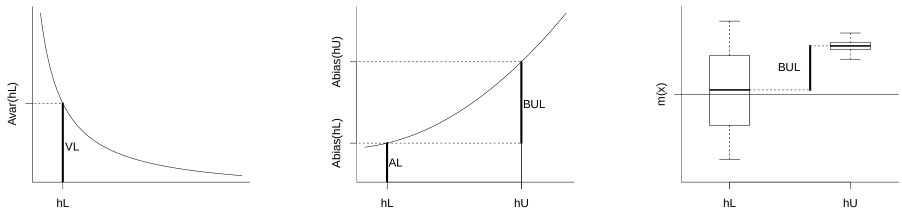
from which the definition of $c(\mathbf{x}, K, \mathbf{H}_{(j)})$ can be clearly deduced. The behaviour of the functions in (13) and (14) are shown in figure 1, plots (a) and (b). The marginal asymptotic mean square error becomes

$$AMSE\{h_j | \mathbf{H}_{(j)}\} = \left[a(\mathbf{x}, K, \mathbf{H}_{(j)}) + b_j(\mathbf{x}, K) h_j^2 \right]^2 + c(\mathbf{x}, K, \mathbf{H}_{(j)}) \frac{1}{nh_j} \tag{15}$$

and the optimal value of the bandwidth h_j is

$$h_j^{opt} = \arg\min_{h_j} AMSE\{h_j | \mathbf{H}_{(j)}\}, \qquad j = 1, \dots, d.$$

There are three different cases, two of which have a trivial solution. The first one is when



(a) Asymptotic variance (b) Asymptotic bias (c) BID mechanism

Figure 1: Asymptotic variance (a) and bias (b) of the local linear estimator $\hat{m}(\mathbf{x};\mathbf{H})$, as a function of the marginal bandwidth h_j. Plot (c) shows the BID mechanism, which derives from the comparison between the LP estimations obtained with the two bandwidths $h_j^L << h_j^U$.

the covariate $X(j)$ is a *linear covariate*, that is when $\partial m(\mathbf{x})/\partial x_j = C_1$, for some value $C_1 \neq 0$ not dependent on $X(j)$. In such case $b_j(\mathbf{x}, K) \equiv 0$ and the (15) is minimized for h_j infinitely large. The second case is when the covariate $X(j)$ is irrelevant. Note that an irrelevant covariate is a special *linear covariate*, for which $\partial m(\mathbf{x})/\partial x_j = C_1 \equiv 0$, so the optimal bandwidth is again infinitely large. The last case is when the variable $X(j)$ is a *nonlinear covariate*, that is for $j = 1, \dots, k$. In such a case (and only in such a case), the optimal bandwidth must be estimated by solving the following equation

$$\frac{\partial\, Abias^2\{h_j|\mathbf{H}_{(j)}\}}{\partial h_j} = -\frac{\partial\, Avar\{h_j|\mathbf{H}_{(j)}\}}{\partial h_j}. \tag{16}$$

In order to solve the (16), we need to approximate in some way the asymptotic bias and variance of the LLE. For example, the cross-validation methods approximate the mean square error by estimating it on a grid of bandwidths and then find the optimal value by minimizing such estimated curve with respect to $\mathbf{H}$. This method is impracticable in multivariate regression, both theoretically and computationally. On the other side, we propose a method which is easily applicable to high dimensional models.

First of all, consider the variance functional in (14). As a function of h_j, its behaviour is depicted in plot (a) of figure 1. Given the (14), we can approximate a generic point of the curve by knowing the value of the function for a given bandwidth h_j. In particular, if we fix a low value of the bandwidth h^L, the asymptotic variance will be

$$Avar\{h^L|\mathbf{H}_{(j)}\} = c(\mathbf{x}, K, \mathbf{H}_{(j)})\frac{1}{nh^L}. \tag{17}$$

From the (17) we can derive a value for $c(\mathbf{x}, K, \mathbf{H}_{(j)})$ and reformulate the (14) as

$$Avar\{h_j|\mathbf{H}_{(j)}\} = Avar\{h^L|\mathbf{H}_{(j)}\}\,\frac{h^L}{h_j} = \overline{V_j^L}\,\frac{h^L}{h_j}. \tag{18}$$

Now consider the bias functional in (13). As a function of h_j, its behaviour is depicted in plot (b) of figure 1. We follow the same arguments as before in order to approximate the asymptotic bias function. Suppose to fix an upper value of the bandwidth $h^U >> h^L$ and to evaluate the function for the two values of bandwidths h^L and h^U. We have

$$Abias\{h^U|\mathbf{H}_{(j)}\} - Abias\{h^L|\mathbf{H}_{(j)}\} = b_j(\mathbf{x}, K)\left[(h^U)^2 - (h^L)^2\right] = \overline{B_j^{UL}}. \tag{19}$$

We can reformulate the (13) as follows

$$\begin{aligned} Abias\{h_j|\mathbf{H}_{(j)}\} &= a(\mathbf{x}, K, \mathbf{H}_{(j)}) + \frac{Abias\{h^U|\mathbf{H}_{(j)}\} - Abias\{h^L|\mathbf{H}_{(j)}\}}{(h^U)^2 - (h^L)^2} h_j^2 \\ &= \overline{A_j^L} + \frac{\overline{B_j^{UL}}}{(h^U)^2 - (h^L)^2} h_j^2. \end{aligned} \tag{20}$$

From the (16), (18) and (20) we obtain

$$4\left[\frac{\overline{B_j^{UL}}}{(h^U)^2 - (h^L)^2}\right]^2 h_j^5 + \frac{4\overline{A_j^L}\,\overline{B_j^{UL}}}{(h^U)^2 - (h^L)^2} h_j^3 - \overline{V_j^L} h_j^L = 0 \tag{21}$$

which represents the estimation equation for the optimal bandwidth h_j^{opt}. The following Lemma is shown in the appendix.

Lemma 1 (Optimal bandwidth matrix)**.** *There is a unique real positive solution to the following system of equations*

$$\begin{cases} 4\left[\frac{\overline{B_1^{UL}}}{(h^U)^2-(h^L)^2}\right]^2 h_1^5 + \frac{4\overline{A_1^L}\,\overline{B_1^{UL}}}{(h^U)^2-(h^L)^2} h_1^3 - \overline{V_1^L} h^L = 0 \\ \vdots \\ 4\left[\frac{\overline{B_k^{UL}}}{(h^U)^2-(h^L)^2}\right]^2 h_k^5 + \frac{4\overline{A_k^L}\,\overline{B_k^{UL}}}{(h^U)^2-(h^L)^2} h_k^3 - \overline{V_k^L} h^L = 0 \end{cases}$$

with respect to the variables $h_1, \ldots, h_k$, where $k < d$ is the number of nonlinear covariates *in model (1). Such a solution identifies the multivariate optimal bandwidth matrix $\mathbf{H}^{opt} = diag(h_1^{opt}, \ldots, h_k^{opt})$.*

5. Estimation of the bias-variance functionals

Following the idea of the plug-in method, we can estimate the marginal optimal bandwidth plugging into the (21) an estimation of the unknown functionals, and then solving the equation with respect to h_j. We can note that the unknown quantities, which are graphically evidenced in figure 1, are

$$\overline{B_j^{UL}} = Abias\{h^U|\mathbf{H}_{(j)}\} - Abias\{h^L|\mathbf{H}_{(j)}\} \tag{22}$$

$$\overline{A_j^L} = Abias\{h^L|\mathbf{H}_{(j)}\} = \sum_{i \neq j} b_i(\mathbf{x}, K) h_i^2 \tag{23}$$

$$\overline{V_j^L} = Avar\{h^L|\mathbf{H}_{(j)}\} \tag{24}$$

and these functionals can also be used to derive the bias-corrected multivariate confidence bands for the function $m(\mathbf{x})$, using the asymptotic normality of the LLE shown in Lu (1996). So, using our BID smoothing procedure, the estimation of the multivariate bandwidth $\mathbf{H}$ and the estimation of the multivariate bias-corrected confidence bands of $m(\mathbf{x})$ have a common solution: the estimation of the functionals in (22)-(24).

Figure 1 explains the idea underlying our proposal for the estimation of these functionals. Plots (a) and (b) show a typical behavior of the asymptotic mean square error of the local linear estimator, for a given axis $1 \leq j \leq k$. For a large value of the bandwidth ($h_j = h^U$), the variance is low but there is much bias (so this is a situation of *bias-inflation*). On the other side, when the bandwidth is low ($h_j = h^L$), a large variance of the estimator is compensated by its low bias (so this is a situation of *bias-deflation*). This is more clearly evidenced by the two box-plots shown in plot (c) of Figure 1, which summarize the typical distributions of the local linear estimations of $m(\mathbf{x})$ for a relevant number of Monte Carlo replications, considering respectively the two bandwidths h^L and h^U. Note that the difference between the medians of the two boxplots reflects the increment in the expected value of the bias observed when increasing the bandwidth h_j from h^L to h^U. Therefore it is proportional to $\overline{B_j^{UL}}$. So, the comparison between the two estimations determines what we call the *Bias-Inflation-Deflation* mechanism.

Following this idea, for the estimation of $\overline{B_j^{UL}}$ we consider the two bandwidth matrices

$$\begin{array}{rcl} \mathbf{H}^U &=& diag(h^U, \ \ldots, \ h^U, \ \ldots, \ h^U) \\ \mathbf{H}_j^L &=& diag(h^U, \ \ldots, \ h^L, \ \ldots, \ h^U) \end{array}$$

which differ only for the value in position (j, j), with $j \in C$, respectively equal to h^U and h^L. Given (19) and Lemma 2 (see the appendix), we can show that

$$\begin{array}{rl} & E\left[\hat{m}(\mathbf{x};\mathbf{H}^U) - \hat{m}(\mathbf{x};\mathbf{H}_j^L)|\mathbf{X}_1,\ldots,\mathbf{X}_n\right] \\ = & Abias\{\hat{m}(\mathbf{x};\mathbf{H}^U)\} - Abias\{\hat{m}(\mathbf{x};\mathbf{H}_j^L)\} + O_p(n^{-1/2}) \\ \approx & b_j(\mathbf{x},K)\left[(h^U)^2 - (h^L)^2\right] = \overline{B_j^{UL}} \end{array}$$

therefore we propose the following estimator of the bias $b_j(\mathbf{x}, K)$ for the axis j

$$\widehat{\overline{B_j^{UL}}} = \hat{m}(\mathbf{x};\mathbf{H}^U) - \hat{m}(\mathbf{x};\mathbf{H}_j^L), \qquad j = 1,\ldots,k \tag{25}$$

$$\hat{b}_j(\mathbf{x},K) = \frac{\widehat{\overline{B_j^{UL}}}}{[(h^U)^2 - (h^L)^2]}. \tag{26}$$

Following the suggestion in Fan and Gijbels (1995), we use the following estimator of the functional $\overline{V_j^L}$

$$\widehat{\overline{V_j^L}} = \hat{\sigma}_\varepsilon^2 \mathbf{e}_1^T(\mathbf{\Gamma}^T\mathbf{W}_j^L\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{W}_j^L\mathbf{W}_j^L\mathbf{\Gamma}(\mathbf{\Gamma}^T\mathbf{W}_j^L\mathbf{\Gamma})^{-1}\mathbf{e}_1, \quad j = 1,\ldots,k \tag{27}$$

where $\mathbf{W}_j^L = diag\left(K_{\mathbf{H}_j^L}(\mathbf{X}_1 - \mathbf{x}),\ldots,K_{\mathbf{H}_j^L}(\mathbf{X}_n - \mathbf{x})\right)$ and $\hat{\sigma}_\varepsilon^2$ is some consistent estimator of σ_ε^2 (see, for example, the estimator proposed in Lafferty and Wasserman (2008)).

Finally, we note from the (23) that the functional $\overline{A_j^L}$ depends on the values of the bandwidths h_i and the biases $b_i(\mathbf{x}; K)$ generated on the axes $i \neq j$. For this reason, we must consider a preliminary step for the estimation of such values. These are obtained considering k univariate bandwidth estimation problems, setting $\overline{A_j^L} = 0$ in the (21)

$${}_0\hat{h}_j^{opt} = \left\{\frac{\widehat{\overline{V_j^L}}h^L}{4}\left[\frac{(h^U)^2 - (h^L)^2}{\widehat{\overline{B_j^{UL}}}}\right]^2\right\}^{1/5} \qquad j = 1,\ldots,k.$$

Now we solve the system of equations in Lemma 1 after plugging the previous estimation of $\overline{B_j^{UL}}$ and $\overline{V_j^L}$ proposed in (25) and (27), and the following estimation of $\overline{A_j^L}$

$${}_s\widehat{\overline{A_j^L}} = \sum_{i \neq j}^{k} {}_{s-1}\hat{h}_i^2 \frac{\widehat{\overline{B_i^{UL}}}}{(h^U)^2 - (h^L)^2}, \qquad s = 1,2,\ldots \tag{28}$$

and then we iterate (increasing s) until convergence. This represents a numerical step which does not require any further kernel estimations, so it is very fast. Our practical experience from the simulation study shows that the convergence is reached in few steps. Note that the component $\overline{A_j^L}$ implies a correction of the optimal bandwidth, by means of the second term in the (21), to take account of the interconnections among the variables. When this component is equal to zero, the formula of the optimal bandwidth h_j^{opt} is equivalent to the one derived in the univariate regression.

Remark 1: the bandwidth estimation procedure proposed here follows a marginalized approach. Anyway, note that the estimated bandwidths are consistent with the optimal multivariate bandwidth $\mathbf{H}^{opt}$ derived with no marginalization. In other words, the estimation procedure, described in this section, is consistent in the sense that it gives the same solutions as in Lemma 1 (see Theorem 1). Moreover, this procedure suggests a fast algorithm to solve the non-linear system in Lemma 1.

6. Theoretical results

In this section we present the theoretical results which justify the BID procedure. In particular, Theorems 1 and 2 together with Remark 2, show the consistency and the rate of convergence in the case of uniform covariates, while Theorems 3 and 4 (see section 7) will consider the non-uniform case.

Considering model (1), let $\tilde{\mathbf{H}}$ be the matrix with the final estimated bandwidths and $\tilde{\mathbf{H}}_k$ be the submatrix with the final estimated bandwidths for the *nonlinear covariates*, assumed to be (for simplicity) the first k on the diagonal of $\tilde{\mathbf{H}}$, that is $\tilde{\mathbf{H}}_k = diag(\tilde{h}_1, \dots, \tilde{h}_k)$. Moreover, assume that $\mathbf{H}_k^{opt}$ is the diagonal matrix with the optimal bandwidths for such k *nonlinear covariates*. The following two Theorems hold.

Theorem 1 (consistency in selection)**.** *Assume that the assumptions (A1)-(A6), reported in the appendix, hold with* $s = 4$. *Then we have*

$$P\left(\tilde{h}_j = h^U, \textit{for all } j > k\right) \to 1 \qquad n \to \infty$$

$$(\tilde{h}_1, \dots, \tilde{h}_k) = O_p\left(n^{-1/(4+k)}\right).$$

Furthermore, $\tilde{\mathbf{H}}_k(\mathbf{H}_k^{opt})^{-1} \xrightarrow{p} \mathbf{I}_k$*, where the convergence in probability is componentwise with* $\mathbf{I}_k$ *the identity matrix of order k.*

Remark 2: (*Oracle property: consistency in selection*) By Theorem 1 we have that $P(\hat{C} = C) \to 1$ as $n \to \infty$, with $\hat{C}$ and C the sets of estimated nonlinear covariates and true ones, respectively. Using the same assumptions as in Theorem 1 with $s = 5$, it follows that

$$P(\hat{A} = A) \to 1, \qquad n \to \infty,$$

where $\hat{A}$ and A are the sets of estimated linear covariates and true ones, respectively (see Table 1, step 4b). This result follows straightforward by applying Lemma 2 and Corollary 1 and the same arguments as in the proofs of Lemmas 7.1, 7.4 and 7.5 in Lafferty and Wasserman (2008). Therefore, Theorem 1 and this Remark lead to the first part of the Oracle property (consistency in selection).

Theorem 2 (Oracle property: rate of convergence)**.** *Under the assumptions of Theorem (1), we have*

$$\left[\hat{m}(\mathbf{x};\tilde{\mathbf{H}}) - m(\mathbf{x})\right]^2 = O_p\left(n^{-4/(4+k)}\right).$$

Note from Theorem 2 that only the *nonlinear covariates* in C have a strong influence on the rate of convergence, since they represent the only dimensions for which the bandwidths shrink the support of the local regression, reducing the number of usable observations. The other bandwidths remain large, so the efficiency of the estimation procedure remarkably improves. As a consequence, in order to avoid the *curse of dimensionality*, we must assume that the number of *nonlinear covariates* k is fixed and relatively small, while the number of relevant variables r can diverge. Note that if $m(\mathbf{x})$ is a linear model ($k = 0$), then the BID estimator reaches the rate $O_p(n^{-1})$ which is equivalent to the parametric case. Moreover, this result is valid for general models, including the *mixed effect* terms.

Remark 3: (*diverging number of covariates*) The proposed selection method as in Theorem 1 and Remark 2 can be extended to the case when the number of covariates, $d \to \infty$. Suppose that $k = O(1)$, $r \leq d$ and $d = O\left(\frac{\log n}{\log\log n}\right)$, the results of Theorem 1 and Remark 2 are still valid using the same arguments as in the proof of Lemma 7.1 in Lafferty and Wasserman (2008). It implies that the number of *linear* covariates can diverge at the same order as d.

7. Extension to non-uniform designs

The test used in step 3a is based on the fundamental assumption that the covariates are uniformly distributed. The reason why this assumption is so crucial can be understood from Lemma 7.1 and Remark 7.2 in Lafferty and Wasserman (2008). In few words, the uniformity assumption simplifies the asymptotic analysis of the local linear estimator, which is particularly hard if one takes into account the unusual assumption that, for some dimensions, the bandwidth may not tend to zero when $n \to \infty$. In order to overcome this problem, and to extend the applicability of both the BID and the RODEO procedures, we propose to consider a transformation of model (1).

Let F_j denote the univariate marginal distribution function of $X(j)$, let F_j^{-1} be its inverse and f_j its density function. Consider the transformed vector of covariates $\mathbf{U}_i = \mathbf{F}_X(\mathbf{X}_i)$, where the function $\mathbf{F}_X(\mathbf{X}_i) : \mathbb{R}^d \to \mathbb{R}^d$ is defined as follows

$$\mathbf{F}_X(\mathbf{X}_i) = (F_1(X_i(1)), \ldots, F_d(X_i(d))).$$

Model (1) can be rewritten as

$$Y_i = m(\mathbf{F}_X^{-1}(\mathbf{U}_i)) + \varepsilon_i = g(\mathbf{U}_i) + \varepsilon_i, \tag{29}$$

where $g = m \cdot \mathbf{F}_X^{-1}$ and $\mathbf{U}_i$ is uniformly distributed on the unit cube. Consider the transformed point of estimation $\mathbf{u} = \mathbf{F}_X(\mathbf{x})$ and note that

$$\frac{\partial g(\mathbf{u})}{\partial u_j} = \frac{\partial m(\mathbf{x})}{\partial x_j}\frac{1}{f_j(u_j)}, \qquad j = 1, \ldots, d. \tag{30}$$

Now, supposing that $f_j(u_j) > 0$, the partial derivatives in (30) are equal to zero if and only if $\partial m(x)/\partial x_j = 0$, that is when the covariate $X(j)$ is irrelevant for m in the point

x. So, $g(u) \equiv m(x)$ and the function g depends on the same covariate as the function m. Therefore, both the problems of multivariate function estimation and variable selection can be solved equivalently using models (1) and (29), but the fundamental difference is that model (29) satisfies the assumptions (A1)-(A6) reported in the appendix. So, the basic BID procedure can be applied consistently replacing the covariates $X(j)$ with $U(j) = F_j(X(j))$, for $j = 1, \ldots, d$, and considering model (29). We call the last case the general BID procedure. There are some drawbacks to be taken into account. The first one is that the transformed model (29) does not have the same structure as the original model, since, for example, the *linear covariates* of model (1) generally become *nonlinear covariates* in model (29). So the two models cannot be used equivalently for model selection. The other drawback is that the distribution functions F_j, needed to transform model (1), are unknown. We can estimate them through the corresponding empirical distribution functions $\hat{F}_j$, but this introduces additional variability in the final estimations, due to the estimation error of F_j. Now, we present the theoretical results that justify the generalised version of the BID procedure.

Considering model (29), let $\tilde{\mathbf{H}}$ be the matrix with the final estimated bandwidths and $\tilde{\mathbf{H}}_r$ be the diagonal matrix with the final estimated bandwidths for the relevant covariates, assumed to be the first r on the diagonal of $\tilde{\mathbf{H}}$, that is $\tilde{\mathbf{H}}_r = diag(\tilde{h}_1, \ldots, \tilde{h}_r)$. Finally, assume that $\mathbf{H}_r^{opt}$ is the diagonal matrix with the optimal bandwidths for such r relevant covariates. The following two Theorems hold.

Theorem 3 (consistency). *Assume that the assumptions (A1)-(A5), reported in the appendix, hold with $s = 5$. Moreover, assume that the Kernel function has a bounded first derivative and the density functions, $f_j(\cdot)$, $j = 1, \ldots, d$, have a bounded fourth derivative. Then we have*

$$P\left(\tilde{h}_j = h^U, \textit{for all } j > r\right) \to 1 \qquad n \to \infty$$

$$(\tilde{h}_1, \ldots, \tilde{h}_r) = O_p\left(n^{-1/(4+r)}\right).$$

Furthermore, $\tilde{\mathbf{H}}_r(\mathbf{H}_r^{opt})^{-1} \xrightarrow{p} \mathbf{I}_r$, where the convergence in probability is componentwise with $\mathbf{I}_r$ the identity matrix of order r.

Theorem 4 (rate of convergence). *Under the assumptions of Theorem (3), we have*

$$\left[\hat{m}(\mathbf{x}; \tilde{\mathbf{H}}) - m(\mathbf{x})\right]^2 = O_p\left(n^{-4/(4+r)}\right).$$

Remark 4: The conditions in Theorem 3 for the first derivative of the Kernel function to be bounded and the assumption (A4) for $s = 5$ are only sufficient, to simplify the proof. It is possible to relax these conditions with the assumption for the Kernel function to be Holder-continuous and assumption (A4) with $s = 4$.

Remark 5: Looking at the proof of Theorem 3, we can still use the BID algorithm with the same threshold λ_j, $j = 1 \ldots, d$, as in (8), in the case of non-uniform covariates.

8. Results from a simulation study

In this section we investigate the empirical performance of the BID procedure. In the first example, we generate datasets from six different models.

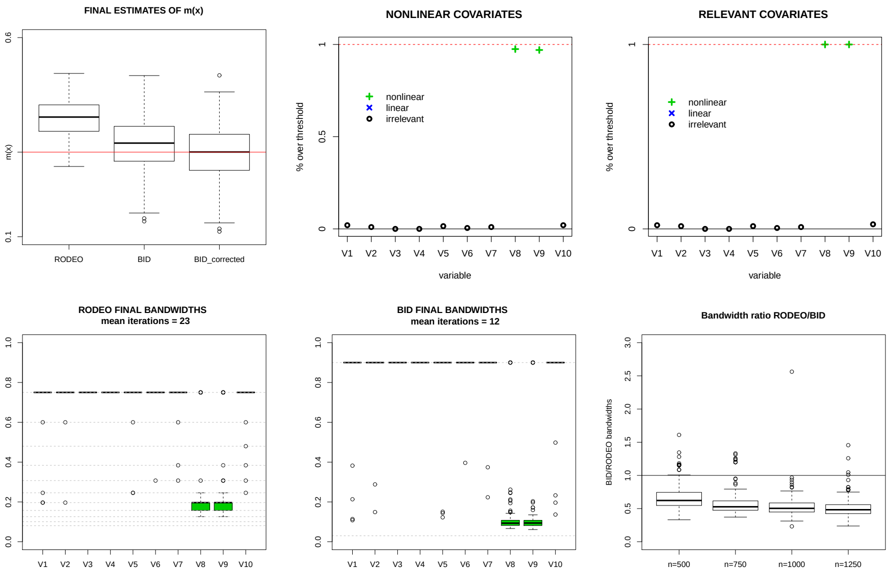


Figure 2: Results for model 1, when $d = 10$ and $n = 750$. On the top: (*left*) the final estimates of $m(\mathbf{x})$ obtained using the RODEO, the BID and the bias-corrected BID methods, respectively; (*center*) for each covariate, the percentage of times that the nonlinearity threshold is exceeded (only the covariates 8 and 9 are nonlinear); (*right*) the percentage of times that the relevance threshold is exceeded (only the covariates 8 and 9 are relevant). On the bottom: results of the final bandwidths, estimated by the RODEO method (*left*) and the BID method (*center*); the box-plots of the ratios between the BID estimated bandwidths and the RODEO estimated bandwidths, for increasing values of n (*right*).

Model	$m(\mathbf{x})$	r	k	Model	$m(\mathbf{x})$	r	k
1	$5x_8^2x_9^2$	2	2	4	$2x_{10}x_1x_2x_3x_4 + 5x_8^2x_9^2$	7	2
2	$2x_{10} + 5x_8^2x_9^2$	3	2	5	$2x_{10}x_1 + x_2x_3x_4 + 5x_8^2x_9^2$	7	2
3	$2x_{10}x_2 + 5x_8^2x_9^2$	4	2	6	$5x_8^2x_9^2 + 2x_{10}^2x_1$	4	3

Model 1 has been used by Lafferty and Wasserman (2008), and it is considered here for comparison with the RODEO results. The other models are variants of the first one, with the addition of some mixed effect terms. We simulate 200 Monte Carlo replications for each model, considering different configurations of settings: the number of covariates equal to $d = (10, 15, 20, 25)$ and the number of observations equal to $n = (500, 750, 1000)$. The number of relevant covariates varies from $r = 2$ to $r = 7$. The remaining $d - r$ covariates are irrelevant, so they are generated independently from Y. Note that the *linear*, the *nonlinear* and the irrelevant covariates are not sequentially sorted, but they are inserted randomly in the models. Finally, all the covariates are uniformly distributed, $f_X \sim U(0, 1)$, and the errors are normally distributed, $f_\varepsilon \sim N(0, 0.5^2)$.

We implement the basic BID procedure. For comparison, we implement the RODEO method as well, using the same settings as in Lafferty and Wasserman (2008). In particular, we use the same value for the β parameter and the same Kernel function $K(u) = (5 - u^2)\mathbb{I}(|u| < \sqrt{5})$. The point of estimation is $\mathbf{x} = (1/2, 1/2, \ldots, 1/2)$.

Figures 2 and 3 show the results of the estimations, for model 1 and 3 respectively, when $d = 10$ and $n = 750$. The plot on the top-left reports three box-plots, which summarize the 200 final estimates of the regression function $m(\mathbf{x})$ obtained using the RODEO method, the BID method and the bias-corrected BID method, respectively. Remember that the RODEO method does not estimate the bias of the LLE, so only the first two box-plots are directly

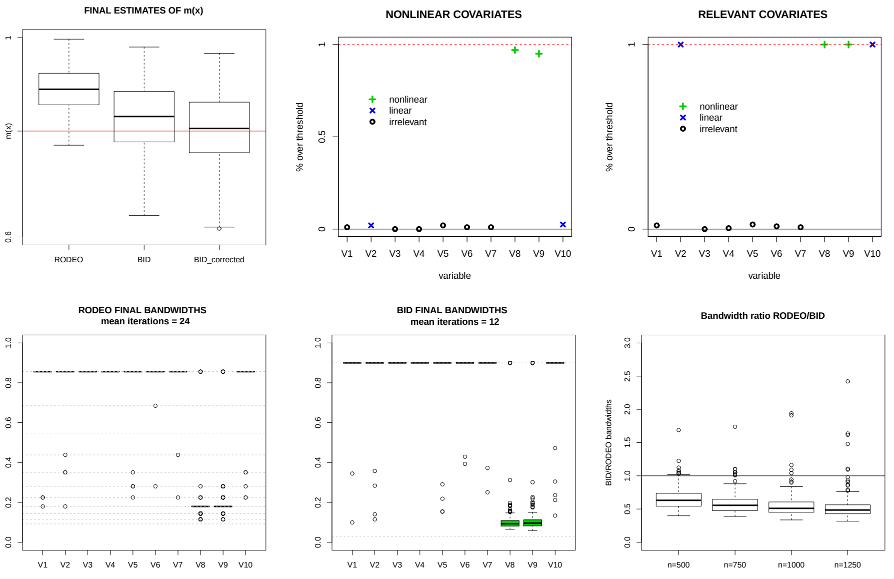


Figure 3: As in figure 2, but for model 3 (for this model, only the covariates 2, 8, 9, and 10 are relevant, among which the covariates 8 and 9 are *nonlinear*).

comparable. The third box-plot, which shows the final bias-corrected estimations obtained with our BID procedure, is reported for completeness. It is evident from these results that the BID method produces better estimations, because it uses an unbiased estimation of the optimal bandwidth matrix, contrary to the RODEO method, which only identifies the correct order of such bandwidth. Moreover, it is evident from the third box-plot, that the bias correction stage is determinant in order to produce good inferential results. The other two plots on the top of each figure show, for each one of the 10 covariates, the percentage of times that the nonlinearity threshold and the relevance threshold are exceeded (steps 3b and 4b of the algorithm, respectively). Note that, for the hard-threshold linearity test (on the top-center of the figures), we desire to hit the one line in the case of *nonlinear covariates* (denoted with the $+$ symbol), and the zero line in the opposite case. So this test works satisfactorily, since it correctly identifies the covariates 8 and 9 as *nonlinear*. On the other side, for the relevance test (on the top-right of the figures), we desire to hit the one line in the case of relevant covariates (which include the *nonlinear covariates*, denoted with the $+$ symbol, and the *linear covariates*, denoted with the $\times$ symbol), and the zero line otherwise.

An important difference between the RODEO and the BID algorithms, which influences the computing time of the two procedures, concerns the total number of iterations. The RODEO method works through a double cycle, since it iterates along the d covariates and then, for each *nonlinear covariate*, along a grid of bandwidths. The width of the grid depends on the parameter $\beta = O(\log n)$ of the RODEO procedure. So, the total number of iterations depends on n and d. On the other side, the BID method iterates only along the d covariates, so the number of iterations depends only on d. As a result, the BID procedure is faster than the RODEO procedure. A comparison between the number of iterations of the two procedures is made in the bottom of figures 2 and 3. Here we compare the results of the final bandwidths estimated with the RODEO method (on the left) and the BID method (in the center). The grid of bandwidths used by RODEO, which is determined by the parameter β, is evidenced by means of the light dashed lines in the plot on the bottom-left (note that such a grid has been fixed as in the paper of Lafferty and Wasserman (2008)). On the other side, the BID method does not use any grid (the two light dashed lines in the central plot indicate the bandwidths h^L and h^U used in the BID mechanism). So, the average number of iterations is different for the two methods: they are reported in the main title of the respective plots.

Finally, the plot on the bottom-right of figures 2 and 3 shows the box-plots of the ratios between the BID bandwidths and the RODEO bandwidths, for increasing values of n. As $n \to \infty$, the ratio tends to a constant value less than one, showing that the order of the two estimated bandwidths is the same, although their values are systematically different. There is an intuitive explanation for this: the BID procedure guarantees an unbiased estimation of the optimal bandwidth matrix, while the RODEO procedure does not have such an objective. So the RODEO estimated bandwidths do not include the constant of optimality but only the order of the optimal bandwidth matrix.

Table 2 reports the mean square error (multiplied by 10), for models 1-6, obtained with the RODEO, the BID and the bias corrected BID methods, for different values of n and d. The results in the table show the consistence of the BID procedure and give evidence of the advantage of bias-correction, especially for small datasets.

		RODEO				BID				BID-corrected			
Model	$d =$	10	15	20	25	10	15	20	25	10	15	20	25
1	$n = 500$	.117	.148	.143	.145	.062	.100	.135	.210	.057	.085	.111	.177
	$n = 750$	.091	.087	.094	.102	.050	.041	.044	.057	.048	.040	.039	.048
	$n = 1000$	.074	.073	.074	.082	.036	.029	.036	.041	.035	.027	.036	.037
2	$n = 500$	.119	.147	.142	.143	.062	.100	.135	.210	.057	.085	.111	.177
	$n = 750$	.088	.085	.089	.097	.050	.041	.044	.057	.048	.040	.039	.048
	$n = 1000$	.072	.069	.075	.079	.036	.029	.036	.041	.035	.027	.036	.037
3	$n = 500$	.124	.152	.147	.144	.071	.102	.136	.209	.065	.087	.112	.179
	$n = 750$	.089	.084	.095	.100	.055	.046	.054	.065	.052	.045	.048	.057
	$n = 1000$	.073	.070	.075	.081	.037	.033	.041	.043	.035	.031	.042	.038
4	$n = 500$	.118	.148	.141	.142	.068	.099	.131	.209	.062	.085	.107	.177
	$n = 750$	.088	.084	.091	.098	.051	.038	.046	.060	.051	.036	.040	.051
	$n = 1000$	.071	.069	.075	.079	.037	.030	.036	.041	.037	.027	.035	.036
5	$n = 500$	.117	.159	.143	.144	.075	.106	.142	.212	.072	.093	.119	.182
	$n = 750$	.090	.085	.093	.010	.056	.046	.055	.064	.055	.046	.051	.056
	$n = 1000$	.071	.071	.075	.079	.042	.039	.042	.042	.042	.037	.042	.039
6	$n = 500$	.300	.348	.332	.332	.195	.227	.301	.428	.169	.188	.259	.374
	$n = 750$	.224	.222	.245	.260	1.140	.146	.140	.155	1.109	.133	.114	.127
	$n = 1000$	.202	.186	.200	.210	.129	.120	.185	.106	.120	.110	.176	.089

Table 2: Mean square error ($\times 10$), for models 1-6, for different values of n and d.

The second experiment considers the case when the covariates are not uniformly distributed. We use the following model

$$m(\mathbf{x}) = \frac{1}{20} x_8^2 x_9^2, \qquad X(j) \sim Exp(2) \qquad j = 1, \dots, d \tag{31}$$

which is equivalent to model 1, but with the covariates exponentially distributed. We replace the coefficient 5 of model 1 with 1/20 in order to have the same signal/error ratio and so to make the results comparable with those of model 1. For model (31), the hard-threshold linearity test of Lafferty and Wasserman (2008) is not consistent. For such a model, table 3 shows the percentage of times that the nonlinearity threshold is exceeded, for different values of n. The three rows on the top refer to the application of the test to the original model in (31), using the non-uniform covariates $X(j)$ and the basic BID procedure, for $n = 500, 750, 1000$ respectively. The three rows on the bottom refer to the application of the test to the transformed model $g(\mathbf{u}) = m(\mathbf{F}_X^{-1}(\mathbf{u}))$ obtained from (31) as explained in section 7. Note that only the covariates 8 and 9 are nonlinear. The hard-threshold nonlinearity test misses the detection of such nonlinearities for the original model, as expected,

		V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Original model	$n = 500$	0	0	0	0	0	0	0	0.005	0.000	0
	$n = 750$	0	0	0	0	0	0	0	0.000	0.005	0
	$n = 1000$	0	0	0	0	0	0	0	0.000	0.005	0
Transformed model	$n = 500$	0	0	0	0	0	0	0	0.705	0.750	0
	$n = 750$	0	0	0	0	0	0	0	0.860	0.875	0
	$n = 1000$	0	0	0	0	0	0	0	0.915	0.935	0

Table 3: Percentages of rejection of the linearity hypothesis in the hard-threshold test of Lafferty and Wasserman (2008), for model (31). The rows on the top refer to the original model in (31), with non-uniform covariates $X(j)$. The rows on the bottom refer to the transformed model, with uniform covariates $\hat{U}(j)$.

while it correctly identifies the nonlinearities for the transformed model. Of course, the percentages for covariates 8th and 9th are lower than those observed in figure 2, since the transformed model is obtained by estimating the distribution function, so an additional estimation error is involved. The other results, for the estimation of the bandwidths and the regression function, are equivalent to those reported in figure 2.

A. Assumptions and proofs

(A1) The bandwidth $\mathbf{H}$ is a diagonal and strictly positive definite matrix.

(A2) The multivariate Kernel function $K(\mathbf{u})$ is a product kernel, based on a univariate kernel function $K(u)$ with compact support, which is non negative, symmetric and bounded; this implies that all the moments of the Kernel exist and that the odd-ordered moments of K and K^2 are zero, that is

$$\int u_1^{i_1} u_2^{i_2} \cdots u_d^{i_d} K^l(\mathbf{u}) d(\mathbf{u}) = 0 \qquad \text{if some } i_j \text{ is odd, for } l = 1, 2. \tag{32}$$

(A3) The second derivatives of $m(\mathbf{x})$ are $|m_{jj}(\mathbf{x})| > 0$, for each $j = 1, \ldots, k$.

(A4) All derivatives of $m(\cdot)$ are bounded up to and including order s.

(A5) $(h^U, \ldots, h^U) \in \mathbb{B} \subset \mathbb{R}^d$ and $(h^L, \ldots, h^L) \in \mathbb{B} \subset \mathbb{R}^d$ with $h^U > h^L > 0$.

(A6) The density function $f_X(\mathbf{x})$ of $(X_1, \ldots, X_d)$ is Uniform on the unit cube.

Proof of Theorem 1. The first part of Theorem 1, $P(h_j = h^U) \to 1$ for $j > k$, follows straightforward by using Lemmas 3 and 4 and the same arguments as in the proof of Lemma 7.5 in Lafferty and Wasserman (2008).

Now, we suppose that $j \leq k$ (nonlinear covariate), for which we have to estimate the optimal multivariate bandwidth. By (17) and (18) we have that $\overline{V_j^L} = O(n^{-1}), \forall j$. Moreover, by Lemma 1, there exists one and only one multivariate optimal bandwidth, say $\{h_1^*, \ldots, h_k^*\}$. It can be shown that $h_i^* = O\left(n^{-\alpha}\right)$, $i = 1, \ldots, k$, with $\alpha > 0$. But $\{h_i^*\}$, $i = 1, \ldots, k$, is the solution of the system in Lemma 1. So, we have to satisfy the following condition

$$O(n^{-5\alpha}) + O(n^{-5\alpha}) = O(n^{-1+(k-1)\alpha}).$$

Note that $\overline{V_j^L} h^L$ contains only $k-1$ bandwidths which tend to zero. In this way, $\alpha = \frac{1}{k+4}$.

Without loss of generality, we can write $h_j^{*U} = h_j^U n^{-\alpha}$ and $h_j^{*L} = h_j^L n^{-\alpha}$, with some $h_j^U > h_j^L > 0, j = 1, \dots, k$. By Lemma 3,

$$E\left(\widehat{\overline{B_j^{UL}}}\right) = \overline{B_j^{UL}} + O(n^{-2\alpha}), \qquad j \leq k.$$

Since the solutions of the system in Lemma 1 are continuous functions of $\widehat{\overline{B_j^{UL}}}$ and $\widehat{\overline{V_j^L}}$, we have that $\tilde{h}_j = O_p(n^{-1/(4+k)})$, $j = 1, \dots, k$. Since $P(j \text{ is nonlinear}) \to 1$, $\forall j \leq k$, when $n \to \infty$, the result follows. □

Proof of Theorem 2. The proof follows the same lines as in Corollary 5.2 in Lafferty and Wasserman (2008) using the results of Theorem 1. □

Proof of Theorem 3. Let $\mathbf{U}_i := \mathbf{F}_X(\mathbf{X}_i) := (F_1(X_i(1)), \dots, F_d(X_i(d)))$, where $F_j(\cdot)$ is the univariate marginal distribution function, $j = 1, \dots, d$. Let $\hat{\mathbf{U}}_i := \hat{\mathbf{F}}_X(\mathbf{X}_i) := \left(\hat{F}_1(X_i(1)), \dots, \hat{F}_d(X_i(d))\right) = \left(\hat{U}_{i1}, \dots, \hat{U}_{id}\right)$, where $\hat{F}_j(\cdot)$ is the empirical distribution function, $j = 1, \dots, d$.

We use the idea of Choi et al. (2000). Let $W_i := \prod_{j=1}^d \frac{1}{h_j} K\left(\frac{x_j - U_{ij}}{h_j}\right)$ as in (7.21) of Lafferty and Wasserman (2008) and $\hat{W}_i := \prod_{j=1}^d \frac{1}{h_j} K\left(\frac{x_j - \hat{U}_{ij}}{h_j}\right)$.

Now we consider the first element in the matrix (7.20a) of Lafferty and Wasserman (2008). Using the Taylor's expansion about U_i, we have

$$\frac{1}{n}\sum_{i=1}^n \hat{W}_i = \frac{1}{n}\sum_{i=1}^n W_i + \frac{1}{n}\sum_{i=1}^n (\tilde{\mathbf{W}}'_i)^T (\hat{\mathbf{U}}_i - \mathbf{U}_i)$$

where $(\tilde{\mathbf{W}}'_i)$ is a d dimension vector of the first derivatives of W_i with respect to U_{ij}, $j = 1, \dots, d$ evaluated in a point, say $\boldsymbol{\eta}_i$, which belongs to a neighborhood of $\mathbf{U}_i$ such that $\|\boldsymbol{\eta}_i\| \leq \|\hat{\mathbf{U}}_i - \mathbf{U}_i\|$, with $\|\cdot\|$ the Euclidean norm.

Let $\hat{A}_{11} := \frac{1}{n}\sum_{i=1}^n \hat{W}_i$ and $A_{11} := \frac{1}{n}\sum_{i=1}^n W_i$. It follows that

$$P\left(\left|\hat{A}_{11} - E(A_{11})\right| > \epsilon s_j(h)\right) \leq P\left(|A_{11} - E(A_{11})| > \epsilon s_j(h)/2\right) + \quad (I)$$

$$+P\left(\left|\frac{1}{n}\sum_{i=1}^n (\tilde{\mathbf{W}}'_i)^T (\hat{\mathbf{U}}_i - \mathbf{U}_i)\right| > \epsilon s_j(h)/2\right) \quad (II),$$

where $s_j^2(h) = \frac{C}{nh_j^2}\prod_{i=1}^d \frac{1}{h_i}$ as in Lemma 7.1 of Lafferty and Wasserman (2008). The constant C is defined in (7.10) of Lafferty and Wasserman (2008). We put $\epsilon = \sqrt{\delta \log n}$ as in Lemma 3. For the part (I), using the proof of Lemma 7.1 in Lafferty and Wasserman (2008), we have that

$$P\left(|A_{11} - E(A_{11})| > \epsilon s_j(h)/2\right) \leq \left(\frac{1}{n}\right)^{c_1}$$

where $0 < c_1 < \infty$ and it is independent of n.

For the second expression in (II), since the dimension of vectors is finite, d, it is sufficient to bound a component of position j, that is

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{W'}_{ij}(\hat{U}_{ij}-U_{ij}) \leq \sup_{x\in\mathbb{R}}\left|\hat{F}_j(x)-F_j(x)\right|\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{W'}_{ij}\right|.$$

Since the Kernel function is defined on a compact set and its first derivative is bounded, it follows that $\frac{1}{n}\sum_{i=1}^{n}\left|\tilde{W'}_{ij}\right| = O_p(1)$, $j = 1,\ldots,d$. Using the Hoeffding's inequality we have that

$$P\left(\sup_{x\in\mathbb{R}}\left|\hat{F}_j(x)-F_j(x)\right| > \epsilon s_j(h)\right) \leq n^{-c_2} \qquad j = 1,\ldots,d,$$

where $0 < c_2 < \infty$ and it is independent of n.

Put $c := \min\{c_1, c_2\}$. Finally, it follows that $I + II \leq n^{-c}$. So, we have the same kind of bound as in Lafferty and Wasserman (2008) and Lemma 3.

Using the arguments above, we can show that the other elements of the matrices in (7.20a), (7.35) and (7.39) of Lafferty and Wasserman (2008) have the same order of convergence as in Lemma 7.1 of Lafferty and Wasserman (2008) and Lemma 3. The results of Lemma 7.4 in Lafferty and Wasserman (2008) and Lemma 4 hold again.

Using the assumptions of this Theorem we can write $m(\mathbf{X}_i) = m\cdot\mathbf{F}_X^{-1}(\mathbf{U}_i) := g(\mathbf{U}_i)$. So that, the assumption (A6) is still valid. Moreover, the arguments above show that we can use the approximation $g(\hat{\mathbf{U}}_i)$. In general, when we consider a *linear covariate* with F_j which is not uniform, the function $g(\cdot)$ becomes non linear. In this way, we can apply Theorem 1 with r non linear covariates. The result follows. □

Proof of Theorem 4. It is sufficient to apply Theorem 2 replacing Theorem 1 with Theorem 3. □

B. Lemmas and Corollaries

To be simple, we arrange the covariates as follows: *nonlinear covariates* for $j = 1,\ldots,k$, *linear covariates* for $j = k+1,\ldots,r$ and irrelevant variables for $j = r+1,\ldots,d$. Moreover, the set of linear covariates A can be further partitioned into two disjoint subsets: the covariates from $k+1$ to $k+s_c$ belong to the subset A_c, which includes those *linear covariates* which are multiplied to other *nonlinear covariates*, introducing *nonlinear mixed effects* in model (1); the covariates from $k+s_c+1$ to r belong to the subset A_u, which includes those *linear covariates* which have a linear additive relation in model (1) or which are multiplied to other *linear covariates*, introducing *linear mixed effects* in model (1). Therefore, $A = A_c \cup A_u$ and $C \cup A \cup U = \{1,\ldots,d\}$. In such a framework, the gradient and the Hessian matrix of the function m become

$$\mathbb{D}_m(\mathbf{x}) = \begin{pmatrix} \mathbb{D}_m^C(\mathbf{x}) \\ \mathbb{D}_m^{A_c}(\mathbf{x}) \\ \mathbb{D}_m^{A_u}(\mathbf{x}) \\ \mathbf{0} \end{pmatrix} \qquad \mathbb{H}_m(\mathbf{x}) = \begin{pmatrix} \mathbb{H}_m^C(\mathbf{x}) & \mathbb{H}_m^{CA_c}(\mathbf{x}) & \mathbf{0} & \mathbf{0} \\ \mathbb{H}_m^{CA_c}(\mathbf{x})^T & \mathbb{H}_m^{A_c}(\mathbf{x}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{H}_m^{A_u}(\mathbf{x}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{33}$$

where $\mathbf{0}$ is a vector or matrix with all elements equal to zero. Note that the matrices $\mathbb{H}_m^C(\mathbf{x})$, $\mathbb{H}_m^{A_c}(\mathbf{x})$ and $\mathbb{H}_m^{A_u}(\mathbf{x})$ are symmetric, whereas the matrix $\mathbb{H}_m^{CA_c}(\mathbf{x})$ is not. Moreover, for additive models without mixed effects, all the sub-matrices in $\mathbb{H}_m(\mathbf{x})$ are zero, except for $\mathbb{H}_m^C(\mathbf{x})$, which is diagonal.

In our analysis, it is also necessary to take account of those terms in the Taylor's expansion of $m(\mathbf{x})$ involving the partial derivatives of order 3 (see the proof of Lemma 2 for the details). To this end, we define the following matrix

$$\mathbb{G}_m(\mathbf{x}) = \begin{pmatrix} \frac{\partial^3 m(\mathbf{x})}{\partial x_1^3} & \frac{\partial^3 m(\mathbf{x})}{\partial x_1 \partial x_2^2} & \cdots & \frac{\partial^3 m(\mathbf{x})}{\partial x_1 \partial x_d^2} \\ \frac{\partial^3 m(\mathbf{x})}{\partial x_2 \partial x_1^2} & \frac{\partial^3 m(\mathbf{x})}{\partial x_2^3} & \cdots & \frac{\partial^3 m(\mathbf{x})}{\partial x_2 \partial x_d^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^3 m(\mathbf{x})}{\partial x_d \partial x_1^2} & \frac{\partial^3 m(\mathbf{x})}{\partial x_d \partial x_2^2} & \cdots & \frac{\partial^3 m(\mathbf{x})}{\partial x_d^3} \end{pmatrix} = \begin{pmatrix} \mathbb{G}_m^C(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbb{G}_m^{A_c C}(\mathbf{x}) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}. \quad (34)$$

Note that the matrix $\mathbb{G}_m(\mathbf{x})$ is not symmetric. Note also that, for additive models, matrix $\mathbb{G}_m^{A_c C}(\mathbf{x})$ is null while matrix $\mathbb{G}_m^C(\mathbf{x})$ is diagonal.

In the same way, let the bandwidth matrix be $\mathbf{H} = diag(\mathbf{H}_C, \mathbf{H}_{A_c}, \mathbf{H}_{A_u}, \mathbf{H}_U)$. Remember that $\|\mathbf{H}_C\| \to 0$ for $n \to \infty$.

Lemma 2. *Under model (1) and assumptions (A1)-(A6), with $s = 5$, the conditional bias of the local linear estimator given by (4) is equal to*

$$E\left\{ \left. \begin{pmatrix} \hat{m}(\mathbf{x}; \mathbf{H}) \\ \hat{\mathbb{D}}_m^C(\mathbf{x}; \mathbf{H}) \\ \hat{\mathbb{D}}_m^{A_c}(\mathbf{x}; \mathbf{H}) \\ \hat{\mathbb{D}}_m^{A_u}(\mathbf{x}; \mathbf{H}) \\ \hat{\mathbb{D}}_m^U(\mathbf{x}; \mathbf{H}) \end{pmatrix} - \begin{pmatrix} m(\mathbf{x}) \\ \mathbb{D}_m^C(\mathbf{x}) \\ \mathbb{D}_m^{A_c}(\mathbf{x}) \\ \mathbb{D}_m^{A_u}(\mathbf{x}) \\ \mathbf{0} \end{pmatrix} \right| \mathbf{X}_1, \ldots, \mathbf{X}_n \right\} = B_m(\mathbf{x}, \mathbf{H}_C) + O_p\left(n^{-1/2}\right),$$

where

$$B_m(\mathbf{x}, \mathbf{H}_C) = \frac{1}{2}\mu_2 \begin{pmatrix} tr\{\mathbb{H}_m^C(x)\mathbf{H}_C^2\} + \nu_1(\mathbf{H}_\mathbf{C}^\mathbf{4}) \\ \mathbb{G}_m^C(\mathbf{x})\mathbf{H}_C^2\mathbf{1} + \left(\frac{\mu_4}{3\mu_2^2} - 1\right) diag\{\mathbb{G}_m^C(\mathbf{x})\mathbf{H}_C^2\}\mathbf{1} + \nu_2(\mathbf{H}_\mathbf{C}^\mathbf{4}) \\ \mathbb{G}_m^{A_c C}(\mathbf{x})\mathbf{H}_C^2\mathbf{1} + \nu_3(\mathbf{H}_\mathbf{C}^\mathbf{4}) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix},$$

where the functions $\nu_1(\cdot) : \mathbb{R}^k \to \mathbb{R}$, $\nu_2(\cdot) : \mathbb{R}^k \to \mathbb{R}^k$ and $\nu_3(\cdot) : \mathbb{R}^k \to \mathbb{R}^{s_c}$ are such that $\nu_1(\mathbf{0}) = 0$, $\nu_2(\mathbf{0}) = \mathbf{0}$ and $\nu_2(\mathbf{0}) = \mathbf{0}$.

Proof: In general, we follow the classic approach used in Ruppert and Wand (1994) and Lu (1996), a part from one substantial difference, *i.e.* we do not assume that the bandwidths tend to zero for $n \to \infty$. This implies that we must bound all the terms of the Taylor expansion with respect to $m(\mathbf{x})$ and with respect to $f_X(\mathbf{x})$, given that the size of the interval around the point x does not vanish with $n \to \infty$. Anyway, assumption (A6) imply that the Taylor expansion is exact with respect to f_X. This simplifies remarkably the proof.

The conditional bias of the LLE is given by

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{H})|\mathbf{X}_1, \ldots, \mathbf{X}_n) - \boldsymbol{\beta}(\mathbf{x}) &= (\boldsymbol{\Gamma}^T \mathbf{W} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \mathbf{W}(\mathbf{M} - \boldsymbol{\Gamma} \boldsymbol{\beta}(\mathbf{x})) \\ &= diag(1, \mathbf{H}^{-1})\mathbf{S}_n^{-1}\mathbf{R}_n \qquad (35) \end{aligned}$$

where $\mathbf{M} = (m(\mathbf{X}_1), \ldots, m(\mathbf{X}_n))$ and, given $\mathbf{u}_t = \mathbf{H}^{-1}(\mathbf{X}_t - \mathbf{x})$, we have

$$\begin{aligned}\mathbf{S}_n &= \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} 1 & \mathbf{u}_t^T \\ \mathbf{u}_t & \mathbf{u}_t\mathbf{u}_t^T \end{pmatrix}|\mathbf{H}|^{-1}K(\mathbf{u}_t)\\ \mathbf{R}_n &= \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} 1 \\ \mathbf{u}_t \end{pmatrix}\left[m(\mathbf{X}_t) - m(\mathbf{x}) - \mathbb{D}_m^T(\mathbf{x})\mathbf{H}\mathbf{u}_t\right]|\mathbf{H}|^{-1}K(\mathbf{u}_t).\end{aligned}$$

For $\mathbf{S}_n$, using Taylor's expansion and assumptions $(A2)$ and (A6), we have

$$\begin{aligned}\mathbf{S}_n &= \int\begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \mathbf{u}\mathbf{u}^T \end{pmatrix}K(\mathbf{u})f_X(\mathbf{x}+\mathbf{H}\mathbf{u})d\mathbf{u} + O_p(n^{-1/2})\\ &= \int\begin{pmatrix} 1 & \mathbf{u}^T \\ \mathbf{u} & \mathbf{u}\mathbf{u}^T \end{pmatrix}K(\mathbf{u}) + O_p(n^{-1/2}) = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mu_2 I_d \end{pmatrix} + O_p(n^{-1/2}). \quad (36)\end{aligned}$$

For the analysis of $\mathbf{R}_n$, we need to introduce some further notation. Suppose that the function $m(\mathbf{x})$ has at least up to order 3 continuous partial derivatives in an open neighborhood of $\mathbf{x} = (x_1, \ldots, x_d)^T$. Let define the kth-order differential $D_m^k(\mathbf{x};\mathbf{y})$ as

$$D_m^k(\mathbf{x},\mathbf{y}) = \sum_{i_1,\ldots,i_d}\frac{k!}{i_1!\times\ldots\times i_d!}\frac{\partial^k m(\mathbf{x})}{\partial x_1^{i_1}\ldots\partial x_d^{i_d}}y_1^{i_1}\times\ldots\times y_d^{i_d},$$

where the summation is over all distinct nonnegative integers $i_1, \ldots, i_d$ such that $i_1 + \ldots + i_d = k$. Using the Taylor's expansion to approximate the function $m(\mathbf{X}_t)$, and assumption (A6), we can write

$$\begin{aligned}R_n &= \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} 1 \\ u_t \end{pmatrix}\left[\frac{1}{2!}D_m^2(\mathbf{x},\mathbf{H}\mathbf{u}_t) + \frac{1}{3!}D_m^3(\mathbf{x},\mathbf{H}\mathbf{u}_t)\right]|\mathbf{H}|^{-1}K(\mathbf{u}_t) + \mathbf{R}_n^*\\ &= \int\begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix}\left[\frac{1}{2!}D_m^2(\mathbf{x},\mathbf{H}\mathbf{u}) + \frac{1}{3!}D_m^3(\mathbf{x},\mathbf{H}\mathbf{u})\right]K(\mathbf{u})f_X(\mathbf{x}+\mathbf{H}\mathbf{u})d\mathbf{u} + \mathbf{R}_n^*\\ &+ O_p(n^{-1/2})\\ &= \int\begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix}\left[\frac{1}{2!}D_m^2(\mathbf{x},\mathbf{H}\mathbf{u}) + \frac{1}{3!}D_m^3(\mathbf{x},\mathbf{H}\mathbf{u})\right]K(\mathbf{u})d\mathbf{u} + \mathbf{R}_n^* + O_p(n^{-1/2})\end{aligned}$$

where $\mathbf{R}_n^*$ represents the residual term, which depends on higher order derivatives of the function $m(\mathbf{x})$. Now, given assumption $(A2)$, some of the terms in the k-th order differentials cancel. We have

$$\begin{aligned}\mathbf{R}_n &= \int\begin{pmatrix} \frac{1}{2!}D_m^2(\mathbf{x},\mathbf{H}\mathbf{u}) \\ \frac{1}{3!}\mathbf{u}D_m^3(\mathbf{x},\mathbf{H}\mathbf{u}) \end{pmatrix}K(\mathbf{u})d\mathbf{u} + \mathbf{R}_n^* + O_p(n^{-1/2})\\ &= \begin{pmatrix} r_1 + r_1^* \\ \mathbf{r}_2 + \mathbf{r}_2^* \end{pmatrix} + O_p(n^{-1/2}), \quad (37)\end{aligned}$$

where the terms r_1^* and $\mathbf{r}_2^*$ comes from $\mathbf{R}_n^*$. Solving the integrals and applying the properties of the Kernel function we have

$$\begin{aligned}r_1 &= \int\frac{1}{2}D_m^2(\mathbf{x},\mathbf{H}\mathbf{u})K(\mathbf{u})d\mathbf{u} = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial^2 m(\mathbf{x})}{\partial x_i\partial x_j}h_ih_j\int u_iu_jK(\mathbf{u})d\mathbf{u}\\ &= \frac{1}{2}\mu_2\sum_{i=1}^{k}\frac{\partial^2 m(\mathbf{x})}{\partial x_i\partial x_i}h_i^2 = \frac{1}{2}\mu_2\, tr\{\mathbb{H}_m^C(\mathbf{x})\mathbf{H}_C^2\};\end{aligned}$$

in the same way, the element of position j of the vector $\mathbf{r}_2$ is

$$\begin{aligned}
r_2^{(j)} &= \int \frac{1}{6} u_r D_m^3(\mathbf{x}, \mathbf{Hu}) K(\mathbf{u}) d\mathbf{u} \\
&= \sum_{i_1,\ldots,i_d} \frac{h_1^{i_1} \cdots h_d^{i_d}}{i_1! \times \ldots \times i_d!} \frac{\partial^3 m(\mathbf{x})}{\partial x_1^{i_1} \cdots \partial x_d^{i_d}} \int u_1^{i_1} \cdots u_r^{i_r+1} \cdots u_d^{i_d} K(\mathbf{u}) d\mathbf{u} \\
&= \left[\sum_{s \neq r} \frac{1}{2} \mu_2^2 \frac{\partial^3 m(\mathbf{x})}{\partial x_r \partial x_s^2} h_r h_s^2 + \frac{1}{6} \mu_4 \frac{\partial^3 m(\mathbf{x})}{\partial x_r^3} h_r^3 \right],
\end{aligned}$$

while the whole vector $\mathbf{r}_2$ is equal to

$$\mathbf{r}_2 = \frac{1}{2} \mu_2^2 \left[\mathbf{H} \mathbb{G}_m(\mathbf{x}) \mathbf{H}^2 + \left(\frac{\mu_4}{3\mu_2^2} - 1 \right) diag\{ \mathbf{H} \mathbb{G}_m(\mathbf{x}) \mathbf{H}^2 \} \right] \mathbf{1}.$$

Following the same arguments, it is easy to show that $r_1^* = \nu_1(\mathbf{H}_C^4)$. Combining the (35), (36) and (37), we obtain

$$E(\hat{\boldsymbol{\beta}}(\mathbf{x}; \mathbf{H}) | \mathbf{X}_1, \ldots, \mathbf{X}_n) - \boldsymbol{\beta}(\mathbf{x}) = diag\{1, \mathbf{H}^{-1}\} \mathbf{S}_n^{-1} \mathbf{R}_n$$

$$\begin{aligned}
&= \begin{pmatrix} r_1 + r_1^* \\ \frac{1}{\mu_2} \mathbf{H}^{-1} (\mathbf{r}_2 + \mathbf{r}_2^*) \end{pmatrix} + O_p(n^{-1/2}) \\
&\approx \begin{pmatrix} \frac{1}{2} \mu_2 \, tr(\mathbb{H}_m^C \mathbf{H}_C^2) + \nu_1(\mathbf{H}_C^4) \\ \frac{1}{2} \mu_2 \mathbb{G}_m \mathbf{H}^2 \mathbf{1} + \left(\frac{\mu_4}{6\mu_2} - \frac{1}{2}\mu_2 \right) diag(\mathbb{G}_m \mathbf{H}^2) \mathbf{1} + \frac{1}{\mu_2} \mathbf{H}^{-1} \mathbf{r}_2^* \end{pmatrix}.
\end{aligned}$$

The result follows after some algebra and splitting the last row in four components, C, A_c, A_u and U, respectively. □

Corollary 1. *Under the assumptions (A1)-(A6), with $s = 5$, the conditional asymptotic bias and the asymptotic variance of the partial derivative estimators $\hat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \mathbf{H})$, defined in (4), are*

$$Abias\{\hat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \mathbf{H})\} = \nu_4(\mathbf{H}_C^2), \quad Avar\{\hat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \mathbf{H})\} = \frac{\sigma_\varepsilon^2 \rho_2}{n |\mathbf{H}| h_j^2}$$

for $j \in C \cup A_C$, with $\nu_4(\cdot) : \mathbb{R}^k \to \mathbb{R}$, $\nu_4(\mathbf{0}) = 0$ and

$$Abias\{\hat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \mathbf{H})\} = 0, \quad Avar\{\hat{\mathbb{D}}_m^{(j)}(\mathbf{x}; \mathbf{H})\} = \frac{\sigma_\varepsilon^2 \rho_2}{n |\mathbf{H}| h_j^2}$$

for $j \in \overline{C \cup A_C}$.

Proof: It is a direct consequence of Lemma 2, using (33) and (34). In fact, using assumptions (A1)-(A6), with $s = 5$, the asymptotic conditional covariance matrix is

$$Cov\left\{ \left. \begin{pmatrix} \hat{\mathbb{D}}_m^C(\mathbf{x}; \mathbf{H}) \\ \hat{\mathbb{D}}_m^{A_c}(\mathbf{x}; \mathbf{H}) \\ \hat{\mathbb{D}}_m^{A_u}(\mathbf{x}; \mathbf{H}) \\ \hat{\mathbb{D}}_m^U(\mathbf{x}; \mathbf{H}) \end{pmatrix} \right| X_1, .., X_n \right\} = \frac{\sigma_\varepsilon^2 \rho_2}{n |\mathbf{H}|} \begin{pmatrix} \mathbf{H}_C^{-2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{A_c}^{-2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_{A_u}^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{H}_U^{-2} \end{pmatrix} (1 + o_p(1)).$$

□

Proof of Lemma 1. Define $d_j := \frac{\overline{B_j^{UL}}}{(h^U)^2-(h^L)^2}$, $A_j\left(\mathbf{h}\right) := \sum_{i\neq j}^{k} d_i h_i^2$ and $V_j := \overline{V_j^L} h^L$, $j = 1, \dots, k$. In this way we can rewrite the equations of Lemma 1 as

$$4d_j^2 h_j^5 + 4A_j\left(\mathbf{h}^{(0)}\right) d_j h_j^3 - V_j = 0, \quad j = 1, \dots, k,$$

for some initial vector $\mathbf{h}^{(0)}$. Fix a j. It is easy to verify that the function $f_j(h_j) := 4d_j^2 h_j^5 + 4A_j\left(\mathbf{h}^{(0)}\right) d_j h_j^3 - V_j$ has one and only one real positive solution, that is $h_j^{(1)}$ such that $f_j(h_j^{(1)}) = 0$. Considering $j \in \{1, \dots, k\}$, we can build the vector $\mathbf{h}^{(1)}$ with the elements $h_j^{(1)}$, given the vector $\mathbf{h}^{(0)}$. So, there exist continuously differentiable functions, g_j, such that $h_j^{(v)} = g_j(\mathbf{h}^{(v-1)})$, $j = 1, \dots, k$, $v \in \mathbb{N}$. Now, using Dini's Theorem, we have

$$\frac{\partial h_j^{(v)}}{\partial h_i^{(v-1)}} = \frac{-2d_i h_i^{(v-1)} h_j^{(v)}}{5d_j (h_j^{(v)})^2 + 3A_j(\mathbf{h}^{(v-1)})} \qquad i \neq j.$$

Note that $\frac{\partial h_j^{(v)}}{\partial h_i^{(v-1)}} = 0$ for $i = j$. But, for increasing values of $v \in \mathbb{N}$, $\{\mathbf{h}^{(v)}\}$ forms a sequence in a compact space of $\mathbb{R}^k$, say S. So, we can extract a subsequence from $\{\mathbf{h}^{(v)}\}$ which is convergent in S, that is $\lim_{n\to\infty} \mathbf{h}^{(v_n)} = \mathbf{h}^* \in S$, with $v_n \to \infty$ when $n \to \infty$.

Without loss of generality, we can consider $S = \{\mathbf{h} : \|\mathbf{h}\| = 1\}$ where $\|\cdot\|$ is the Euclidean norm. Consider the $sign(x)$ function, equal to 1 for positive x and to -1 for negative x. If $sign(d_j A_j(\cdot)) > 0$ then $5d_j(h_j^{(v)})^2 + 3A_j(\mathbf{h}^{(v-1)})$ has a minimum at $h_j^{(v)} = 0$. It follows that

$$\left|\sum_{i=1}^{k} \frac{\partial h_j^{(v)}}{\partial h_i^{(v-1)}} h_i^{(v-1)}\right| = \left|\frac{-2h_j^{(v)} A_j(\mathbf{h}^{(v-1)})}{5d_j(h_j^{(v)})^2 + 3A_j(\mathbf{h}^{(v-1)})}\right| \leq \frac{2}{3}.$$

If $sign(d_j A_j(\cdot)) < 0$, using $h_j^{(v)} > \sqrt{-3A_j(\cdot)/(5d_j)}$, we obtain

$$\sum_{i=1}^{k} \frac{\partial h_j^{(v)}}{\partial h_i^{(v-1)}} h_i^{(v-1)} = \frac{-2h_j^{(v)} A_j(\mathbf{h}^{(v-1)})}{5d_j(h_j^{(v)})^2 + 3A_j(\mathbf{h}^{(v-1)})} > 0 \qquad \forall v.$$

So, in this case, $\{h_j^{(v)}\}$ is a monotone sequence with respect to v. Since $\lim_{n\to\infty} h_j^{(v_n)} = h_j^*$, a component of the vector $\mathbf{h}^*$, it follows that $\lim_{v\to\infty} h_j^{(v)} = h_j^*$. Using these arguments, we can conclude that there exists one and only one solution, $\mathbf{h}^*$. □

We have to state some technical lemmas to prove the Theorems 1 and 2. First of all, we introduce the following quantities. Consider a vector $\mathbf{h} = (h_1, \dots, h_k)$. Let $q(\mathbf{h}) := \sum_{j=1}^{k} p_j h_j^2$ and $q(\mathbf{h}; y_i) := \sum_{j\neq i}^{k} p_j h_j^2 + p_i y_i^2$, for some $k < d$ and $0 < p_j < \infty, \forall j$. Let $R_t(K) := \int K(u)K(w_t u)du$, with $w_0 = 1$ and $w_1 = h^L/h^U$. Define

$$s_1^2(h^U) := \frac{1}{n(h^U)^d}\sigma_\epsilon^2 R_0(K),\ s_2^2(h^U; h^L) := s_1^2(h^U)/w_1,\ s_3^2(h^U; h^L) := s_1^2(h^U)\frac{R_1(K)}{R_0(K)}.$$

Lemma 3. *For every $\boldsymbol{h}^U = (h^U, \ldots, h^U) \in \mathbb{B}$ and $\boldsymbol{h}^L = (h^L, \ldots, h^L) \in \mathbb{B}$, vectors of dimension d, if the assumptions (A1)-(A6) hold, with $s = 4$, then*

$$\begin{aligned} E\left(\widehat{\overline{B_j^{UL}}}\right) &= \overline{B_j^{UL}} + O\left(q(\boldsymbol{h}^U)(h^U)^2 - q(\boldsymbol{h}^U; h^L)(h^L)^2\right) & j \leq k, \\ E\left(\widehat{\overline{B_j^{UL}}}\right) &= 0 & j > k. \end{aligned}$$

Moreover, $\forall \delta > 0$

$$P\left(\left|\frac{\widehat{\overline{B_j^{UL}}} - E\left(\widehat{\overline{B_j^{UL}}}\right)}{s_B(\boldsymbol{h}^U; h^L)}\right| > \sqrt{\delta \log n}\right) \leq 2n^{-\delta\sigma_\epsilon^2/(16c^2)},$$

where $s_B^2(\boldsymbol{h}^U; h^L) := s_1^2(\boldsymbol{h}^U) + s_2^2(\boldsymbol{h}^U; h^L) - 2s_3^2(\boldsymbol{h}^U; h^L)$ *and* $c^2 := \max\{c_1^2; c_2^2\}$ *with* $c_1^2 := s_1^2(\boldsymbol{h}^U)/s_B^2(\boldsymbol{h}^U; h^L)$ *and* $c_2^2 := s_2^2(\boldsymbol{h}^U; h^L)/s_B^2(\boldsymbol{h}^U; h^L)$.

Proof of Lemma 3. Using the assumptions and Lemma (7.1) in Lafferty and Wasserman (2008), $E\left(\widehat{\overline{B_j^{UL}}}\right)$ can be easily derived. Note that p_j in the quantities $q(\mathbf{h}^U)$ and $q(\mathbf{h}^U; h^L)$ depend on the fourth order derivatives of the function $m(\cdot)$.

Now, we can write

$$\frac{\widehat{\overline{B_j^{UL}}} - E\left(\widehat{\overline{B_j^{UL}}}\right)}{s_B(\mathbf{h}^U; h^L)} =$$

$$= \frac{\hat{m}\left(x; H^U\right) - E\left(\hat{m}\left(\mathbf{x}; \mathbf{H}^U\right)\right)}{s_1(\mathbf{h}^U)} c_1 - \frac{\hat{m}\left(\mathbf{x}; \mathbf{H}_j^L\right) - E\left(\hat{m}\left(\mathbf{x}; \mathbf{H}_j^L\right)\right)}{s_2(\mathbf{h}^U; h^L)} c_2.$$

So, we have that

$$P\left(\left|\frac{\widehat{\overline{B_j^{UL}}} - E\left(\widehat{\overline{B_j^{UL}}}\right)}{s_B(\mathbf{h}^U; h^L)}\right| > \sqrt{\delta \log n}\right)$$

$$\leq P\left(\left|\frac{\hat{m}\left(\mathbf{x}; \mathbf{H}^U\right) - E\left(\hat{m}\left(\mathbf{x}; \mathbf{H}^U\right)\right)}{s_1(\mathbf{h}^U)}\right| > \sqrt{\frac{\delta \log n}{4c_1^2}}\right) +$$

$$+P\left(\left|\frac{\hat{m}\left(\mathbf{x}; \mathbf{H}_j^L\right) - E\left(\hat{m}\left(\mathbf{x}; \mathbf{H}_j^L\right)\right)}{s_2(h^U; h_j^L)}\right| > \sqrt{\frac{\delta \log n}{4c_2^2}}\right).$$

Using the Bernstein's inequality as in the proof of Lemma 7.1 in Lafferty and Wasserman (2008), the result follows. □

Now we consider the estimator in (27), for $j = 1, \ldots, d$, as

$$\widehat{\overline{V_j^L}} = \sigma_\epsilon^2 \mathbf{e}_1^T (\mathbf{\Gamma}^T \mathbf{W}_j^L \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{W}_j^L \mathbf{W}_j^L \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{W}_j^L \mathbf{\Gamma})^{-1} \mathbf{e}_1.$$

Lemma 4. *For every $\boldsymbol{h}^U = (h^U, \ldots, h^U) \in \mathbb{B}$ and $\boldsymbol{h}^L = (h^L, \ldots, h^L) \in \mathbb{B}$, vectors of dimension d, if the assumptions (A1)-(A6) hold, with $s = 4$, then $\forall \epsilon > 0$*

$$P\left(\left|\frac{\widehat{V_j^L}}{s_2^2(\boldsymbol{h}^U; h^L)} - 1\right| > \epsilon\right) \to 0 \qquad n \to \infty.$$

Proof of Lemma 4. The result follows by means of Theorem 2.1 in Ruppert and Wand (1994) and Lemma 7.4 in Lafferty and Wasserman (2008). It is sufficient to use Lemma 7.4 in Lafferty and Wasserman (2008) without taking the derivative with respect to the bandwidth h_j. □

References

Bertin, K. and Lecué, G. (2008) Selection of variables and dimension reduction in high-dimensional non-parametric regression, *Electronic Journal of Statistics*, 2, 1224–1241.

Choi, B., Hall, P. and Rousson, V. (2000) Data sharpening methods for bias reduction in nonparametric regression, *The Annals of Statistics*, 28, 1339–1355.

Comminges, L. and Dalalyan, A.S. (2012) Tight conditions for consistency of variable selection in the context of high dimensionality, *The Annals of Statistics*, 40, 2667–2696.

Dai, Y. and Ma, S. (2012) Variable selection for semiparametric regression models with iterated penalisation, *J. of Nonparametric Statistics*, 24, 283–298.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation, *J. of the R. Statistical Society, series B*, 57, 371–394.

Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002) *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, Heidelberg.

Lafferty, J. and Wasserman, L. (2008) Rodeo: sparse, greedy nonparametric regression, *The Annals of Statistics*, 36, 28–63.

La Rocca, M. and Perna, C. (2005) Variable selection in neural network regression models with dependent data: a subsampling approach, *Computational Statistics and Data Analysis*, 48, 415–429.

Li, R. and Liang, H. (2008) Variable selection in semiparametric regression modeling, *The Annals of Statistics*, 36, 261–286.

Lin, L. and Lin, F. (2008) Stable and bias-corrected estimation for nonparametric regression models, *J. of Nonparametric Statistics*, 20, 283–303.

Lin, B. and Zhang, H. (2006) Component selection and smoothing in multivariate nonparametric regression, *The Annals of Statistics*, 34, 2272–2297.

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

Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation, *J. of the American Statistical Association*, 92, 1049–1062.

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

Storlie, C., Bondell, H., Reich, B. and Zhang, H. (2011) Surface estimation, variable selection, and the nonparametric oracle property, *Statistica Sinica*, 21, 679–705.

Variyath, A., Chen, J. and Abraham, B. (2010) Empirical likelihood based variable selection, *J. of Statistical Planning and Inference*, 140, 971–981.

Yang, L. and Tschernig, R. (1999) Multivariate bandwidth selection for local linear regression, *J. of the Royal Statistical Society, Series B*, 61, 793–815.

Zhang, H., Cheng, G. and Liu, Y. (2011) Linear or nonlinear? automatic structure discovery for partially linear models, *J. of the American Statistical Association*, 106, 1099–1112.