DiSES Working Papers

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

GRID for model structure discovering in high dimensional regression

Francesco Giordano
Soumendra Nath Lahiri
Maria Lucia Parrella

Working Paper No. 3.231

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

GRID FOR MODEL STRUCTURE DISCOVERING IN HIGH DIMENSIONAL REGRESSION

Francesco Giordano* Soumendra Nath Lahiri† Maria Lucia Parrella‡

Abstract. Given a nonparametric regression model, we assume that the number of covariates $d \to \infty$ but only some of these covariates are relevant for the model. Our goal is to identify the relevant covariates and to obtain some information about the structure of the model. We propose a new nonparametric procedure, called GRID, having the following features: (a) it automatically identifies the relevant covariates of the regression model, also distinguishing the nonlinear from the linear ones (a covariate is defined *linear/nonlinear* depending on the marginal relation between the response variable and such a covariate); (b) the interactions between the covariates (mixed effect terms) are automatically identified, without the necessity of considering some kind of stepwise selection method. In particular, our procedure can identify the mixed terms of any order (two way, three way, ...) without increasing the computational complexity of the algorithm; (c) it is completely data-driven, so being easily implementable for the analysis of real datasets. In particular, it does not depend on the selection of crucial regularization parameters, nor it requires the estimation of the nuisance parameter σ^2 (self scaling). The acronym GRID has a twofold meaning: first, it derives from Gradient Relevant Identification Derivatives, meaning that the procedure is based on testing the significance of a partial derivative estimator; second, it refers to a graphical tool which can help in representing the identified structure of the regression model. The properties of the GRID procedure are investigated theoretically.

Keywords: Variable selection, model selection, nonparametric model regression.

AMS 2010 classifications: 46A03, 62A01, 60F05.

JEL classifications: C14, C15, C18, C88.

1. Introduction

Nonparametric methods are particularly useful in the preliminary stage of data analysis, for example to make variable selection, model structure discovering and goodness-of-fit tests. In fact, while a correctly specified parametric model is characterized by precise inference, a badly misspecified one leads to inconsistent results. On the other side, nonparametric modelling is associated with greater robustness and less precision. But a criticism often made to the nonparametric procedures is that they are time-consuming and not "user-friendly", because their performance depends crucially on some regularization parameters which are difficult to set. This remarkably affects the potentialities of such procedures. To promote the use of nonparametric approaches, the procedures should be automatic and

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

†NCSU Statistics Department, Campus Box 8203 Raleigh, NC 27695-8203, USA, snlahiri@ncsu.edu

‡UNISA - DISES, Via Ponte Don Melillo, 84084, Fisciano (SA), Italy, mparrella@unisa.it

easy to implement. At the same time, they should assure the oracle property under general assumptions. Therefore, these goals will represent the main priority of our analysis.

We consider the following nonparametric regression model

$$Y_t = m(X_t) + \varepsilon_t, \tag{1}$$

where the X_t represents the $\mathbb{R}^d$-valued covariates and the errors ε_t are *i.i.d.* with zero mean and variance σ^2. Here $m(X_t) = E(Y_t|X_t) : \mathbb{R}^d \to \mathbb{R}$ is the multivariate conditional mean function. The errors ε_t are supposed to be independent of X_t. We use the notation $X_{(j)}$ to refer to the single covariates, for $j = 1 \ldots, d$. We indicate with $f_X(\cdot)$ the multivariate density function of the covariates, having support $supp(f_X) \subseteq \mathbb{R}^d$, and with $f_\varepsilon(\cdot)$ the density of the errors.

We assume that the number of covariates $d \to \infty$ but only some of these covariates are relevant for model (1). Given that the parametric form of the function m is completely unknown, our goal is to identify the relevant covariates and to obtain some information about the structure of model (1). We propose a nonparametric procedure having the following features:

(a) it automatically identifies the relevant covariates of model (1), also distinguishing the nonlinear from the linear ones (a covariate is defined *linear/nonlinear* depending on the marginal relation between the response variable and such a covariate, which corresponds to a relative constant/nonconstant gradient, respectively);

(b) the interactions between the covariates (mixed effect terms) are automatically identified, without the necessity of considering some kind of stepwise selection method. In particular, our procedure can identify the mixed terms of any order (two way, three way, ...) without increasing the computational complexity of the algorithm; moreover, the mixed effect terms are classified as *nonlinear mixed effect*, if they involve some *nonlinear covariates*, or as *linear mixed effect*, if they involve only *linear covariates*;

(c) it is completely data-driven, so being easily implementable for the analysis of real datasets. In particular, it does not depend on the selection of crucial regularization parameters, nor it requires the estimation of the nuisance parameter σ^2 (self scaling). The multiple test selection procedure is based on the Empirical Likelihood approach. Under suitable assumptions, our procedure can be applied to high dimension datasets.

A screening of the available statistical methods proposed so far can help to highlight the main contributions of our work.

Most of the work has been made in the context of variable selection. There are two main approaches to this problem. Both these approaches consider the estimation of the multivariate regression function contextually to relevant variable selection. The first one is based on the idea of LASSO, using some penalized regressions within additive models (see Radchenko & James (2010), Zhang *et al.* (2011), Storlie & alt. (2011), among others). The appeal of this approach is the fast rate of convergence, which essentially derives from the imposition of an additive model and other crutial assumptions. On the other side, a serious drawback is given by the computational complexity and the difficulty of implementation on real datasets. The second approach, which has inspired this work, is based on a general regression function of dimension d, which do not impose any additive restrictions

1^{st} stage: Variable selection		2^{nd} stage: Identification of interactions		
		Mixed with other nonlinear covariates	Mixed with other linear covariates	Pure additive
Nonlinear covariates	C	C_c	C_a	C_p
Linear covariates	A	A_c	A_a	A_p

Table 1: Schematic representation of the GRID procedure. The two dimensions of the table refer to the two stages. The body of the table shows the partition of the regressors $\Xi = \{1, \ldots, d\}$ obtained at the end of the procedure.

on the model (Lafferty & Wasserman (2008)). The main advantage of this approach is its flexibility and simplicity of implementation on real datasets. At the same time, it suffers from a low rate of convergence that makes it unsuitable for the analysis of high dimensional datasets.

Very few papers consider the problem of model selection contextually to variable selection, among which Radchenko & James (2010) and Zhang *et al.* (2011). As far as we know, our procedure is the only one that gives a complete idea of the structure of model (1), without assuming an additive form a priori. As can be seen from points (a)-(c) above, we can derive approximately the exact functional form of the true regression function, which can be used in order to estimate a semiparametric model efficiently.

Our method can be seen as a non trivial extension of the RODEO of Lafferty & Wasserman (2008), in the sense that we use the same framework and some of the ideas and results presented in their paper, but here we propose a new procedure which also fix some of its drawbacks. Moreover, we perform model selection in addition to variable selection. The acronym GRID has a twofold meaning: first, it derives from Gradient Relevant Identification Derivatives, meaning that the procedure is based on testing the significance of a partial derivative estimator; second, it refers to a graphical tool which can help in representing the identified structure of model (1). The estimation procedure used in GRID is based on the conjoint implementation of two nonparametric tools: the local linear estimator (LLE) and the empirical likelihood (EL). The peculiarity of our proposal is that it takes advantage of both the strenghts and weaknesses of the two nonparametric tecniques, and it harmonically integrates them in order to pursue the final aim of the estimation.

The rest of this section is devoted to explain how the GRID method works.

Our aim is to classify the covariates of model (1) into disjoint sets: 1) the set of *nonlinear covariates*, which includes those variables $X_{(j)}$ having a nonlinear effect on the dependent variable Y (*i.e.*, those with non constant gradient); 2) the set of *linear covariates*, which includes the variables $X_{(j)}$ having a linear effect on the response variable Y (*i.e.*, constant gradient); 3) the set of irrelevant covariates, collecting the variables for which the gradient is equal to zero. Denote with C, A and U, respectively, the correspondent index sets and let $\Xi = C \cup A \cup U$ represent the set of regressors $\{1, \ldots, d\}$. Secondly, we point to automatically detect the interactions among the covariates, identifying exactly which mixed effects are 'active' in model (1). Therefore, each index set can be further partitioned as shown in table 1.

The two dimensions of the table refer to the two stages of the GRID procedure: the first one focuses on variable selection while the second looks at the interaction terms. More

specifically, the information on the interaction terms is given in the following way. Let I^j be the set of covariates mixed with the j-th covariate, for $j \in \Xi$. A convention used here is that $j \notin I^j$, which means that self-interaction is excluded in practice. The GRID procedure gives a consistent estimation of the sets C and A in the first stage, and the sets I^j in the second stage. The other sets can be derived easily by known relationships. In particular, $I_C^j = I^j \cap C$ is the set of *nonlinear covariates* which are mixed with the j-th covariate and $I_A^j = I^j \cap A$ is the set of *linear covariates* mixed with the j-th covariate. Then $I^j = I_C^j \cup I_A^j$. But also $C_c = \cup_{j \in C} I_C^j$, $C_a = \cup_{j \notin C} I_C^j$, $A_c = \cup_{j \in C} I_A^j$ and $A_a = \cup_{j \notin C} I_A^j$. All this permits to do variable selection and model structure discovering simultaneously. For example, when $d = 10$ and the model is

$$Y_t = 2X_{t1} + X_{t2}^2 X_{t3} + 10X_{t4}X_{t5}X_{t6} + \exp(X_{t7}) + \varepsilon_t, \tag{2}$$

then the first stage of the procedure is devoted to identify the following sets of covariates

$$C = \{2, 7\}, \quad A = \{1, 3, 4, 5, 6\}, \quad U = \{8, 9, 10\},$$

while the second stage of the procedure identifies the following sets of interactions

$$I_A^2 = \{3\}, \quad I_C^3 = \{2\}, \quad I_A^4 = \{5, 6\}, \quad I_A^5 = \{4, 6\}, \quad I_A^6 = \{4, 5\}.$$

To make the GRID procedure 'user-friendly', the method is presented in section 5 through a detailed algorithm and the results of the estimation are shown through an intuitive plot which points out clearly both the relevant covariates and their interactions, and helps to write down the (estimated) functional form of the regression function $m(x)$.

The details are presented in the following sections. Section 2 gives the notation. In section 3 we give the main idea at the basis of the selection procedure. Then, in section 5, we present the test-statistic, the algorithm and the GRID plot. Section 4 describes the multiple testing method, based on the Empirical Likelihood inference. All the proofs are collected in the appendix.

2. Basics of the multivariate local linear estimator

The local linear estimator is a nonparametric tool whose properties have been studied deeply. See Ruppert & Wand (1994), among others. Let $x = (x_1, \ldots, x_d)$ be the target point at which we estimate m. The LLE performs a locally weighted least squares fit of a linear function, being equal to

$$\arg\min_{\beta_0, \beta_1} \sum_{t=1}^{n} \left\{ Y_t - \beta_0(x) - \beta_1^T(x)(X_t - x) \right\}^2 K_H(X_t - x) \tag{3}$$

where $(\cdot)^T$ denotes the transpose operator, the function $K_H(u) = |H|^{-1} K(H^{-1}u)$ gives the local weights and $K(u)$ is the Kernel function, a d-variate function. The $d \times d$ matrix H contains the smoothing parameters, and it is 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 regression function. Denote with $\beta(x) = (\beta_0(x), \beta_1^T(x))^T$ the vector of coefficients to estimate. Using the matrix notation, the solution of the minimization problem in the (3) can be written in closed form:

$$\hat{\beta}(x; H) = (\Gamma^T W \Gamma)^{-1} \Gamma^T W \Upsilon, \tag{4}$$

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

$$\Upsilon = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix},$$

$$W = \begin{pmatrix} K_H(X_1 - x) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & K_H(X_n - x) \end{pmatrix}.$$

Let $\mathbb{D}_g(x)$ denote the gradient and $\mathbb{H}_g(x)$ the Hessian matrix of a d-variate function g. Note from (3) that $\hat{\beta}(x;H)$ gives an estimation of the function $m(x)$ and its gradient. In particular,

$$\hat{\beta}(x;H) = \begin{pmatrix} \hat{\beta}_0(x;H) \\ \hat{\beta}_1(x;H) \end{pmatrix} \equiv \begin{pmatrix} \hat{m}(x;H) \\ \hat{\mathbb{D}}_m(x;H) \end{pmatrix}. \tag{5}$$

Despite its conceptual and computational simplicity, the practical implementation of the LLE is not trivial, especially in the multivariate case, where it is subject to many drawbacks. First of all, its consistence depends on the correct identification of the bandwidth matrix H. An asymptotically optimal bandwidth exists and can be derived taking account of the bias-variance trade-off, but the estimation of it is difficult in the multivariate framework. Secondly, the resulting estimator of the regression function is biased, even when using the optimal bandwidth matrix, and this makes the inference based on it unreliable. Finally, the LLE is strongly affected by the *curse of dimensionality* problem, so these estimators become impracticable for high-dimensional problems.

Anyway, the use of the LLE made here is non-standard from several points of view, as we will see in the following sections. The advantage of our approach is that we manage to use the Local Linear approximation technique avoiding all the drawbacks listed above. To this end, we work with a variant of the classic estimator. Basically, we are not interested in the function estimation itself, but only its bias, from which we can obtain a clear information about the structure of the underlying regression model.

We consider the following assumptions.

A1) The bandwidth H is a diagonal and strictly positive definite matrix with diagonal elements $h_j = O(1)$ for $j = 1, \dots, d$.

A2) The d-variate Kernel function K is a product kernel, with compact support and zero odd moments. Therefore, the following moments exist bounded (we assume that $\mu_0 = 1$)

$$\mu_r = \int u_1^r K(u_1) du_1, \qquad \nu_r = \int u_1^r K^2(u_1) du_1 \qquad r = 0, 1, \dots, 4.$$

Moreove, we assume that $K \in C^1[-a, a]$ for some $a > 0$.

A3) All the partial derivatives of the function $m(x)$ up to and including fifth order are bounded.

A4) The density f_X is uniform on the unit cube.

Remark 1.1: The assumption $A1$ is different from the typical assumptions made on the bandwidth matrix H. As a consequence, all the theorems available in the statistical literature concerning the properties of the multivariate LLE are invalidated and cannot be applied to our framework. Anyway, in addition to the theoretical derivations shown in this paper, a confirmation of the reasonableness of our choice lies in Bertin & Lecue (2008).

Remark 1.2: The assumptions $A3$ and $A4$ are needed in order to bound the Taylor expansion of the function $m(x)$, as shown in the proofs. We relax condition $A4$, although only in part, in Theorem 2.

3. The main idea for model structure discovering

Assume that there are k nonlinear covariates in C, $r-k$ linear covariates in A and $d-r$ irrelevant variables in the complementary set $U = \overline{A \cup C}$. So, r is the number of relevant covariates of model (1). Without loss of generality, we assume that the predictors are ordered 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 is furtherly partitioned into disjoint subsets: the covariates from $k+1$ to $k+s$ belong to the subset A_c, which includes those linear covariates which are multiplied to nonlinear covariates, introducing *nonlinear mixed effects* in model (1); the covariates from $k+s+1$ to $k+r$ belong to the subset $A_u = A_a \cup A_p$, which includes those linear covariates which have a linear additive relation in model (1) or which are mixed to other linear covariates (*linear mixed effects*). We want to stress here that the GRID procedure automatically identifies such sets of indices, so the assumptions made here have the only purpose of gaining clarity in the exposition.

In such a framework, using $x = (x_C, x_{A_c}, x_{A_u}, x_U)$, the gradient and the Hessian matrix of the function m become

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

where 0 is a vector or matrix with all elements equal to zero, $\mathbb{D}_m^C(x) = \partial m(x)/\partial x_C$, $\mathbb{D}_m^{A_c}(x) = \partial m(x)/\partial x_{A_c}$, $\mathbb{D}_m^{A_u}(x) = \partial m(x)/\partial x_{A_u}$ and

$$\mathbb{H}_m^C(x) = \begin{pmatrix} \frac{\partial^2 m(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 m(x)}{\partial x_1 \partial x_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 m(x)}{\partial x_k \partial x_1} & \cdots & \frac{\partial^2 m(x)}{\partial x_k \partial x_k} \end{pmatrix}, \quad \mathbb{H}_m^{CA_c}(x) = \begin{pmatrix} \frac{\partial^2 m(x)}{\partial x_1 \partial x_{k+1}} & \cdots & \frac{\partial^2 m(x)}{\partial x_1 \partial x_{k+s}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 m(x)}{\partial x_k \partial x_{k+1}} & \cdots & \frac{\partial^2 m(x)}{\partial x_k \partial x_{k+s}} \end{pmatrix} \tag{6}$$

$$\mathbb{H}_m^{A_c}(x) = \begin{pmatrix} 0 & \frac{\partial^2 m(x)}{\partial x_{k+1} \partial x_{k+2}} & \cdots & \cdots & \frac{\partial^2 m(x)}{\partial x_{k+1} \partial x_{k+s}} \\ \frac{\partial^2 m(x)}{\partial x_{k+2} \partial x_{k+1}} & 0 & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & 0 & \frac{\partial^2 m(x)}{\partial x_{k+s-1} \partial x_{k+s}} \\ \frac{\partial^2 m(x)}{\partial x_{k+s} \partial x_{k+1}} & \cdots & \cdots & \frac{\partial^2 m(x)}{\partial x_{k+s} \partial x_{k+s-1}} & 0 \end{pmatrix}.$$

The matrix $\mathbb{H}_m^{A_u}(x)$ is defined similarly to the matrix $\mathbb{H}_m^{A_c}(x)$, with a zero diagonal, and the matrix $\mathbb{H}_m^{A_cA_u}(x)$ is defined similarly to $\mathbb{H}_m^{CA_c}(x)$. Note that the matrices $\mathbb{H}_m^C(x)$,

$\mathbb{H}_m^{A_c}(x)$ and $\mathbb{H}_m^{A_u}(x)$ are symmetric, whereas the matrices $\mathbb{H}_m^{CA_c}(x)$ and $\mathbb{H}_m^{A_cA_u}(x)$ are not. Moreover, for additive models without mixed effects, all the sub-matrices in $\mathbb{H}_m(x)$ are zero, except for $\mathbb{H}_m^C(x)$ which is diagonal.

In our analysis, it is also necessary to take account of those terms in the Taylor's expansion of $m(x)$ involving the partial derivatives of order 3 (see the proof of Proposition 1 for the details). To this end, we define the following matrix

$$\mathbb{G}_m(x) = \begin{pmatrix} \frac{\partial^3 m(x)}{\partial x_1^3} & \frac{\partial^3 m(x)}{\partial x_1 \partial x_2^2} & \cdots & \frac{\partial^3 m(x)}{\partial x_1 \partial x_d^2} \\ \frac{\partial^3 m(x)}{\partial x_2 \partial x_1^2} & \frac{\partial^3 m(x)}{\partial x_2^3} & \cdots & \frac{\partial^3 m(x)}{\partial x_2 \partial x_d^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^3 m(x)}{\partial x_d \partial x_1^2} & \frac{\partial^3 m(x)}{\partial x_d \partial x_2^2} & \cdots & \frac{\partial^3 m(x)}{\partial x_d^3} \end{pmatrix} = \begin{pmatrix} \mathbb{G}_m^C(x) & 0 & 0 & 0 \\ \mathbb{G}_m^{A_cC}(x) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \quad (7)$$

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

In the same way, let us partition the bandwidth matrix as $H = diag(H_C, H_{A_c}, H_{A_u}, H_U)$ and the gradient of a function $g(x)$ as $\mathbb{D}_g(x) = (\mathbb{D}_g^C(x)^T, \mathbb{D}_g^{A_c}(x)^T, \mathbb{D}_g^{A_u}(x)^T, \mathbb{D}_g^U(x)^T)^T$.

3.1 Identifying the nonlinear effects

The rationale of our proposal lies in Proposition 1 and Theorem 1. In Proposition 1 we derive the conditional bias of the LLE in (5), under the specific assumptions considered here. In Theorem 1, we introduce a variant of the previous estimator, which has similar properties but is more suitable for our specific needs.

Let $\mathbf{1}$ be a vector of ones. The $O_p(M)$ and $O(M)$ terms must be intended for each element of a matrix/vector M. Here and in the proofs, the notation $\delta(\cdot)$ is used to denote a finite quantity – scalar, vector or matrix – whose elements depend on the arguments of $\delta(\cdot)$. In particular, it is equal to zero if at least one of its arguments is zero. Moreover, it can be used several times in the same proposition to denote different quantities, all finite.

Proposition 1. *Under model (1) and assumptions (A1)-(A4), the conditional bias of the local linear estimator given by (5) is equal to*

$$E\left\{ \begin{pmatrix} \hat{m}(x;H) \\ \hat{\mathbb{D}}_m(x;H) \end{pmatrix} - \begin{pmatrix} m(x) \\ \mathbb{D}_m(x) \end{pmatrix} \middle| X_1, \ldots, X_n \right\} = \begin{pmatrix} b_m(x;H_C) \\ B_{\mathbb{D}}(x,H_C) \end{pmatrix} + O_p(n^{-\frac{1}{2}}), \quad (8)$$

where

$$\begin{aligned} b_m(x;H_C) &= \frac{1}{2}\mu_2\, tr\{\mathbb{H}_m^C(x)H_C^2\} + \delta(H_C^4) \\ B_{\mathbb{D}}(x,H_C) &= \begin{pmatrix} B_{\mathbb{D}}^C \\ B_{\mathbb{D}}^{A_c} \\ B_{\mathbb{D}}^{A_u} \\ B_{\mathbb{D}}^U \end{pmatrix} \end{aligned}$$

$$= \frac{1}{2}\mu_2 \begin{pmatrix} \mathbb{G}_m^C(x)H_C^2\mathbf{1} + \left(\frac{\mu_4}{3\mu_2^2} - 1\right) diag\{\mathbb{G}_m^C(x)H_C^2\}\mathbf{1} + \delta(H_C^4) \\ \mathbb{G}_m^{A_cC}(x)H_C^2\mathbf{1} + \delta(H_C^4) \\ 0 \\ 0 \end{pmatrix}. \quad (9)$$

Our results in Proposition 1, on the biases of the estimators $\hat{m}(x;H)$ and $\hat{\mathbb{D}}_m(x;H)$, are congruent with the results in Theorem 2.1 of Ruppert & Wand (1994) (but note that our bandwidth matrix H corresponds to their $H^{1/2}$) and Theorem 1 of Lu (1996). Anyway, there are substantial differences in the proofs, because of the different assumptions made here and because we keep trace of the different influences of the bandwidth matrices H_C, H_{A_c}, H_{A_u} and H_U on the bias of the local linear estimator.

The main result of Proposition 1 is that it shows some interesting relationships between the bias of $\hat{\beta}(x;H)$ and the bandwidth matrix $H = diag(H_C, H_{A_c}, H_{A_u}, H_U)$, which can be exploited in order to analyze the structure of model (1). Generalizing the idea proposed in the paper of Lafferty & Wasserman (2008), we can make these relationships emerge through the derivative of $\hat{\beta}(x;H)$ with respect to H. In fact, note that for $n \to \infty$

$$\frac{\partial}{\partial H} E\left\{ \left(\begin{array}{c} \hat{m}(x;H) \\ \hat{\mathbb{D}}_m(x;H) \end{array} \right) \middle| X_1, \ldots, X_n \right\}$$

$$\begin{aligned} \equiv \quad & \frac{\partial}{\partial H} E\left\{ \left(\begin{array}{c} \hat{m}(x;H) \\ \hat{\mathbb{D}}_m(x;H) \end{array} \right) - \left(\begin{array}{c} m(x) \\ \mathbb{D}_m(x) \end{array} \right) \middle| X_1, \ldots, X_n \right\} \\ \longrightarrow \quad & \left(\begin{array}{c} \frac{\partial}{\partial H} b_m(x;H_C) \\ \frac{\partial}{\partial H} B_{\mathbb{D}}(x,H_C) \end{array} \right) \end{aligned}$$

where

$$\begin{aligned} \frac{\partial}{\partial H} b_m(x;H_C) &= \left(\frac{\partial b_m(x;H_C)}{\partial H_C}, \frac{\partial b_m(x;H_C)}{\partial H_{A_c}}, \frac{\partial b_m(x;H_C)}{\partial H_{A_u}}, \frac{\partial b_m(x;H_C)}{\partial H_U} \right) \\ &= (\delta(H_C), 0, 0, 0) \end{aligned} \tag{10}$$

and

$$\frac{\partial}{\partial H} B_{\mathbb{D}}(x,H_C) = \left(\begin{array}{c} \partial B_{\mathbb{D}}^{C}/\partial H \\ \partial B_{\mathbb{D}}^{A_c}/\partial H \\ \partial B_{\mathbb{D}}^{A_u}/\partial H \\ \partial B_{\mathbb{D}}^{U}/\partial H \end{array} \right) = \left(\begin{array}{cccc} \delta(H_C) & 0 & 0 & 0 \\ \delta(H_C) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right). \tag{11}$$

So matrix in (11) has a sparse structure similar to $\mathbb{G}_m$. From the (10) and (11) we have an overview of what are the influences of the bandwidths on the local linear estimations of $m(x)$ and $\mathbb{D}_m(x)$. Some stylized facts can be outlined. In particular,

(i) the derivatives $\partial E\{\hat{m}(x;H)\}/\partial H$ in the (10) are considered in the RODEO method as a tool to identify the relevant covariates of model (1). Anyway, there is a problem: the relevant linear covariates in A and the irrelevant variables in U become indistinguishable. So only the *nonlinear covariates* in C can be identified basing on (10). To overcome this, Lafferty & Wasserman (2008) suggest to identify first the linearities through a LASSO or to change the degree of the local polynomial estimator to zero (i.e. to use the Nadaraya-Watson estimator). Both these solutions seem to be suboptimal;

(ii) the elements of matrix (11) give additional important information on the structure of model (1); in fact, the element of position i, j of such matrix reflects the sensitivity

of the i-partial derivative estimator to variations of the bandwidth h_j. By Proposition 1, we have

$$\frac{\partial B_{\mathbb{D}}^{(i)}(x, H_C)}{\partial h_j} \approx \begin{cases} h_j \mu_2 \frac{\partial^3 m(x)}{\partial x_i \partial x_j^2} & \text{if } i \neq j \\ h_j \frac{\mu_4}{6\mu_2} \frac{\partial^3 m(x)}{\partial x_j^3} & \text{if } i = j \end{cases} \tag{12}$$

and such value can be zero depending on the value of the derivative $\frac{\partial^3 m(x)}{\partial x_i \partial x_j^2}$. Therefore, given i and j, the formula in the (12) is different from zero if there are mixed effects in model (1) between two *nonlinear covariates* or between a *linear covariate* X_i and a *nonlinear covariate* X_j, in the case $i \neq j$; or if the covariate is a *nonlinear covariate* of order ≥ 3, in the case $i = j$. So, this derivatives can help to identify the *linear covariates* in A_c and the *nonlinear mixed effect* terms.

(iii) Of course, the result of the formula in the (12) is always zero if $j \in U$, as desired. Anyway, also the pure *linear covariates* in A_u and the *linear mixed effects* become "transparent", so they are confused with the covariates in U. We will address the problem of identifying such linearities in section 3.2.

In order to improve the rate of convergence of d shown in the RODEO method, we propose to base our identification procedure on a variant of the estimator (4). In fact, if we desire to consider the case when $d > n$, the estimator (4) is not well defined because the rank of Γ is the smallest number between $d + 1$ and n. To avoid this constrain, due to the necessity of inverting the regression matrix, we introduce the following estimator

$$M(x; H) = \frac{1}{n} diag(1, H^{-2}) \Gamma^T W \Upsilon \equiv \begin{pmatrix} M_0(x; H) \\ M_1(x; H) \end{pmatrix}. \tag{13}$$

The estimator (13) is a simplified version of estimator (5), which uses the assumption $A4$. Its properties in terms of bias are similar to those reported in Proposition 1, as shown in Theorem 1, so it can be used for variable selection basing on the previous ideas.

Therefore, we need to consider the derivatives of (13) w.r.t. the different bandwidths. We compute

$$\begin{aligned} \dot{M}_{0j} &= \frac{\partial M_0(x; H)}{\partial h_j} \qquad j = 1, \ldots, d \\ \dot{M}_{1j} &= \frac{\partial M_1(x; H)}{\partial h_j} \equiv \{\dot{M}_{1j}^{(i)}\}_{i=1,\ldots,d}, \end{aligned} \tag{14}$$

whose explicit expressions derive from

$$\begin{aligned} \frac{\partial M(x; H)}{\partial h_j} &= \frac{\partial}{\partial h_j} \left[\frac{1}{n} \begin{pmatrix} 1 & 0 \\ 0 & H^{-2} \end{pmatrix} \Gamma^T W \Upsilon \right] \\ &= \frac{1}{n} O_j \Gamma^T W \Upsilon + \frac{1}{n} \begin{pmatrix} 1 & 0 \\ 0 & H^{-2} \end{pmatrix} \Gamma^T \frac{\partial}{\partial h_j} W \Upsilon, \end{aligned}$$

where O_j is a matrix with $d + 1$ rows and $d + 1$ columns, with all zeros except the element in position $(j + 1, j + 1)$ which is equal to $-\frac{2}{h_j^3}$.

Since W is a diagonal matrix with elements

$$K_H(X_t - x) = \frac{1}{|H|} \prod_{k=1}^{d} K\left(\frac{X_{tk} - x_k}{h_k} \right),$$

its derivative with respect to h_j is

$$\frac{\partial}{\partial h_j} K_H\left(X_t - x\right) = K_H\left(X_t - x\right)\left(-\frac{1}{h_j} + \frac{\partial}{\partial h_j}\log K\left(\frac{X_{tj} - x_j}{h_j}\right)\right).$$

So

$$\frac{\partial}{\partial h_j} W = W L_j$$

where $L_j = diag\left(\frac{\partial \log K((X_{1j}-x_j)/h_j)}{\partial h_j} - \frac{1}{h_j}, \ldots, \frac{\partial \log K((X_{nj}-x_j)/h_j)}{\partial h_j} - \frac{1}{h_j}\right)$. Finally, we propose the following estimator

$$\frac{\partial M(x;H)}{\partial h_j} = \frac{1}{n}\left[O_j \Gamma^T W + \begin{pmatrix} 1 & 0 \\ 0 & H^{-2} \end{pmatrix}\Gamma^T W L_j\right]\Upsilon \equiv \begin{pmatrix} \dot{M}_{0j} \\ \dot{M}_{1j} \end{pmatrix}. \tag{15}$$

Theorem 1. *Under model (1) and assumptions (A1)-(A4), the following result holds*

$$E\left\{\dot{M}_{0j}\right\} = \begin{cases} \theta_{0j}^m \neq 0 & \textit{if and only if } j \in C \\ \theta_{0j}^m = 0 & \textit{otherwise} \end{cases} \tag{16}$$

$$E\left\{\dot{M}_{1j}^{(i)}, i \neq j\right\} = \begin{cases} \theta_{ij}^m \neq 0 & \textit{if and only if } i \in I^j, j \in C \\ \theta_{ij}^m = 0 & \textit{otherwise} \end{cases} \tag{17}$$

where the exact expressions for θ_{ij}^m, $i = 0, \ldots, d$ and $j = 1, \ldots, d$, $i \neq j$, are (35) and (36) in the appendix.

Remark 3.1: Theorem 1 can be used to detect the nonlinear effects in model (1). In fact, basing on the (16), the derivatives $\dot{M}_{0j}$ can be used in order to identify the *nonlinear covariates*, obtaining C. Basing on (17), the derivatives $\dot{M}_{1j}^{(i)}$ can be used in order to identify the interactions for the *nonlinear covariates*, obtaining I^j, for $j \in C$. As a consequence, we also obtain the sets $I_C^j = I^j \cap C$, for $j \in C$, then $C_c = \cup_{j\in C} I_C^j$, $C_a = (\cup_{j\in C} I^j)\backslash C_c$ and $C_p = C\backslash(C_c \cup C_a)$. But also $I_A^j = I^j \backslash I_C^j$, for $j \in C$, and $A_c = \cup_{j\in C} I_A^j$. Looking at table 1, the only sets which cannot be identified using Theorem 1 are the sets A_a and A_p, including the *pure linear effects*.

Remark 3.2: The values of the bandwidths are not crucial in our procedure, because we are not interested in the exact estimation of the function $m(x)$. So, given that the identification of the covariates is based on evaluating the bias of the LLE, we prefer to use a bandwidth matrix which produces a very high bias. This means to take very large bandwidths, for example $h = 0.9$, which has benefits on the efficiency of the estimator in (13).

3.2 Identifying the linear effects

Basing on the expression (12), the pure *linear covariates* in $A_u = A_a \cup A_p$ and the *linear mixed effects* in I_A^j, for $j \in A$, would be transparent to our identification procedure. Anyway, a convenient solution is to consider an auxiliary regression with some of the covariates transformed, so that the *linear covariates* of the original model become *nonlinear* in the auxiliary model. In particular, if we think at model (1) under the partition $\{C, A_c, A_u, U\}$, it must necessarily be

$$m(x) = m_1(x_C, x_{A_c}) + m_2(x_{A_c}, x_{A_u}).$$

Now, let us define a transformation $z = \phi(x)$ and its inverse $x = \phi^{-1}(z)$ as follows (componentwise)

$$z = \phi(x) = (x_C, x_{A_c}^{1/2}, x_{A_u}^{1/2}, x_U^{1/2}), \quad x = \phi^{-1}(z) = (x_C, z_{A_c}^2, z_{A_u}^2, z_U^2). \tag{18}$$

We can consider the following auxiliary regression

$$Y_t = m(\phi^{-1}(Z_t)) + \varepsilon_t = g(Z_t) + \varepsilon_t, \qquad t = 1, \ldots, n,$$

where the new regression function can be written as

$$g(z) = g_1(x_C, z_{A_c}) + g_2(z_{A_c}, z_{A_u}).$$

Note once again that we use the same index partition considered in the first regression. Thanks to the transformation in (18), the function $g_2(\cdot)$ depends only on the covariates in A. Moreover, we are sure that these covariates have a nonlinear effect in the auxiliary regression model $g(z)$. In fact,

$$z_j = \phi(x_j) = x_j^{1/2} \qquad \Longrightarrow \qquad x_j = \phi^{-1}(z_j) = z_j^2 \qquad \forall j \in A \cup U$$

so the partial derivatives are

$$\begin{aligned} \frac{\partial g(z)}{\partial z_j} &= \frac{\partial m(\phi^{-1}(z))}{\partial z_j} = \frac{\partial m}{\partial x_j}\frac{\partial x_j}{\partial z_j} = \begin{cases} 2a_j z_j \neq 0 & \text{for } j \in A \\ 0 & \text{for } j \in U \end{cases} \\ \frac{\partial^2 g(z)}{\partial z_j \partial z_j} &= \begin{cases} 2a_j \neq 0 & \text{for } j \in A \\ 0 & \text{for } j \in U \end{cases}, \end{aligned}$$

where $a_j = \partial m(x)/\partial x_j$ is constant with respect to x_j, $\forall j \in A$. Therefore, the linear covariates in A behaves nonlinearly in the auxiliary regression, while the irrelevant covariates remain still so.

Given that we are not interested in the exact estimation of the function $g(z)$, we can exclude the nonlinear covariates in C in the auxiliary regression. Note that, when we consider the auxiliary regression with the transformed covariates $Z_t = \phi(X_t)$, the density f_Z does not satisfy the assumption $A4$, so Proposition 1 and Theorem 1 cannot be applied.

The following theorem cover this case.

Theorem 2. *Using model (1), assumptions (A1)-(A4) and the transformed random variables*

$$Z_t = \{\phi(X_{(s)}), s \in A\}$$

with ϕ defined in (18), then the following result holds for the estimator defined in (13)

$$E\left\{\dot{M}_{0j}\right\} = \begin{cases} \theta_{0j}^g \neq 0 & \textit{if and only if } j \in A \\ \theta_{0j}^g = 0 & \textit{otherwise} \end{cases} \tag{19}$$

$$E\left\{\dot{M}_{1j}^{(i)}, i \neq j\right\} = \begin{cases} \theta_{ij}^g \neq 0 & \textit{if } i \in I^j, j \in A \\ \theta_{ij}^g = 0 & \textit{if } j \in U \end{cases} \tag{20}$$

where the exact expressions for θ_{ij}^g, $i = 0, \ldots, d$ and $j = 1, \ldots, d$, are (41) and (42) in the appendix. Moreover, using model (1), assumptions (A1)-(A4) and the transformed random variables

$$Z_t = \{X_{(i)}\} \cup \{\phi(X_{(s)}), s \in A, s \neq i\}$$

with ϕ defined in (18), then the following result holds for the estimator defined in (13)

$$E\left\{\dot{M}_{1j}^{(i)}, i \neq j\right\} = \begin{cases} \theta_{ij}^* \neq 0 & \textit{if and only if } i \in I^j, j \in A \\ \theta_{ij}^* = 0 & \textit{otherwise} \end{cases}, \tag{21}$$

where the exact expression of the θ_{ij}^* are (43) in the appendix.

Remark 3.3: Basing on the (19), the derivatives $\dot{M}_{0j} = \partial M_0(z; H)/\partial h_j$, calculated with the transformed covariates Z, can be used in order to identify the *linear covariates*, obtaining the set A. Anyway, we cannot use the (20) in order to identify the *linear mixed effects* in I^j, for $j \in A$, given that it is not a one to one relationship. On the other side, we can identify correctly such effects using the (21), which is derived under the assumption of ϕ-transformation for all the linear covariates in A except the i-th.

Now, for completeness, we can derive the variances for estimators in (15).

Proposition 2. *Under assumptions (A1)-(A4) the estimators $\dot{M}_{0j}$ and $\dot{M}_{1j}$ have the the following mean conditional variances*

(i) $n\,E\left(Var(\dot{M}_{0j}\,|X_1,\ldots,X_n)\right) = \sigma^2 \frac{\nu_0^d}{4|H|h_j^2}$.

(ii) $n\,E\left(Var(\dot{M}_{1j}\,|X_1,\ldots,X_n)\right) = \sigma^2 \frac{\nu_2 \nu_0^{d-1}}{4|H|h_j^2} H^{-2} I_j$, *where I_j is an identity matrix of order d except that the element in position (j,j) is* 9.

Remark 3.4 If we consider the transformation in (18), the covariates, Z, have a non Uniform distribution but the density function is still bounded on $[0,1]^d$. So, Expanding $f_Z(z+Hu)$ by Taylor's series, one can show that $n\,E\left(Var(\dot{M}_{0j}\,|X_1,\ldots,X_n)\right) = \sigma^2 \frac{\nu_0^d}{4|H|h_j^2} f_Z(z)$ using, in particular, assumption (A2). On the other side, the mean conditional variance matrix, $n\,E\left(Var(\dot{M}_{1j}\,|X_1\ldots,X_n)\right)$, exists but it is not diagonal. Moreover, one can show that $n\,E\left(Var(\dot{M}_{1j}^{(i)}\,|X_1,\ldots,X_n)\right) = \sigma^2 \frac{c_{ij}\nu_2\nu_0^{d-1}}{4|H|h_j^2 h_i^2} f_Z(z)$, $\forall i$, where $c_{ij} = 1$ if $i \neq j$ and $c_{ij} = 9$ for $i = j$.

Remark 3.5 Without loss of generality, we can suppose that $E(\dot{M}_{0j}) = 0$ and $E(\dot{M}_{1j}) = \mathbf{0}$. Since, in this case, we have

$$n\,Var(\dot{M}_{0j}) = n\,E\left(Var(\dot{M}_{0j}\,|X_1,\ldots,X_n\right) + n\,E\left(E\left(\dot{M}_{0j}\,|X_1,\ldots,X_n\right)^2\right),$$

one can show that $n\,Var(\dot{M}_{0j}) \leq (C_1 + \sigma^2)\frac{\nu_0^d}{4|H|h_j^2}$, with $C_1 = \sup_{X\in[0,1]^d} m^2(X)$, using the same arguments as in the proof of Proposition 2. In the same way, it follows that $n\,Var(\dot{M}_{1j}^{(i)}) \leq (C_1 + \sigma^2)\frac{c_{ij}\nu_2\nu_0^{d-1}}{4|H|h_j^2 h_i^2}$ where c_{ij} are defined in the previous remark.

4. Inference by Empirical Likelihood

Variable selection is usually done through some multiple testing procedure. We propose to use one based on the Empirical Likelihood (EL) technique. The main advantage of this choice is that we do not need to estimate the nuisance parameter σ^2, which would be difficult in the multivariate high dimensional context. This represents a big improvement over the RODEO method of Lafferty & Wasserman (2008). Another advantage is that we can relax the assumption of gaussianity for f_ε.

A peculiarity of our proposal which deserves attention is the particular implementation of the empirical likelihood technique to the LLE. There are two innovative aspects, compared with the other papers appeared in the statistical literature combining EL and LLE. Firstly, it is known that the use of the EL for the analysis of the kernel based estimators is affected by the bias problem, so that a correction is necessary and usually performed through the undersmoothing technique. In our procedure, this problem is avoided because we use the EL to analyse a local polynomial estimator which is unbiased under the null hypothesis. Secondly, the analysis of the asymptotic statistical properties of the EL procedure must consider that the bandwidths in H are fixed (not tending to zero as $n \to \infty$), making such analysis non standard and the EL estimator more efficient.

Without loss of generality, suppose that $E(\dot{M}_{0j}) = \theta_{0j}$ and $E(\dot{M}_{1j}^{(i)}) = \theta_{ij}$, $i = 1, \ldots, d$ and $j = 1, \ldots, d$, according to Theorem (1) and / or Theorem (2). Now, we need to rewrite the univariate estimators in (15) as:

$$\dot{M}_{0j} = \frac{1}{n} \sum_{k=1}^{n} q_{1,j}(X_k; K, H)(Y_k - \theta_{0j}) \tag{22}$$

$$\dot{M}_{1j}^{(i)} = \frac{1}{n} \sum_{k=1}^{n} q_{i+1,j}(X_k; K, H)(Y_k - \theta_{ij}) \tag{23}$$

where $q_{1,j}(X_k; K, H)$ is the first row of matrix in (15), $q_{i+1,j}(X_k; K, H)$ is the row $i+1$ of matrix in (15), X_k is the d-dimensional vector of covariates, Y_k is the dependent variable and K, H are the Kernel function and the bandwidth matrix, respectively, for $i = 1, \ldots, d$, $j = 1, \ldots, d$ and $k = 1, \ldots, n$. For brevity, we do not write K and H in $q_{.,.}(\cdot)$. So $q_{1,j}(X_k; K, H) \equiv q_{1,j}(X_k)$ and $q_{i+1,j}(X_k; K, H) \equiv q_{i+1,j}(X_k)$.

By theorems (1) and (2) we are interested to consider the cases when the estimators in (22) and (23) are unbiased, *i.e.* $\theta_{0j} = 0$ and $\theta_{ij} = 0$. Therefore, we can build the $-2\log$ Empirical Likelihood Ratio for $\dot{M}_{0j}$ as:

$$-2\log R_{0j}(\theta_{0j}) = -2\sum_{k=1}^{n} \log n p_k^{(j)}, \qquad p_k^{(j)} = \frac{1}{n}\frac{1}{1 + \lambda Z_{k,j}^{(0)}} \tag{24}$$

s.t.

$$\sum_{k=1}^{n} p_k^{(j)} = 1, \qquad \sum_{k=1}^{n} p_k^{(j)} Z_{k,j}^{(0)} = 0,$$

where $Z_{k,j}^{(0)} := q_{1,j}(X_k)(Y_k - \theta_{0j})$. In the same way, it follows the $-2\log$ Empirical Likelihood Ratio for $\dot{M}_{1j}^{(i)}$ as:

$$-2\log R_{1j}^{(i)}(\theta_{ij}) = -2\sum_{k=1}^{n} \log n p_k^{(i,j)}, \qquad p_k^{(i,j)} = \frac{1}{n}\frac{1}{1 + \lambda Z_{k,j}^{(i)}} \tag{25}$$

s.t.

$$\sum_{k=1}^{n} p_k^{(i,j)} = 1, \qquad \sum_{k=1}^{n} p_k^{(i,j)} Z_{k,j}^{(i)} = 0,$$

where $Z_{k,j}^{(i)} := q_{i+1,j}(X_k)(Y_k - \theta_{ij})$. The following proposition gives the consistency of (24) and (25). In the following results we consider assumption $(A4)$ which can be replaced by the distribution function in section (3.2) as in Theorem (2). We can state the following proposition.

Proposition 3. *Suppose that $E(\varepsilon_t^2) < \infty$ and assumptions (A1) - (A4) hold. If $\theta_{0j} = 0$ and $\theta_{ij} = 0$, $i = 1, \dots, d$, $j = 1, \dots, d$, then*

$$-2 \log R_{0j}(0) \xrightarrow{d} \chi^2_{(1)} \qquad -2 \log R_{1j}^{(i)}(0) \xrightarrow{d} \chi^2_{(1)} \qquad n \to \infty,$$

for every $d > 0$ and $d \to \infty$.

Furthermore, If $\theta_{0j} \neq 0$ and $\theta_{ij} \neq 0$, $i = 1, \dots, d$, $j = 1, \dots, d$, then

$$P\left(-2 \log R_{0j}(0) > M\right) \to 1 \qquad P\left(-2 \log R_{1j}^{(i)}(0) > M\right) \to 1 \qquad n \to \infty,$$

for every $d > 0$, $d \to \infty$ and $\forall M > 0$.

5. The GRID procedure

In this section we present the algorithm for estimating and testing the values of θ_{ij}, in order to classify the covariates of model (1). As said in the introduction, the acronym GRID has a twofold meaning: first, it derives from *Gradient Relevant Identification Derivatives*, meaning that the procedure is based on testing the significance of a partial derivative estimator; second, it refers to a graphical tool which can help in representing the identified structure of model (1). This is explained in the next section.

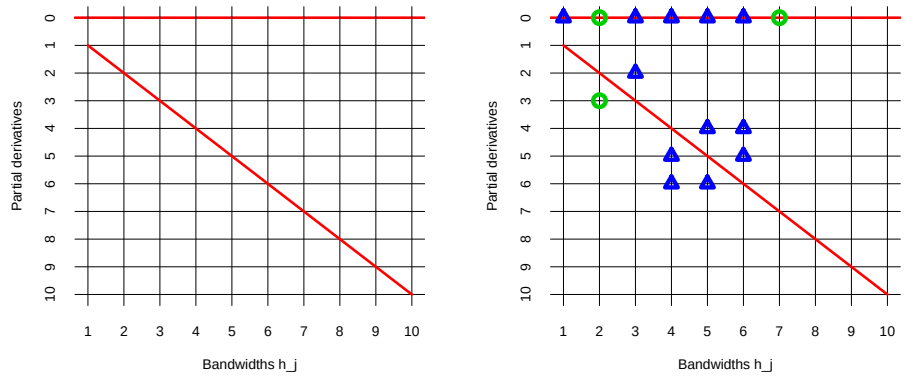
5.1 The GRID plot

Using the values θ_{ij}^m defined in the Theorem 1, we derive the following matrix

$$\Theta^m = E\left\{\frac{\partial M(x;H)}{\partial H}\right\} = \begin{pmatrix} \theta_{01}^m & \dots & \theta_{0d}^m \\ \theta_{11}^m & \dots & \theta_{1d}^m \\ \vdots & \ddots & \vdots \\ \theta_{d1}^m & \dots & \theta_{dd}^m \end{pmatrix}. \tag{26}$$

The matrix Θ_m joins the derivatives in (10), in the first row $i = 0$, with the derivatives in (11), in the other rows $i = 1, \dots, d$. So matrix Θ_m has dimension $(d+1) \times d$. Its values can be derived easily from expression (12), but they are also reported explicitly in the appendix. We can derive the matrix Θ^g in a similar way, using the values θ_{ij}^* defined in Theorem 2. The elements of these matrix are estimated through the (15).

A schematic representation of (the estimated) matrix Θ is made through the GRID-plot in figure 5.1, part a). The horizontal red line on the top denotes the position of the derivatives $\dot{M}_{0j}$, for $j = 1, \dots, d$. It shows the relevant variables classified in C and A (green dots for



(a) GRID plot when $d = 10$

(b) GRID representation of model (2)

Figure 1: A schematic representation of matrix Θ, by means of a grid of dimension $(d + 1, d)$ equivalent to θ, which is used to summarize the structure of model $m(x)$.

the nonlinear covariates in C and blu triangles for the linear covariates in A). The diagonal red line shows the positions of the cases $i = j$, which are excluded from our analysis. This is highlighted to help reading the other points. The other points of the GRID-plot refer to the derivatives $\dot{M}_{1j}^{(i)}$, for the cases $i \neq j$. They will indicate the presence of the mixed effect terms. In fact, the interactions between covariates come out reading the plot by row or by column. So, this part of the GRID-plot (*i.e.*, the whole matrix excluding row 0) is symmetric in terms of positions, but it can be asymmetric in terms of symbols (*i.e.*, when a linear variable is mixed to a nonlinear variable).

To give an idea about the GRID representation, part b) of figure 5.1 shows the GRID-plot for model (2). Here we see from row zero that there are 7 relevant covariates, among which 2 are *nonlinear*. Looking at rows 0 and 4, we can see that the covariate X_4 is linear and is mixed with other two linear covariates (X_5 and X_6). This is a *linear mixed effect*, given that it involves only *linear covariates*. There is also a *nonlinear mixed term*, which is represented by the couple circle-tringle involving the $2nd$ covariate (*nonlinear*) and the $3rd$ one (*linear*). We also see from rows 0 and 1 that the covariate X_1 is linear additive (no mixing effects), since the triangle is present in line zero but there are no points of interactions in line 1.

From a practical point of view, a point in position (i, j) of the GRID-plot indicates a positive test for the relative entry value of matrix Θ^m (or equivalently Θ^g), which means rejecting the null hypothesis $H_0 : \theta_{ij} = 0$, for $i = 0, \ldots, d$ and $j = 1, \ldots, d$, in a multiple testing fashion, as explained in section 4.

5.2 The algorithm

Let $X_{(j)}$ represent a Uniform covariate while $Z_{(j)}$ stands for the same covariate applying the transformation (18). The GRID procedure runs the following steps.

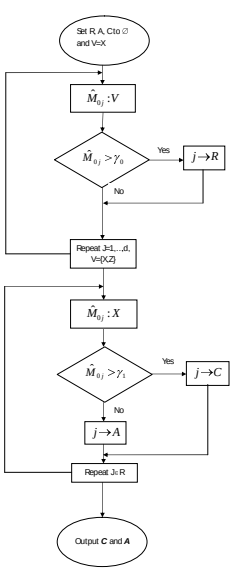
O. Set the bandwidth matrix to a high value ($H = h_d^* I_d$). Let $R = C \cup A$ be the set of relevant covariates. Initialize all the sets ($C, A, R, R_X, R_Z, \ldots$) to the empty set $\emptyset$.

I. **First stage** (*variable selection*):

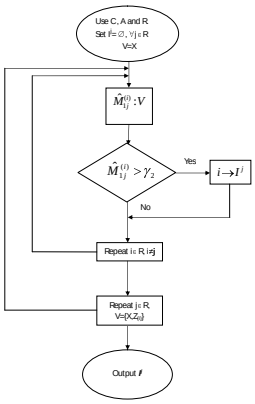
- For $j = 1, \ldots, d$, do:
  - using the covariates $X_{(j)}$, $j \in \Xi$, compute the (univariate) statistic $\dot{M}_{0j}$ defined in (15)
  - using EL, compute the threshold γ_0, as explained in section 4
  - if $\dot{M}_{0j} > \gamma_0$ then (*relevant covariate*)
  - insert the index j in the set R_X
  - using the covariates $Z_{(j)}$, $j \in \Xi$, compute the (univariate) statistic $\dot{M}_{0j}$ defined in (15)
  - using EL, compute the threshold γ_0, as explained in section 4
  - if $\dot{M}_{0j} > \gamma_0$ then (*relevant covariate*)
  - insert the index j in the set R_Z
- $R = R_X \cup R_Z$.
- For $j \in R$, do:
  - using the covariates $X_{(j)}$, $j \in R$, compute the (univariate) statistic $\dot{M}_{0j}$ defined in (15)
  - using the EL, compute the threshold γ_1, as explained in section 4
  - if $\dot{M}_{0j} > \gamma_1$ then (*nonlinear covariate*) then insert the index j in the set C and mark a green point on the GRID-plot, in position $(0, j)$
  - otherwise (*linear covariate*) insert the index j in the set A and mark a blue point on the GRID-plot, in position $(0, j)$.
- Output C, A

II. **Second stage** (*identifying the mixing terms*):

- For $j \in R$, do:
  - using the covariates $X_{(j)}$, $j \in R$, compute the (vectorized) statistic $\dot{M}_{1j}$ defined in (15)
  - For $i \in C$, $i \neq j$ do
    - using the EL for $\dot{M}_{1j}^{(i)}$, compute the thresholds γ_2 as explained in section 4
    - if $\dot{M}_{1j}^{(i)} > \gamma_2$ then (*interaction*) insert the index i in the set I_X^j
    - mark one point on the GRID-plot in positions (i, j), green if $j \in C$ and blue otherwise
  - For $i \in R$, $i \neq j$, do:
    - using the covariates $X_{(i)} \cup Z_{(j)}$, $j \in R$ and $j \neq i$, compute the (vectorized) statistic $\dot{M}_{1j}$ defined in (15)
    - using the EL for $\dot{M}_{1j}^{(i)}$, compute the thresholds γ_2 as explained in section 4
    - if $\dot{M}_{1j}^{(i)} > \gamma_2$ then (*interaction*) insert the index i in the set I_Z^j
    - mark one point on the GRID-plot in positions (i, j), green if $j \in C$ and blue otherwise
  - $I^j = I_X^j \cup I_Z^j$
- Output , I^j for $j \in R$.



(a) Variable Selection



(b) Model Selection

Figure 2: Flow-chart for the GRID procedure. Note that X stands for Uniform random variables while Z is the set of transformed random variables using (18). In particular, in figure (d), there is $Z_{(i)}$ which denotes the set of random variables Z except the covariate i which is Uniform, as described in the algorithm.

A. Proofs

In general, in the proofs of the LLE's properties we follow the classic approach used in Lu (1996) and Ruppert & Wand (1994), a part from three substantial differences. The first is that here the bandwidths do not tend to zero for $n \to \infty$ (see assumption $A1$). This implies that we must bound all the terms of the Taylor expansion with respect to $m(x)$, given that the size of the interval around the point x does not vanish with $n \to \infty$. For the same reason, we must also bound the terms of the Taylor expansion with respect to $f_X(x)$, the density function. To this aim, in Proposition 1 we consider assumption $A4$, so that the Taylor expansion is exact with respect to f_X. Then we relax assumption $A4$ in Theorem 2. Finally, we want to analyze carefully the influences of the bandwidths associated to the different covariates on the bias of $\hat{\beta}(x;H)$. As a consequence, we will partition all the involved matrices along the index sets $\{C, A_c, A_u, U\}$.

Proof of Proposition 1: The conditional bias of the LLE is given by

$$E(\hat{\beta}(x;H)|X_1,\ldots,X_n) - \beta(x) = (\Gamma^T W \Gamma)^{-1}\Gamma^T W(M - \Gamma\beta(x))$$

where $M = (m(X_1),\ldots,m(X_n))^T$ and $\beta(x) = (m(x), \mathbb{D}_m^T(x))^T$. Note that, given $u_t = H^{-1}(X_t - x)$, we have

$$n^{-1}\Gamma^T W\Gamma = diag(1,H)S_n\, diag(1,H) \quad (27)$$

$$n^{-1}\Gamma^T W(M - \Gamma\beta(x)) = diag(1,H)R_n \quad (28)$$

with

$$\begin{aligned} S_n &= \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} 1 & u_t^T \\ u_t & u_t u_t^T \end{pmatrix}|H|^{-1}K(u_t) \\ R_n &= \frac{1}{n}\sum_{t=1}^{n}\begin{pmatrix} 1 \\ u_t \end{pmatrix}\left[m(X_t) - m(x) - \mathbb{D}_m^T(x)Hu_t\right]|H|^{-1}K(u_t), \end{aligned}$$

so the bias can be simply written as

$$E(\hat{\beta}(x;H)|X_1,\ldots,X_n) - \beta(x) = diag(1, H^{-1})S_n^{-1}R_n. \tag{29}$$

For S_n, using Taylor's expansion and assumption $A4$, we have

$$\begin{aligned} S_n &= \int\begin{pmatrix} 1 & u^T \\ u & uu^T \end{pmatrix}K(u)f_X(x+Hu)du + O_p(n^{-1/2}) \\ &= f_X(x)\int\begin{pmatrix} 1 & u^T \\ u & uu^T \end{pmatrix}K(u)du + \int\begin{pmatrix} 1 & u^T \\ u & uu^T \end{pmatrix}\left[\mathbb{D}_f^T(x)Hu\right]K(u)du \\ &+ O_p(n^{-1/2}) \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 I_d \end{pmatrix} + O_p(n^{-1/2}). \end{aligned} \tag{30}$$

For the analysis of R_n, we need to introduce some further notation. Given assumption $A3$, let define the vth-order differential $D_m^v(x;y)$ as

$$D_m^v(x,y) = \sum_{i_1,\ldots,i_d}\frac{v!}{i_1!\times\ldots\times i_d!}\frac{\partial^v m(x)}{\partial x_1^{i_1}\ldots\partial x_d^{i_d}}y_1^{i_1}\times\ldots\times y_d^{i_d}, \tag{31}$$

where the summation is over all distinct nonnegative integers $i_1,\ldots,i_d$ such that $i_1 + \ldots + i_d = v$. Using the Taylor's expansion to approximate the function $m(X_t)$, and the assumption $A4$, we can write

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

where R_n^* represents the residual term, which depends on the higher order derivatives of the function $m(x)$. This element will be discussed later. Now remember that the odd-order moments of the kernel product are null, so some of the terms in the v-th order differentials cancel. We have

$$R_n = \int\begin{pmatrix} \frac{1}{2!}D_m^2(x,Hu) \\ \frac{1}{3!}uD_m^3(x,Hu) \end{pmatrix}K(u)du = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} + R_n^* + O_p(n^{-1/2}) \tag{32}$$

where γ_1 is a scalar, while γ_2 is a d-dimensional vector. Solving the integrals and applying the properties of the kernel we have

$$\begin{aligned}\gamma_1 &= \int \frac{1}{2} D_m^2(x, Hu) K(u) du = \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\partial^2 m(x)}{\partial x_i \partial x_j} h_i h_j \int u_i u_j K(u) du \\ &= \frac{1}{2} \mu_2 \sum_{i=1}^{d} \frac{\partial^2 m(x)}{\partial x_i \partial x_i} h_i^2 = \frac{1}{2} \mu_2 \, tr\{H \mathbb{H}_m(x) H\}.\end{aligned}$$

The component γ_2 is a vector of length d. Its element of position r is

$$\begin{aligned}\gamma_2^{(r)} &= \int \frac{1}{6} u_r D_m^3(x, Hu) K(u) du \\ &= \sum_{i_1, \ldots, i_d} \frac{h_1^{i_1} \cdots h_d^{i_d}}{i_1! \times \ldots \times i_d!} \frac{\partial^3 m(x)}{\partial x_1^{i_1} \cdots \partial x_d^{i_d}} \int u_1^{i_1} \cdots u_r^{i_r+1} \cdots u_d^{i_d} K(u) du \\ &= \left[\sum_{s \neq r} \frac{1}{2} \mu_2^2 \frac{\partial^3 m(x)}{\partial x_r \partial x_s^2} h_r h_s^2 + \frac{1}{6} \mu_4 \frac{\partial^3 m(x)}{\partial x_r^3} h_r^3 \right],\end{aligned}$$

while the whole vector γ_2 is equal to

$$\gamma_2 = \frac{1}{2} \mu_2^2 \left[H \mathbb{G}_m(x) H^2 + \left(\frac{\mu_4}{3\mu_2^2} - 1 \right) diag\{H \mathbb{G}_m(x) H^2\} \right] \mathbf{1}.$$

concerning the residual term R_n^*, using the assumption $A3$ and remembering the (31), we can define

$$\delta(D_m^v, H_C^v) \leq \sum_{i_1, \ldots, i_k} \frac{v!}{i_1! \cdots i_k!} \sup_{x \in [0,1]^d} \frac{\partial^v m(x)}{\partial x_1^{i_1} \ldots \partial x_k^{i_k}} h_1^{i_1} \times \ldots \times h_k^{i_k} < \infty,$$

where the sum is taken for all the combination of indexes $i_1 + \ldots + i_k = v$. Note that $\delta(D_m^v, H_C)$ depends on the derivatives of total order v, which are bounded given assumption $A3$. Moreover, it depends only on the bandwidths in H_C. Then

$$R_n^* = \begin{pmatrix} \delta(D_m^4, H_C^4) \\ \delta(D_m^5, H_C^5) \end{pmatrix}.$$

Combining the (29), (30) and (32), we obtain

$$\begin{aligned}&E(\hat{\beta}(x; H) | X_1, \ldots, X_n) - \beta(x) = diag\{1, H^{-1}\} S_n^{-1} R_n \\ &\approx \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\mu_2} H^{-1} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \\ &= \frac{1}{2} \mu_2 \begin{pmatrix} tr(H \mathbb{H}_m H) \\ \mathbb{G}_m(x) H^2 \mathbf{1} + \left(\frac{\mu_4}{3\mu_2^2} - 1 \right) diag\{\mathbb{G}_m(x) H^2\} \mathbf{1} \end{pmatrix}. \qquad (33)\end{aligned}$$

Then we can further detail these expressions remembering the (6) and (7), and noting that $\forall v, w$

$$H^v \mathbb{H}_m(x) H^w = \begin{pmatrix} H_C^v \mathbb{H}_m^C(x) H_C^w & H_C^v \mathbb{H}_m^{CA_c}(x) H_{A_c}^w & 0 & 0 \\ H_{A_c}^v \mathbb{H}_m^{A_c C}(x) H_C^w & H_{A_c}^v \mathbb{H}_m^{A_c}(x) H_{A_c}^w & 0 & 0 \\ 0 & 0 & H_{A_u}^v \mathbb{H}_m^{A_u}(x) H_{A_u}^w & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$H^v\mathbb{G}_m(x)H^w = \begin{pmatrix} H_C^v\mathbb{G}_m^C(x)H_C^w & 0 & 0 & 0 \\ H_{A_c}^v\mathbb{G}_m^{A_cC}(x)H_C^w & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

After some algebra, we have the result of the Proposition. □

Proof of Theorem 1:
It is sufficient to use the results shown in Proposition 1 w.r.t. the estimators $\dot{M}_{0j}$ and $\dot{M}_{1j}$ defined in (15). Remembering the (27) and (28), we can write

$$\begin{aligned} & E(M(x;H)|X_1,\ldots,X_n) - \beta(x) = \frac{1}{n}\, diag(1,H^{-2})\Gamma^T W M - \beta(x) \\ = \;& \frac{1}{n}\, diag(1,H^{-2})\Gamma^T W\,(M - \Gamma\beta(x)) + \frac{1}{n}\, diag(1,H^{-2})\Gamma^T W\Gamma\beta(x) - \beta(x) \\ = \;& diag(1,H^{-1})R_n + \left[diag(1,H^{-1})S_n\, diag(1,H) - I_{d+1}\right]\beta(x). \end{aligned}$$

Now, using assumption $A4$ and the results of Proposition 1, the bias of the estimator (15) is

$$E(M(x;H)) - \beta(x) = \begin{pmatrix} b_m(x;H_C) \\ \mu_2 B_{\mathbb{D}}(x;H_C) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & (\mu_2-1)I_d \end{pmatrix}\beta(x). \tag{34}$$

So we obtain

$$E\left(M_0(x;H)\right) - m(x) = b_m(x;H_C)$$

where the estimator $M_0(x;H)$ is defined in (13) and $b_m(x;H_C)$ is the bias of LLE as in Proposition (1). Taking the derivative w.r.t. h_j, at both sides, we have

$$\theta_{0j}^m = E\left(\dot{M}_{0j}\right) = \frac{\partial}{\partial h_j} b_m(x;H_C).$$

Since $b_m(x;H_C)$ depends on the bandwidths of the covariates in C, the first part of theorem is shown. The detailed expressions of the expected derivatives θ_{0j}^m are

$$\theta_{0j}^m = \begin{cases} h_j\mu_2\frac{\partial^2 m(x)}{\partial^2 x_j} + \delta(D_{m_j}^4;H_C^4) & \text{if } j \in C \\ 0 & \text{otherwise} \end{cases}. \tag{35}$$

Note that $\delta(D_{m_j}^4;H_C^4)$ depends on the partial derivatives of order 4 involving the j-th covariate, where $j \in C$, which are bounded given assumption $A3$. So, it is equal to zero when $\partial^2 m(x)/\partial x_j^2 = 0$.

Now we consider the estimator $M_1(x;H)$ in (13). Using again the (34) and the same arguments as in the proof of Proposition (1), we have

$$E\left(M_1^{(i)}(x;H)\right) - \mathbb{D}_m^{(i)}(x) = \mu_2 B_{\mathbb{D}}^{(i)}(x;H_C) + (\mu_2-1)\mathbb{D}_m^{(i)}(x)$$

where (i) stands for the component of position (i) in the vectors $M_1(x;H)$, $\mathbb{D}_m(x)$ and $B_{\mathbb{D}}(x;H_C)$, $i = 1\ldots,d$. The quantity $B_{\mathbb{D}}(x;H_C)$ is defined in Proposition (1) and μ_2 is the second moment of the Kernel K. As before, taking the derivative w.r.t. h_j, at both sides, it follows

$$\theta_{ij}^m = E\left(\dot{M}_{1j}^{(i)}\right) = \frac{\partial}{\partial h_j}\mu_2 B_{\mathbb{D}}^{(i)}(x;H_C), \qquad \forall i,j = 1\ldots,d;\; i \neq j.$$

Using (9) in Proposition (1) and remembering the (7), we know that $\frac{\partial}{\partial h_j} B_{\mathbb{D}}^{(i)}(x; H_C) \neq 0$ if and only if $i \in I^j$ and $j \in C$. Note that, for $i \neq j$, I^j stands for set of covariates (linear or nonlinear) which are mixed with the covariate j.

The formula of the expected derivatives θ_{ij}^m are

$$\theta_{ij}^m = \begin{cases} h_j \mu_2^2 \frac{\partial^3 m(x)}{\partial x_i \partial x_j^2} + \delta(D_{m_{ij}}^5; H_C^5) & \text{if } i \in I^j, j \in C \\ 0 & \text{otherwise} \end{cases} . \tag{36}$$

It can be shown that $\delta(D_{m_{ij}}^5; H_C^5)$ includes the partial derivatives of order 5 involving both the i-th and j-th covariates. They are bounded given the assumption $A3$, and they are all equal to zero when $\partial^3 m(x)/\partial x_i \partial x_j^2 = 0$. □

Proof of Theorem 2: The key aspect of this proof is to show that the higher order terms of the Taylor expansion of R_n do not compromise the results of our procedure. In fact, when $A4$ is not assumed, the derivatives of the density function f_X are different from zero, introducing further components in the Taylor expansion of R_n. Moreover, given assumption $A1$, such higher order terms may not vanish, contrary to what happens in the classic framework of local linear estimators, where the bandwidths tend to zero for $n \to \infty$.

In particular, the transformation $Z_t = \phi(X_t)$ defined in (18) applied to the uniform covariates X_t implies that the marginal density of each transformed covariate is linear, being equal to

$$f_Z(z_i) = 2z_i, \qquad i \in A_u \cup U.$$

Because of the linearity, the gradient and the Hessian matrix of the density function $f_Z(z)$ are the following

$$\mathbb{D}_f(z) = \begin{pmatrix} 0 \\ \mathbb{D}_f^{A_c} \\ \mathbb{D}_f^{A_u} \\ \mathbb{D}_f^{U} \end{pmatrix} \qquad \mathbb{H}_f(z) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \mathbb{H}_f^{A_c} & \mathbb{H}_f^{A_c A_u} & \mathbb{H}_f^{A_c U} \\ 0 & \mathbb{H}_f^{A_u A_c} & \mathbb{H}_f^{A_u} & \mathbb{H}_f^{A_u U} \\ 0 & \mathbb{H}_f^{U A_c} & \mathbb{H}_f^{U A_u} & \mathbb{H}_f^{U} \end{pmatrix}, \tag{37}$$

with the diagonal of $\mathbb{H}_f$ equal to zero. This implies that we need to consider the Taylor expansion of R_n w.r.t. the derivatives of f_z up to order 4.

Now remember that, using the transformed variables $Z_t = \phi(X_t)$ defined in (18), the auxiliary regression function becomes

$$g(z) = g_1(x_C, z_{A_c}) + g_2(z_{A_c}, z_{A_u}).$$

Therefore, given that the aim here is to identify the covariates in $A = A_c \cup A_u$, we can focus on function g_2. This is further justified by the structure of $\mathbb{D}_f(z)$ and $\mathbb{H}_f(z)$ shown in (37). So, without loss of generality, we can use $z = (z_{A_c}, z_{A_u}, z_U)$ of $d - k$ dimension. The gradient and the Hessian matrix of function g become

$$\mathbb{D}_g(z) = \begin{pmatrix} \mathbb{D}_g^{A_c} \\ \mathbb{D}_g^{A_u} \\ 0 \end{pmatrix} \qquad \mathbb{H}_g(z) = \begin{pmatrix} \mathbb{H}_g^{A_c} & \mathbb{H}_g^{A_c A_u} & 0 \\ \mathbb{H}_g^{A_u A_c} & \mathbb{H}_g^{A_u} & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{38}$$

where the submatrices have changed compared with the first regression (in particular, note that $\mathbb{H}_g^{A_c}$ and $\mathbb{H}_g^{A_u}$ have not a zero diagonal). Moreover, the matrix $\mathbb{G}_g(z)$ becomes

$$\mathbb{G}_g(z) = \begin{pmatrix} \mathbb{G}_g^{A_c} & \mathbb{G}_g^{A_c A_u} & 0 \\ \mathbb{G}_g^{A_u A_c} & \mathbb{G}_g^{A_u} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{39}$$

where the submatrices are defined as usual. In particular, note that $\mathbb{G}_g^{A_c}$ and $\mathbb{G}_g^{A_u}$ are full matrices with zero diagonal, as a consequence of the ϕ-tranformation of the covariates in A. Moreover we define with H_Z, Γ_Z, W_Z and M_Z the corresponding quantities w.r.t. H, Γ, W and M using z whose dimension is $d-k$. Finally, let $\mathbb{D}_f^Z$ and $\mathbb{H}_f^Z$ be the same quantities as in (37) without the zeros. So that $\mathbb{D}_f^Z$ is a vector of dimension $d-k$ and $\mathbb{H}_f^Z$ is a matrix with $d-k$ rows and $d-k$ columns.

Using again the (27) and (28), we can write

$$\begin{aligned} & E(M(z;H_Z)|Z_1,\ldots,Z_n) - \beta(z) \\ = \; & \frac{1}{n}\, diag(1, H_Z^{-2})\Gamma_Z{}^T W_Z M_Z - \beta(z) \\ = \; & diag(1, H_Z^{-1}) R_n + \left[diag(1, H_Z^{-1}) S_n\, diag(1, H_Z) - I_{d+1-k} \right] \beta(z), \end{aligned}$$

but now we have to consider the higher order terms of S_n and R_n induced by f_Z.

Let us consider the vector R_n in the (32) with the additional nonzero terms of the Taylor expansion w.r.t. f_Z, using the assumptions $A1-A4$ and the transformed covariates. We have

$$\begin{aligned} E(R_n) &= f_Z(z) \int \begin{pmatrix} \frac{1}{2!} D_g^2(z, H_Z u) \\ \frac{1}{3!} u D_g^3(z, H_Z u) \end{pmatrix} K(u) du \\ &+ \int \begin{pmatrix} \frac{1}{3!} D_g^3(z, H_Z u)[(\mathbb{D}_f^Z(z))^T H_Z u] \\ \frac{1}{2!} u D_g^2(z, H_Z u)[(\mathbb{D}_f^Z(z))^T H_Z u] \end{pmatrix} K(u) du + \\ &\quad + \int \begin{pmatrix} \frac{1}{2!} D_g^2(z, H_Z u)[\frac{1}{2} u^T H_Z \mathbb{H}_f^Z(z) H_Z u] \\ \frac{1}{3!} u D_g^3(z, H_Z u)[\frac{1}{2} u^T H_Z \mathbb{H}_f^Z(z) H_Z u] \end{pmatrix} K(u) du \\ &\quad + \int \begin{pmatrix} \frac{1}{2!} D_g^2(z, H_Z u)[\frac{1}{4!} D_f^4(z, H_Z u)] \\ \frac{1}{3!} u D_g^3(z, H_Z u)[\frac{1}{4!} D_f^4(z, H_Z u)] \end{pmatrix} K(u) du \\ &= R_0 + R_1 + R_2 + R_3, \end{aligned}$$

where the four terms are equal to

$$R_0 = \frac{1}{2}\mu_2 f_Z(z) \begin{pmatrix} tr\left(H_Z \mathbb{H}_g H_Z\right) \\ \mu_2 \left[H_Z \mathbb{G}_g H_Z^2\right] \mathbf{1} \end{pmatrix}$$

$$R_1 = \frac{1}{2}\mu_2^2 \begin{pmatrix} (\mathbb{D}_f^Z)^T \left[H_Z^2 \mathbb{G}_g H_Z^2\right] \mathbf{1} \\ \left[2 H_Z \mathbb{H}_g H_Z^2 + H_Z\, tr\left(H_Z \mathbb{H}_g H_Z\right) + \left(\frac{\mu_4}{\mu_2^2} - 3\right) diag\left(H_Z \mathbb{H}_g H_Z^2\right)\right] \mathbb{D}_f^Z \end{pmatrix}$$

$$R_2 = \frac{1}{2}\mu_2 \begin{pmatrix} \mu_2\, tr\left(\mathbb{H}_f^Z H_Z^2 \mathbb{H}_g H_Z^2\right) \\ \mu_2^2 H_Z \mathbb{H}_f^Z H_Z^2 \mathbb{G}_g H_Z^2 \mathbf{1} + (\mu_4 - \mu_2^2)\, diag\left(H_Z \mathbb{H}_f^Z H_Z^2 \mathbb{G}_g H_Z^2\right) \mathbf{1} + H_Z \mathbb{J}_g \end{pmatrix}$$

$$R_3 = \mu_2^4 \begin{pmatrix} 0 \\ H_Z \mathbb{L}_g \end{pmatrix},$$

and $\mathbb{J}_g$ and $\mathbb{L}_g$ are vectors whose i-th element is equal to

$$\begin{aligned} \mathbb{J}_g^{(i)} &= 3\mu_2^2 \sum_{s \neq i} \sum_{j \neq \{s,i\}} \frac{\partial^2 f_Z(z)}{\partial z_s \partial z_j} h_s^2 \frac{\partial^3 g(z)}{\partial z_i \partial z_s \partial z_j} h_j^2 \\ \mathbb{L}_g^{(i)} &= \sum_{j \neq i} \sum_{s \neq \{j,i\}} \sum_{u \neq \{j,s,i\}} \frac{\partial^3 g(z)}{\partial z_j \partial z_s \partial z_u} \frac{\partial^4 f_Z(z)}{\partial z_i \partial z_j \partial z_s \partial z_u} h_j^2 h_s^2 h_u^2. \end{aligned}$$

Note that the vector R_2 has been derived exploiting the simple structure of $\mathbb{H}_f^Z$ and $\mathbb{G}_g$ (both with zero diagonal). For simplicity, we do not consider here the residual term R_n^*, which can be analyzed following the same arguments as in Proposition 1.

For S_n, using Taylor's expansion and assumptions $A1 - A4$ with the transformed covariates Z, we have

$$\begin{aligned} E(S_n) &= \int \begin{pmatrix} 1 & u^T \\ u & uu^T \end{pmatrix} K(u) f_Z(z + H_Z u) du \\ &= f_Z(z) \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 I_{d-k} \end{pmatrix} + \begin{pmatrix} 0 & \mu_2 (\mathbb{D}_f^Z)^T H_Z \\ \mu_2 H_Z \mathbb{D}_f^Z & 0 \end{pmatrix} \\ &+ \begin{pmatrix} 0 & 0 \\ 0 & \mu_2 H_Z \mathbb{H}_f^Z H_Z \end{pmatrix} \\ &= \begin{pmatrix} f_Z(z) & \mu_2 (\mathbb{D}_f^Z)^T H_Z \\ \mu_2 H_Z \mathbb{D}_f^Z & f_Z(z) \mu_2 I_{d-k} + \mu_2 H_Z \mathbb{H}_f^Z H_Z \end{pmatrix}. \end{aligned}$$

Note, again, that we have derived the previous result using $tr(H_Z \mathbb{H}_f^Z H_Z) = 0$, given the linearity of f_Z.

The bias of the estimator (15), using the transformed covariates, becomes

$$\begin{aligned} E(M(z; H_Z)) - \beta(z) &= diag(1, H_Z^{-1})(R_0 + R_1 + R_2 + R_3) \\ &+ \left[diag(1, H_Z^{-1})\, E(S_n)\, diag(1, H_Z) - I_{d+1-k} \right] \beta(z), \end{aligned} \tag{40}$$

where

$$diag((1, H_Z^{-1}))R_0 = \frac{1}{2}\mu_2 f_Z(z) \begin{pmatrix} tr\left(H_Z \mathbb{H}_g H_Z\right) \\ \mu_2 \left[\mathbb{G}_g H_Z^2\right] \mathbf{1} \end{pmatrix}$$

$$diag((1, H_Z^{-1}))R_1 = \frac{1}{2}\mu_2^2 \begin{pmatrix} (\mathbb{D}_f^Z)^T \left[H_Z^2 \mathbb{G}_g H_Z^2\right] \mathbf{1} \\ \left[2\mathbb{H}_g H_Z^2 + tr\left(H_Z \mathbb{H}_g H_Z\right) I_{d-k} + \left(\frac{\mu_4}{\mu_2^2} - 3\right) diag\left(\mathbb{H}_g H_Z^2\right)\right] \mathbb{D}_f^Z \end{pmatrix}$$

$$diag((1, H_Z^{-1}))R_2 = \frac{1}{2}\mu_2 \begin{pmatrix} \mu_2\, tr\left(\mathbb{H}_f^Z H_Z^2 \mathbb{H}_g H_Z^2\right) \\ \mu_2^2 \mathbb{H}_f^Z H_Z^2 \mathbb{G}_g H_Z^2 \mathbf{1} + (\mu_4 - \mu_2^2)\, diag\left(\mathbb{H}_f^Z H_Z^2 \mathbb{G}_g H_Z^2\right) \mathbf{1} + \mathbb{J}_g \end{pmatrix}$$

$$diag((1, H_Z^{-1}))R_3 = \mu_2^4 \begin{pmatrix} 0 \\ \mathbb{L}_g \end{pmatrix}$$

and

$$\begin{aligned} &\left[diag(1, H_Z^{-1})\, E(S_n)\, diag(1, H_Z) - I_{d+1-k} \right] \beta(z) \\ &= \begin{pmatrix} [f_Z(z) - 1] g_2(z) + \mu_2 (\mathbb{D}_f^Z)^T H_Z \mathbb{D}_g \\ \mu_2 H_Z \mathbb{D}_f^Z g_2(z) + [\mu_2 f_Z(z) - 1] \mathbb{D}_g + \mu_2 H_Z \mathbb{H}_f^Z H_Z \mathbb{D}_g \end{pmatrix}. \end{aligned}$$

Now, the first part of the theorem, in the (19), can be easily shown observing the first component of each vector. Note that the bandwidth matrix H_Z appears always multiplied by $\mathbb{D}_g(z)$, $\mathbb{H}_g(z)$ or $\mathbb{G}_g(z)$. So, given v and w, we have

$$\begin{aligned} H_Z^v \mathbb{H}_g(z) H_Z^w &= \begin{pmatrix} H_{A_c}^v \mathbb{H}_g^{A_c}(z) H_{A_c}^w & H_{A_c}^v \mathbb{H}_g^{A_c A_u}(z) H_{A_c}^w & 0 \\ H_{A_u A_c}^v \mathbb{H}_g^{A_c}(z) H_{A_c}^w & H_{A_u}^v \mathbb{H}_g^{A_u}(z) H_{A_u}^w & 0 \\ 0 & 0 & 0 \end{pmatrix} = \delta(H_{A_c}, H_{A_u}) \\ H_Z^v \mathbb{G}_g(z) H_Z^w &= \delta(H_{A_c}, H_{A_u}) \\ H_Z \mathbb{D}_g(z) &= \delta(H_{A_c}, H_{A_u}). \end{aligned}$$

Therefore, we obtain

$$E\left(M_0(z;H_Z)\right) - g_2(z) = \delta(H_{A_c}, H_{A_u}) + [f_Z(z) - 1]g_2(z)]$$

where the estimator $M_0(z;H_Z)$ is defined in (13) and uses the transformed covariates $Z_{(j)}, j \in A$. Taking the derivative w.r.t. h_j, at both sides, we have

$$\theta^g_{0j} = E\left(\partial M_0(z;H_Z)/\partial h_j\right) = 0 \qquad \forall j \notin A.$$

In oder to prove the second part of the theorem, in the (20), we must consider the second element of each vector in the (40). Following the same arguments as before, it can be shown that

$$E\left(M_1(z;H_Z)\right) - \mathbb{D}_g(z) = \delta(H_{A_c}, H_{A_u}) + [\mu_2 f_Z(z) - 1]\mathbb{D}_g(z) + \mu_2 H_Z \mathbb{D}^Z_f g_2(z)$$

which implies that

$$\theta^g_{ij} = E\left(\partial M_1^{(i)}(z;H_Z)/\partial h_j\right) = 0 \qquad \forall i \in \Xi, j \in U \qquad i \neq j. \tag{41}$$

Anyway, we need to analyze θ^g_{ij} for $i \neq j$ and $j \in A$, given that these values can be used to identify the mixed effect terms. So, we derive the exact formula of θ^g_{ij}, for $i \neq j$, equal to

$$\begin{aligned}
\theta^g_{ij} &= \mu_2 f_Z(z) h_j \left[\mu_2 \frac{\partial^3 g_2(z)}{\partial z_i \partial z_j^2} + 2\mu_2 \frac{\partial^2 g_2(z)}{\partial z_i \partial z_j} \frac{\partial \log f_Z(z)}{\partial z_j}\right] + \qquad (42)\\
&+ 6\mu_2^4 h_j \sum_{s \neq \{i,j\}} \frac{\partial^3 g_2(z)}{\partial z_s \partial z_i \partial z_j} \frac{\partial^2 f_Z(z)}{\partial z_s \partial z_j} + \mu_2 h_j \left[(\mu_4 - \mu_2^2) h_i^2 \frac{\partial^3 g_2(z)}{\partial z_j \partial z_i^2} \frac{\partial^2 f_Z(z)}{\partial z_i \partial z_j}\right] +\\
&+ \mu_2 h_j \left[\frac{\partial^2 g_2(z)}{\partial z_j \partial z_j} \frac{\partial f_Z(z)}{\partial z_i} + \mu_2^2 \sum_{s \neq i} \frac{\partial^3 g_2(z)}{\partial z_s \partial z_j^2} \frac{\partial^2 f_Z(z)}{\partial z_i \partial z_s} h_s^2 + \frac{\partial g_2(z)}{\partial z_j} \frac{\partial^2 f_Z(z)}{\partial z_i \partial z_j}\right] +\\
&+ 2\mu_2^4 h_j \sum_{s \neq \{j,i\}} \sum_{u \neq \{j,s,i\}} \frac{\partial^3 g_2(z)}{\partial z_j \partial z_s \partial z_u} \frac{\partial^4 f_Z(z)}{\partial z_i \partial z_j \partial z_s \partial z_u} h_s^2 h_u^2.
\end{aligned}$$

We can see from the previous formula that θ^g_{ij} is different from zero if there are mixed effects between the two covariates $Z_{(i)}$ and $Z_{(j)}$, that is when $i \in I_j$. Anyway, it can be different from zero also when $i \notin I^j$, given that there are the terms in the third and fourth rows of the previous formula which depend on the partial derivatives of g_2 w.r.t. covariates different from (i,j). In order to make this as a one to one relation, it is necessary to force to zero the third and fourth rows of the formula. This can be done considering a uniform density for $Z_{(i)}$, because in such a case all the terms in the third and fourth rows (but also the term in the second row) would be canceled by the derivative $\partial f_Z(z)/\partial z_i$, equal to zero. Therefore, we suggest not to transform the i-th covariate in the auxiliary regression. This proves the third part of the theorem, in the (21).

The formula of the expected derivative θ^*_{ij}, for $i \neq j$, is

$$\theta^*_{ij} = \mu_2^2 f_Z(z) h_j \left[\frac{\partial^3 g_2(z)}{\partial z_i \partial z_j^2} + 2\frac{\partial^2 g_2(z)}{\partial z_i \partial z_j} \frac{\partial \log f_Z(z)}{\partial z_j}\right] + 6\mu_2^4 h_j \sum_{s \neq i,j} \frac{\partial^3 g_2(z)}{\partial z_s \partial z_i \partial z_j} \frac{\partial^2 f_Z(z)}{\partial z_s \partial z_j}. \tag{43}$$

□

Proof of Proposition 2:

First, we have

$$n\,E\left(Var(\dot{M}_{0j}\,|X_1,\ldots,X_n)\right) = \sigma^2\,E\left[\frac{\partial}{\partial h_j}\frac{1}{|H|^2}K^2\left(H^{-1}\,(X-x)\right)\right].$$

By assumptions (A1)-(A4) we can change the mean operator w.r.t. the derivative operator. Therefore,

$$n\,E\left(Var(\dot{M}_{0j}\,|X_1,\ldots,X_n)\right) = \sigma^2\left[\frac{\partial}{\partial h_j}\left(|H|^{-1/2}\right)\right]^2 E\left[|H|^{-1}K^2\left(H^{-1}\,(X-x)\right)\right].$$

Changing the variable $u = H^{-1}\,(X-x)$ we have the result in (*i*).

For point (*ii*), we have

$$n\,E\left(Var(\dot{M}_{1j}\,|X_1,\ldots,X_n)\right)$$

$$= \sigma^2\,E\left[\frac{\partial}{\partial h_j}\left(H^{-2}(X-x)\frac{1}{|H|^2}K^2\left(H^{-1}\,(X-x)\right)(X-x)^T H^{-2}\right)\right].$$

Changing the variable $u = H^{-1}\,(X-x)$ and using the same arguments as in point (*i*), it follows

$$n\,E\left(Var(\dot{M}_{1j}\,|X_1,\ldots,X_n)\right)$$

$$= \sigma^2\left[\frac{\partial}{\partial h_j}\left(H^{-1}|H|^{-1/2}\right)\right]^2 E\left(uu^T K^2(u)\right) = \sigma^2\frac{\nu_2\nu_0^{d-1}}{4|H|h_j^2}H^{-2}I_j.$$

□

References

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

Chen X.S., Peng L. and Qin Y.L. (2009) Effects of data dimension on empirical likelihood. *Biometrika*, 96, 711–722.

DiCiccio T., Hall P. and Romano J. (1991) Empirical Likelihood is Bartlett-Correctable. *The Annals of Statistics*, 19, 1053–1061.

Hall P. (1992) *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hall P., La Scala B. (1990) Methodology and Algorithms of Empirical Likelihood. *International Statistical Review*, 58, 109–127.

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

Lahiri S.N., Mukhopadhyay S. (2012) Supplement to "'A penalized empirical likelihood method in high dimensions"'. DOI:10.1214/12-AOS1040SUPP .

Lahiri S.N., Mukhopadhyay S. (2012) A penalized empirical likelihood method in high dimensions. *The Annals of Statistics*, 40, 2511–2540.

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

Masry E. (1996) Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17, 571–599.

Owen A. (1990) Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18, 90–120.

Radchenko P., James G.M. (2010) Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of American Statistical Association*, 105, 1541–1553.

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

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

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