The Annals of Statistics
2020, Vol. 48, No. 3, 1848–1874
https://doi.org/10.1214/19-AOS1846

GRID: A VARIABLE SELECTION AND STRUCTURE DISCOVERY METHOD FOR HIGH DIMENSIONAL NONPARAMETRIC REGRESSION

BY FRANCESCO GIORDANO[1,*], SOUMENDRA NATH LAHIRI[2] AND MARIA LUCIA PARRELLA[1,†]

[1]*Department of Economics and Statistics, University of Salerno,* [*]*giordano@unisa.it;* [†]*mparrella@unisa.it*
[2]*Department of Statistics, North Carolina State University, snlahiri@ncsu.edu*

We consider nonparametric regression in high dimensions where only a relatively small subset of a large number of variables are relevant and may have nonlinear effects on the response. We develop methods for variable selection, structure discovery and estimation of the true low-dimensional regression function, allowing any degree of interactions among the relevant variables that need not be specified *a-priori*. The proposed method, called the GRID, combines empirical likelihood based marginal testing with the local linear estimation machinery in a novel way to select the relevant variables. Further, it provides a simple graphical tool for identifying the low dimensional nonlinear structure of the regression function. Theoretical results establish consistency of variable selection and structure discovery, and also Oracle risk property of the GRID estimator of the regression function, allowing the dimension d of the covariates to grow with the sample size n at the rate $d = O(n^a)$ for any $a \in (0, \infty)$ and the number of relevant covariates r to grow at a rate $r = O(n^\gamma)$ for some $\gamma \in (0, 1)$ under some regularity conditions that, in particular, require finiteness of certain absolute moments of the error variables depending on a. Finite sample properties of the GRID are investigated in a moderately large simulation study.

1. Introduction. Extraction of low dimensional structures in high dimensional data is a challenging task. It requires selection of relevant variables as well as the estimation of the resulting low dimensional structure. For high dimensional sparse linear regression models, important methodological advancement has been made in recent years where different variable selection and screening methods have been proposed and accurate estimation of the nonzero regression coefficients is accomplished. Some of the most popular methods include the LASSO and its variants (cf. Tibshirani (1996), GLASSO in Friedman, Hastie and Tibshirani (2007) and Meinshausen and Bühlmann (2006), COSSO in Lin and Zhang (2006), DASSO in James, Radchenko and Lv (2009)), the SCAD (cf. Fan and Li (2001)), the Adaptive LASSO of Zou (2006) and the MCP (cf. Zhang (2010)), among others. Variable selection and screening in the high dimensional linear model context include the SIS method of Fan and Lv (2008) and the marginal test based method of Chang, Tang and Wu (2013).

A more challenging and often a more realistic problem is to select the relevant variables that possibly have nonlinear interactions with the response. For nonlinear variables, the methods developed for the linear case do not perform satisfactorily. The literature on nonlinear variable selection and estimation is relatively sparse. Hall and Miller (2009) developed a nonlinear variable selection method based on a generalized correlation measure. Simultaneous selection and estimation of the nonlinear effects of covariates on the response variable is a significantly more difficult problem. Even when the exact set of relevant variables are given, estimation of the nonlinear mean structure is notoriously difficult in high dimensions,

Received June 2017; revised March 2019.
MSC2010 subject classifications. Primary 62G08, 62G20; secondary 62G10, 62H15.
Key words and phrases. Empirical likelihood, marginal testing, variable selection consistency.

primarily due to the curse of dimensionality (cf. Stone (1982), Stone et al. (1997)): For a nonparametric regression model,

$$Y_t = m(X_t) + \varepsilon_t, \quad t = 1, \ldots, n,$$

with d-dimensional covariates X_t and independent and identically distributed (*i.i.d.*) error variables $\varepsilon_t \sim N(0, \sigma^2)$, the minimax rate of estimation of the regression function $m(\cdot)$ under the L^2-loss function over an order-2 Sobolev ball in $\mathbb{R}^d$ is only $O(n^{4/[4+d]})$. Also, results of Fan et al. (1997) show that the conditional minimax rate (given the X_ts) has a similar order, namely, $O_p(n^{4/[4+d]})$. As a result, the accuracy of the estimated nonparametric regression function deteriorates quite rapidly as d increases. In an important paper, Lafferty and Wasserman (2008) developed a method, called the RODEO, for simultaneous variable selection and sparse nonparametric regression function estimation, allowing the dimension d to grow with the sample size n. The RODEO is based on multiple testing with a Studentized pivot (which, in particular, requires estimation of the error variance $\sigma^2 = E\varepsilon_1^2$). Some of the main advantages of this approach are its flexibility and simplicity of computation, as well as the full structural generality of the regression function $m(\cdot)$. However, because of the latter, it also suffers from the curse of dimensionality that makes it unsuitable for the analysis of high dimensional regressions beyond $d = O(\log n / \log\log n)$.

An alternative approach is based on penalized regression methods, albeit under some simplifying structural restrictions on the regression function, for example, an additive structure; see Radchenko and James (2010), Zhang, Cheng and Liu (2011), Storlie et al. (2011) and the references therein. For the case of additive models, where the regression function $m(\cdot)$ can be represented as $m(x_1, \ldots, x_d) = m_1(x_1) + \cdots + m_d(x_d)$ for some functions $m_1(\cdot), \ldots, m_d(\cdot)$ on $\mathbb{R}$ and a given estimation point $x = (x_1, \ldots, x_d)$, Lin and Zhang (2006), Ravikumar et al. (2009) and Meier, van de Geer and Bühlmann (2009) consider estimation of $m(\cdot)$ under suitable sparsity conditions on the component functions $m_j(\cdot)$. Zhang, Cheng and Liu (2011) consider a semiparametric formulation where the regression function has a parametric linear regression component and an additive nonlinear component (but NO interactions among the covariates). They develop the LAND method of penalized regression based on basis expansions which, in addition to selecting the relevant variables, can successfully identify the set of linear and nonlinear variables, and thus, provide important information about the structure of the regression function $m(\cdot)$. Extensions to models that allow for *two-way* interactions among the d covariates in $m(\cdot)$, that is, $m(x_1, \ldots, x_d) = \sum_{j=1}^d m_j(x_j) + \sum_{j>k} m_{jk}(x_j, x_k)$ for functions $m_j(\cdot)$, $m_{jk}(\cdot)$, have been considered by Choi, Li and Zhu (2010) and Radchenko and James (2010). Both of these papers require the user to prespecify the order of interaction (2-way, 3-way, ...), essentially scaling up the effective dimension to $O(d^k)$ for the kth order interaction terms among d covariates (for $k \geq 2$), which quickly makes them unfeasible for applications in practice, even for a moderately large d.

Building on the work of Lafferty and Wasserman (2008), in this paper we develop a new method, called the *Gradient relevant identification of derivatives* (or the *GRID*) method, for simultaneous nonlinear variable selection and estimation of the low dimensional structure of the regression function under sparsity. The type of sparsity we consider allows for interactions of arbitrarily high (but bounded) order that need not be specified *a priori*. To briefly describe the methodology and the findings of the paper, consider the nonparametric regression model

$$Y_t = m(X_t) + \varepsilon_t, \quad t = 1, \ldots, n, \tag{1.1}$$

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. The errors ε_t are independent of X_t, and are *not* assumed to be Gaussian. Here, $m(X_t) = \mathrm{E}(Y_t | X_t) : \mathbb{R}^d \to \mathbb{R}$ is the conditional mean function. We use the notation $X_t = (X_{t1}, \ldots, X_{td})$ to refer to the vector of covariates (but sometimes X_j will also denote

the jth covariate; the difference will be clear from the context). We assume that the number of covariates $d \to \infty$ but only r of these covariates are relevant for model (1.1), where $r \ll d$ can be bounded or unbounded. For structure discovery, we call a relevant covariate j (non)linear if $m(\cdot)$ is a (non)linear function of the jth component of X_t. The same issue of linear and nonlinear covariates has also been addressed by Zhang, Cheng and Liu (2011), but in a somewhat simpler setup of partially linear models. To be precise, here we say that X_{tj} is an irrelevant covariate if the jth partial derivative $\partial m/\partial x_j$ is equal to zero everywhere on the support of the function; otherwise it is a relevant covariate. And among the relevant covariates, we consider X_{tj} as a linear covariate if the second-order partial derivative $\partial^2 m/\partial x_j^2$ is zero everywhere on the support; otherwise it is a nonlinear covariate. See (1.2) below for an illustrative example. Next, denote the sets of nonlinear, linear, relevant and irrelevant covariates in (1.1) by C, A, R and U, respectively. Thus, $R = C \cup A$ and $U = \{1, \ldots, d\} \setminus R$. The proposed GRID method has the following features:

(a) It automatically identifies the set R of relevant covariates of model (1.1), also distinguishing the nonlinear ones from the linear ones, with probability tending to 1. The number of either type of covariates can be either bounded or unbounded.

(b) It can automatically identify the interaction terms of any order (two way, three way, ...) without increasing the computational complexity of the algorithm.

(c) It is completely data-driven and is easy to implement in practice. In particular, it neither requires selection of regularization parameters nor the estimation of the nuisance parameter σ^2.

(d) It is based on a marginal test based selection procedure and under appropriate moment assumptions, it can be applied to high dimensions of order $d = O(n^a)$ *for any* $a \in (0, \infty)$, with $r = O(n^\gamma)$ for some suitable $\gamma \in (0, 1)$.

The GRID method combines two well-developed nonparametric tools in an effective way: (1) the local linear estimation (LLE) technique of Fan (1992), and (2) the empirical likelihood (EL) method of Owen (1988), and it can be seen as a nontrivial extension of the RODEO of Lafferty and Wasserman (2008). The GRID makes use of the same framework and some of the ideas of RODEO, but differs from it by the use of EL methodology and by careful construction of some *new* estimating functions that help to elicit the structural properties of the relevant variables (e.g., linear versus nonlinear and their interaction patterns). In contrast to the RODEO, the GRID employs the EL technique to conduct a series of hypotheses tests to identify the relevant variables and the model structure. An important advantage derived from the use of the EL methodology is that the GRID completely avoids the difficult issue of estimating the variance parameter σ^2 in the high dimensional nonlinear model (1.1). In addition, it does not require the Gaussian assumption on the error variables.

The acronym GRID has a two-fold meaning: First, it derives from *Gradient Relevant Identification of Derivatives*, meaning that the procedure is based on testing the significance of partial derivative estimators (motivated by the LLE methodology); Second, it refers to a graphical tool which can help in identifying the structure of model (1.1). We now illustrate the latter using Figure 1 with the following example: Let $d = 10$ and let the true model be given by

$$Y_t = 2X_{t1} + X_{t2}^2 X_{t3} + 10X_{t4}X_{t5}X_{t6} + \exp(X_{t7})X_{t2} + \varepsilon_t. \tag{1.2}$$

The first stage of the GRID procedure identifies (the indices of) the following sets of covariates:

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

which are given by the first row (marked with row-label 0) in Figure 1. The selected variables are automatically classified as linear (denoted by $\circ$) and nonlinear (denoted by $\triangle$). The

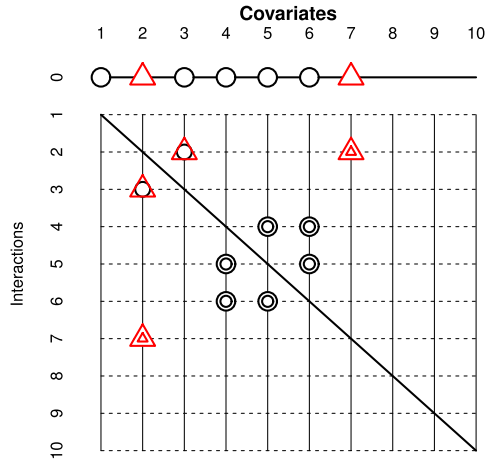

FIG. 1. *The GRID representation of model* (1.2). *The first row represents the selected linear* (*denoted by* ∘) *and nonlinear* (*denoted by* △) *variables. The entries in the rest of the grid represent interactions among the selected variables.*

unmarked variables constitute the set U of irrelevant variables. This step requires applying the EL testing $O(d)$-many times with specific choices of estimating functions derived from a modified LLE method (cf. Section 2 below).

In the second stage of the procedure, a different set of estimating functions are employed which help to identify the following sets of interactions:

$$I^1 = \{1\}, \qquad I^2 = \{2, 3, 7\}, \qquad I^3 = \{3, 2\}, \qquad I^4 = \{4, 5, 6\}, \qquad I^5 = \{5, 4, 6\},$$
$$I^6 = \{6, 4, 5\}, \qquad I^7 = \{7, 2\},$$

where I^j includes the interactions of variable j with other covariates, for $j \in R$. By default, each set I^j (for $j \in R$) automatically includes the index j (self-interaction). Therefore, if the set I^j has the only component j, then X_{tj} appears in the model as an isolated additive covariate, like X_{t1} in model (1.2). These index sets can be determined by successive scans of the columns of the GRID plot. This step requires applying the EL testing $O(r^2)$-times, where r is the number of relevant covariates, that is, $r = |A \cup C|$, the size of $A \cup C$. Thus, the entire GRID procedure requires $O(d + r^2)$-many multiple tests based on the EL, which scales *linearly* in the total number of covariates and as a quadratic function of the number of relevant variables that is typically of a much smaller order than d. Further, unlike existing methods, it identifies the interaction terms of any order among the relevant variables without having to prespecify the highest order *a priori*.

Next, note that using the sets A, C and I^j from Figure 1, we can identify the actual low dimensional structure of the regression function $m(\cdot)$ as

$$m(x_1, \ldots, x_{10}) = \beta_1 x_1 + m_1(x_2, x_7) + m_2(x_2)x_3 + \ell_1(x_4, x_5, x_6)$$

for some (nonzero) constants β_1, some nonlinear functions $m_1(\cdot)$ and $m_2(\cdot)$ and some (multi)linear function $\ell_1(\cdot)$ of x_4, x_5, x_6. To see this, it is sufficient to scan the GRID plot of Figure 1 by columns, reading each column from the top to the bottom. For example, in the first column (position $j = 1$), we can note a circle at the top (in row 0), showing that X_{t1} is a linear covariate, and no symbols in the column below, indicating that x_1 appears in $m(\cdot)$ as a linear term by itself, with no interaction. Next, consider the second column (position $j = 2$), which shows a triangle at the top row and two symbols in the column below it. This means that covariate X_{t2} is nonlinear and interacts with other two covariates. The two symbols in

the column below are in positions 3 and 7, which correspond to a linear and a nonlinear covariate, respectively (the two different symbols, circle-triangle and triangle-triangle, show the kind of interactions). Thus, the interaction term involving variable X_{t2} must be of the form $m_1(x_2, x_7) + m_2(x_2)x_3$, as X_{t7} does not interact with X_{t3}. Columns $j = 3$ and $j = 7$ of the plot reaffirm the interaction term involving the linear variable X_{t3} and the nonlinear variables X_{t2} and X_{t7}. Similarly, the columns 4, 5 and 6 jointly yield the second (linear interaction) term of $m(\cdot)$ above. Note that the maximum order of the interaction terms corresponds to the maximum number of symbols appearing in the columns of the GRID plot (including position zero on the top). It is three for the case shown in Figure 1 for model (1.2).

Once the low dimensional structure of the high dimensional regression function $m(\cdot)$ has been identified, estimation of $m(\cdot)$ reduces to a relatively simple task of estimating the additive components of the low dimensional regression function. This can be done using any of the standard nonparametric function estimation methods (cf. Fan and Gijbels (1996), Prakasa Rao (1983), Tsybakov (2009)) with better accuracy than estimating the high dimensional regression function in full generality. Indeed, known mean squared error (MSE) results show that MSE-consistent estimation of the d-dimensional regression function itself breaks down whenever the dimension $d \gg \log n$ (so that $n^{-\frac{4}{d+4}} \nrightarrow 0$) while the low dimensional function is consistently estimable through the GRID method for $d = O(n^a)$ for any $a > 0$, provided suitable moment conditions on ε_1 and growth conditions on r are satisfied. The growth rate of d in our work may also be compared with the existing methods for high dimensional nonparametric regression. As pointed out earlier, the RODEO has a natural limit of $d = O(\log n)$ on the growth rate of d. For the linear and additive case with two-way interactions, Choi, Li and Zhu (2010) allow $d = o(n^{1/10})$. For nonlinear additive models with *two-way* interactions and *Gaussian* errors, Radchenko and James (2010) establish sparsistency results for VANISH, allowing d to grow at a subexponential rate. In comparison, GRID considers a general nonlinear nonparametric regression model with an *arbitrary* (but fixed) order of interaction among the relevant variables, allowing *non-Gaussian* error distribution and allowing the dimension to grow at the rate $d = O(n^a)$ for any $a \in (0, \infty)$. It also allows the number r of relevant variables to grow at the rate $r = O(n^\gamma)$ with some suitable $0 \le \gamma < 1$ (depending on the moment condition; see Theorem 3). This rate of r is comparable with the cases of additive models (cf. Ravikumar et al. (2009)) while significantly superior to available rates for nonadditive and nonparametric models, namely, $r = O(1)$ in Lafferty and Wasserman (2008) and Bertin and Lecué (2008)) and $r = O(\log n)$ in Comminges and Dalalyan (2012). The higher rates of d and r in GRID for nonadditive and nonparametric model result from its construction which completely separates out the selection procedure from the estimation task. In addition, the computational cost associated with the implementation of the GRID scales linearly in d and as a quadratic of r, making it viable in practice for a large d; see Section 6.4 for more details on the computational times for the GRID. In summary, the GRID seems to provide a significant advancement over existing methods of simultaneous nonlinear variable selection and sparse nonparametric regression function estimation in presence of any given order (not depending on n) of complex interactions among the relevant covariates in high dimensions d that may grow as an arbitrary polynomial power of the sample size.

The rest of the paper is organized as follows. In Section 2, we review some background concepts that are useful for describing the GRID method. In Section 3, we construct new estimating functions that are needed in the formulation of the GRID for nonlinear variable selection and structure identification. We also describe the GRID method and the GRID plot in Section 3. In Section 4, we report theoretical properties of the proposed method for variable selection and estimation for covariates $X_1 \in \mathbb{R}^d$ having the uniform distribution on $(0, 1)^d$. An extension to the case of correlated covariates having nonuniform distributions on $(0, 1)^d$ is given in Section 5. Results from a moderately large simulation study are presented in

Section 6. Additional simulation results, proofs of the theoretical results and an algorithmic representation of the GRID method are relegated to a supplementary materials file, hereafter referred to as Giordano, Lahiri and Parrella (2020).

2. Background and motivation. Let $\mathbb{D}_g(x)$ denote the gradient and $\mathbb{H}_g(x)$ the Hessian matrix of a d-variate function g, evaluated at the point x. We shall use the notation $\delta(A, B, \ldots)$ to denote a *generic real valued function*, depending on one or more arguments where the arguments are matrices. Depending on the context, the operator $|\cdot|$ may denote the determinant of a matrix, the absolute value of a number, or the cardinality of a set. Given a set A, the set A^c is the complement of A. We also use c_j, C_j to denote generic constants with values in $(0, \infty)$ that do not depend on n and d. Finally, let $(\cdot)^T$ denote the transpose operator.

2.1. *Empirical likelihood based tests.* The EL method of Owen (1988) is a nonparametric method that defines a likelihood function for certain parameters of the underlying distribution, albeit without a parametric model. Qin and Lawless (1994) extended the scope of the EL to parameters that are specified through estimating equations (again, without the specification of a parametric model). To describe it briefly, let $T_1, T_2, \ldots$ be *iid* random vectors with common distribution G and let $\theta = \theta(G)$ be a d-dimensional parameter of interest specified by

$$E\psi_j(T_1; \theta) = 0, \quad j = 1, \ldots, d \tag{2.1}$$

for some functions $\psi_j(\cdot; \cdot)$. The EL for θ, for $j = 1, \ldots, d$, is defined as

$$L_n(\theta) = \sup\left\{\prod_{t=1}^{n} p_t : p_t \in [0, 1], \sum_{t=1}^{n} p_t = 1, \sum_{t=1}^{n} p_t \psi_j(T_t; \theta) = 0\right\}. \tag{2.2}$$

The EL ratio test statistic for testing $\mathcal{H}_0 : \theta = \theta_0$ against $\mathcal{H}_0 : \theta \neq \theta_0$ is given by $R_n(\theta_0) = n^n L_n(\theta_0)$. It is known (cf. Owen (1988, 2001), Qin and Lawless (1994)) that under some regularity conditions (in particular, with d fixed),

$$-2\log R_n(\theta_0) \xrightarrow{d} \chi^2_d \quad \text{as } n \to \infty, \tag{2.3}$$

where $\xrightarrow{d}$ denotes convergence in distribution. Thus, a version of Wilk's theorem (cf. Wilks (1938)) holds just as in the case of the likelihood ratio test in classical parametric models. Note that the limit distribution of $-2\log R_n(\theta_0)$ is free of any unknown parameters and, hence, it can be readily used to calibrate the EL ratio test statistic. For developing the GRID method, we shall use the EL with a suitable choice of the estimating functions ψ_j's for variable selection and structure discovery through marginal testing. In the next section, we describe the background and motivation behind the construction of these special estimating functions.

2.2. *Multivariate local linear estimation.* Local linear estimation (LLE) is a nonparametric method for estimating the regression function $m(\cdot)$ in (1.1) (cf. Fan (1992), Ruppert and Wand (1994)). To estimate $m(\cdot)$ at $x = (x_1, \ldots, x_d)$, the LLE performs a locally weighted least squares fit of a linear function. Let

$$\hat{\boldsymbol{\beta}}(x; H) \equiv \arg\min_{\beta_0, \beta_1} \sum_{t=1}^{n} \{Y_t - \beta_0 - \beta_1^T (X_t - x)\}^2 K_H(X_t - x), \tag{2.4}$$

where the function $K_H(u) = |H|^{-1} K(H^{-1}u)$ gives the local weights with a d-variate product Kernel function $K(u) = \prod_{j=1}^{d} K_1(u_j)$. The bandwidth matrix H controls the bias and

the variance of the resulting LLE of $m(x)$. For simplicity, we shall suppose that $H = \operatorname{diag}(h_1, \ldots, h_d)$ is a diagonal matrix with strictly positive entries. The estimator $\hat{\boldsymbol{\beta}}(x; H)$ can be written in a closed form as

$$\hat{\boldsymbol{\beta}}(x; H) = \left(\Gamma^T W \Gamma\right)^{-1} \Gamma^T W \Upsilon, \tag{2.5}$$

where $\Upsilon = (Y_1, \ldots, Y_n)^T$ and

$$\Gamma = \begin{pmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix}, \qquad W = \begin{pmatrix} K_H(X_1 - x) & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & K_H(X_n - x) \end{pmatrix}.$$

Note from (2.4) that $\hat{\boldsymbol{\beta}}(x; H)$ gives estimators of the function $m(x)$ and its gradient $\mathbb{D}_m(x)$:

$$\hat{\boldsymbol{\beta}}(x; H) = \begin{pmatrix} \hat{\beta}_0(x; H) \\ \hat{\boldsymbol{\beta}}_1(x; H) \end{pmatrix} \equiv \begin{pmatrix} \hat{m}(x; H) \\ \hat{\mathbb{D}}_m(x; H) \end{pmatrix}. \tag{2.6}$$

Despite its conceptual and computational simplicity, accurate estimation of $m(\cdot)$ by the LLE in the multivariate case requires choosing a suitable bandwidth matrix H. Although asymptotically optimal bandwidth can be derived taking account of the bias-variance trade-off, like many other nonparametric methods, the LLE is also strongly affected by the *curse of dimensionality* problem. As a result, the LLE of $m(\cdot)$, considered in the construction of RODEO in Lafferty and Wasserman (2008), is impractical for $d \gg \log n$. However, here we adopt a somewhat *different* approach and do *not* aim to estimate the function $m(\cdot)$ as a function of all d covariates. Instead, we focus on low dimensional structure discovery, which no longer requires explicit estimation of the function itself, by using a large bandwidth (that does *not* go to zero with the sample size) at a suitable point $x^* \in (0, 1)^d$. This is a very *nonstandard* point of view, and is a salient feature of our approach that plays a critical role in scaling up the proposed GRID method to $d = O(n^a)$, for any $a \in (0, \infty)$.

2.3. *Motivation and main ideas behind model structure discovery.* To identify the model structure of the sparse nonparametric regression function $m(\cdot)$, we shall use the EL methodology with some "suitable" estimating functions based on a variant of LLE with a nonstandard choice of the bandwidth. Recall that A, C, R and U, respectively, denote the sets of linear, nonlinear, relevant and irrelevant covariates and that $|R| = r$, where $|B|$ denotes the size of a finite set B. Thus, $R = A \cup C$ and $U = \{1, \ldots, d\} \setminus R$. Set $|C| = k$ and without loss of generality (w.l.g.), suppose that $C = \{1, \ldots, k\}$, $A = \{k + 1, \ldots, r\}$ and $U = \{r + 1, \ldots, d\}$. Next, partition $A = A_c \cup A_l$, where $A_c = \{k + 1, \ldots, k + s\}$ consists of those linear covariates which interact with nonlinear covariates, leading to *nonlinear mixed effects*, like covariate X_3 of model (1.2), and where $A_l = \{k + s + 1, \ldots, r\}$ corresponds to isolated or mixed linear covariates, like covariates X_1, X_4, X_5 and X_6 of model (1.2). The GRID procedure will automatically identify such sets of indices.

Next, for $D_1, D_2 \in \{C, A_c, A_l, U\}$, let x_{D_1} denote the subvector $(x_j : j \in D_1)$ of a vector $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ and similarly, $\mathbb{F}^{D_1, D_2}$ denote the submatrix of a $d \times d$ matrix $\mathbb{F}$ with row indices in D_1 and column indices in D_2. For example, with this notation, using $x = (x_C^T, x_{A_c}^T, x_{A_l}^T, x_U^T)^T$, the gradient and the Hessian matrix of the function $m(\cdot)$ become

$$\mathbb{D}_m(x) = \begin{pmatrix} \mathbb{D}_m^C(x) \\ \mathbb{D}_m^{A_c}(x) \\ \mathbb{D}_m^{A_l}(x) \\ 0 \end{pmatrix}, \qquad \mathbb{H}_m(x) = \begin{pmatrix} \mathbb{H}_m^{CC}(x) & \mathbb{H}_m^{CA_c}(x) & 0 & 0 \\ \mathbb{H}_m^{A_cC}(x) & \mathbb{H}_m^{A_cA_c}(x) & \mathbb{H}_m^{A_cA_l}(x) & 0 \\ 0 & \mathbb{H}_m^{A_lA_c}(x) & \mathbb{H}_m^{A_lA_l}(x) & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{2.7}$$

where 0 is a matrix of suitable dimensions with all elements equal to zero. Similarly, let $\mathbf{1}$ denote a vector of ones. Note that the diagonal submatrices $\mathbb{H}_m^{CC}(x)$, $\mathbb{H}_m^{A_cA_c}(x)$ and $\mathbb{H}_m^{A_lA_l}(x)$ are symmetric, but the off-diagonal matrices are not. Finally, define the $d \times d$ matrix of third-order partial derivatives $\mathbb{G}_m = (\frac{\partial^3 m(x)}{\partial x_i \partial x_j^2})$ and consider the diagonal bandwidth matrix $H = \operatorname{diag}(H_C, H_{A_c}, H_{A_l}, H_U)$. Using the definitions of the sets C, A_l, etc., it is easy to verify that the *only* possible nonzero submatrices of $\mathbb{G}_m$ are $\mathbb{G}_m^{CC}$ and $\mathbb{G}_m^{A_cC}$; all the remaining submatrices are 0.

While the function $m(\cdot)$ is unknown, a *key observation* we make in here is that such structural information about $m(\cdot)$ can also be similarly identified from the bias of $\hat{\beta}_0(x; H)$ of (2.6) with an asymptotically *nonvanishing bandwidth*. Further, estimators of the bias and its derivatives can also be generated through the LLE technology itself. We make these connections precise in the following result. Let $\mathcal{X}_n = \{X_t : t = 1, \ldots, n\}$ and define the moments of the univariate Kernel $K_1(\cdot)$ as

$$\mu_l = \int u_1^l K_1(u_1)\, du_1, \qquad \nu_l = \int u_1^l K_1^2(u_1)\, du_1 \quad l = 0, 1, \ldots, 4. \tag{2.8}$$

PROPOSITION 1. *Suppose that model* (1.1) *holds and that Assumptions* A1–A4 *of Section* 4 *hold for some* $x^* \in (0,1)^d$. *Then, for the local linear estimator* (2.6), *its conditional bias at the point* x^* *is given by*

$$\mathrm{E}\left\{\begin{pmatrix}\hat{m}(x^*; H)\\ \hat{\mathbb{D}}_m(x^*; H)\end{pmatrix} - \begin{pmatrix}m(x^*)\\ \mathbb{D}_m(x^*)\end{pmatrix}\Big|\mathcal{X}_n\right\} = \begin{pmatrix}b_m(x^*; H_C)\\ B_{\mathbb{D}}(x^*, H_C)\end{pmatrix} + O_p(n^{-\frac{1}{2}}), \tag{2.9}$$

where the order symbol is valid componentwise, and where $b_m(x^*; H_C) = \frac{1}{2}\mu_2 \operatorname{tr}\{\mathbb{H}_m^{CC}(x^*) \times H_C^2\} + \delta(H_C)$, $B_{\mathbb{D}}(x^*, H_C)^T = ([B_{\mathbb{D}}^C]^T, [B_{\mathbb{D}}^{A_c}]^T, [B_{\mathbb{D}}^{A_l}]^T, [B_{\mathbb{D}}^U]^T)$ *with* $B_{\mathbb{D}}^U = 0 = B_{\mathbb{D}}^{A_l}$, $B_{\mathbb{D}}^{A_c} = \frac{1}{2}\mu_2 \mathbb{G}_m^{A_cC}(x^*) H_C^2 \mathbf{1} + \delta(H_C)$ *and* $B_{\mathbb{D}}^C = \frac{1}{2}\mu_2[\mathbb{G}_m^{CC}(x^*) H_C^2 \mathbf{1} + (3^{-1}\mu_4\mu_2^{-2} - 1) \times \operatorname{diag}\{\mathbb{G}_m^{CC}(x^*) H_C^2\}\mathbf{1} + \delta(H_C)]$.

Proposition 1 gives biases of the estimators $\hat{m}(x; H)$ and $\hat{\mathbb{D}}_m(x; H)$. The leading terms are similar to those in Theorem 2.1 of Ruppert and Wand (1994) (but note that our bandwidth matrix H corresponds to theirs $H^{1/2}$). However, there are substantial differences in the proofs due to the nonstandard assumption that the bandwidths are fixed as a function of n and do not go to zero. *As a consequence of this, the residual term* $\delta(H_C)$ *does not vanish as* $n \to \infty$; The important fact that we shall make use of herein is that the term $\delta(H_C)$ depends only on the bandwidths h_j, $j \in C$, so that $\frac{\partial}{\partial h_j}\delta(H_C) = 0$ for all $j \notin C$.

Proposition 1 reveals some interesting relationships between the bias of $\hat{\boldsymbol{\beta}}(x^*; H)$ (cf. (2.6)) and the bandwidth matrix $H = \operatorname{diag}(H_C, H_{A_c}, H_{A_l}, H_U)$. Generalizing the ideas proposed in Lafferty and Wasserman (2008) (which considers the LLE under the standard assumption that bandwidths go to zero suitably with the sample size), we can make these relationships emerge through the derivatives of $\hat{\boldsymbol{\beta}}(x^*; H)$ with respect to H. Note that

$$\frac{\partial}{\partial H} E\{\hat{m}(x^*; H)|\mathcal{X}_n\} = \frac{\partial}{\partial H} E\{(\hat{m}(x^*; H) - m(x))|\mathcal{X}_n\} \approx \frac{\partial}{\partial H} b_m(x^*; H_C), \tag{2.10}$$

where $\frac{\partial}{\partial H} b_m(x^*; H_C) = (\frac{\partial b_m(x^*;H_C)}{\partial H_C}, \frac{\partial b_m(x^*;H_C)}{\partial H_{A_c}}, \frac{\partial b_m(x^*;H_C)}{\partial H_{A_l}}, \frac{\partial b_m(x^*;H_C)}{\partial H_U}) = (\delta(H_C), 0, 0, 0)$. Similarly,

$$\frac{\partial}{\partial H} E\{\hat{\mathbb{D}}_m(x^*; H)|\mathcal{X}_n\} \approx \frac{\partial}{\partial H} B_{\mathbb{D}}(x^*, H_C) = \begin{pmatrix} \partial B_{\mathbb{D}}^{C}/\partial H \\ \partial B_{\mathbb{D}}^{A_c}/\partial H \\ \partial B_{\mathbb{D}}^{A_l}/\partial H \\ \partial B_{\mathbb{D}}^{U}/\partial H \end{pmatrix} = \begin{pmatrix} \delta(H_C) & 0 & 0 & 0 \\ \delta(H_C) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{2.11}$$

The quantities in (2.10) and (2.11) have a sparse structure that can be exploited to extract useful information about the covariates and their interactions. For example, the derivatives in (2.10) can be used for *nonlinear* variable selection (as considered in the RODEO method) and the elements of the matrix (2.11) can be used for model structure exploration (not considered in RODEO). Specifically, for $i \neq j$, the (i, j)th element equals

$$\frac{\partial B_{\mathbb{D}}^{(i)}(x^*, H_C)}{\partial h_j} = h_j \mu_2 \frac{\partial^3 m(x)}{\partial x_i \partial x_j^2}\bigg|_{x=x^*} + \delta(H_C) \tag{2.12}$$

which is different from zero if there are mixed effects in model (1.1) between two *nonlinear covariates* or between a *linear covariate* X_{ti} and a *nonlinear covariate* X_{tj} (see the proof of Proposition 1 in the supplement Giordano, Lahiri and Parrella (2020) for more details). So, these derivatives can help to identify the *nonlinear covariates* in C, the *linear covariates* in A_c and the *nonlinear mixed effect* terms in C and A_c. Using these observations as *motivation*, we will describe the construction of the GRID method next (which will use a new set of estimating equations that are different from those described above).

3. The GRID method. In this section, we will build on insights from the last section and develop the GRID method in high dimensions. It turns out that the LLE itself is not suitable for this purpose. Indeed, in its construction, the RODEO used the LLE with a *vanishing* bandwidth, allowing $d = d_n = O(\log n / \log\log n)$ where n is the sample size. In order to allow for growth rates $d \gg n$, here we need to modify the basic LLE estimator and base our identification procedure on a variant of (2.5). In Section 3.1 below, we develop a new set of estimating equations for this purpose. Using these, we formulate the GRID algorithm and describe the GRID plot, respectively, in Section 1 of Giordano, Lahiri and Parrella (2020) and in Section 3.2 below.

3.1. *Construction of new estimating functions.*

3.1.1. *Estimating functions for identifying the nonlinear effects.* Note that for $d > n$, the estimator (2.5) is not well-defined because $\Gamma^T W \Gamma$ is a singular matrix. To avoid this constraint and also to reduce the computational burden associated with inverting the $(d + 1) \times (d + 1)$ matrix $(\Gamma^T W \Gamma)$ in (2.5), we consider the following statistic:

$$M(x; H) = \frac{1}{n} \operatorname{diag}(1, H^{-2}) \Gamma^T W \Upsilon \equiv \begin{pmatrix} M_0(x; H) \\ M_1(x; H) \end{pmatrix}, \tag{3.1}$$

where H is the diagonal bandwidth matrix. The estimator (3.1) is a modified version of (2.6). Note that

$$\frac{\partial M(x; H)}{\partial h_j} = \frac{1}{n} \mathbb{O}_j \Gamma^T W \Upsilon + \frac{1}{n} \begin{pmatrix} 1 & 0 \\ 0 & H^{-2} \end{pmatrix} \Gamma^T \frac{\partial}{\partial h_j} W \Upsilon,$$

where $\mathbb{O}_j$ is a $(d+1)\times(d+1)$ matrix, with all zeros except the $(j+1, j+1)$ element, which is equal to $-\frac{2}{h_j^3}$. Further, it is easy to check that $\frac{\partial}{\partial h_j}W = WL_j$ where $L_j = \operatorname{diag}(\frac{\partial \log K_1((X_{1j}-x_j)/h_j)}{\partial h_j} - \frac{1}{h_j}, \ldots, \frac{\partial \log K_1((X_{nj}-x_j)/h_j)}{\partial h_j} - \frac{1}{h_j})$ and $K_1(\cdot)$ is the univariate kernel function. Hence,

$$\frac{\partial M(x;H)}{\partial h_j} = \frac{1}{n}\left[\mathbb{O}_j\Gamma^T W + \begin{pmatrix}1 & 0\\ 0 & H^{-2}\end{pmatrix}\Gamma^T WL_j\right]\Upsilon \equiv \begin{pmatrix}\dot{M}_{0j}\\ \dot{M}_{1j}\end{pmatrix} \quad \text{(say)}, \tag{3.2}$$

where $\dot{M}_{0j} = \frac{\partial M_0(x;H)}{\partial h_j}$ and $\dot{M}_{1j} = \frac{\partial M_1(x;H)}{\partial h_j} \equiv \{\dot{M}_{1j}^{(i)}\}_{i=1,\ldots,d}$. With this, we have the following result.

THEOREM 1. *Suppose that model* (1.1) *holds and that Assumptions* A1–A4 *of Section* 4 *hold for some* $x^* \in (0,1)^d$. *Then the following results hold with* $x = x^*$:

$$E\{\dot{M}_{0j}\} = \begin{cases}\theta_{0j}^m & \text{if } j \in C,\\ 0 & \text{otherwise},\end{cases} \tag{3.3}$$

$$E\{\dot{M}_{1j}^{(i)}, i \neq j\} = \begin{cases}\theta_{ij}^m & \text{if } i \in I^j,\ j \in C,\\ 0 & \text{otherwise},\end{cases} \tag{3.4}$$

where the exact expressions for θ_{ij}^m, $0 \le i \le d$ *and* $1 \le j \le d$, $i \neq j$ *are given in* (S.9) *and* (S.10) *of the Giordano, Lahiri and Parrella* (2020).

Theorem 1 can be used to detect the nonlinear effects in model (1.1). In fact, by (3.3) and (3.4), the derivatives $\partial M_0(x^*;H)/\partial h_j$ and $\partial M_1(x^*;H)/\partial h_j$ can be used to identify the *nonlinear covariates* (obtaining the set C) and the interactions for the *nonlinear covariates*. An important difference between the two estimators in (2.6) and (3.1) is that the second can be analyzed without conditioning on the set of observed values, so that the $O_p(n^{-1/2})$ term does not appear in the bias. Further, as pointed out before, the modified statistic does not require inversion of a high dimensional matrix and is computationally much simpler. Both of these factors are crucial for validity of our method in very high dimensions where $d \gg n$.

REMARK 3.1. In contrast to the critical effects of the bandwidth choice on the performance of the LLE in nonparametric regression, the value of the bandwidth is not very crucial in our procedure, because we are not interested in the estimation of the function at this stage. As specified in Assumption A1 below, we do not even require the bandwidths to vanish asymptotically. In fact, given that the identification of the covariates is based on evaluating the bias of the modified LLE, we need a bandwidth matrix which produces a very high bias. This suggests using relatively large bandwidths in practice. See Sections 4 and 6 for more details.

3.1.2. *Estimating functions for identifying the linear effects.* Note that the estimating functions suggested by Theorem 1 are effective in identifying only the nonlinear covariates and their interactions, but not the linear covariates. A similar problem exists with the RODEO which considers the derivatives $\partial E\{\hat{m}(x;H)\}/\partial H$ of the LLE in (2.10) and can only identify the *nonlinear covariates* in C (but not those in the sets A or U). To overcome this, Lafferty and Wasserman (2008) suggest identifying first the linear variables through LASSO or to change the degree of the local polynomial estimator to zero (i.e., to use the Nadaraya–Watson estimator). Both of these solutions make use of extraneous methods and require further attention to the choice of critical tuning parameters (the choice of the regularization parameter in the case of LASSO and the choice of the bandwidth matrix for the

Nadaraya–Watson estimator) which are not well studied in the present framework. In comparison, here we propose a simple solution to the problem that allows the user to apply the same algorithm, but to a modified regression problem. Specifically, we consider an auxiliary regression where all those covariates that have not been selected in the first pass, are to be transformed, so that the *linear covariates* of the original model become *nonlinear* in the auxiliary model. Note that the model (1.1) under the partition $\{C, A_c, A_l, U\}$ must necessarily be of the form: $m(x) = m_1(x_C, x_{A_c}) + m_2(x_{A_c}, x_{A_l})$. Define the transformation $z = \phi(x)$ and its inverse $x = \phi^{-1}(z)$ as follows:

$$z = \phi(x) = \big(x_C, x_{A_c}^{1/2}, x_{A_l}^{1/2}, x_U^{1/2}\big), \qquad x = \phi^{-1}(z) = \big(x_C, z_{A_c}^2, z_{A_l}^2, z_U^2\big), \tag{3.5}$$

and let $Z_t = \phi(X_t)$ be the transformed random variables. Next, consider the following auxiliary regression:

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

where the new regression function can be written as $g(z) = g_1(x_C, z_{A_c}) + g_2(z_{A_c}, z_{A_l})$. Note once again that we use the same index partition as in the first regression. By (3.5), the function $g_2(\cdot)$ depends only on the covariates in A and these covariates have a nonlinear effect in the auxiliary regression model $g(z)$. In fact, $z_j = \phi(x_j) = x_j^{1/2} \Longrightarrow x_j = \phi^{-1}(z_j) = z_j^2$ for all $j \in A \cup U$, so the partial derivatives are

$$\frac{\partial g(z)}{\partial z_j} = \begin{cases} 2c_j z_j \neq 0 & \text{for } j \in A, \\ 0 & \text{for } j \in U \end{cases} \quad \text{and} \quad \frac{\partial^2 g(z)}{\partial z_j \partial z_j} = \begin{cases} 2c_j \neq 0 & \text{for } j \in A, \\ 0 & \text{for } j \in U, \end{cases}$$

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

To select the transformed variables in A through EL testing, we need a result similar to Theorem 1 under the auxiliary regression model. However, this can no longer be obtained directly from Theorem 1, as the density f_Z of the transformed covariates $Z_t = \phi(X_t)$ is nonuniform. The following result gives the required estimating functions in this new case.

THEOREM 2. *Suppose that Assumptions* A1–A4 *of Section 4 hold for some* $x^* \in (0, 1)^d$. *Then, at the transformed point* $z^* = \phi(x^*)$, *(cf.* (3.5)), *the following results hold for the estimator* (3.1):

$$E\left\{\frac{\partial M_0(z^*; H)}{\partial h_j}\right\} = \begin{cases} \theta_{0j}^g & \text{if } j \in A, \\ 0 & \text{otherwise}, \end{cases} \tag{3.6}$$

$$E\left\{\frac{\partial M_1^{(i)}(z^*; H)}{\partial h_j}, i \neq j\right\} = \begin{cases} \theta_{ij}^g & \text{if } i \in I^j, j \in A, \\ 0 & \text{if } j \in U, \end{cases} \tag{3.7}$$

where the exact expressions for θ_{ij}^g *are reported in* (S.16) *and* (S.17) *in Giordano, Lahiri and Parrella* (2020). *Moreover, with the transformation* $z_{-i} = \{x_C\} \cup \{x_i\} \cup \{x_s^{1/2}, s \in C^c, s \neq i\} \equiv \phi_{-i}(x)$ (*where the* i*th covariate is NOT transformed*),

$$E\left\{\frac{\partial M_1^{(i)}(z_{-i}^*; H)}{\partial h_j}, i \neq j\right\} = \begin{cases} \theta_{ij}^* & \text{if } i \in I^j, j \in A, \\ 0 & \text{otherwise}, \end{cases} \tag{3.8}$$

where θ_{ij}^* *is defined in* (S.18) *in Giordano, Lahiri and Parrella* (2020) *and* $z_{-i}^* = \phi_{-i}(x^*)$.

REMARK 3.2. Note that by using the (3.6), the derivatives $\partial M_0(z^*; H)/\partial h_j$ can be used for identifying the *linear covariates*, obtaining the set A. However, we cannot identify the *linear mixed effects* in I^j, for $j \in A$, using (3.7) alone, as θ^g_{ij} can be nonzero also for $i \notin I^j$, $j \in A$. The problem is resolved by using (3.8) which is derived by taking the ϕ-transformation for all the covariates in the complement set C^c *except* for the ith, which allows us to correctly identify such effects using the (3.8).

REMARK 3.3. There are theoretical reasons that justify the choice of the square root transformation in the (3.5). In fact, the square root transformation allows one to obtain a linear density function of the transformed covariates Z_t, which is $f_Z(z) = 2z$. The linearity of the density function f_Z makes the higher order terms in the Taylor expansion of the estimator equal to zero, resulting in simpler expressions for θ^g_{0j}, θ^g_{ij} and θ^*_{ij}.

3.2. *The GRID plot.* We now describe the details of the GRID plot, a graphical tool representing the model structure discovery part of the GRID algorithm, as illustrated in Section 1. Using the values θ^m_{ij} of Theorem 1, define the $(d+1) \times d$ Boolean matrix

$$\Theta^m = \begin{pmatrix} \theta^m_{01} \neq 0 & \theta^m_{02} \neq 0 & \dots & \theta^m_{0d} \neq 0 \\ \hline 1 & \dots & \dots & \theta^m_{1d} \neq 0 \\ \theta^m_{21} \neq 0 & 1 & \dots & \theta^m_{2d} \neq 0 \\ \vdots & \vdots & \ddots & \vdots \\ \theta^m_{d1} \neq 0 & \theta^m_{d2} \neq 0 & \dots & 1 \end{pmatrix}, \tag{3.9}$$

where the first row considers the θ_{0j}'s from (3.3) and the subsequent rows with $i = 1, \dots, d$ consider the θ_{ij}'s from (3.4). We can also derive the matrix Θ^g in a similar way, using the values θ^g_{0j} and θ^*_{ij} from Theorem 2. The elements θ_{ij} in these matrices are estimated by $\dot{M}_{ij}$ through (3.2). The diagonal positions in the lower part of the matrix, where $i = j$, are excluded from the analysis because they correspond to self-interactions.

Tests are done using the EL procedure with the corresponding estimating function for one parameter at a time, for the null hypothesis $\mathcal{H}_0 : \theta_{ij} = 0$, for $i = 0, \dots, d$ and $j = 1, \dots, d$, with $i \neq j$, as explained in Section 2.1, in a multiple testing fashion. A point in the (i, j)th position of the GRID plot indicates a positive test result (i.e., rejecting $\mathcal{H}_0$) for the (i, j)th entry value of matrix Θ^m to identify nonlinearities (or Θ^g to identify linearities). (See Giordano, Lahiri and Parrella (2020) for more details of the individual steps and for a flow-chart description of the GRID algorithm.) Thus, positive tests in matrix Θ^m identify nonlinearities (= triangles), and the first row gives $\hat{C}$. Positive tests in matrix $\Theta^g \setminus \Theta^m$ identify linearities (= circles), and the first row gives $\hat{A}$. Finally, the nonzero values in the jth column of the matrix $\Theta^m \cup \Theta^g$ identify the set $\hat{I}^j$, for $j = 1, \dots, r$. A schematic representation of the estimated and tested matrix Θ is made through the GRID-plot as in Figure 1.

In Section 4 below, we describe the theoretical properties of the proposed GRID method. For brevity, we will often drop the index g or m from θ_{0j} and θ_{ij} when there is no chance of confusion.

4. Theoretical properties of the GRID.

4.1. *Assumptions.* We shall use the following assumptions:

A1 The bandwidth H is a diagonal matrix with strictly positive diagonal entries: $H = \text{diag}(h_1, \dots, h_d)$, with $c_1 \leq h_j$ for $j = 1, \dots, d$ for some $c_1 \in (0, \infty)$.

A2 The d-variate kernel function K is a product kernel, based on a nonnegative and symmetric univariate kernel density function $K_1 \in \mathcal{C}^1[-c_2, c_2]$ for some $c_2 > 0$ such that for some $x^* = (x^*_1, \dots, x^*_d)' \in (0, 1)^d$, $0 < x^*_j - c_2 h_j < x^*_j + c_2 h_j < 1$ for all $j = 1, \dots, d$.

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

A4 X_1 is uniformly distributed on the unit cube $(0, 1)^d$.

We now briefly comment on the assumptions. Assumption A1 requires the bandwidth matrix to be a diagonal matrix, which simplifies the development of the GRID method and is adequate for our purpose. The major difference between A1 and the typical assumption made on the bandwidth matrix H is that here the componentwise bandwidths h_j do *not* go to zero with the sample size. As a consequence, all the large sample theorems available in the statistical literature concerning the properties of the multivariate LLE can not be applied to our framework (but see Bertin and Lecué (2008)). The conditions on the d-variate kernel K in Assumption A2 and on the uniform distribution of the design points in Assumption A4 are the same as in Lafferty and Wasserman (2008). We shall relax Assumption A4 in Section 5 to allow for nonuniform and dependent covariates. Assumption A3 on the existence of the derivatives of $m(\cdot)$ is crucial for developing the estimating functions for identifying patterns of nonlinear interactions. For the other parts of Assumption A2, a typical choice of the point x^* is $x^* = (1/2, \ldots, 1/2)'$. Also note that Assumption A2 implies that all the moments of the kernel K exist and that the odd-ordered moments of K and $(K)^2$ are zero: For $l = 1, 2$,

$$\int u_1^{i_1} u_2^{i_2} \cdots u_d^{i_d} (K)^l(u)\, d(u) = 0 \quad \text{if } i_j \text{ is odd, for some } j. \tag{4.1}$$

4.2. *Consistency of variable selection and model discovery.* As described earlier, for variable selection and model structure discovery under GRID, here we propose to use the EL method of Owen (1988). The main advantage of this choice is that we do not need to estimate the nuisance parameter σ^2, which is itself a difficult problem in the high dimensional context. The EL methodology has been used earlier in the LLE literature in the low dimensional setting in different inference problems (cf. Chen and Van Keilegom (2009), Chen and Qin (2000) and Zhang and Liu (2003)), again using varying degrees of smoothing. In contrast, here we employ the EL method for testing an *unbounded* number of marginal hypotheses using the specially constructed *unbiased* estimating functions of Section 3, under the nonstandard condition that the bandwidths in H are fixed (and do *not* tend to zero as $n \to \infty$).

To describe the details of the EL step, first we rewrite the univariate estimators in (3.2) as

$$\dot{M}_{0j} = \frac{1}{n} \sum_{t=1}^{n} q_{1,j}(X_t) Y_t, \tag{4.2}$$

$$\dot{M}_{1j}^{(i)} = \frac{1}{n} \sum_{t=1}^{n} q_{i+1,j}(X_t) Y_t, \tag{4.3}$$

where, for $1 \le i, j \le d$ and $1 \le t \le n$, $q_{1,j}(X_t) \equiv q_{1,j}(x^*, X_t; K, H)$ is the first row of matrix in (3.2), $q_{i+1,j}(X_t) \equiv q_{i+1,j}(x^*, X_t; K, H)$ is the row $i+1$ of matrix in (3.2) (with the x^* given by Assumption A2), X_t is the d-dimensional vector of covariates, and Y_t is the dependent variable. (For brevity, we suppress the dependence on the point x^*, the Kernel function K and the bandwidth matrix H). Note that by Theorem 1, $\mathrm{E}(\dot{M}_{0j}) = \theta_{0j}$ and $\mathrm{E}(\dot{M}_{1j}^{(i)}) = \theta_{ij}$ for all i, j. We define the EL for θ_{ij} as $L_n^{(ij)}(\theta_{ij}) = \sup\{\prod_{t=1}^n p_t^{(ij)} : 0 \le p_t^{(ij)} \le 1, \sum_{t=1}^n p_t^{(ij)} = 1, \sum_{t=1}^n p_t^{(ij)}[q_{i+1,j}(X_t)Y_t - \theta_{ij}] = 0\}$, $0 \le i \le d$ and $1 \le j \le d$. The EL ratio statistic for testing $\mathcal{H}_0 : \theta_{ij} = 0$ is then given by $R_n^{(ij)}(\theta_{ij}) = n^n L_n^{(ij)}(\theta_{ij})$ for all i, j. Standard arguments using Lagrange multipliers show (cf. Owen (2001)) that

$$-2 \log R_n^{(ij)}(\theta_{ij}) = -2 \sum_{t=1}^{n} \log n p_t^{(ij)}, \tag{4.4}$$

with $p_t^{(ij)} = [n\{1 + \lambda_{ij}\mathbb{Z}_t^{(ij)}\}]^{-1}$ where $\sum_{t=1}^n p_t^{(ij)} = 1$, $\sum_{t=1}^n p_t^{(ij)}\mathbb{Z}_t^{(ij)} = 0$, $\mathbb{Z}_t^{(ij)} := q_{i+1,j}(X_t)Y_t - \theta_{ij}$, $0 \le i \le d$ and $1 \le j \le d$. For consistency of variable selection and structure discovery, we carry out the marginal EL tests that reject $\mathcal{H}_0 : \theta_{ij} = 0$ when $-2\log R_n^{(ij)}(0) > \eta_n$ for some suitable cut-off point η_n. Here, η_n controls the probability of a wrong decision, for both $\theta_{ij} = 0$ (e.g., for irrelevant variables) and $\theta_{ij} \neq 0$ (e.g., for identifying non/linear variables). The choice of η_n is specified in the theorem below.

THEOREM 3. *Suppose that Assumptions* A1–A4 *hold*, $d = O(n^a)$ *and* $|R| = r = O(n^\gamma)$ *for some* $a \in (0, \infty)$ *and* $\gamma \in [0, 1)$. *Let* $\eta_n = 2a(\log n)^2$, $E|\varepsilon_1|^{2\omega} < \infty$ *for some* $\omega > a + 1$ *and*

$$E K_H^{2\omega}(X_1 - x^*)/\big[E K_H^2(X_1 - x^*)\big]^\omega = O(1). \tag{4.5}$$

(a) (VARIABLE SELECTION CONSISTENCY): *If* $\min_{j\in C}|\theta_{0j}^m| \ge C_1 n^{-k_1}$ *and* $\min_{j\in A}|\theta_{0j}^g| \ge C_1 n^{-k_1}$ *for some* $C_1 > 0$, $k_1 \in [0, 1/2)$ *and* $\gamma < a$, *then* $P(\hat{R} = R) \to 1$ *as* $n \to \infty$.

(b) (CONSISTENCY OF MODEL STRUCTURE DISCOVERY): *If* $\min_{j\in C, i\in I^j}|\theta_{ij}^m| \ge C_2 n^{-k_2}$ *and* $\min_{j\in A, i\in I^j}|\theta_{ij}^*| \ge C_2 n^{-k_2}$ *for some* $C_2 > 0$, $k_2 \in [0, 1/2)$ *and* $\gamma < a/2$, *then*

$$P\big(\hat{C} = C, \hat{A} = A, \text{and } \hat{I}^j = I^j \text{ for all } j = 1, \ldots, r\big) \to 1 \quad \text{as } n \to \infty.$$

Thus, Theorem 3 shows that the GRID has the consistency of variable selection property in very high dimensions, with $d = O(n^a)$, provided the 2ωth order absolute moment of the error variables is finite for some $\omega \in (a + 1, \infty]$. It is also able to identify the correct model structure (cf. part (b)) with probability tending to one under the same moment condition, provided the number r of relevant covariates grow at a suitable rate, among other conditions. For the validity of either part of Theorem 3, the error variables ε_i need not be Gaussian. Further, from the proof of Theorem 3, it also follows that the variable selection consistency and consistency of the model structure parts hold in ultrahigh dimensions with $\log d = o(n)$, provided ε_1 has a finite moment generating function in some neighborhood of the origin. However, the bounds on r are primarily determined by the strength of the signal condition on θ_{ij}'s and do not improve beyond the hard threshold $\gamma < 1$ with the finiteness of higher order moments of ε_1.

REMARK 4.1. The minimal signal strength assumptions, for example, $\min_{j\in C}|\theta_{0j}| \ge C_1 n^{-k_1}$, with $C_1 > 0$ and $k_1 \in [0, 1/2)$ is important to control the overall false positive rate in the multiple testing done as a part of the GRID. It is worth noting that by virtue of using an asymptotically nonvanishing bandwidth, the componentwise convergence rates of the estimating functions satisfy the bound $|\dot{M}_{0j} - \theta_{0j}| = O_p(n^{-1/2})$, just as in the parametric case. As a result, it is adequate to require the minimum signal strength to satisfy the (parametric) lower bounds $\min\{|\theta_{0j}^m| : j \in C\} \ge C_1 n^{-k_1}$ and $\min\{|\theta_{0j}^g| : j \in A\} \ge C_1 n^{-k_1}$ for some $C_1 > 0$ and $k_1 \in [0, 1/2)$. The same is true for the minimum signal strength assumption for model selection.

REMARK 4.2. Note that the choice of the asymptotically nonvanishing common bandwidth h must also satisfy the condition (4.5) when $d = O(n^a)$. This is required to ensure that the moments of (normalized) $q_{1,j}(X_t)$ and $q_{i+1,j}(X_t)$ appearing in the estimating functions of the marginal EL tests are finite for all i, j. This condition is always satisfied with any h for the case where the kernel $K_1(\cdot)$ is the probability density function of a uniform distribution.

Further, as in the proof of Theorem 3, it also follows that for a nonuniform kernel, a choice of h is also permissible. A simple sufficient condition is given in Lemma 1 in the Giordano, Lahiri and Parrella (2020).

Once the model structure discovery is done, it is possible to write down a low dimensional representation of the regression function $m(\cdot)$ using the GRID plot. In the next section, we consider post-model selection estimation of the true regression function.

4.3. *Estimation of the low dimensional regression function.* Note that the true regression function $m(\cdot)$ can be expressed as

$$m(x) = m_0 + \sum_{J_i \in \mathcal{J}} m_{J_i}(x_{J_i}), \tag{4.6}$$

where $\mathcal{J}$ is the collection of disjoint index sets J_i that are subsets of $\{1, \ldots, r\}$ having size one (for isolated covariates) or more (for mixed covariates) such that $J_i \cap J_j = \varnothing$ for $i \neq j$, $\bigcup_i J_i = R$. In other words, a set J_i derives from the union of all the "linked" sets I^j, that is, those which share (directly or indirectly) some common units. For example, for model (1.2), we have $J_1 = I^1$, $J_2 = I^2 \cup I^3 \cup I^7$ and $J_3 = I^4 \cup I^5 \cup I^6$. For identifiability purposes, we shall assume that $E(m_{J_i}) = 0$, for all J_i, so that $m_0 = E(Y_t)$. For simplicity, we assume that the Y_t have been centered around their mean and will ignore the intercept m_0 in the estimation task. We shall suppose that the maximum order of the interaction terms, that is, $d_1 \equiv \max\{|J_i| : J_i \in \mathcal{J}\}$, is fixed and does not change with the sample size n. This restriction gives a sparse additive structure of the regression function $m(\cdot)$ where the number d_2 (say) of the component functions may be unbounded but the maximum order of interactions is a possibly arbitrarily large but bounded number. The formulation here also generalizes the standard additive function formulation (cf. Opsomer (2000), Fan and Jiang (2005) and references therein) by allowing interactions among the relevant variables. This class of functions also appears in a recent work on estimation of sparse high dimensional nonparametric regression (cf. Yang and Tokdar (2015)) and seems to be the natural extension to consider beyond the additive models in high dimensional regressions. In fact, the last paper derives the minimax optimal rate for the estimators in nonparametric regression as in (4.6). In particular, if the unknown function is completely nonadditive then $r = o(\log n)$, where r is the number of relevant covariates. Instead, if one knows the low-dimensional additive structure of the unknown function, then $r = O(n^\gamma)$, for some $0 \leq \gamma < 1$ (cf. Remark 3.4 of Yang and Tokdar (2015)). In this section, we establish the error of estimation of the true (unknown) additive structure in the (4.6) following the structure discovery step of the GRID.

Note that the GRID procedure gives an estimate of the set $\mathcal{J}$ which is obtained by considering the columns of the GRID plot that summarizes the isolated terms and the interaction terms. Note also that the GRID procedure gives estimates of the sets A and C directly, which we shall denote by $\hat{A}$ and $\hat{C}$, respectively. Using this structural information from the GRID plot, the estimation of $m(x)$ can be done using any suitable nonparametric additive function estimation technique. For example, basing on the LLE methodology, we can generalize the backfitting method of Opsomer (2000) and define a GRID-based (preliminary) estimator of the regression function in (4.6) as

$$\widehat{m}(x) = \sum_{J_i \in \hat{\mathcal{J}}} \widehat{m}_{J_i}(x_{J_i}; H_{J_i}), \tag{4.7}$$

where the estimated additive components $\widehat{m}_{J_i}(x_{J_i}; H_{J_i}) \equiv e_1^T \widehat{\mathbf{m}}_{J_i}$ derive from the solution to the following system of equations:

$$\widehat{\mathbf{m}}_{J_1} = S_1\Big(\Upsilon - \sum_{i \neq 1} \widehat{\mathbf{m}}_{J_i}\Big),$$

$$\begin{aligned}
\widehat{\mathbf{m}}_{J_2} &= S_2\Big(\Upsilon - \sum_{i\neq 2}\widehat{\mathbf{m}}_{J_i}\Big),\\
&\vdots\\
\widehat{\mathbf{m}}_{J_{d_2}} &= S_{d_2}\Big(\Upsilon - \sum_{i\neq d_2}\widehat{\mathbf{m}}_{J_i}\Big),
\end{aligned} \tag{4.8}$$

where $\Upsilon = (Y_1, \ldots, Y_n)^T$, $\widehat{\mathbf{m}}_{J_i} = (\widehat{m}_{J_i}(X_{1J_i}; H_{J_i}), \ldots, \widehat{m}_{J_i}(X_{nJ_i}; H_{J_i}))^T$, $S_i = (\mathbf{s}_{J_i, X_{1J_i}}, \ldots, \mathbf{s}_{J_i, X_{nJ_i}})^T$, $\mathbf{s}^T_{J_i, X_{nJ_i}} = e_1^T(\Gamma^T_{J_i} W_{J_i} \Gamma_{J_i} + n^{-1}\mathbb{I}_{J_i})^{-1}\Gamma^T_{J_i} W_{J_i}$. Here, $\mathbb{I}_k$ is the identity matrix of order $k \geq 1$ and e_i is a vector with a one in position i and zeros elsewhere. The $|J_i|$-dimensional quantities Γ_{J_i}, W_{J_i} and X_{tJ_i} are defined as in Section 2.2, with respect to the subset J_i of covariates, but with *new* choices of the bandwidths H_{J_i} as indicated below. The inclusion of n^{-1} in the normalizing matrix S_i does not change the order of the bias and the variance terms, but makes the estimator stable when $\Gamma^T_{J_i} W_{J_i} \Gamma_{J_i}$ is (nearly) singular (cf. Fan (1993)). Lemma 2.1 of Opsomer (2000) gives the assumptions under which system (4.8) has a unique solution.

For better clarity, we now highlight the key considerations in the choices of the bandwidths for the variable selection and model structure discovery tasks of the GRID and for its estimation task. For estimating the regression function $m(x)$, we first apply the identification step of Section 4.2 where all the bandwidths (denoted by h_j) are bounded away from zero as in Assumption A1. In the second step, we carry out the low-dimensional estimation procedure in which the bandwidths are chosen according to identification of linear or nonlinear covariates from the first step. More specifically, in the second step, we choose bandwidths, $h_j^\diamond$, such that they go to zero when n goes to infinity *only for the nonlinear covariates* but not for the linear ones. Thus, in (4.7), for $j \in \hat{A}$ (i.e., for linear covariates), we shall use a bandwidth $h_j^\diamond$ that remains bounded away from zero to ensure root-n consistency, while we require the other $h_j^\diamond$, $j \in \hat{C}$ to die out at a suitable rate with the sample size to ensure consistency. In fact, recall from Proposition 1 that the LLE is unbiased for linear functions (even using asymptotically nonvanishing bandwidths) and, therefore, the only bandwidths which asymptotically contribute to the leading term of the MSE of the GRID regression function estimator are those associated with the nonlinear covariates (i.e., those included in $H^\diamond_{J_i \cap C}$, for $J_i \in \mathcal{J}$). Accordingly, for the preliminary estimator $\widehat{m}(x^0)$ at a point $x^0 \in (0,1)^d$, we set $h_j^\diamond = c_1$ for all $j \in \hat{A}$ for some $c_1 > 0$ and pick vanishing bandwidths $h_j^\diamond$ for $j \in J_i \cap \hat{C}$ such that

$$\sum_{J_i \in \hat{\mathcal{J}}} \left[\frac{1}{n|H^\diamond_{J_i}|} + \operatorname{tr}\big((H^\diamond)^2_{J_i \cap \hat{C}}\big)\right] \leq C_1 n^{-\delta_0} \tag{4.9}$$

for some $\delta_0 \in (0,1)$, $C_1 \in (0, \infty)$ (not depending on j, n). To define the GRID estimator $m^\dagger(x^0)$ of the regression function $m(x)$ at $x = x^0$, note that $m(x) = O(1)$ for all $x \in (0,1)^d$. Hence, we truncate the preliminary estimator $\widehat{m}(x^0)$ at a suitable threshold and define the GRID estimator as

$$m^\dagger(x^0) = \begin{cases} \widehat{m}(x^0) & \text{if } |\widehat{m}(x^0)| \leq t_n, \\ \dfrac{\widehat{m}(x^0)}{|\widehat{m}(x^0)|} t_n & \text{if } |\widehat{m}(x^0)| > t_n, \end{cases}$$

where t_n is a positive constant. Conditions on t_n will be specified in the statement of Theorem 4 below. For stating Theorem 4, define $d_{*i} = |J_i| - |J_i \cap C|$ and $B_n(x) = \sum_{J_i \in \mathcal{J}, J_i \cap C \neq \varnothing} \sum_{j \in J_i \cap C} (h_j^\diamond)^2 \frac{\partial^2 m(x_{J_i})}{\partial x_j^2}$. Also, recall that $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and that μ_2 and

ν_0 are as defined in (2.8). For $x_1 \in (0,1)$ and $c > 0$, set $\nu_0(x_1, c) = \int_{-x_1/c}^{(1-x_1)/c} K_1^2(v)\,dv$ and $\tilde{\nu}(c) = \int_0^1 \nu_0(x_1, c)\,dx_1$. Then, we have the following result:

THEOREM 4. *Suppose that the assumptions and the conditions of Theorem 3 hold. Further, let* $E|\varepsilon_1|^{2\omega} < \infty$ *with* $\omega = a + 2$ *for some* $a \in [1, \infty)$, $d_2 = |\mathcal{J}| = O(r)$ *and* $r = O(n^{\gamma})$ *with* $0 \le \gamma < 1/2 - \delta_0$, *with* δ_0 *of* (4.9), *and* $t_n \asymp \log n$.

(a) (LOCAL RATE). *For any* $x^0 \in (0,1)^d$,

$$
\begin{aligned}
&E\big[(m^{\dagger}(x^0) - m(x^0))^2 | \mathcal{X}_n\big] \\
&= \Bigg[\sum_{\substack{J_i \in \mathcal{J} \\ J_i \cap C = \varnothing}} \frac{\sigma^2 \prod_{j \in J_i} \nu_0(x_j^0, c_1)}{n c_1^{|J_i|}} + \sum_{\substack{J_i \in \mathcal{J} \\ J_i \cap C \neq \varnothing}} \frac{\sigma^2 \nu_0^{|J_i \cap C|} \prod_{j \in J_i \cap C^c} \nu_0(x_j^0, c_1)}{n c_1^{d_{*i}} |H^{\diamond}_{J_i \cap C}|} \\
&\quad + d_2(d_2 - 1) \cdot O(n^{-1}) + \left(\frac{\mu_2^2}{4} \cdot B_n^2(x^0) \right) \Bigg] (1 + o_p(1)).
\end{aligned} \tag{4.10}
$$

(b) (GLOBAL RATE).

$$
\begin{aligned}
&E\left[\int \big(m^{\dagger}(x) - m(x)\big)^2 \, dx \,\middle|\, \mathcal{X}_n \right] \\
&= \Bigg[\sum_{\substack{J_i \in \mathcal{J} \\ J_i \cap C = \varnothing}} \frac{\sigma^2 \tilde{\nu}(c_1)^{|J_i|}}{n c_1^{|J_i|}} + \sum_{\substack{J_i \in \mathcal{J} \\ J_i \cap C \neq \varnothing}} \frac{\sigma^2 \nu_0^{|J_i \cap C|} \tilde{\nu}(c_1)^{d_{*i}}}{n c_1^{d_{*i}} |H^{\diamond}_{J_i \cap C}|} \\
&\quad + d_2(d_2 - 1) \cdot O(n^{-1}) + \left(\frac{\mu_2^2}{4} \int B_n^2(x)\,dx \right) \Bigg] (1 + o_p(1)).
\end{aligned} \tag{4.11}
$$

Thus, it follows that the GRID based estimator of the true regression function is MSE/MISE consistent if we choose asymptotically vanishing bandwidths $h_j^{\diamond}$ for $j \in \hat{C}$, satisfying the hypotheses of Theorem 4. As explained earlier, for the linear covariates, the bandwidths are chosen to be bounded away from zero. Further, the number of additive components is allowed to grow at a rate that is given by a fractional power of the sample size n.

REMARK 4.3 (Optimal bandwidth matrix). It is possible to determine the optimal bandwidths for estimating the function $m(\cdot)$ using the expansion result (4.10) and (4.11), and following the marginalized approach suggested in Giordano and Parrella (2016).

REMARK 4.4 (NP-Oracle property). Suppose that the number of additive components $d_2 = O(1)$ and that each nonlinear function in (4.6) is k-times continuously differentiable with a α-Hölder continuous k-th derivative; Note that, by Assumption A3, $k \ge 4$. Next, write $\alpha_m = k + \alpha$ which specifies the common smoothness coefficient for the nonlinear unknown functions in (4.6). Then, the rates for the optimal bandwidths are $h_j^{\diamond} = O(n^{-\frac{1}{2\alpha_m + r_i}})$ for $j \in C \cap J_i$ for where $r_i = |J_i \cap C|$. Then, using the arguments in the proof of Theorem 4, it can be shown that

$$
E\big[(m^{\dagger}(x^0) - m(x^0))^2 | \mathcal{X}_n\big] = O_p(n^{-\frac{2\alpha_m}{2\alpha_m + r_J^*}}),
$$

where $r_J^* = \max_{1 \le i \le d_2} |J_i \cap C|$. Note that this rate matches the optimal rate of estimating the nonparametric regression function $m(x)$ by an Oracle that has the knowledge of the true low dimensional structure (4.6) of the regression function *a priori*. As a result, the GRID with optimal bandwidths has the NP-Oracle property (i.e., nonparametric Oracle property) as defined in Storlie et al. (2011).

REMARK 4.5 (Advantages of model structure discovery). For the task of variable selection (part (a) of Theorem 3) only, we require the number of relevant covariates to satisfy $r = O(n^\gamma)$, with $0 \le \gamma < \min\{1, a\} \equiv \gamma^*$. For the model structure discovery (cf. part (b) of Theorem 3), we need $r = O(n^\gamma)$ with $0 \le \gamma < \min\{1, a/2\} \equiv \gamma^{**}$. And we require $0 \le \gamma < 1/2 - \delta_0$ for the estimation property (Theorem 4). In comparison, for a general nonadditive regression function, existing methods that only have the variable selection property (but not the structure discovery capabilities) can only estimate a single nonlinear component of $r = o(\log n)$ covariates (cf. Comminges and Dalalyan (2012), Yang and Tokdar (2015)).

To summarize, the GRID based estimator can be used for consistent estimation of the nonparametric regression function $m(\cdot)$ in very high dimensions, provided the regularity conditions of Theorem 4 hold. In comparison to existing methods where the dimension d is at the best restricted to a logarithmic rate of growth as a function of the sample size n, the GRID can identify the correct set of linear and nonlinear variables and their interactions and it can simultaneously provide a consistent estimator of the true regression function without any need to specify the maximum order of interactions *a priori*, allowing the dimension d to grow at the rate $d = O(n^a)$ for any $a \in (0, \infty)$.

5. An extension of GRID to dependent and nonuniform covariates. In this section, we present an extension of the GRID method allowing for nonuniform and dependent covariates taking values in $(0, 1)^d$. From the proof of Theorem 3, it follows that the variable selection consistency and consistency of model structure discovery properties of the GRID critically depend on the behavior of certain "moments" involving the covariate vector $X_1 \in \mathbb{R}^d$ and the Kernel function $K(\cdot)$. For the case of nonuniform and dependent covariates $X_1 \in \mathbb{R}^d$, the GRID continues to have these properties, provided the corresponding moments are close to their values in the uniform case. To state these, let $f(\cdot)$ denote the joint pdf of $X_1 = (X_{11}, \ldots, X_{1d})'$ and let $f^{(j)}$ denote the jth partial derivative of f, $j = 1, \ldots, d$. Also, for any $p \ge 1$ and $1 \le j, i_1, \ldots, i_p \le d$, define the product moments

$$\mu_p^*(i_1, \ldots, i_p) = \int \left(\prod_{l=1}^{p} u_{i_l} \right) K(u) f(x^* + Hu)\, du,$$

$$\mu_{pj}^*(i_1, \ldots, i_p) = \int \left(\prod_{l=1}^{p} u_{i_l} \right) K(u) f^{(j)}(x^* + Hu)\, du,$$

where x^* is as in Assumption A2. Also, define the analogs $\{\tilde{\mu}_p^*(\cdot), \tilde{\mu}_{pj}^*(\cdot)\}$ and $\{\tilde{\mu}_{p,-i}^*(\cdot), \tilde{\mu}_{pj,-i}^*(\cdot)\}$, obtained by replacing $f(\cdot)$ in above with the pdfs of the transformed variables $\phi(x)$ and $\phi_{-i}(x)$ from Theorem 2, respectively. To prove the variable selection consistency and consistency of model structure discovery properties of the GRID under dependence and nonuniformity, instead of Assumption A4, we shall use the following assumption on the X_i:

A4′: There exist constants $k_3, C_2 \in (0, \infty)$ such that for all $i \in R$, $j \in \{1, \ldots, d\}$ and for all $i_1, \ldots, i_p, j_1, \ldots, j_{q-1} \in R$ with $1 \le p, q - 1 \le 4$,

$$|\mu_p(i_1, \ldots, i_p)| + |\mu_{qj}(j, j_1, \ldots, j_{q-1})| \le C_2 n^{-k_3},$$

for $\mu_p \in \{\mu_p^*, \tilde{\mu}_p^*, \tilde{\mu}_{p,-i}\}$ and $\mu_{qj} \in \{\mu_{qj}^*, \tilde{\mu}_{qj}, \tilde{\mu}_{qj,-i}\}$.

The moment conditions and their orders (namely, p and q above) are directly determined by the Taylor's expansions of the biases of the modified LLEs given in Section 3. It is easy to check that under Assumption A2, the left hand side of the inequality above is identically zero when X_1 has the uniform distribution on $(0, 1)^d$. Thus, Assumption A4′ essentially puts

a bound on potential deviations of the distribution of X_1 from the uniform distribution. To appreciate the extent of nonuniformity and dependence among the covariates that is allowed by Assumption A4′, consider the following example:

EXAMPLE 5.1. Suppose that the pdf of $X_1 \in \mathbb{R}^d$ is given by

$$f(x) = 1 + \prod_{j=1}^{d} \xi_j(x_j), \quad x = (x_1, \ldots, x_d)' \in (0,1)^d,$$

where $\xi_j : (0,1) \to \mathbb{R}$ are integrable functions such that $f(x) \geq 0$ for all x. Note that $f(\cdot)$ is a proper pdf if $\int_0^1 \xi_{j_0}(x_{j_0})\, dx_{j_0} = 0$ for some $j_0 \in \{1, \ldots, d\}$. Assuming that this condition holds and that $\int_0^1 |\xi_j(x_j)|\, dx_j \neq 0$ for all $j = 1, \ldots, d$ (i.e., none of the functions $\xi_j(\cdot)$ are identically equal to zero, a.e.), it follows that under $f(\cdot)$, the d covariates in X_1 are neither independent nor uniformly distributed over $(0,1)^d$. Next suppose that $x^* = (1/2, \ldots, 1/2)'$ and the functions $\xi_j(\cdot)$ are differentiable on $(0,1)$ for all j. Then, Assumption A4′ is satisfied if at least one of the marginal moments under individual $\xi_j(\cdot)$ is zero for each of the moments in A4′. A simple sufficient condition for this holds if a small number of the functions $\xi_j(\cdot)$ satisfy some symmetry property. Specifically, suppose that there exist $j_1, j_2 \in C$ such that $\xi_{j_1}(\cdot)$ is symmetric about $1/2$ and $\xi_{j_2}(\cdot)$ is antisymmetric about $1/2$. Then it is not difficult to verify that all the moments in Assumption A4′ are zero and hence, Assumption A4′ holds. The upper bound $C_2 n^{-k_3}$ on the moments leaves additional room for allowing further deviations of the joint distribution of the covariates from uniformity and independence.

We next point out that for the validity of the GRID in the dependent case, k_3 need not depend on d, the dimension of the regression model, which is allowed to have the same growth rate $d = O(n^a)$ with $a \in (0, \infty)$ as in the independent (covariates) case, under the same moment conditions on the error variable ε_1. However, we do require k_3 to depend on the order of r, the number of relevant variables. We also require a stronger condition on the signal strength. More precisely, we have the following result on consistency of variable selection and model structure discovery in the dependent and nonuniform case.

THEOREM 5. *Suppose that Assumptions* A1–A3, A4′ *and* (4.5) *hold*, $d = O(n^a)$, *and* $|R| = r = O(n^\gamma)$ *for some* $a \in (0, \infty)$ *and* $\gamma \in [0, 1)$. *Let* $\eta_n = \eta_n^\star \equiv 2a\sqrt{n}/(\log n)$ *and let* $E|\varepsilon_1|^{2\omega} < \infty$ *for some* $\omega > a + 1$.

(a) (VARIABLE SELECTION CONSISTENCY): *Suppose that* $\min_{j \in C} |\theta_{0j}| \geq c_1$ *and* $\min_{j \in A} |\theta_{0j}^g| \geq c_1$ *for some* $c_1 \in (0, \infty)$ *and* $2\gamma < k_3 - 1/4$. *Then*,

$$P(\hat{R} = R) \to 1 \quad \text{as } n \to \infty.$$

(b) (CONSISTENCY OF MODEL STRUCTURE DISCOVERY): *If* $\min_{j \in C, i \in I^j} |\theta_{ij}| \geq c_2$ *and* $\min_{j \in A, i \in I^j} |\theta_{ij}^*| \geq c_2$ *with* $c_2 \in (0, \infty)$ *and* $3\gamma < k_3 - 1/4$, *then*

$$P\big(\hat{C} = C, \hat{A} = A, \text{and } \hat{I}^j = I^j \text{ for all } j = 1, \ldots, r\big) \to 1 \quad \text{as } n \to \infty.$$

Thus, under the conditions of Theorem 5, the GRID method continues to perform consistent variable selection and model structure discovery even when the covariates are possibly nonuniform and dependent. For either of the tasks, the conditions on the strength of dependence are not affected by model dimension $d = Cn^a$. However, the dependence must be suitably smaller when r, the number of relevant variables, is larger. Further, the requirements on the level of dependence is weaker for variable selection consistency of the GRID as compared to consistency of model structure discovery. The difference in requirements on k_3 in

parts (a) and (b) of Theorem 5 arises from the forms of the structural parameters θ_{ij} where under part (a), the effect of dependence on the values of θ_{0j} comes from $O(r^2)$ many interaction terms while under part (b), it comes from $O(r^3)$ many terms. See the proof of Theorem 5 for more details.

In Theorem 5, the condition on the minimum nonzero signal strength can be weakened. It is straightforward to modify the proof of Theorem 5 to allow the minimum signal strengths to go to zero at a certain rate depending on k_3 and γ, but we do not pursue such refinements here in order to keep the statement of Theorem 5 simple. In the same vein, an analog of Theorem 4 on estimation of the true low dimensional regression function holds under dependence of the covariates, where the variance part, given by the first two terms of the MSE/ MISE expansions in Theorem 4, now involves cross product moments of $K_{H_J}(x_J^0 - X_{1,J})\varepsilon_i$, $J \in \mathcal{J}$ (where the subscript J denotes the sub-vectors/matrices corresponding to row/column indices in J), and the term '$d_2(d_2-1)O(n^{-1})$' is replaced by '$d_2(d_2-1)O(n^{-k_3})$', requiring k_3 to be suitably small.

In the next section, we consider finite sample performance of the GRID method for both independent and dependent covariates and also compare the performance of GRID with the existing methods.

6. Simulation. The Monte Carlo simulation is based on 200 iterations. To begin with, we shall consider uniformly distributed covariates. For each model, the additive components are standardized so that they all have variance equal to one, to make them comparable each other. We consider 8 different models, summarized in the following table.

Model	$m(x)$	Error density
1	$X_6^3 X_7^3 + X_{10}$	$\varepsilon \sim N(0,1)$
2	$\sin(10X_2) + X_3X_4 + X_5$	$\varepsilon \sim N(0,1)$
3	$X_1 + X_2 + X_3 + X_4 + X_5$	$\varepsilon \sim N(0,1)$
4	$\exp(10X_1X_6X_8)$	$\varepsilon \sim N(0,0.5^2)$
5	$X_1X_2 + X_1X_7^3$	$\varepsilon \sim N(0,1)$
6	$(X_1+1)^3 + \sin(10X_2)$	$\varepsilon \sim N(0,1)$
7	$X_1 + \frac{1}{1+X_2} + \sin(X_3) + \exp(X_4) + X_5^2$	$\varepsilon \sim N(0,1)$
8	$5X_1^2X_2^2$	$\varepsilon \sim N(0,0.5^2)$

For all simulation results, the kernel is $K_1(u) = 1/C_1(5-u^2)\mathbb{I}_{\{|u|\le\sqrt{5}\}}$, as in Lafferty and Wasserman (2008), where C_1 is a scale factor to make the integral equal to one.

The results of the simulations are shown in the following Tables 1–2, and in Tables 9–11 given in the supplementary material Giordano, Lahiri and Parrella (2020), for different values of dimensions d and sample sizes n. For each model, we report the proportion of times that a given covariate X_i is classified as a relevant covariate (i.e., it is classified in the set R) or as a nonlinear covariate (i.e., it is classified in the set C), and as part of an interaction term (i.e., it is classified in the set I). Note that the dimension d varies from a minimum value 20 to a maximum value 2000, and note also that in the right-hand columns of the tables the dimension is such that $d > n$. The performance of GRID is satisfying. Of course, the performance deteriorates for very high dimensions, but it always improves as long as the sample size increases, showing the consistency.

In order to analyze the effects of the selection on function estimation, for each run of the simulation study we estimate the structure of the model by GRID and then we pass it to a generalized additive model nonparametric estimator to have the estimated function (we use

TABLE 1

Simulation results for model 1 *and different dimensions* d *and sample sizes* n. *The values show the proportion of times that a given covariate* X_i *is classified as a relevant covariate* (R), *as a nonlinear covariate* (C), *and as part of an interaction term* (I). *The symbol* (*) *denotes a value* ≤ 0.025 *while the symbol* (–) *means zero*

		Model 1								
		$d = 20$			$d = n/2$			$d = 2n$		
	n	R	C	$I(6,7)$	R	C	$I(6,7)$	R	C	$I(6,7)$
X_6	300	0.975	0.330	0.900	0.855	0.335	0.720	0.630	0.330	0.365
	500	1.000	0.610	1.000	0.990	0.595	0.985	0.810	0.480	0.620
	1000	1.000	0.910	1.000	1.000	0.915	1.000	0.910	0.835	0.815
X_7	300	0.940	0.325	0.900	0.875	0.335	0.720	0.580	0.250	0.365
	500	1.000	0.370	1.000	0.995	0.635	0.985	0.765	0.515	0.620
	1000	1.000	0.935	1.000	1.000	0.890	1.000	0.835	0.815	0.815
X_{10}	300	1.000	*	–	1.000	*	–	0.995	0.035	–
	500	1.000	*	–	1.000	*	–	1.000	*	–
	1000	1.000	*	–	1.000	*	–	1.000	*	–

the function `gam` of the R package `mgcv`). In Table 3, we report the average of the ratio of the MSEs for the estimated model by GRID and for the true model. The values in brackets are the standard deviations. We consider three different distributions for the errors, that is, Normal, Exponential and Pareto, whose parameters are set so that the variances are the same (cf. Tables 1, 2, 9 and 10).

6.1. *Results for additive models*. First, we analyze two additive models, named 6 and 7 in the previous table. In particular, model 6 has been taken from Lafferty and Wasserman (2008), while model 7 has been used by Radchenko and James (2010). As in Radchenko and James (2010), we standardize the components in order to make them of equivalent magnitude.

TABLE 2

Simulation results for model 2 *with different dimensions* d *and sample sizes* n. *The values show the proportion of times that a given covariate* X_i *is classified as a relevant covariate* (R), *as a nonlinear covariate* (C), *and as part of an interaction term* (I). *The symbol* (*) *denotes a value* ≤ 0.025 *while the symbol* (–) *means zero*

		Model 2								
		$d = 20$			$d = n/2$			$d = 2n$		
	n	R	C	$I(3,4)$	R	C	$I(3,4)$	R	C	$I(3,4)$
X_2	300	0.805	0.805	–	0.530	0.530	–	0.300	0.303	–
	500	0.970	0.970	–	0.835	0.835	–	0.520	0.520	–
	1000	1.000	1.000	–	0.990	0.990	–	0.715	0.715	–
X_3	300	0.920	*	0.275	0.710	*	0.170	0.450	*	0.050
	500	0.990	*	0.535	0.950	*	0.505	0.690	*	0.240
	1000	1.000	*	0.920	1.000	*	0.920	0.795	*	0.590
X_4	300	0.915	*	0.275	0.790	*	0.170	0.430	*	0.050
	500	0.990	*	0.535	0.950	*	0.505	0.610	*	0.240
	1000	1.000	*	0.920	1.000	0.035	0.920	0.815	*	0.590
X_5	300	1.000	*	–	1.000	*	–	0.990	*	–
	500	1.000	*	–	1.000	*	–	1.000	*	–
	1000	1.000	*	–	1.000	0.030	–	0.990	*	–

TABLE 3

Average ratios of the MSEs of the estimated function by the GRID and the Oracle estimator, with different error densities and $n = 500$. The values in brackets are the standard deviations

	Error density	$d = 20$	$d = n/2$	$d = 2n$
Model 1	$N(0, 1)$	0.976 (0.05)	0.977 (0.07)	0.850 (0.17)
	Exp(1)	0.973 (0.07)	0.977 (0.06)	0.873 (0.17)
	Pa(11.04; 13)	0.975 (0.06)	0.967 (0.08)	0.875 (0.17)
Model 2	$N(0, 1)$	0.964 (0.07)	0.909 (0.13)	0.707 (0.20)
	Exp(1)	0.965 (0.07)	0.901 (0.14)	0.733 (0.21)
	Pa(11.04; 13)	0.968 (0.07)	0.907 (0.14)	0.712 (0.19)
Model 3	$N(0, 0.25)$	0.729 (0.37)	0.534 (0.39)	0.308 (0.32)
	Exp(2)	0.768 (0.36)	0.579 (0.40)	0.289 (0.33)
	Pa(5.52; 13)	0.822 (0.32)	0.553 (0.41)	0.279 (0.32)
Model 4	$N(0, 1)$	0.958 (0.06)	0.956 (0.06)	0.906 (0.11)
	Exp(1)	0.964 (0.05)	0.958 (0.06)	0.918 (0.10)
	Pa(11.04; 13)	0.963 (0.05)	0.957 (0.06)	0.913 (0.11)

Note that models 6 and 7 have no interaction terms. Anyway, to make comparisons among the three methods fair (remember that RODEO and GRID use a non additive structure, so they take always into account the interaction terms), we add in VANISH only the two-order interaction terms.

In Table 4, *FP* is the false positive rate, *FN* is the false negative rate and *LS* is the ratio between the estimated variance of residuals for different methods w.r.t. the same variance using the true model (both are estimated with generalized additive nonparametric estimator, GAM). Moreover, the dash (–) means that the method cannot be applied or is unfeasible, for the given dimensionality.

The results in Table 4 confirm the simulation results in Table 2 of Radchenko and James (2010). For case $d = 20$, the most efficient procedure is VANISH because Models 6 and 7 satisfy the additivity assumption. Remember that GRID is more general, so it loses efficiency since it is not based on the additivity assumption. Anyway, we can note from the tables that

TABLE 4

Simulation results for the additive models 6 and 7, with sample size $n = 500$ and different dimensions d. The values show the FP = False Positive and the FN = False Negative rates. The value LS is the ratio between the estimated variance of residuals for different methods w.r.t. the same variance using the true model (both are estimated through a Generalized Additive Model nonparametric estimator). The symbol (–) means the method is unfeasible

		Model 6			Model 7		
d		GRID	RODEO	VANISH	GRID	RODEO	VANISH
20	FP	0.077	0.118	0.036	0.090	0.005	0
	FN	0.015	0.814	0	0	1.000	0.005
	LS	1.030	1.890	1.000	1.112	5.757	1.001
250	FP	0.086	0.096	–	0.095	0	–
	FN	0.071	0.990	–	0.005	1.000	–
	LS	1.082	2.677	–	1.112	6.046	–
1000	FP	0.130	–	–	0.155	–	–
	FN	0.285	–	–	0.315	–	–
	LS	1.303	–	–	1.404	–	–

TABLE 5
Simulation results for the non additive model 8, *with sample size* $n = 500$ *and dimension* $d = 10$. *Here, FP, FN LS are as in table* 4

	Model 8		
	GRID	RODEO	*VANISH*
FP	0.075	0.0165	0.075
FN	0	0.550	0.015
LS	1.047	2.247	1.326

GRID is comparable with VANISH and works better than RODEO, which also does not assume an additive model. As the dimension d increases, the results for the GRID method are confirmed, while the other two methods become unfeasible. In particular, RODEO cannot be applied for $d > n$ while VANISH would require adding 31,125 interaction terms when $d = 250$ and even more when $d > 250$, which is impossible.

Besides, there are further considerations to make in favor of the GRID method. First of all, it also takes into account the type of covariate (linear or nonlinear) and the interaction terms of any order, contrary to what the other two methods do. Secondly, it is important to note that the results reported for the RODEO and VANISH methods have been produced after a deep search for the best setting values of their tuning/nuisance parameters (when available, we use the true values, otherwise we search for the best values numerically). Note that this cannot be done in the real data applications. GRID, instead, has no tuning parameters to initialize.

6.2. *Results for non additive models*. We consider now model 8, a non additive model used in Lafferty and Wasserman (2008). In this case we do not standardize because we have only one component. We choose the same parameters used in their paper, so $n = 500$, $d = 10$ and $\sigma = 0.5$.

For this example, the ORACLE model is $f(X_1, X_2)$. Therefore, the variance of residuals is very similar w.r.t. the ORACLE model, $f(X_1, X_2)$, if the identification works correctly. But note that VANISH always assume a model with the main effects, $f_1(X_1) + f_2(X_2) + f_{12}(X_1, X_2)$, which is different from the true one. So, this is a case in which the VANISH method performs poorly.

The results are reported in Table 5. We consider the same indicators as before: *FP* is the false positive rate, *FN* is the false negative rate and *LS* is the ratio between the estimated variance of residuals for different methods w.r.t. the same variance using the true model (both are estimated with GAM, a generalized additive model). For VANISH, we compute *FP* and *FN* on the main effects.

6.3. *Results for dependent covariates*. In this section, we report the results of a simulation study when the covariates are not independent. First, we consider the case when the irrelevant covariates have some correlation structure but the relevant ones are independent. We define the following:

$$\text{Model 9:} \quad Y = X_6^2 + X_{10} + \varepsilon.$$

The covariates have this structure: $X_i = U_i$, $i = 1, \ldots, 10$, $X_i = \frac{U_i + cW}{1+c}$, $i = 11, \ldots, d$ where $U_i \sim U(0; 1)$, $i = 1, \ldots, d$, are i.i.d. variables and $W \sim U(0; 1)$ is independent of U_i. For $c = 1$, the correlation is 0.5 and all the irrelevant covariates are correlated but the relevant ones are not. We report the results for this case in Table 6. As one expects, the results confirm

TABLE 6

Simulation results for Model 9 *with different dimensions* d *and sample sizes* n. *The covariates follows the structure above with* $c = 1$. *The values show the proportion of times that a given covariate* X_i *is classified as a relevant covariate* (R), *as a nonlinear covariate* (C). *The symbol* (*) *denotes a value* ≤ 0.05

		Model 9					
		$d = 20$		$d = n/2$		$d = 2n$	
	n	R	C	R	C	R	C
X_6	300	1.000	0.920	1.000	0.920	0.955	0.900
	500	1.000	0.990	1.000	1.000	0.895	0.895
	1000	1.000	1.000	1.000	1.000	0.985	0.985
X_{10}	300	1.000	*	1.000	*	0.795	*
	500	1.000	*	0.995	*	0.630	*
	1000	1.000	*	1.000	*	0.850	*

a good performance for the variable selection (nonlinear, linear and irrelevant covariates).

The second case we analyze refers to a nonlinear *screening selection*, that is a screening selection procedure only for the nonlinear covariates, in much larger dimensions. The covariates have the structure: $X_i = U_i + cW$, $i = 1, \ldots, d$ where W is independent of U_i. For $c = 0.65$ the correlation is 0.3 and for $c = 1$ the correlation is 0.5. Note that this correlation structure satisfies the condition 2 of Fan and Lv (2008). The model we consider is

$$\text{Model 10:} \quad Y = 2X_6^2 + X_{10} + \varepsilon,$$

where the nonlinear component is two times the noise variability. We fix a length of $\{8, 12\}$ for relevant nonlinear covariates according to $n \in \{300, 500\}$. The results in Table 7 confirm that the GRID procedure can detect the true nonlinear covariates. For another example where all covariates are correlated, see Giordano, Lahiri and Parrella (2020).

6.4. *Computational time*. In this section, we report the time for the GRID procedure. As pointed out earlier, there are two built-in features of the GRID algorithm that makes the computation fast. First, the estimating functions $\dot{M}_{0j}$ and $\dot{M}_{1j}$ derived from the LLEs are always available in *closed* form and are easy to compute. Second, the EL procedure is applied to 1-dimensional parameters, one at a time, for which very fast computational algorithms are available (cf. Owen (2001)). Table 8 shows the computational times to perform the GRID procedure, using a PC with Intel Core i7, Quad-Core and 3.40 GHz. We perform both variable

TABLE 7

Simulation results for Model 10 *with different sample sizes* n, *dimensions* d *and* c. *The covariates follow the dependence structure above with* $c = 0.65$ *and* $c = 1$. *The table shows the proportions for the True nonlinear covariate in the sets of lengths* $\{8, 12\}$ *for* $n \in \{300, 500\}$, *respectively*

	Model 10			
	$c = 0.65$		$c = 1$	
n	$d = 4n$	$d = 10n$	$d = 4n$	$d = 10n$
300	1	1	0.99	0.91
500	1	0.99	1	0.96

TABLE 8
Average computational time (seconds) for one iteration of the GRID procedure for Model 1 with different sample sizes n and different dimensions d

	d		
n	20	$n/2$	$2n$
300	0.10	0.13	0.33
500	0.14	0.28	0.74
1000	0.47	0.97	2.52

and model selection steps and also estimate the "reduced" regression function using the *gam* function in the R package *mgcv*. The times are in seconds and refer to the average values over 50 iterations of GRID procedure on a simulated data from Model 1 with Uniform covariates and Normal errors. Computational times are found to be significantly shorter than those of the competing methods considered in Section 6.1. See Giordano, Lahiri and Parrella (2020) for more details.

Acknowledgements. The authors thank two anonymous referees, the Associate Editor and the Editor, Prof. Tailen Hsing, for their constructive comments that led to significant improvements to an earlier draft of the manuscript, including the addition of Section 5.

Research of the second author was partially supported by NSF Grants DMS-1310068 and DMS-1613192.

SUPPLEMENTARY MATERIAL

Supplement to "GRID: A variable selection and structure discovery method for high dimensional nonparametric regression" (DOI: 10.1214/19-AOS1846SUPP; .pdf). Proofs and additional simulation results.

REFERENCES

BERTIN, K. and LECUÉ, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.* **2** 1224–1241. MR2461900 https://doi.org/10.1214/08-EJS327

CHANG, J., TANG, C. Y. and WU, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.* **41** 2123–2148. MR3127860 https://doi.org/10.1214/13-AOS1139

CHEN, S. X. and QIN, Y. S. (2000). Empirical likelihood confidence intervals for local linear smoothers. *Biometrika* **87** 946–953. MR1813987 https://doi.org/10.1093/biomet/87.4.946

CHEN, S. X. and VAN KEILEGOM, I. (2009). A review on empirical likelihood methods for regression. *TEST* **18** 415–447. MR2566404 https://doi.org/10.1007/s11749-009-0159-5

CHOI, N. H., LI, W. and ZHU, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.* **105** 354–364. With supplementary material available online. MR2656056 https://doi.org/10.1198/jasa.2010.tm08281

COMMINGES, L. and DALALYAN, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40** 2667–2696. MR3097616 https://doi.org/10.1214/12-AOS1046

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004. MR1209561

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216. MR1212173 https://doi.org/10.1214/aos/1176349022

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications. Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

FAN, J. and JIANG, J. (2005). Nonparametric inferences for additive models. *J. Amer. Statist. Assoc.* **100** 890–907. MR2201017 https://doi.org/10.1198/016214504000001439

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

FAN, J., GASSER, T., GIJBELS, I., BROCKMANN, M. and ENGEL, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Ann. Inst. Statist. Math.* **49** 79–99. MR1450693 https://doi.org/10.1023/A:1003162622169

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2007). Sparse inverse covarinace estimation with the graphical lasso. *Biostatistics* **9** 432–441.

GIORDANO, F., LAHIRI, S. N. and PARRELLA, M. L. (2020). Supplement to "GRID: A variable selection and structure discovery method for high dimensional nonparametric regression." https://doi.org/10.1214/19-AOS1846SUPP

GIORDANO, F. and PARRELLA, M. L. (2016). Bias-corrected inference for multivariate nonparametric regression: Model selection and oracle property. *J. Multivariate Anal.* **143** 71–93. MR3431420 https://doi.org/10.1016/j.jmva.2015.08.016

HALL, P. and MILLER, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18** 533–550. MR2751640 https://doi.org/10.1198/jcgs.2009.08041

JAMES, G. M., RADCHENKO, P. and LV, J. (2009). DASSO: Connections between the Dantzig selector and lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 127–142. MR2655526 https://doi.org/10.1111/j.1467-9868.2008.00668.x

LAFFERTY, J. and WASSERMAN, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.* **36** 28–63. MR2387963 https://doi.org/10.1214/009053607000000811

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500 https://doi.org/10.1214/009053606000000722

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443 https://doi.org/10.1214/09-AOS692

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

OPSOMER, J. D. (2000). Asymptotic properties of backfitting estimators. *J. Multivariate Anal.* **73** 166–179. MR1763322 https://doi.org/10.1006/jmva.1999.1868

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049 https://doi.org/10.1093/biomet/75.2.237

OWEN, A. B. (2001). *Empirical Likelihood*. CRC Press/CRC, Boca Raton, FL.

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation. Probability and Mathematical Statistics*. Academic Press, New York. MR0740865

QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325. MR1272085 https://doi.org/10.1214/aos/1176325370

RADCHENKO, P. and JAMES, G. M. (2010). Variable selection using adaptive nonlinear interaction structures in high dimensions. *J. Amer. Statist. Assoc.* **105** 1541–1553. MR2796570 https://doi.org/10.1198/jasa.2010.tm10130

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255 https://doi.org/10.1111/j.1467-9868.2009.00718.x

RUPPERT, D. and WAND, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22** 1346–1370. MR1311979 https://doi.org/10.1214/aos/1176325632

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

STONE, C. J., HANSEN, M. H., KOOPERBERG, C. and TRUONG, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.* **25** 1371–1470. With discussion and a rejoinder by the authors and Jianhua Z. Huang. MR1463561 https://doi.org/10.1214/aos/1031594728

STORLIE, C. B., BONDELL, H. D., REICH, B. J. and ZHANG, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statist. Sinica* **21** 679–705. MR2829851 https://doi.org/10.5705/ss.2011.030a

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9** 60–62.

YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139 https://doi.org/10.1214/14-AOS1289

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, H. H., CHENG, G. and LIU, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *J. Amer. Statist. Assoc.* **106** 1099–1112. MR2894767 https://doi.org/10.1198/jasa.2011.tm10281

ZHANG, J. and LIU, A. (2003). Local polynomial fitting based on empirical likelihood. *Bernoulli* **9** 579–605. MR1996271 https://doi.org/10.3150/bj/1066223270

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735