Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Component identification and estimation in nonlinear high-dimensional regression models by structural adaptation

Samarov, Alexander, Spokoiny, Vladimir and Vial, Celine

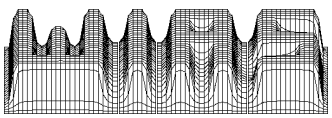
University of Massachusetts-Lowell
and Massachusetts Institute of Technology.
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, U.S.A.
E-Mail: samarov@mit.edu

Weierstrass Institute
and Humboldt University Berlin,
Mohrenstr. 39, 10117 Berlin, Germany
E-Mail: spokoiny@wias-berlin.de
URL: http://www.wias-berlin.de/~spokoiny

ENSAI, Campus Ker Lann,
rue B. Pascal, 35170 Bruz, France
E-Mail: vial@ensai.fr

No. 828

Berlin 2003



1991 *Mathematics Subject Classification.* 62H30, 62J02.

Key words and phrases. Structural adaptation, partially linear model, component analysis.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
D — 10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

This article proposes a new method of analysis of a partially linear model whose nonlinear component is completely unknown. The target of analysis is identification of the set of regressors which enter in a nonlinear way in the model function, and the complete estimation of the model including slope coefficients of the linear component and the link function of the nonlinear component. The procedure also allows for selecting the significant regression variables. As a by-product, we develop a test of linear hypothesis against a partially linear alternative, or, more generally, a test that the nonlinear component is M-dimensional for $M = 0, 1, 2, \ldots$.

The approach proposed in this article is fully adaptive to the unknown model structure and applies under mild conditions on the model. The only important assumption is that the dimensionality of nonlinear component is relatively small. The theoretical results indicate that the procedure provides a prescribed level of the identification error and estimates the linear component with the accuracy of order $n^{-1/2}$. A numerical study demonstrates a very good performance of the method even for small or moderate sample sizes.

1 Introduction

We consider the model

$$y = f(\boldsymbol{x}) + \varepsilon, \qquad f(\boldsymbol{x}) = \theta^\top \boldsymbol{x}_1 + G(\boldsymbol{x}_2), \tag{1.1}$$

where $\boldsymbol{x}^\top = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top)$, $dim(\boldsymbol{x}_2) = M$, $dim(\boldsymbol{x}_1) = d - M$, and $M \ll d$. Function $G(\cdot)$, vector of coefficients θ, and the distribution of the noise ε are unknown. And most importantly, we do not know with respect to which $d_1 = d - M$ variables $\boldsymbol{x}_1$ the model is linear.

The model (1.1) naturally generalizes the linear model and are called a *partially linear* model. Such models can be used in analysis of high dimensional data when the assumption of linearity is too restrictive. They can also be used as a natural alternative to a linear model in the problem of testing the linearity assumption. A general case with a high dimensional nonlinear component makes the analysis complicated because of the "curse of dimensionality" problem. In this paper we consider the situation in which the nonlinear component is low dimensional, that is, M is relatively small.

Hristache, Juditsky and Spokoiny (2001) and Hristache, Juditsky, Polzehl and Spokoiny (2001) (referred to as HJS and HJPS, respectively, in the rest of the paper) proposed a new method of exploring a high-dimensional regression model with the help of a general structural adaptation approach. The aim of the present article is to apply this approach to the estimation and inference in the partially linear model (1.1). The analysis includes, in particular, estimation of the degree of nonlinearity M, identifying with respect to which $d - M$ variables $\boldsymbol{x}_1$ the model is linear or equivalently which M variables enter in f in a nonlinear fashion, estimation of the vector θ and of the nonlinear link function G.

It is important to note that the approach proposed here provides also a new method of selecting significant variables in nonparametric regression in case when the dimensionality of the nonlinear component is relatively small. More specifically, after selecting M (significant) nonlinear variables, one can further select variables among the linear ones using standard methods of linear regression analysis or by testing significance of linear variable slopes as suggested in Section 4.3 below.

As a by-product of our analysis, we develop a test of the hypothesis of linearity against a partially linear alternative, and, more generally, a test of the hypothesis that the dimensionality of the nonlinear component does not exceed the prescribed value M.

Following the work of Engle, Granger, Rice and Weiss (1986), much attention has been directed to estimating model (1.1). See, for example, Heckman (1986), Rice (1986), Chen (1988), Robinson (1988), Speckman (1988), Gao (1995), Schick (1996a,b), Bhattacharya and Zhao (1997), Mammen and Van der Geer (1997), Hamilton and Truong (1997), Eubank, Kambour, Kim, Klipple, Reese and Schimek (1998), Schimek (2000), Golubev and Härdle (2000, 2001). Further references and applications of partially linear models could be found in the recent book by Härdle, Liang, and Gao (2000). This literature addressed the problem of estimation of the parametric and nonparametric components of the model (1.1) under the assumption that the "nonlinear" variables $\boldsymbol{x}_2$ are specified and, in fact, most papers assume that $M = 1$. Various estimators have been proposed which achieve root-n rate or are semiparametrically efficient for estimating the parametric component θ as well as those which achieve the usual nonparametric rates for estimating $G(\boldsymbol{x}_2)$.

To our knowledge, the only paper which addressed the problem of selecting which variables $\boldsymbol{x}_2$ enter nonlinearly in the model (1.1) was Chen and Chen (1991). That paper proposed a model-selection-type rule and showed that the probability of the correct identification by this method goes to one as the sample size goes to infinity. Härdle and Korostelev (1996) showed for the similar problem of selecting the significant variables in an additive model that the error of classification can be made exponentially small. In this paper we consider

another setup which seems to be more appealing for practical applications. Namely, we develop a nonlinear component identification method which guarantees a prescribed level of model misspecification uniformly over the class of models whose nonlinear component is separated away from the linear one by the squared distance of order $n^{-1}\log n$ or larger. Our results are essentially nonasymptotic and apply for a small or moderate sample size.

Härdle, Spokoiny and Sperlich (2001) considered a similar problem of identifying the linear component for an additive model, using a wavelet (Haar) expansion of every additive component. The advantage of the structure adaptive procedure proposed here is that the additive structure is not required and is not used in the method.

There also exists a large literature on testing a parametric, in particular linear, regression model against nonparametric alternative. See, for example, Eubank and Spiegelman (1990), Eubank and Hart (1992), Ledwina (1994), Härdle and Mammen (1993), Fan (1996), Hart (1997), Stute (1997), Horowitz and Spokoiny (2001) and references therein. Our testing results are stated in the spirit of Spokoiny (2001) focusing on the minimal separation distance between the null and the alternative providing test consistency.

The paper is organized as follows. Section 2 contains the description of the structure adaptive estimation algorithm. Accuracy of estimation by the proposed method is described in Section 3. Further problems of identification of the nonlinear component and of estimation of slope coefficients of the linear component are discussed in Section 4. Section 5 presents results of a simulation study and an application to real data. Conclusion and some extensions of the method are presented in Section 6. The proofs are collected in the Appendix.

2 Structure adaptive procedure

This section explains the adaptive estimation procedure starting with a short heuristic discussion.

2.1 Preliminaries

The idea of structural adaptation from HJPS can be summarized as follows.

(i) knowing the structural information helps better estimate the model function;

(ii) a good pilot estimator of the model function helps recover some structural information about the model.

These two observations lead to the following iterative procedure: we start with a purely nonparametric estimator of the model function; then the above two steps (estimation of the model and estimation of the structure) are iterated several times increasing the amount of structural information and improving the quality of model estimation during iteration.

HJPS considered the problem of estimation for a multi-index model in which the regression function is of the form $f(\boldsymbol{x}) = g(\theta_1^\top \boldsymbol{x}, \ldots \theta_M^\top \boldsymbol{x})$, where $\theta_1, \ldots, \theta_M$ are unknown index vectors in $I\!R^d$. The partially linear model (1.1) can be regarded as a special case of the multi-index model with $M+1$ indices. Indeed, $f(x)$ depends on x only through $\theta^\top \boldsymbol{x}_1$ and the coordinate vectors corresponding to the nonlinear component. So, one can formally apply the procedure from HJPS in the considered case. However, the special structure of the model (1.1) allows to considerably simplify the procedure and further analysis that justifies a separate treatment of the partially linear models.

Here the structure of the model (1.1) is described by the set $\mathcal{J}$ of indices corresponding to the nonlinear component $\boldsymbol{x_2}$. An alternative description can be done by using the average gradient idea. Namely, if the function $f(x)$ is linear with respect to the mth coordinate function x_m, then the partial derivative $\partial f / \partial x_m$ is a constant, and therefore, the variance V_m of the mth partial derivative can be used to measure the degree of nonlinearity of the mth coordinate. Suppose that some information about the set $\mathcal{J}$ or, equivalently, about the values V_m is available. Now we explain how this information can be used for improving the quality of estimation of the model function f. A local linear estimator of the function f and its gradient ∇f at a point X_i is given by

$$\begin{pmatrix} \widehat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} = \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top K\Big(\frac{|X_{ij}|^2}{b^2}\Big) \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K\Big(\frac{|X_{ij}|^2}{b^2}\Big)$$

where $X_{ij} = X_j - X_i$, b is a *bandwidth* and $K(\cdot)$ is a univariate kernel K on the positive semiaxis. supported on $[0,1]$. The bandwidth b should be selected in a way that the ball with the radius b and the center the point of estimation X_i contains at least $d+1$ design points which for large value of d leads to a the bandwidth b of order one and to a huge estimation bias. This phenomenon is called the "curse of dimensionality". Observe now that the function f has anisotropic smoothness properties: smoothness of G in direction of the nonlinear component, and infinite smoothness (corresponding to a linear function) in other directions. This suggests to apply an anisotropic bandwidth for estimating the model function and its gradient. So, the 'ideal' estimator which utilizes the knowledge of the set $\mathcal{J}$ can be defined by using the different bandwidths for different components of the vector x. Let $\boldsymbol{b} = \mathrm{diag}(b_1, \ldots, b_d)$ be a diagonal matrix with the diagonal entries

$b_1, \ldots, b_d$. Define the local linear estimator with the *anisotropic bandwidth* $\boldsymbol{b}$ by

$$\begin{pmatrix} \widehat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} = \left\{ \sum_{j=1}^{n} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^{\top} K_d(X_{ij}, \boldsymbol{b}) \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} K_d(X_{ij}, \boldsymbol{b}) \ , \qquad (2.1)$$

where $K_d(\boldsymbol{x}, \boldsymbol{b}) = K(|\boldsymbol{b}^{-1}\boldsymbol{x}|^2)$. Knowing that the value V_m is significantly positive (that is, m is presumably in $\mathcal{J}$) leads to the choice of smaller bandwidth values b_m for such m and possibly larger bandwidths for the other regressors. This would help to avoid the "curse of dimensionality" problem if the dimension M of the nonlinear component is not too large, cf. Carroll, Fan, Gijbels and Wand (1997) or HJPS.

Next we explain how the structural information can be extracted from the pilot estimator (2.1) of the model function. Define for every coordinate x_m of $\boldsymbol{x} \in \mathbb{R}^d$ a set of functions $\psi_{1m}, \ldots, \psi_{Lm}$ satisfying the conditions:

$$\sum_{i=1}^{n} \psi_{lm}(X_{i,m}) = 0, \qquad n^{-1} \sum_{i=1}^{n} \psi_{lm}(X_{i,m}) \psi_{l'm}(X_{i,m}) = \delta_{ll'} .$$

In other words, $\{\psi_{lm}, l = 1, \ldots, L\}$ is a orthonormal set of functions with respect to the design of mth coordinate. Each of ψ_{lm} is also orthogonal to the constant function. The latter property implies that if f is linear with respect to x_m, then

$$\beta^*_{lm} = n^{-1} \sum_{i=1}^{n} \nabla f_m(X_i) \psi_{lm}(X_{i,m}) \equiv 0 \qquad (2.2)$$

for every $l = 1, \ldots, L$, where $\nabla f_m(\boldsymbol{x}) = \partial f / \partial x_m(\boldsymbol{x})$. Thus, the sum

$$v^*_m = \sum_{l=1}^{L} (\beta^*_{lm})^2$$

can be used as the measure of nonlinearity of f with respect to x_m. Having estimated the gradient of f for all X_i, we can also estimate the coefficients β_{lm} with

$$\widehat{\beta}_{lm} = n^{-1} \sum_{i=1}^{n} \widehat{\nabla f}_m(X_i) \psi_{lm}(X_{i,m}) \qquad (2.3)$$

and use the sum $\widehat{v}_m = \widehat{\beta}^2_{1m} + \ldots + \widehat{\beta}^2_{Lm}$ as the estimated degree of nonlinearity of f with respect to the mth regression variable.

Next, the quantities $\widehat{v}_m$ can be used to define new anisotropic bandwidth $\boldsymbol{b}$ taking smaller bandwidths for the regressors with large $\widehat{v}_m$.

Remark 2.1. Similarly to HJPS, we use here the estimation method based on the Fourier expansion of the gradient $\nabla f(\boldsymbol{x})$. Alternatively, one can estimate V_m directly using the

average of $|\nabla f_m(X_i)|^2$. However, a detailed calculation (not given in the paper) shows that the procedure based on such a direct estimation of the quadratic functionals V_m leads to worse estimation results. At the same time, the loss of information from replacing V_m with v_m^* as a measure of nonlinearity is not significant if L is chosen sufficiently large, see more on the choice of L in Section 6.

2.2 Iterative procedure

We now present the description of the method. The procedure involves input parameters h_1, a_h, $\rho_1, \rho_{\min}, a_\rho$ and η. The parameter of ellipticity ρ decreases geometrically from ρ_1 to $\rho_{\min}$ by the factor $a_\rho < 1$ while the bandwidth h increases geometrically from h_1 by the factor $a_h > 1$ during iterations. The value η can be interpreted as the "memory parameter" of the procedure. The choice of these parameters, as well as of the set of basis functions $\{\psi_{lm}\}$ will be discussed in Section 2.3. The algorithm reads as follows:

1. Select h_1. Set $\widehat{v}_1^{(0)} = \ldots = \widehat{v}_d^{(0)} = 0$, and $k = 1$. Compute for $i = 1, \ldots, n$

$$\widehat{V}_i^{(0)} = \sum_{j=1}^n \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top, \qquad \widehat{S}_i^{(0)} = \sum_{j=1}^n \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} Y_j,$$

where $X_{ij} = X_j - X_i$.

2. Compute

$$b_m^{(k)} = h_k \Big(1 + \rho_k^{-2} \widehat{v}_m^{(k-1)}\Big)^{-1/2}, \qquad m = 1, \ldots, d. \tag{2.4}$$

Define $\boldsymbol{b}^{(k)} = \mathrm{diag}(b_1^{(k)}, \ldots, b_d^{(k)})$.

3. For every X_i compute

$$V_i^{(k)} = \eta V_i^{(k-1)} + (1-\eta) \sum_{j=1}^n \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top K_d\Big(X_{ij}, \boldsymbol{b}^{(k)}\Big), \tag{2.5}$$

$$S_i^{(k)} = \eta S_i^{(k-1)} + (1-\eta) \sum_{j=1}^n \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} Y_j \, K_d\Big(X_{ij}, \boldsymbol{b}^{(k)}\Big), \tag{2.6}$$

and

$$\begin{pmatrix} \widehat{f}^{(k)}(X_i) \\ \widehat{\nabla f}^{(k)}(X_i) \end{pmatrix} = \Big(V_i^{(k)}\Big)^{-1} S_i^{(k)}. \tag{2.7}$$

Compute $\widehat{\varepsilon}_i^{(k)} = Y_i - \widehat{f}^{(k)}(X_i)$.

4. Compute for $m = 1, \ldots, d$ and $l = 1, \ldots, L$

$$\widehat{\beta}_{lm}^{(k)} = n^{-1} \sum_{i=1}^n \widehat{\nabla f}_m^{(k)}(X_i) \psi_{lm}(X_{i,m}), \qquad \widehat{v}_m^{(k)} = \sum_{l=1}^L \Big|\widehat{\beta}_{lm}^{(k)}\Big|^2. \tag{2.8}$$

If $\widehat{v}_m^{(k)} > 1$, then set $\widehat{v}_m^{(k)} = 1$.

5. Increase k by 1. Set $\rho_{k+1} = a_\rho \rho_k$, $h_{k+1} = a_h h_k$. If $\rho_{k+1} \geq \rho_{\min}$, then set $k = k+1$ and continue with Step **2**; otherwise terminate.

Denote by k_n the number of iterations and by $\widehat{\beta}_{lm} = \widehat{\beta}_{lm}^{(k_n)}$ (resp. $\widehat{v}_m = \widehat{v}_m^{(k_n)}$) the last step estimators of β_m^* (resp. v_m^*). Similarly, $\widehat{f}(X_i)$ denotes the last step estimator of $f(X_i)$ and $\widehat{\varepsilon}_i = \widehat{\varepsilon}_i^{(k_n)} = Y_i - \widehat{f}(X_i)$.

Remark 2.2. At every step k of the algorithm the bandwidth $b_m^{(k)}$ for the mth regressor is selected between h_k and $h_k\rho_k$ depending on the value $\widehat{v}_m^{(k-1)}$. For the linear component, the values $\widehat{v}_m^{(k-1)}$ should be small leading to a bandwidth about h_k, while for the nonlinear regressors with a large value v_m^*, the estimator $\widehat{v}_m^{(k-1)}$ will be also large leading to a bandwidth about $h_k\rho_k$. During iteration the parameter h grows to $h_{\text{final}} \geq 1$ while $h_k\rho_k$ decreases to $\rho_{\min}$ leading to the adaptive anisotropic bandwidth at the last step.

Remark 2.3. We cut $\widehat{v}_m^{(k)}$ at one at step 4 in order to avoid too strong shrinkage in direction of mth regressor which may occur for too large values of $\widehat{v}_m^{(k)}$.

2.3 Choice of parameters

It is obvious that the quality of estimation by the proposed method strongly depends on the rule for changing the parameters h and ρ, and, in particular, on their values at the initial and final iteration. Some related discussion about this choice can be found in HJPS. The general idea is to ensure that the parameter h grows to one and $h\rho$ decreases under the constraint that at every iteration k there exist enough design points in every or almost every local ellipsoidal neighborhoods $E^{(k)}(X_i) = \left\{\boldsymbol{x} : \left|\left(\boldsymbol{b}^{(k)}\right)^{-1}(\boldsymbol{x} - X_i)\right|^2 \leq 1\right\}$.

Assuming that every $b_m^{(k)}$ is close to the 'ideal bandwidth' $b_m^{*(k)} = h_k(1 + \rho_k^{-2} v_m^*)^{-1/2}$ we observe, that neighborhood $E^{(k)}(X_i)$ is stretched at each iteration step by factor a_h in all directions and is shrunk by a factor about a_ρ in directions of the M-dimensional nonlinear component $\mathcal{J}$ where a_h and a_ρ are parameters of the procedure. Therefore, the Lebesgue measure of every such neighborhood is changed each time by a factor about $a_h^d a_\rho^M$. This leads to the constraint $a_h^d a_\rho^M \geq 1$, cf. Assumption 3 in Section 3 below. Under the assumption of a random design with a positive density, this would result in an increase of the mean number of design points inside each $E^{(k)}(X_i)$. Our theoretical results will be stated for the choice $h_1 \asymp n^{-1/\max\{4,d\}}$, $h_{\max} \asymp 1$, $\rho_1 = 1$, $\rho_{\min} \asymp \left(n^{-1}\log n\right)^{1/3}$, see Section 3 for more details. Similarly to HJPS, such a choice under the constraint $a_h^d a_\rho^M > 1$ is possible only for $M \leq 3$.

We recommend to define for every $m = 1, \ldots, d$ the set of functions ψ_{lm}, $l = 1, \ldots, L$

by orthogonalizing the set of polynomials $x_m, x_m^2, \dots x_m^L$ with respect to the design of the m th regressor under the constraint $\sum_{i=1}^n \psi_{lm}(X_{i,m}) = 0$. A model or variable dependent choice of the basis $\{\psi_{lm}\}$ is possible as well. The "memory parameter" η used in (2.5) and (2.6) can be taken between 0.1 and 0.5 . The number L can be taken between 5 and 10, see Section 6 for more discussion.

Remark 2.4. Similarly to HJS and HJPS we apply in our numerical study a slightly modified procedure. The only difference is in the definition of the estimated vectors $\widehat{\beta}_{lm}$. Namely we define

$$\widehat{\beta}_{lm}^{(k)} = \left(\sum_{i=1}^n w_i^{(k)} \right)^{-1} \sum_{i=1}^n w_i^{(k)} \widehat{\nabla f}_m^{(k)}(X_i) \psi_{lm}(X_{i,m}),$$

where $w_i^{(k)}$ is square root of the smallest eigenvalue of the matrix $V_i^{(k)}$, that is, $w_i^{(k)} = \lambda_{\min}^{1/2}(V_i^{(k)})$. In addition, the basis functions ψ_{lm} should be modified as each step to satisfy the condition $\sum_{i=1}^n w_i^{(k)} \psi_{lm}(X_i) = 0$.

2.4 Estimation of the noise variance

The variance σ^2 of the noise ε does not enter in the description of the method. However, it will be used for defining the stopping rule of the algorithm and the resampling procedure in Section 4. Here we briefly discuss how this variance can be estimated under the assumption of the noise homogeneity at every step of the algorithm.

A natural variance estimator can be defined on the base of residuals squared after each the step k : $\left|\widetilde{\sigma}^{(k)}\right|^2 = n^{-1} \sum_{i=1}^n \left|\widehat{\varepsilon}_i^{(k)}\right|^2$. This simple crude estimator can be refined, see e.g. Gasser, Sroka and Jennen-Steinmetz (1986) or Spokoiny (2002) and reference therein. Namely, the residuals $\widehat{\varepsilon}_i^{(k)}$ can obviously be represented in the form $\widehat{\varepsilon}_i^{(k)} = \sum_{j=1}^n c_{ij}^{(k)} Y_j$ where $c_{ij}^{(k)}$ are known coefficients. These coefficients are random and dependent on the Y_i 's through the random bandwidths $\widehat{b}_m^{(k)}$. However, our theoretical results indicate that one can ignore this dependence and proceed as if the coefficients $c_{ij}^{(k)}$ were deterministic and correspond to the "ideal" bandwidths $b_m^{*(k)}$.

Next, if the function f is sufficiently smooth, then the distribution of the residuals $\widehat{\varepsilon}_i$ only weakly depends on this function f and can be effectively evaluated for $f \equiv 0$. In the last case, $\boldsymbol{E}\left|\widehat{\varepsilon}_i^{(k)}\right|^2 = \sigma^2 \sum_{j=1}^n \left|c_{ij}^{(k)}\right|^2$ that leads to the estimator

$$\left|\widehat{\sigma}^{(k)}\right|^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^n \left|c_{ij}^{(k)}\right|^2 \right)^{-1} \left|\widehat{\varepsilon}_i^{(k)}\right|^2. \tag{2.9}$$

The properties of this estimator are briefly discussed in Section 3.4 below.

3 Accuracy of estimation

In this section we present the results about the accuracy of estimation of the functionals β^*_{lm} and v^*_m by the proposed iterative procedure.

3.1 Assumptions

As in HJPS, it is useful to proceed with the renormalized link function. In what follows we consider the model

$$f(\boldsymbol{x}) = \boldsymbol{x}_1^\top \theta + g(\mathcal{R}^* \boldsymbol{x}_2) \tag{3.1}$$

where $\mathcal{R}^*$ is the diagonal $M \times M$ matrix with diagonal entries $\sqrt{v^*_m}$, $m \in \mathcal{J}$ and g is a nonlinear link function.

Our main results will be stated under the following assumptions.

Assumption 1. (Kernel) The kernel $K(\cdot)$ is continuously differentiable decreasing function on $\mathbb{R}_+$ with $K(0) = 1$ and $K(x) = 0$ for all $|x| \geq 1$.

Assumption 2. (Errors) The random variables ε_i in (1.1) are independent and normally distributed with zero mean and variance σ^2.

Assumption 3. (Range of parameters h_k, ρ_k) The parameters of the procedure fulfill $\rho_1 = 1$, $\rho_{\min} = (\sigma^2 n^{-1} L \log n)^{1/3}$, $h_1 = C_0 n^{-\frac{1}{4\vee d}}$ with a constant $C_0 \geq 1$, $h_{\max} \geq 1$ and $a_h^d a_\rho^M \geq 1$.

Assumption 4. (Link function) The function g from (3.1) is twice differentiable with a bounded second derivative, so that, for some constant C_g and for all $u, v \in \mathbb{R}^M$

$$|g(v) - g(u) - (v - u)\,\nabla g(u)| \leq C_g\,|u - v|^2;$$

Our last assumption concerns the design properties. In what follows we assume that the design is deterministic. That is, $X_1, \ldots, X_n$ are non-random points in $\mathbb{R}^d$. Note, however, that the case of a random design can be considered as well, supposing $X_1, \ldots, X_n$ independent and identically distributed random points in $\mathbb{R}^d$ with a design density $p(x)$. Then all the results should be understood conditionally on the design.

In order for the procedure to work, we have to suppose that the design points (X_i) are "well diffusedänd, as a consequence, at k th iteration of the algorithm, all local gradient estimators from (2.7) corresponding to the anisotropic bandwidth $\boldsymbol{b}^{(k)} = \mathrm{diag}(b_1^{(k)}, \ldots, b_d^{(k)})$ from (2.4) are well defined. The latter is equivalent to the condition that all the matrices $V_i^{(k)}$ from (2.5) are non-singular. We also define for the k th iteration the "ideal

anisotropic bandwidth" $\boldsymbol{b}^{*(k)}$ having the diagonal entries $b_m^{*(k)} = \left(1+\rho_k^{-2}v_m^*\right)^{-1/2} h_k$. The closeness of $\boldsymbol{b}^{(k)}$ to the "ideal bandwidth" $\boldsymbol{b}^{*(k)}$ can be characterized by the values $U_m^{(k)} = \left(b_m^{(k)}/b_m^{*(k)}\right)^2 = \left(1+\rho_k^{-2}v_m^*\right)/\left(1+\rho_k^{-2}\widehat{v}_m^{(k-1)}\right)$, $m=1,\ldots,d$. If $\widehat{v}_m^{(k-1)} = v_m^*$, then $U_m^{(k)} = 1$. The condition we need means that at the step k of the algorithm, for every anisotropic bandwidth $\boldsymbol{b} = \mathrm{diag}(b_1,\ldots,b_d)$ close to $\boldsymbol{b}^{*(k)}$ in the above sense, the design is regular within the elliptic neighborhood with the center at each point X_i and with the principal semiaxis b_m, $m=1,\ldots,d$.

Define $Z_{ij}^{(k)} = \left(\boldsymbol{b}^{*(k)}\right)^{-1}(X_j - X_i)$ for $i,j=1,\ldots,n$. These vectors describe locations of the design points in the coordinate system shifted by X_i and rescaled by $\boldsymbol{b}^{*(k)}$. For a vector $U=(U_1,\ldots,U_d)^\top \in \mathbb{R}^d$ with $U_m \geq 0$, define $D_U = \mathrm{diag}(U_1,\ldots,U_d)$. Then, for $\boldsymbol{b} = D_U^{-1/2}\boldsymbol{b}^{*(k)}$, it holds $K_d(X_{ij},\boldsymbol{b}) = K\left(\left(Z_{ij}^{(k)}\right)^\top D_U Z_{ij}^{(k)}\right)$. Set

$$
\begin{aligned}
N_i^{(k)}(U) &= \sum_{j=1}^n K\Big(\left(Z_{ij}^{(k)}\right)^\top D_U Z_{ij}^{(k)}\Big), \qquad i=1,\ldots,n,\\
\mathcal{V}_i^{(k)}(U) &= \sum_{j=1}^n \begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}^{(k)}\end{pmatrix}^\top K\Big(\left(Z_{ij}^{(k)}\right)^\top D_U Z_{ij}^{(k)}\Big), \qquad i=1,\ldots,n.
\end{aligned}
$$

In what follows $\|A\|$ stands for the matrix norm associated with the Euclidean vector norm: $\|A\| = \sup_\lambda |A\lambda|/|\lambda|$.

Assumption 5. (Design) There exist constants C_V, C_K, $C_{K'}$ and some $\alpha \in (0,1/2)$, such that for all vectors $U \in \mathbb{R}^d$ satisfying $|U_m - 1| \leq \alpha$, $m=1,\ldots,d$, and for $k \leq k_n$ the inverse matrices $\mathcal{V}_i^{(k)}(U)^{-1}$ are well defined with

$$
N_i^{(k)}(U)\left\|\mathcal{V}_i^{(k)}(U)^{-1}\right\| \leq C_V, \qquad i=1,\ldots,n;
$$

Moreover, for $i,j=1,\ldots,n$,

$$
\begin{aligned}
\sum_{i=1}^n \frac{1}{N_i^{(k)}(U)} K\Big(\left(Z_{ij}^{(k)}\right)^\top D_U Z_{ij}^{(k)}\Big) &\leq C_K,\\
\sum_{i=1}^n \frac{1}{N_i^{(k)}(U)} \left|K'\big(\left(Z_{ij}^{(k)}\right)^\top D_U Z_{ij}^{(k)}\big)\right| &\leq C_{K'},\\
\sum_{j=1}^n \frac{1}{N_i^{(k)}(U)} \left|K'\big(\left(Z_{ij}^{(k)}\right)^\top D_U Z_{ij}^{(k)}\big)\right| &\leq C_{K'}.
\end{aligned}
$$

where K' means the derivative of K.

Remark 3.1. As already mentioned in HJS and HJPS, in the case of random design with a continuous positive density one can fix some constants C_V, C_K and $C_{K'}$ (which depend on the dimension d and the design distribution) such that the bounds in Assumption 5 hold with probability which converges to 1 exponentially fast as n grows to infinity.

3.2 The first step estimator

Our first result describes the quality of the estimators $\widehat{\beta}_{lm}^{(1)}$ obtained at the first step of the algorithm. These estimators correspond to usual nonparametric local linear estimation of the function f and its gradient. We also state the result about the accuracy of estimation of the values v_m^* at the first step.

Let β_m^* denote an L-vector with the components β_{lm}^*, $l = 1, \ldots, L$.

Proposition 3.1. *Let Assumptions 1 through 5 hold. For the first-iteration estimator $\widehat{\beta}_m^{(1)}$ of the vector β_m^*, $m = 1, \ldots, d$, it holds:*

$$\widehat{\beta}_m^{(1)} - \beta_m^* = s_m h_1 + \frac{\eta_m}{h_1\sqrt{n}},$$

where deterministic L-vectors s_m satisfy $|s_m| \le C_g C_V^{1/2} v_{(1)}^*$ with $v_{(1)}^* = \max_{m=1,\ldots,d} v_m^*$, and the η_m's are mean zero Gaussian random L-vectors with components η_{lm} such that $\boldsymbol{E}|\eta_{lm}|^2 \le 2\sigma^2 C_V^2 C_K$, $l = 1, \ldots, L$. Also, it holds

$$\boldsymbol{P}\left(\max_{m=1,\ldots,d}\left|\widehat{\beta}_m^{(1)} - \beta_m^*\right| > \delta_1\right) \le \frac{1}{n}, \tag{3.2}$$

where

$$\delta_1 = C_g\, C_V^{1/2}\, v_{(1)}^*\, h_1 + \frac{\sqrt{2L}\sigma C_V C_K^{1/2} z_n}{h_1\sqrt{n}}, \tag{3.3}$$

and $z_n = (1 + 2\log(nd) + 2\log\log(nd))^{1/2}$.

Moreover, for the first-iteration estimator $\widehat{v}_m^{(1)}$, $m = 1, \ldots, d$, it holds:

$$\boldsymbol{P}\left(\left|\widehat{v}_m^{(1)} - v_m^*\right| \le \delta_1^2 + 2\delta_1\tau_{m,1},\ \forall m = 1, \ldots, d\right) \ge 1 - \frac{1}{n}, \tag{3.4}$$

where $\tau_{m,1} = \sqrt{v_m^*}(1 + v_m^*)^{-1/2} \le \min\{1, \sqrt{v_m^*}\}$.

3.3 The quality of the final estimators

Now we present the result which indicates how the accuracy of estimation can be improved by the iterative algorithm. As in HJS and HJPS, the quality of the final estimators depends on the 'direction'. This quality is of order $n^{-1/2}$ for the linear component and is worse for the nonlinear component. This fact has a very simple explanation: estimation of a nonlinear component is a harder task than that of a linear one; hence, the worse accuracy. To express this fact, we introduce the scaling factor $P_{\rho,m} = (1 + \rho^{-2} v_m^*)^{-1/2}$, where ρ is a running parameter of the procedure. Note that $P_{\rho,m} = 1$ for all linear regressors which have $v_m^* = 0$. If v_m^* is a positive constant, then $P_{\rho,m} \asymp \rho$. We will see that the estimation

error $\widehat{\beta}_{lm} - \beta^*_{lm}$, after being multiplied by $P_{\rho,m}$, can be bounded uniformly over l, m at every step of the algorithm. This implies, in particular, that the quality of estimation of the nonlinear component is about $P_\rho^{-1} \asymp \rho^{-1}$ times worse than the quality for the linear one.

In the next theorem and in Theorem 4.1 below, ρ (resp. h) denotes ρ_{k_n} (resp. h_{k_n}) at the last iteration. Recall that h, ρ satisfy conditions $h \geq 1$ and $\rho = \left(\sigma^2 n^{-1} L \log n\right)^{1/3}$.

Theorem 3.1. *Let Assumptions 1 through 5 hold. Then there exist a random set A with $\boldsymbol{P}(A) \geq 1 - 3k_n/n$ and, for every $m = 1, \ldots, d$, a Gaussian zero mean random vector $\xi^*_m = (\xi^*_{1m}, \ldots, \xi^*_{Lm})^\top \in I\!R^L$ defined as a linear combination of the errors ε_i with deterministic coefficients, which depend on the "ideal" bandwidth $\boldsymbol{b}^* = \boldsymbol{b}^{*(k_n)}$, the design $X_1, \ldots, X_n$, basis functions $\psi_{lm}(\cdot)$, and the kernel K only, and such that*

$$\boldsymbol{E}|\xi^*_{lm}|^2 \leq 2\sigma^2 C_V^2 C_K \qquad l = 1, \ldots, L,\ m = 1, \ldots, d,$$

and on A it holds

$$\begin{aligned} &\max_{m=1,\ldots,d} \left| P_{\rho,m}(\widehat{\beta}_m - \beta^*_m) - n^{-1/2}\xi^*_m \right| \leq C\left(\sigma^2 n^{-1} L \log n\right)^{2/3}, \\ &\max_{m=1,\ldots,d} \left| P_{\rho,m}(\widehat{\beta}_m - \beta^*_m) \right| \leq \delta_n, \\ &\max_{m=1,\ldots,d} |P^2_{\rho,m}(\widehat{v}_m - v^*_m)| \leq \delta_n^2 + 2\delta_n \tau_m, \end{aligned} \tag{3.5}$$

where $C = C(d, M, C_g, C_V, C_K, C_{K'}, \overline{\psi})$, $\overline{\psi} = \max_{i,l,m} |\psi_{lm}(X_i)|$,

$$\delta_n = \sqrt{2C_V^2 C_K \sigma^2 n^{-1} L z_n^2} + C\left(\sigma^2 n^{-1} L \log n\right)^{2/3} \tag{3.6}$$

and $\tau_m = \rho\sqrt{v^*_m}\left(\rho^2 + v^*_m\right)^{-1/2} \leq \min\{\rho, \sqrt{v^*_m}\}$. This implies that on A for every $m \notin \mathcal{J}$, with $\omega_n = C\left(\sigma^2 n^{-1} L \log n\right)^{2/3}$:

$$|\widehat{\beta}_m - n^{-1/2}\xi^*_m| \leq \omega_n, \qquad |\widehat{\beta}_m| \leq \delta_n, \qquad \left||\widehat{\beta}_m|^2 - n^{-1}|\xi^*_m|^2\right| \leq \omega_n^2 + 2\omega_n\delta_n. \tag{3.7}$$

Remark 3.2. The meaning of the random set A appearing in Theorem 3.1 can be understood as follows. The result of every iteration of the algorithm is random. With some probability it may happen that the estimation result at some step of the procedure does not follow the model structure. For instance, with some probability, $\widehat{v}_m$ can be large even if $v^*_m = 0$. Our results indicate that the overall probability of such events is rather small and their complement is precisely the set A (of a dominating probability) on which the procedure 'works', that is, the iterative procedure leads to improvement of the quality of estimation at every iteration. The other results of Theorem 3.1 claim that on the set A, the adaptive estimators $\widehat{\beta}_{lm}$ behave essentially as the 'ideal' estimators $\widehat{\beta}^*_{lm}$ corresponding to the 'ideal' bandwidth $\boldsymbol{b}^*$. Since our further analysis is based on the final step estimators $\widehat{\beta}_{lm}$, all our results that follow will also be stated conditionally on this set A.

Remark 3.3. (Origin of the constraint $M \leq 3$) It follows from the proof of Theorem 3.1 that the bias of the 'ideal' estimators $\widehat{\beta}^*_{lm}$ based on the local linear smoothing with the 'ideal' bandwidth $\boldsymbol{b}^* = \boldsymbol{b}^{*(k_n)}$ is of the order $(n^{-1}\log n)^{-2/3}$ only if the dimensionality M of the nonlinear component does not exceed 3. For $M \geq 4$, the model dependent bias of estimation is of order $n^{-1/2}$ or larger while the stochastic component (which is model free) is of order $n^{-1/2}$. The same applies for the adaptive estimators $\widehat{\beta}_{lm}$. Therefore, the leading term in the estimation loss is model free only for $M \leq 3$, and the estimators $\widehat{\beta}_{lm}$ do not achieve asymptotic normality at root-n rate for $M \geq 4$.

3.4 Variance estimation

The algorithm delivers an estimator $\widehat{\sigma}^2$, see (2.9), of the error variance σ^2. This estimator also utilizes the estimated structural information and improves upon the purely nonparametric variance estimators. Spokoiny (2002) has shown that in a general high dimensional regression model with $d > 8$, a root-n consistent estimation of the variance σ^2 is impossible. Here the use of the structural assumption allows to relax this condition and to get a root-n accuracy for any d.

Theorem 3.2. *Let Assumptions 1 through 5 hold. There exists a constant C_σ, which depends on the constants entering in these assumptions only, such that*

$$\boldsymbol{P}\left(\sqrt{n}\left|\widehat{\sigma}^2 - \sigma^2\right| > C_\sigma \sigma^2 \lambda\right) \leq 2e^{-\lambda^2/4} + 3k_n/n.$$

4 Inference in a partially linear model

This section explains how the model (1.1) can be explored using our iterative procedure and results of Section 3. First we state the important separation result that will be used in the analysis below.

Let some integer $\mathcal{M}$ be fixed. We put the estimated values $\widehat{v}_m$ in the decreasing order $\widehat{v}_{(1)} \geq \widehat{v}_{(2)} \geq \widehat{v}_{(3)} \ldots$ and denote by $\widehat{\mathcal{J}}_{\mathcal{M}}$ the index set corresponding to the $\mathcal{M}$ largest values $\widehat{v}_m$. Theorem 3.1 implies the following separation result.

Theorem 4.1. *Let $u_n = \delta_n/\rho < \sqrt{2} - 1$ with $\rho = \rho_{k_n}$ and δ_n from (3.6). Let r be some number satisfying $r \geq 1$. If $v^*_m > (r s_r \delta_n)^2$ for all $m \in \mathcal{J}$ where*

$$s_r = \frac{1 + \sqrt{1 + (r^2+1)(1 - u_n^2 - 2u_n)}}{1 - u_n^2 - 2u_n},$$

then it holds on the random set A defined in Theorem 3.1 $\widehat{v}_m > r^2\delta_n^2$ for $m \in \mathcal{J}$ and $\widehat{v}_m \leq \delta_n^2$ for $m \notin \mathcal{J}$ and thus, $\mathcal{J} \subseteq \widehat{\mathcal{J}}_{\mathcal{M}}$ for all $\mathcal{M} \geq M$.

Remark 4.1. The result of Theorem 4.1 applied with $r = 1$ yields the sufficient separation condition: if $v_m^* > (s_1\delta_n)^2$, then, with a high probability, $\widehat{v}_m > \delta_n^2$ for $m \in \mathcal{J}$ and $\widehat{v}_m \leq \delta_n^2$ for $m \notin \mathcal{J}$, and therefore $\widehat{\mathcal{J}}_M = \mathcal{J}$. For application of this result to the resampling scheme below in this section, we introduced the factor $r \geq 1$, which ensures a qualified separation between linear and nonlinear component.

The value $u_n = \delta_n/\rho$ is small at least if n is sufficiently large. Hence, s_r defined in Theorem 4.1 is bounded by a constant depending on r only and therefore, the threshold $t^* = (rs_r\delta_n)^2$, providing with a high probability a correct separation between linear and nonlinear components is of order $\delta_n^2 \approx (n^{-1}\log n)$. It can be easily seen that the separation with the prescribed level of the identification error is impossible if the separation distance square is smaller in order than n^{-1}. Therefore, the procedure provides a near optimal rate of separation within a log-factor.

4.1 Testing the hypothesis about M

Here we discuss the problem how the estimators $\widehat{v}_m$ of v_m^* can be used for selecting the nonlinear component and for testing the hypothesis that the dimensionality M of the nonlinear component does not exceed the prescribed value $\mathcal{M}$. As special cases, for $\mathcal{M} = 0$ we get the hypothesis that the original model is linear, and for $\mathcal{M} = 1$, the hypothesis that the nonlinear component is univariate. Then the effective nonlinear dimension of the model can be estimated by the minimal $\mathcal{M}$ such that the hypothesis $M \leq \mathcal{M}$ is not rejected.

The idea of the method is very simple: reject $H_{\mathcal{M}} : M \leq \mathcal{M}$ if the value $\widehat{v}_{(\mathcal{M}+1)}$ is significantly positive. To formalize the procedure, we have to specify, for a given α, the critical value t_α such that the test has the significance level about α. Suppose that the true model satisfies $M \leq \mathcal{M}$ and that the values v_m^* for all $m \in \mathcal{J}$ exceed the value $t^* = (rs_r\delta_n)^2$ for some $r \geq 1$. Then Theorems 3.1 and 4.1 imply that

- under the null hypothesis $M \leq \mathcal{M}$, the index $(\mathcal{M}+1)$ corresponds with a high probability to a linear component;
- for $m \notin \mathcal{J}$, the distributions of the $\widehat{\beta}_{lm}$'s and of $\widehat{v}_m$ only weakly depend on the model function f, see Remark 3.2;
- for every $m \in \mathcal{J}$, if v_m^* is separated from zero by distance of order δ_n^2, then the same is true with a high probability for the estimator $\widehat{v}_m$.

These observations suggest to apply the resampling scheme that mimics only the distri-

bution of the values $\widehat{v}_{(1)}, \ldots, \widehat{v}_{(\mathcal{M}+1)}$. More precisely, we construct an artificial regression function $\widetilde{f}_{\mathcal{M}}$ that has exactly $\mathcal{M}$-dimensional nonlinear component corresponding to $m \in \widehat{\mathcal{J}}_{\mathcal{M}}$ and such that all the functionals of type β^*_{lm} constructed for this function $\widetilde{f}_{\mathcal{M}}$ coincide with the $\widehat{\beta}_{lm}$'s, that is,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \widetilde{f}_{\mathcal{M}}(X_i)}{\partial x_m} \psi_{lm}(X_{i,m}) = \begin{cases} \widehat{\beta}_{lm} & \text{if } m \in \widehat{\mathcal{J}}_{\mathcal{M}}, \\ 0 & \text{otherwise,} \end{cases} \qquad l = 1, \ldots, L.$$

The function $\widetilde{f}_{\mathcal{M}}$ can be selected as a linear combination of the functions x_m^l/l for $l = 2, \ldots, L+1$:

$$\widetilde{f}_{\mathcal{M}}(\boldsymbol{x}) = \sum_{m \in \widehat{\mathcal{J}}_{\mathcal{M}}} \sum_{l=2}^{L+1} c_{lm} x_m^l / l, \tag{4.1}$$

where the coefficients c_{lm} must fulfill

$$\Psi_m c_m = \widehat{\beta}_m\,, \qquad m \in \widehat{\mathcal{J}}_{\mathcal{M}}\,. \tag{4.2}$$

Here c_m (resp. $\widehat{\beta}_m$) denotes the vector in $\mathbb{R}^L$ with the elements c_{lm} (resp. $\widehat{\beta}_{lm}$) and Ψ_m is the $L \times L$ matrix whose elements are the sums

$$\Psi_{m,l'l} = \frac{1}{n}\sum_{i=1}^{n} X_{i,m}^l \psi_{l'm}(X_{i,m}), \qquad l, l' = 1, \ldots, L.$$

We resample from the model

$$\widetilde{Y}_i = \widetilde{f}_{\mathcal{M}}(X_i) + \widehat{\sigma}_i \widetilde{\varepsilon}_i,$$

where $\widetilde{\varepsilon}_i$ are i.i.d. standard normal. The variances $\widehat{\sigma}_i^2$ either fulfill $\widehat{\sigma}_i^2 = \widehat{\sigma}^2$ for the variance estimator $\widehat{\sigma}^2$ from Section 2.4, or they simply are defined by $\widehat{\sigma}_i^2 = \widehat{\varepsilon}_i^2$. The first proposal suits well the case of a homogeneous noise, and the second one is similar to the wild bootstrap idea and should be used if the assumption of noise homogeneity is questionable.

The recommended estimator of the critical value can be computed by using the following simulation procedure:

1. For each $i = 1, \ldots, n$, generate $\widetilde{Y}_i = \widetilde{f}_{\mathcal{M}}(X_i) + \widehat{\sigma}_i \widetilde{\varepsilon}_i$, where $\widetilde{\varepsilon}_i$ is sampled randomly from the standard normal law.

2. Use the data set $\{\widetilde{Y}_i, X_i : i = 1, \ldots, n\}$ to estimate gradient projections β^*_{lm} with estimator (2.3) based on gradient estimator (2.1) with the last step bandwidth $\boldsymbol{b} = \boldsymbol{b}^{(k_n)}$. Denote the resulting estimator by $\widetilde{\beta}_{lm}$. Compute $\widetilde{v}_m = \sum_{l=1}^{L} |\widetilde{\beta}_{lm}|^2$ for every $m = 1, \ldots, d$ and the statistic $\widetilde{T}_{\mathcal{M}}$, that is $\widetilde{v}_{(\mathcal{M}+1)}$.

3. Define t_α as the $(1-\alpha)$-quantile of the empirical distribution of $\widetilde{T}_{\mathcal{M}}$ that is obtained by repeating steps 1-2 many times.

Theorem 4.2. *Let Assumptions 1–5 hold and* $\min_{m\in\mathcal{J}} v_m^* \geq (rs_r\delta_n)^2$ *with* $r = s_1$. *If* $\mathcal{M} = M$, *then*

$$\boldsymbol{P}(H_{\mathcal{M}} \text{ is rejected}) \quad \leq \quad \alpha + 3(k_n+1)/n.$$

4.2 Identification of the nonlinear component

Here we describe how the effective nonlinear dimension M and the index set $\mathcal{J}$ corresponding to the nonlinear component can be estimated using the above testing procedure. Let some positive $\alpha < 1$ be fixed. Starting with $\mathcal{M} = 0$, we consider the model with $\mathcal{M}$-dimensional nonlinear component due to (4.1) and (4.2) and test the hypothesis $M \leq \mathcal{M}$ at the level α as described in Section 4.1. Terminate if the hypothesis $M \leq \mathcal{M}$ is not rejected, otherwise increase $\mathcal{M}$ by one. Finally we set $\widehat{\mathcal{M}} =$ "the first nonrejected $\mathcal{M}$" and $\widehat{\mathcal{J}} = \widehat{\mathcal{J}}_{\widehat{\mathcal{M}}}$.

Theorem 4.3. *Let Assumptions 1 through 5 hold and* $\min_{m\in\mathcal{J}} v_m^* \geq (rs_r\delta_n)^2$ *with* $r = s_1$. *Then*

$$\boldsymbol{P}(\widehat{\mathcal{J}} \neq \mathcal{J}) \quad \leq \quad \alpha + 3(k_n+M)/n.$$

Remark 4.2. It can be easily checked that the results of Theorems 4.2 and 4.3 continue to hold even if the test level α depends on n and goes to zero as n grows. In particular, one can take $\alpha = n^{-a}$ with $a < 1/2$. With such a choice, our method leads to a consistent estimation of the set $\mathcal{J}$.

4.3 Estimation and inference for the linear component

The method described above allows to classify the regressors into linear and nonlinear. Moreover, the result of classification is correct with a dominating probability provided the sample size is large enough. The impact of every linear regression variable in the model function is characterized by the corresponding slope coefficient θ_m from (1.1). Here we discuss how these slope coefficients can be estimated. We use again the observation that $\partial f/\partial x_m \equiv \theta_m$ for every $m \notin \mathcal{J}$. Therefore, the sum

$$\widehat{\theta}_m = \frac{1}{n}\sum_{i=1}^{n} \widehat{\nabla f}_m(X_i) \tag{4.3}$$

is a reasonable estimator of θ_m. Here $\widehat{\nabla f}(X_i)$ is the gradient estimator obtained at the last step of the algorithm. Our next result claims that $\widehat{\theta}_m$ from (4.3) estimates the true value θ_m with the root-n accuracy and that it can be very well approximated by a Gaussian random variable. This result can be viewed as an application of Theorem 3.1 for $m \notin \mathcal{J}$ and $\psi_{lm} \equiv 1$.

Theorem 4.4. *Let Assumptions 1 through 5 hold. Then for every $m \notin \mathcal{J}$, there exists a Gaussian zero mean random variable γ_m^* which is defined as a linear combination of the errors ε_i with deterministic coefficients, depending on the "ideal" bandwidth $\boldsymbol{b}^* = \boldsymbol{b}^{*(k_n)}$, the design $X_1, \ldots, X_n$, the basis functions $\psi_{lm}(\cdot)$ and the kernel K only, and such that*

$$\boldsymbol{E}|\gamma_m^*|^2 \le 2\sigma^2 C_V^2 C_K,$$

and on the random set A from Theorem 3.1 with $\boldsymbol{P}(A) \ge 1 - 3k_n/n$, it holds

$$\max_{m \notin \mathcal{J}} \left| \widehat{\theta}_m - \theta_m - n^{-1/2}\gamma_m^* \right| \le C_1 \left(\sigma^2 n^{-1} L \log n\right)^{2/3},$$

where $C_1 = C_1(d, M, C_g, C_V, C_K, C_{K'}, \overline{\psi})$.

Remark 4.3. The above estimator $\widehat{\theta}_m$ can be slightly refined by explicitly using the estimated structural information about the model. Namely, an application of the anisotropic bandwidth $\widehat{\boldsymbol{b}} = \mathrm{diag}(\widehat{b}_1, \ldots, \widehat{b}_d)$ with $\widehat{b}_m$ from the last iteration for $m \in \widehat{\mathcal{J}}$ and $\widehat{b}_m = \infty$ for $m \notin \widehat{\mathcal{J}}$ leads under condition of the correct identification to the classical partially linear estimator for the case with known $\mathcal{J}$, see e.g. Härdle, Liang and Gao (1999).

Remark 4.4. (Selecting significant regressors) The procedure in Section 4.2 can be also used for identifying the significant components. All the regressors entering in the nonlinear component are automatically significant. The linear regressors can be further analyzed for significance. Theorem 4.4 claims that the normalized estimation error $\sqrt{n}\left(\widehat{\theta}_m - \theta_m\right)$ is asymptotically normal. Moreover, the asymptotic variance of $\widehat{\theta}_m$ can be easily estimated. Indeed, $\widehat{\theta}_m$ is a linear combination of the observations Y_i with the known coefficients c_{im}, that is, $\widehat{\theta}_m = \sum_i c_{im} Y_m$. Then $\widehat{\sigma}_m^2 = \sigma^2 \sum_i c_{im}^2$ is an estimator of $\mathrm{Var}(\widehat{\theta}_m)$. The search of significant regressors can be done by the rule $\left|\widehat{\theta}_m\right|^2 \ge \lambda^2 \widehat{\sigma}_m^2$ for some $\lambda > 0$, see illustration of this procedure in Section 5. We skip the further discussion for the reasons of space.

5 Simulated and real data results

In this section we illustrate the performance of the proposed method on some simulated examples and give a real data application. With the simulated examples we aim to illustrate

how the performance of the proposed method depends on the sample size n, dimension d of the model, the dimensionality of the nonlinear component M and the noise variance σ^2. We especially focus on the component classification results: identification of the nonlinear variables and selection of the significant variables. We also demonstrate how the quality of estimation of the nonlinear components improves during iteration.

In our simulation study we apply the modified procedure (see Remark 2.4) with the following parameter setting:

$$\rho_1 = 1, \qquad \rho_{\min} = n^{-1/3}, \qquad a_\rho = e^{-1/6}, \qquad \eta = 0.2, \qquad a_h = a_\rho^{-1/2}.$$

The initial bandwidth h_1 is selected from the condition $\#\{i : M_h(X_i) \geq d+1\} \geq n/2$, where $M_h(x)$ stand for the number of the design points X_i in the ball of radius h and center x. This condition ensures that for at least a half of the design points the local gradient estimator is well defined. This setting leads to the number of iterations $k(n) \approx \frac{\log(\rho_1/\rho_{\min})}{\log a_\rho} \approx 2 \log n$.

The procedure utilizes the kernel $K(|x|^2) = (1 - |x|^2)_+^2$. For every $m \leq d$, the basis system $\{\psi_{1m}(x_m), \ldots, \psi_{Lm}(x_m)\}$ is obtained using polynomials of x_m of degree from one to L satisfying the conditions $\sum_{i=1}^n w_i \psi_{lm}(X_{i,m})\psi_{l'm}(X_{i,m}) / \sum_{i=1}^n w_i = \delta_{ll'}$ and $\sum_{i=1}^n \psi_{lm}(X_{i,m}) w_i = 0$ where $w_i = w_i^{(k)} = \lambda_{\min}^{1/2}(V_i^{(k)})$ for k th iteration with $k \geq 1$. We apply $L = 6$.

In our simulation study we consider the model $Y_i = \theta^\top X_i + g(X_{i,d-M+1}, \ldots, X_{i,d}) + \varepsilon_i$ for M between 1 and 3. The vector θ is taken of the form $\theta = (1, 2, 3, 4, 0, \ldots, 0)^\top$. The link function g is $g(u) = g_1(u) = e^u + e^{-u}$ for $M = 1$, $g(u_1, u_2) = g_1(u_1)g_1(u_2)$ for $M = 2$ and $g(u_1, u_2, u_3) = g_1(u_1)g_1(u_2)g_1(u_3)$ for $M = 3$. The dimension d is taken $4 + M$ or larger. The errors ε_i are i.i.d. normal with parameters $(0, \sigma^2)$ for $\sigma^2 = 0.1$. The design $X_1, \ldots, X_n$ is modelled randomly so that each X_i follows $\mathrm{Norm}(0.2, 0.8^2)$-distribution, restricted to the $[-1, 1]^d$-cube. The experiments were done for sample size $n = 100, 200, 400$. The results displaying the quality of estimation by the iterative algorithm are summarized in Tables 1 for $M = 1$ and in Table 2 for $M = 2$. We display the mean losses $|\widehat{v}_m|$ for one linear regressor and $|\widehat{v}_m^{(k)} - v_m^*|$ for nonlinear regressors where $v_m^* = |\beta_m^*|^2$ and β_m^* is the vector with the components $\beta_{lm}^* = \sum_{i=1}^n w_i^{(k)} \nabla f_m(X_{i,m}) \psi_{lm}(X_{i,m}) / \sum_{i=1}^n w_i^{(k)}$.

It is interesting to observe that the quality of estimating the linear regressor x_1 improves with growing dimension d.

In Table 2 we demonstrate in addition how the error of estimation depends on the noise variance σ^2. One can see that the estimation risk for the nonlinear components only

Table 1: Case $M = 1$: mean loss $|\widehat{v}_m - v_m^*|$ for the nonlinear regressor for the first, second, fourth, and final iteration and final losses $|\widehat{v}_1|$ for the first linear regressor. Results are obtained from $N = 250$ simulations. The interquartile range of the losses is given in parentheses.

d	n	nonlinear regressor			linear regressor x_1
		1st	4th	final	final
5	100	0.9580 (0.1865)	0.6656 (0.1546)	0.3069 (0.2400)	0.0139 (0.0113)
5	200	0.9395 (0.1378)	0.7711 (0.1300)	0.2424 (0.2024)	0.0072 (0.0057)
6	200	0.9432 (0.1231)	0.7207 (0.1067)	0.1641 (0.1766)	0.0018 (0.0016)
8	200	0.9362 (0.1253)	0.6703 (0.1003)	0.2232 (0.1797)	0.0006 (0.0005)
10	100	0.9574 (0.2064)	0.6743 (0.1526)	0.5822 (0.2756)	0.0005 (0.0004)
10	200	0.9406 (0.1522)	0.6777 (0.1202)	0.3690 (0.2213)	0.0002 (0.0002)
10	400	0.9348 (0.0925)	0.7217 (0.0838)	0.2316 (0.1399)	0.0001 (0.0001)

slightly increases with σ while it is essentially proportional to σ for the linear one. An explanation might be that the estimation error for the nonlinear components is mostly due to the nonparametric bias which disappears in the linear components during iteration process by structural adaptation.

Table 2: Case $M = 2$: mean loss $|\widehat{v}_m - v_m^*|$ for the nonlinear regressors for the first, second, fourth, and final iteration and final losses $|\widehat{v}_1|$ for the first linear regressor. Results are obtained from $N = 250$ simulations. The interquartile range of the losses is given in parentheses.

d	n	σ^2	1st nonlinear regressor			2nd nonlinear regressor			linear regressor x_1
			1st	4th	final	1st	4th	final	final
6	200	0.1	4.6117 (0.6646)	3.7349 (0.5028)	0.4763 (0.5211)	4.6337 (0.6257)	3.7576 (0.5402)	0.4473 (0.4785)	0.0081 (0.0063)
8	200	0.1	4.6397 (0.6683)	3.4423 (0.5431)	0.4244 (0.4108)	4.5942 (0.6646)	3.4085 (0.4621)	0.4058 (0.4607)	0.0025 (0.0019)
10	100	0.1	4.6338 (0.8840)	3.1450 (0.7307)	0.7573 (0.5302)	4.6862 (1.0155)	3.1642 (0.7312)	0.7089 (0.4938)	0.0043 (0.0032)
10	200	0.1	4.5537 (0.7458)	3.2806 (0.5065)	0.5812 (0.3404)	4.5904 (0.7649)	3.2917 (0.5875)	0.5489 (0.4014)	0.0011 (0.0010)
10	400	0.1	4.5198 (0.4850)	3.5276 (0.3566)	0.4457 (0.3121)	4.5584 (0.4168)	3.5594 (0.3562)	0.4319 (0.3023)	0.0004 (0.0003)
10	400	0.2	4.5198 (0.4850)	3.5284 (0.3483)	0.4403 (0.3949)	4.5584 (0.4167)	3.5602 (0.3642)	0.4325 (0.3948)	0.0007 (0.0006)
10	400	0.4	4.5198 (0.4850)	3.5297 (0.3316)	0.4637 (0.4891)	4.5584 (0.4167)	3.5615 (0.3727)	0.4666 (0.5029)	0.0017 (0.0013)

The next figure illustrates the result of Theorem 4.1 about separation between linear and nonlinear component. Let $\mathcal{L}_N(\xi)$ denote the empirical distribution of the random variable ξ based on its sample of size N. A good separation between linear and nonlinear components means that the functions $\mathcal{L}_N(\widehat{v}_m)$ for every $m \in \mathcal{J}$ and $1 - \mathcal{L}_N(\widehat{v}_m)$ for $m \notin \mathcal{J}$ have non-overlapping support.

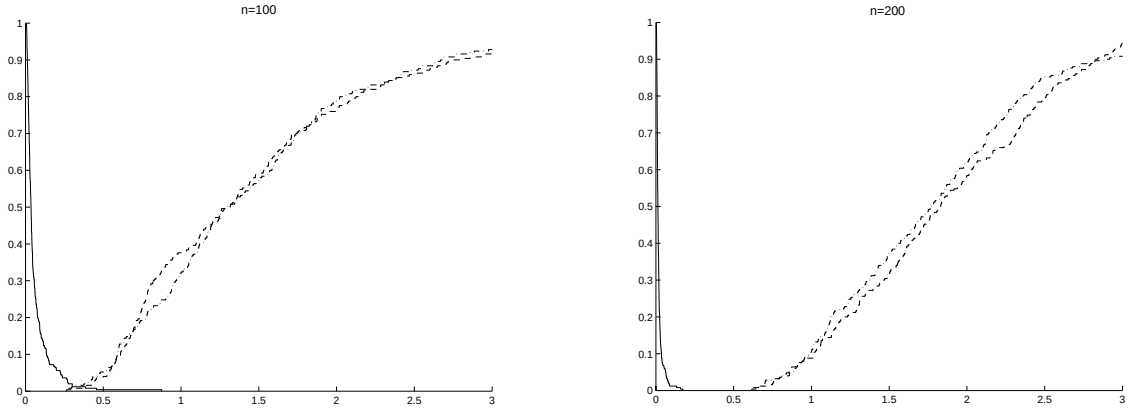


Figure 1: Case $M = 2$, $d = 6$: $\mathcal{L}_N(\widehat{v}_5)$, $\mathcal{L}_N(\widehat{v}_6)$ (dotted lines) and $1 - \mathcal{L}_N(\max_{m=1,\ldots,4} \widehat{v}_m)$ (solid line) for $n = 100, 200$ from 250 realizations.

We observe a very good separability for $n = 100$ and a possibility of perfect separation for $n = 200$.

Next we illustrate how the quality of estimation of the linear component improves with the sample size. Figure 2 shows box-plots of the estimation errors $n^{1/2}\|\widehat{\theta} - \theta^*\|$ of the linear component after the final iteration for $d = 6$, $M = 2$ and different sample sizes n.

Table 3 illustrates the performance of the test of the hypothesis $M \leq \mathcal{M}$ and the quality of the classification rule from Sections 4.2 and 4.3 for different M, d and n. In this table we present the fraction of wrong classifications for every of nonlinear regressors and for the whole model.

One can observe once again that the results (the fraction of wrong classifications) improve as the dimensionality d grows. This can be explained by the fact that the distribution of the test statistic used for classification will be more and more degenerated with growing dimension d.

Another observation is that for $M = 3$, the procedure requires some minimal sample size to start selecting all the three nonlinear components. For $n = 100$ we obtain for almost all the cases $\widehat{M} < M$. For $n = 200$ and $d = 7$ we correctly classify in only about 30% cases but for $d = 10$ the fraction of wrong classifying is already under control.

Figure 3 illustrates the quality of estimation of the noise variance σ^2 by $\widehat{\sigma}^2$ for one example

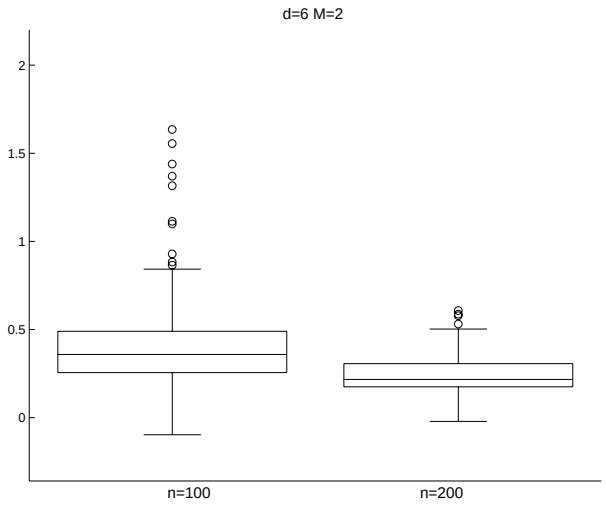


Figure 2: Case $M = 2$: box-plots of the estimation errors $n^{1/2}\|\widehat{\theta} - \theta^*\|$ of the linear component after the final iteration for $d = 6$. Results are obtained from $N = 250$ simulations.

Table 3: Fraction of wrong classifications for every nonlinear regressor and for the whole model. Results are obtained from $N = 250$ simulations and 500 bootstrap replications.

M	d	n	1st n.c.	2nd n.c.	3rd n.c.	$\{\widehat{\mathcal{J}} \neq \mathcal{J}\}$
		100	0.152	–	–	0.18
1	5	200		–	–	
		400	0	–	–	
1	10	200	0	–	–	0.
		400	0	–	–	0.00
1	20	400	0	–	–	0.
		100	0.268	0.308	–	0.38
2	6	200	0.056	0.048	–	0.1
		400	0.004	0.004	–	0.024
2	10	200	0	0	–	0.008
		400	0	0	–	0.0
2	20	400	0	0	–	0
		100	0.976	0.96	0.964	0.992
3	7	200	0.62	0.656	0.656	0.748
		400	0.076	0.06	0.072	0.1
3	10	200	0.004	0	0.004	0.004
		400	0	0	0	0

with $d = 6$, $M = 2$ and different sample size n. The results are in agreement with the root-n consistency of the estimator $\widehat{\sigma}^2$.

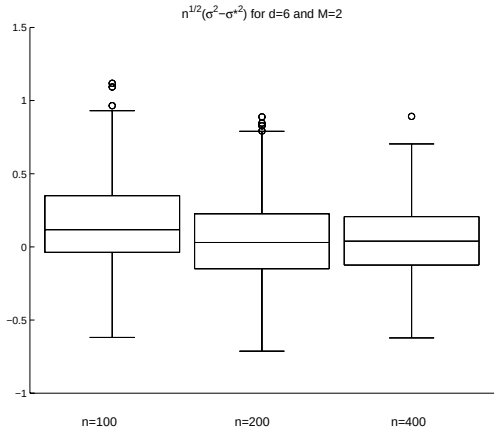


Figure 3: Box-plots of the estimation errors $n^{1/2}\|\widehat{\sigma}^2 - \sigma^{*2}\|$ for $d = 6$, $M = 2$ and different sample size n.

5.1 A real data example

This section presents an application of the procedure to a real data set. We consider the example from Sperlich (1998) and Härdle, Spokoiny and Sperlich (2001) where a subsample of the Socio-Economic Panel of Germany from 1992 was studied. The target of analysis is the weekly number of working hours, Y_i, of 607 women with job and living together with a partner. The following explanatory variables were used: the age of woman, between 25 and 60, X_1; her earning per hour, X_2; the prestige index of her kind of profession (Treiman prestige index), X_3; the monthly rent or redemption for their apartment or house, X_4; the monthly net income of their partner, X_5; the number of years of education, X_6; the unemployment rate at the particular tract they live in, X_7; and the number of children younger than 16 years, X_8.

The estimates $\widehat{v}_m$ obtained by our estimation procedure are given in Table 4. We also got the estimate $\widehat{\sigma}^2 = 0.736$ for error variance σ^2.

Table 4: Estimates $\widehat{v}_m$ of v_m^*.

$\widehat{v}_1$	$\widehat{v}_2$	$\widehat{v}_3$	$\widehat{v}_4$	$\widehat{v}_5$	$\widehat{v}_6$	$\widehat{v}_7$	$\widehat{v}_8$
0.05259	0.00729	0.00441	0.00012	0.00060	0.00142	0.00015	0.00875

Next we identify the linear component starting with $\mathcal{M} = 0$ as described in Section 4.2.

Table 5 gives the p-values $PV_{\mathcal{M}}$ for each test $H_{\mathcal{M}}$, which are obtained during the bootstrap procedure, defined as:

$$PV_{\mathcal{M}} = \frac{1}{B}\sum_{b=1}^{B} 1_{\{\widetilde{v}^{(b)}_{(\mathcal{M}+1)} > \widehat{v}_{(\mathcal{M}+1)}\}}$$

The first three hypotheses H_0, H_1 and H_2 are rejected at 10% level, and there is clearly no rejection of H_3. So, for the considered model, the nonlinear dimension is estimated as three and the nonlinear variables are X_1, X_2, and X_3. Our linear/nonlinear variable classification results coincide with those from Härdle, Sperlich and Spokoiny (2001), but with quite different p-values: in our results X_1 (age) is the most nonlinear and X_2 (earning per hour) is the least nonlinear variable among the three, while in Härdle, Sperlich and Spokoiny (2001) the situation is reversed. Note that while their identification was made under the assumption of additive model structure, our results are obtained for a general situation when such additive structure is not required.

Table 5: p-values for consecutive tests

$\mathcal{M}$	$\widehat{v}_{(\mathcal{M}+1)}$	p-values
0	$\widehat{v}_1$	0.003996
1	$\widehat{v}_2$	0.086913
2	$\widehat{v}_3$	0.01998
3	$\widehat{v}_5$	0.47153

6 Conclusion and outlook

The paper has introduced a new method of exploring a partially linear model based on the idea of structural adaptation. The method applies under mild assumptions on the underlying regression function and the regression design. The procedure is fully adaptive and does not require any prior information. The results claim that the proposed procedure with a high probability correctly identifies the nonlinear component and estimates the linear component with the optimal rate $n^{-1/2}$ provided that the dimension of the nonlinear component is not larger than 3. The simulation results demonstrate an excellent performance of the procedure for all considered situations. An important feature of the method is that it is very stable with respect to high dimensionality and for a non-regular design.

Non-Gaussian or heterogeneous noise. The method and results can be easily extended to models with homogeneous non-Gaussian noise satisfying some exponential moment con-

ditions. Another interesting issue is applicability of the method for a general heterogeneous or dependent noise, in particular, to time series models and financial data. We leave these extensions for further research.

The case with $M \geq 4$. The method continues to apply even if $M \geq 4$ and iterations would lead to improvement of the bias. However, the bound for the bias of order $(n^{-1} \log n)^{2/3}$ can be achieved only for $M \leq 3$. For larger M, the bias will be of order $n^{-1/2}$ of bigger and the procedure does not provide root-n consistent estimation of the functionals β^*_{lm}. So, if the hypothesis $M \leq 3$ is rejected, then we recommend to apply for the choice of M some model selection criteria like cross-validation or Mallows C_p.

Data-driven choice of parameter L. The method depends upon the parameter L describing the number of basis functions for every regressor. In the univariate case, either an n-dependent or data-driven choice of such a parameter is usually applied, see Hart (1997) or Spokoiny (2001) and references therein. An adaptive choice of L in the considered problem is an interesting question for further research.

Semiparametrically efficient estimation of the linear component. Due to the result of Theorem 4.4, the proposed estimator of the parameter θ is root-n consistent and asymptotically normal. However, it is unlikely that this or the refined estimator of θ from Section 4.3 is semiparametrically efficient in the sense of minimization of the asymptotic variance, see e.g. Bickel et al. (1998). A modification of the method leading to the semiparametrically efficient estimation of linear part will be discussed elsewhere.

Estimation of the nonlinear component. After the nonlinear component is identified, it can be estimated using the standard methods of nonparametric statistics. Actually, the algorithm gives an estimator of the whole function f and of the linear component, so that the nonlinear component can be extracted as well. This estimator corresponds to the local linear smoothing of the nonparametric M-dimensional function with the bandwidth about $h\rho \approx \rho_{\min}$, and may not achieve the best rate. To improve the quality of estimation, one can apply the classical cross-validation technique for selecting the bandwidth in the direction of the nonparametric component.

Discrete and categorical data. Note that the assumption of linearity is meaningful for discrete or categorical variables as well. It means that the influence of the corresponding regressor is independent of the other variables and therefore, at least in the binary case, can be modelled linearly. Moreover, the procedure easily applies for the situation with discrete data without any change.

7 Appendix

Here we collect the proofs of the main results. For the ease of exposition, we consider only the main procedure (without weights) and only the case of $\eta = 0$. The general case can be considered in the same way.

7.1 One-step improvement

Suppose that we are given some fixed numbers h and ρ (which mean the current values h_k and ρ_k) and a vector $\boldsymbol{v} = (v_1, \ldots, v_d)^\top \in I\!R^d$ which can be viewed as an approximation of $\boldsymbol{v}^* = (v_1^*, \ldots, v_d^*)^\top$ obtained at the previous step of the algorithm. Set also

$$b_m \;=\; h\left(1 + \rho^{-2} v_m\right)^{-1/2}, \qquad m = 1, \ldots, d, \tag{7.1}$$

and define $\boldsymbol{b} = \mathrm{diag}(b_1, \ldots, b_d)$. Define also $\widehat{f}(X_i)$, $\widehat{\nabla f}(X_i)$ and $\widehat{\beta}_{lm}$ by (2.1) and (2.3) for all $l = 1, \ldots, L$ and $m = 1, \ldots, d$ with the just defined bandwidth $\boldsymbol{b}$. We aim to evaluate the estimation errors $\widehat{\beta}_{lm} - \beta_{lm}^*$. To describe the results, we introduce the shrinking factors $P_{\rho,m} = \left(1 + \rho^{-2} v_m^*\right)^{-1/2}$ and define

$$U_m = P_{\rho,m}^2 (1 + \rho^{-2} v_m) = (1 + \rho^{-2} v_m^*)^{-1} (1 + \rho^{-2} v_m)$$

and similarly $U_m^* = P_{\rho,m}^2 (1 + \rho^{-2} v_m^*) = 1$. Clearly the vector $U = (U_1, \ldots, U_d)^\top \in I\!R^d$ uniquely describes $\boldsymbol{v}$, so that we consider later in this section $v = v(U)$ and similarly $\widehat{\beta}_{lm} = \widehat{\beta}_{lm}(U)$ for the functionals $\widehat{\beta}_{lm}$ in (2.3). Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top$ be a vector in $I\!R^d$ with entries $\alpha_m \in (0, 1)$. Define

$$\mathcal{U}_{\boldsymbol{\alpha}} = \{U = (U_1, \ldots, U_d)^\top \in I\!R^d : |U_m - 1| \le \alpha_m \,,\, m = 1, \ldots, d\}.$$

We also define $\alpha^* = \max_{m=1,\ldots,d} \alpha_m$.

Proposition 7.1. *Let Assumptions 1 through 5 hold. Let $\beta_{lm}(U) = \boldsymbol{E}\widehat{\beta}_{lm}(U)$. Then*

$$\sup_{U \in \mathcal{U}_{\boldsymbol{\alpha}}} \sum_{m=1}^{d} \sum_{l=1}^{L} |P_{\rho,m}\{\beta_{lm}(U) - \beta_{lm}^*\}|^2 \le \left(\frac{C_g\, C_V^{1/2}}{1 - \alpha^*}\, \rho^2 h\right)^2$$

and, for every $l = 1, \ldots, L$ and $m = 1, \ldots, d$, there exists a zero mean Gaussian random variable ξ_{lm} defined as a linear combination of the errors ε_i with deterministic coefficients, which depend on $\boldsymbol{v}^*$, the design $\{X_i\}$, the basis functions $\psi_{lm}(\cdot)$, and the kernel K only, and such that

$$\max_{m,l} \boldsymbol{E}\xi_{lm}^2 \le 2\sigma^2 C_V^2 C_K \tag{7.2}$$

and

$$\boldsymbol{P}\Big(\max_{m,l}\sup_{U\in\mathcal{U}_\alpha}\Big|P_{\rho,m}\{\widehat{\beta}_{lm}(U)-\beta_{lm}(U)\}-\frac{\xi_{lm}}{h\sqrt{n}}\Big|>\frac{\sigma\,\overline{\psi}\,C_{\alpha,n}|\boldsymbol{\alpha}|}{h\sqrt{n}}\Big)\le\frac{2}{n},$$

where the maximum is taken over $m=1,\ldots,d$ *and* $l=1,\ldots,L$, $\overline{\psi}=\max_{i,l,m}|\psi_{lm}(X_i)|$ *and*

$$C_{\alpha,n}=\left(\frac{\sqrt{2}\,C_V\,C_{K'}}{(1-\alpha^*)^{3/2}}+\frac{2^{3/2}C_V^2\,C_{K'}\,C_K}{(1-\alpha^*)^{5/2}}\right)\left(2+\sqrt{2\log(ndL)+d\log(4n)}\right).$$

Let β_m^* denote, as in Proposition 3.1, an L-vector with the components β_{lm}^* and $\widehat{\beta}_m=\widehat{\beta}_m(U)$ its estimator with the components $\widehat{\beta}_{lm}(U)$.

Corollary 7.1. *Let* $z_n=(1+2\log(nd)+2\log\log(nd))^{1/2}$ *and*

$$\delta=\frac{C_g\,C_V^{1/2}}{1-\alpha^*}\,h\rho^2+\frac{\sqrt{2L}\sigma C_V C_K^{1/2}z_n}{h\sqrt{n}}+\frac{\sqrt{L}\sigma\,\overline{\psi}\,C_{\alpha,n}|\boldsymbol{\alpha}|}{h\sqrt{n}}. \tag{7.3}$$

Then under the conditions of Proposition 7.1 it holds

$$\boldsymbol{P}\left(\max_{m=1,\ldots,d}\sup_{U\in\mathcal{U}_\alpha}\left|P_{\rho,m}\left(\widehat{\beta}_m(U)-\beta_m^*\right)\right|>\delta\right)\le 3/n.$$

The corollary helps bound the estimation error $P_{\rho,m}^2\left(\widehat{v}_m(U)-v_m^*\right)$.

Proposition 7.2. *Under the conditions of Proposition 7.1,*

$$\boldsymbol{P}\left(\sup_{U\in\mathcal{U}_\alpha}\left|P_{\rho,m}^2\left(\widehat{v}_m(U)-v_m^*\right)\right|\le\delta^2+2\delta\tau_m\ \textit{for all}\ m=1,\ldots,d\right)\ge 1-3/n$$

where $\tau_m=\rho\sqrt{v_m^*}(\rho^2+v_m^*)^{-1/2}\le\min\left\{\rho,\sqrt{v_m^*}\right\}$.

7.2 Proof of Proposition 7.1

Denote by P_ρ the diagonal $d\times d$-matrix with the diagonal entries $P_{\rho,m}$, that is, $P_\rho=\mathrm{diag}\{P_{\rho,1},\ldots,P_{\rho,d}\}$. Similarly, for $U=(U_1,\ldots,U_d)^\top\in\mathbb{R}^d$, define $D_U=\mathrm{diag}\{U_1,\ldots,U_d\}$. Next, for every $i,j\le n$, define $Z_{ij}=h^{-1}P_\rho^{-1}(X_j-X_i)$, $K_{ij}(U)=K(Z_{ij}^\top D_U Z_{ij})$

$$\begin{aligned}\mathcal{V}_i(U) &= \sum_{j=1}^n\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^\top K_{ij}(U),\\ \widehat{s}_i(U) &= h^{-1}\mathcal{V}_i(U)^{-1}\sum_{j=1}^n\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}Y_j\,K_{ij}(U).\end{aligned}$$

It is easy to check that for the $(m+1)$th component $\widehat{s}_{i,m}(U)$ of $\widehat{s}_i(U)$ it holds $\widehat{s}_{i,m}(U)=P_{\rho,m}\,\widehat{\nabla f}_m(X_i)$ and hence,

$$P_{\rho,m}\widehat{\beta}_{lm}(U)=n^{-1}\sum_{i=1}^n\widehat{s}_{i,m}(U)\psi_{lm}(X_{i,m}).$$

The model equation (1.1) implies $\widehat{s}_i(U) = s_i(U) + \zeta_i(U)$ with

$$\begin{aligned} s_i(U) &= h^{-1}\,\mathcal{V}_i(U)^{-1}\sum_{j=1}^{n}\binom{1}{Z_{ij}} f(X_j)\,K_{ij}(U), \\ \zeta_i(U) &= h^{-1}\,\mathcal{V}_i(U)^{-1}\sum_{j=1}^{n}\binom{1}{Z_{ij}} \varepsilon_j\,K_{ij}(U). \end{aligned}$$

This yields, for each coordinate $m = 1, \ldots, d$,

$$\begin{aligned} P_{\rho,m}\{\boldsymbol{E}\widehat{\beta}_{lm}(U) - \beta^*_{lm}\} &= \frac{1}{n}\sum_{i=1}^{n}\{s_{i,m}(U) - P_{\rho,m}\nabla f_m(X_i)\}\,\psi_{lm}(X_{i,m}), \\ P_{\rho,m}\{\widehat{\beta}_{lm}(U) - \boldsymbol{E}\widehat{\beta}_{lm}\} &= \frac{1}{n}\sum_{i=1}^{n}\zeta_{i,m}(U)\psi_{lm}(X_{i,m}). \end{aligned}$$

Clearly $\zeta_{lm}(U) := n^{-1}\sum_{i=1}^{n}\zeta_{i,m}(U)\psi_{lm}(X_{i,m})$ is for every U a linear combination of the Gaussian errors ε_i and therefore it is also a Gaussian vector in $\mathbb{R}^d$.

Define $\mathcal{E}_d$ is the projection from $\mathbb{R}^{d+1}$ onto $\mathbb{R}^d$ dropping the zero coordinate: $\mathcal{E}_d(x_0, \ldots, x_d)^\top = (x_1, \ldots, x_d)^\top$. It is easy to see that the following three statements imply the claimed result:

$$\sup_{U\in\mathcal{U}_\alpha} |\mathcal{E}_d s_i(U) - P_\rho\nabla f(X_i)| \le \frac{C_g\,C_V^{1/2}}{1-\alpha^*}\,h\rho^2, \qquad i = 1, \ldots, n, \tag{7.4}$$

$$\boldsymbol{P}\left(\max_{l,m}\sup_{U\in\mathcal{U}_\alpha} |\zeta_{lm}(U) - \zeta_{lm}(U^*)| > \frac{\sigma C_{\alpha,n}|\boldsymbol{\alpha}|}{h\sqrt{n}}\right) \le 2/n, \tag{7.5}$$

$$\max_{l,m}\boldsymbol{E}|\zeta_{lm}(U^*)|^2 \le \frac{2\sigma^2 C_V^2 C_K}{h^2 n}. \tag{7.6}$$

where the maximum is taken over $l = 1, \ldots, L$ and $m = 1, \ldots, d$. Indeed, the last two statements of the proposition directly follows from (7.5) and (7.6) for $\xi_{lm} = h\sqrt{n}\,\zeta_{lm}(U^*)$. Next, (7.4) implies

$$n^{-1}\sum_{i=1}^{n}\sum_{m=1}^{d}|s_{i,m}(U) - P_{\rho,m}\nabla f_m(X_i)|^2 \le \left(\frac{C_g\,C_V^{1/2}}{1-\alpha^*}\,h\rho^2\right)^2.$$

Since the vectors $\psi_{lm} \in \mathbb{R}^n$ are orthonormal for different l, it follows for the Bessel inequality for every $m \le d$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}|s_{i,m}(U) - P_{\rho,m}\nabla f_m(X_i)|^2 &\ge \sum_{l=1}^{L}\left|\frac{1}{n}\sum_{i=1}^{n}\Bigl(s_{i,m}(U) - P_{\rho,m}\nabla f_m(X_i)\Bigr)\psi_{lm}(X_{i,m})\right|^2 \\ &= \sum_{l=1}^{L}P_{\rho,m}^2\Bigl(\boldsymbol{E}\widehat{\beta}_{lm}(U) - \beta^*_{lm}\Bigr)^2 \end{aligned}$$

and thus,

$$\sum_{m=1}^{d}\sum_{l=1}^{L}P_{\rho,m}^2\Bigl(\boldsymbol{E}\widehat{\beta}_{lm}(U) - \beta^*_{lm}\Bigr)^2 \le \left(\frac{C_g\,C_V^{1/2}}{1-\alpha^*}\,h\rho^2\right)^2.$$

To check the statements (7.4)–(7.6), the following lemma will be useful.

Lemma 7.1. *Let* $|U_m - 1| \le \alpha_m < 1$ *for all* $m = 1, \ldots, d$. *Then for all* i, j , *the inequality* $|Z_{ij}^\top D_U Z_{ij}| \le 1$ *implies* $|Z_{ij}|^2 \le 1/(1-\alpha^*)$ *and* $1 + |Z_{ij}|^2 \le 2/(1-\alpha^*)$.

Proof. Note that the inequalities $Z_{ij}^\top D_U Z_{ij} \le 1$ and $|U_m - 1| \le \alpha_m$ imply

$$\left| Z_{ij}^\top D_U Z_{ij} - |Z_{ij}|^2 \right| = \left| Z_{ij}^\top (D_U - I) Z_{ij} \right| \le \alpha^* |Z_{ij}|^2$$

and thus, $|Z_{ij}|^2 \le (1-\alpha^*)^{-1} Z_{ij}^\top D_U Z_{ij}$. □

Now we check (7.4). Since

$$\begin{aligned} \begin{pmatrix} h^{-1} f(X_i) \\ P_\rho \nabla f(X_i) \end{pmatrix} &= \mathcal{V}_i(U)^{-1} \sum_{j=1}^n \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix}^\top \begin{pmatrix} h^{-1} f(X_i) \\ P_\rho \nabla f(X_i) \end{pmatrix} K_{ij}(U) \\ &= h^{-1} \mathcal{V}_i(U)^{-1} \sum_{j=1}^n \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \{ f(X_i) + X_{ij}^\top \nabla f(X_i) \} K_{ij}(U) \end{aligned}$$

it follows

$$\begin{aligned} s_i(U) - \begin{pmatrix} h^{-1} f(X_i) \\ P_\rho \nabla f(X_i) \end{pmatrix} &= \frac{1}{h} \mathcal{V}_i(U)^{-1} \sum_{j=1}^n \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} \left\{ f(X_j) - f(X_i) - X_{ij}^\top \nabla f(X_i) \right\} K_{ij}(U) \\ &= \frac{1}{h} \mathcal{V}_i(U)^{-1} \sum_{j=1}^n \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} r_{ij} K_{ij}(U) \end{aligned}$$

where in view of (3.1)

$$r_{ij} = g(\mathcal{R}^* X_{j,2}) - g(\mathcal{R}^* X_{i,2}) - (\mathcal{R}^* X_{j,2} - \mathcal{R}^* X_{i,2})^\top \nabla g(\mathcal{R}^* X_{i,2})$$

with $\mathcal{R}^*$ being the diagonal $M \times M$ matrix with diagonal entries $\sqrt{v_m^*}$, $m \in \mathcal{J}$. It is clear that

$$\left| \sqrt{v_m^*} X_{j,m} - \sqrt{v_m^*} X_{i,m} \right|^2 = h^2 v_m^* (1 + \rho^{-2} v_m^*)^{-1} Z_{ij,m}^2 \le h^2 \rho^2 Z_{ij,m}^2 .$$

Therefore,

$$|\mathcal{R}^* X_{j,2} - \mathcal{R}^* X_{i,2}|^2 \le h^2 \rho^2 |Z_{ij}|^2 .$$

This yields by Lemma 7.1 and Assumption 4 for every pair (i, j) with $Z_{ij}^\top D_U Z_{ij} \le 1$:

$$|r_{ij}| \le C_g h^2 \rho^2 (1-\alpha^*)^{-1} .$$

Using the Cauchy-Schwarz inequality and Assumptions 5 we bound

$$
\begin{aligned}
&|\mathcal{E}_d\, s_i(U) - P_\rho\, \nabla f(X_i)| \le h^{-1} \sup_{\lambda \in I\!R^{d+1}\,:\,|\lambda|=1} \left| \lambda^\top \mathcal{V}_i(U)^{-1} \sum_{j=1}^n \binom{1}{Z_{ij}} r_{ij}\, K_{ij}(U) \right| \\
&\le \sup_{|\lambda|=1} h^{-1} \left[\sum_{j=1}^n \lambda^\top \mathcal{V}_i(U)^{-1} \binom{1}{Z_{ij}} \binom{1}{Z_{ij}}^\top K_{ij}(U)\, \mathcal{V}_i(U)^{-1} \lambda \sum_{j=1}^n r_{ij}^2\, K_{ij}(U) \right]^{1/2} \\
&\le \frac{C_g\, h\, \rho^2}{1-\alpha^*} \sup_{|\lambda|=1} \left(\lambda^\top \mathcal{V}_i(U)^{-1} \lambda \sum_{j=1}^n K_{ij}(U) \right)^{1/2} \\
&\le (1-\alpha^*)^{-1}\, C_g\, h\, \rho^2 \left\| N_i(U) \mathcal{V}_i(U)^{-1} \right\|^{1/2} \le (1-\alpha^*)^{-1}\, C_g\, C_V^{1/2}\, h\, \rho^2
\end{aligned}
$$

and (7.4) follows.

By definition every $\zeta_{lm}(U)$ is a linear combination of the ε_i's, that is, there are coefficients $c_{i,lm}(U)$ such that

$$
\zeta_{lm}(U) = \sum_{i=1}^n c_{i,lm}(U) \varepsilon_i .
$$

The coefficients $c_{i,lm}(U)$ depend on the design $X_1, \ldots, X_n$, the basis function ψ_{lm}, the kernel K and the vector U. Moreover, these coefficients satisfy the following conditions:

Lemma 7.2. *For every $l = 1, \ldots, L$ and $m = 1, \ldots, d$*

(i) $$\sum_{i=1}^n |c_{i,lm}(U^*)|^2 \le \frac{2 C_V^2 C_K}{h^2 n} \;;$$

(ii) $$\sup_{U \in \mathcal{U}_\alpha} \sum_{i=1}^n |c_{i,lm}(U)|^2 \le \frac{2 C_V^2 C_K}{(1-\alpha^*) h^2 n} \;;$$

(iii) $$\sup_{U \in \mathcal{U}_\alpha} \left| \frac{dc_{i,lm}(U)}{dU} \right| \le \frac{\kappa_\alpha}{nh}, \quad \textit{where}$$

$$
\kappa_\alpha = \sqrt{2}(1-\alpha^*)^{-3/2} C_V C_{K'} \overline{\psi} + 2^{3/2} (1-\alpha^*)^{-5/2} C_V^2 C_K C_{K'} \overline{\psi} \,.
$$

Proof. Define for $i, j = 1, \ldots, n$

$$
N_i(U) = \sum_{j=1}^n K_{ij}(U), \qquad v_{ij}(U) = \mathcal{V}_i(U)^{-1} \binom{1}{Z_{ij}}.
$$

It follows from Lemma 7.1 and Assumption 5 that $U \in \mathcal{U}_\alpha$ implies for every i, j with $Z_{ij}^\top D_U Z_{ij} \le 1$

$$
|N_i(U)\, v_{ij}(U)| \le C_V (1 + |Z_{ij}|^2)^{1/2} \le C_V \sqrt{2} (1-\alpha^*)^{-1/2}. \tag{7.7}
$$

Next, for a fixed $m \le d$, denote by $v_{ij,m}(U)$ the $(m+1)$ th component of $v_{ij}(U)$. Then

$$\begin{aligned}\zeta_{lm}(U) &= \frac{1}{nh}\sum_{i=1}^{n}\psi_{lm}(X_{i,m})\sum_{j=1}^{n}v_{ij,m}(U)\,K_{ij}(U)\,\varepsilon_j \\ &= \sum_{j=1}^{n}\left(\frac{1}{nh}\sum_{i=1}^{n}\psi_{lm}(X_{i,m})v_{ij,m}(U)\,K_{ij}(U)\right)\varepsilon_j = \sum_{j=1}^{n}c_{j,lm}(U)\,\varepsilon_j.\end{aligned}$$

Clearly $\boldsymbol{E}|\zeta_{lm}(U)|^2 = \sigma^2\sum_{j=1}^{n}c_{j,lm}^2(U)$. The Cauchy-Schwarz inequality, (7.7) and Assumption 5 imply

$$\begin{aligned}\sum_{j=1}^{n}c_{j,lm}^2(U) &= \frac{1}{n^2h^2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\psi_{lm}(X_{i,m})v_{ij,m}(U)\,K_{ij}(U)\right)^2 \\ &\le \frac{1}{n^2h^2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\psi_{lm}^2(X_{i,m})v_{ij,m}(U)\,K_{ij}(U)\right)\left(\sum_{i=1}^{n}v_{ij,m}(U)\,K_{ij}(U)\right) \\ &\le \frac{2C_V^2}{(1-\alpha^*)n^2h^2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\psi_{lm}^2(X_{i,m})\frac{K_{ij}(U)}{N_i(U)}\right)\left(\sum_{i=1}^{n}\frac{K_{ij}(U)}{N_i(U)}\right) \\ &\le \frac{2C_V^2C_K}{(1-\alpha^*)n^2h^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\psi_{lm}^2(X_{i,m})\frac{K_{ij}(U)}{N_i(U)} \\ &= \frac{2C_V^2C_K}{(1-\alpha^*)n^2h^2}\sum_{i=1}^{n}\psi_{lm}^2(X_{i,m}) = \frac{2C_V^2C_K}{(1-\alpha^*)nh^2}.\end{aligned}$$

As a particular case, with $D_U = D_{U^*} = I$ and $\alpha^* = 0$, this yields

$$\sum_{j=1}^{n}c_{j,lm}^2(U^*) \le \frac{2C_V^2C_K}{nh^2}$$

and the first two assertions of the lemma follows.

Now we bound the derivative of each coefficient $c_{jl,m}(U)$ with respect to U. For every pair i,j such that $Z_{ij}^\top D_U Z_{ij} \le 1$, Lemma 7.1 implies

$$\left|\frac{d}{dU}K_{ij}(U)\right| = \left|K'(Z_{ij}^\top D_U Z_{ij})\right||Z_{ij}|^2 \le (1-\alpha^*)^{-1}\left|K'(Z_{ij}^\top D_U Z_{ij})\right|.$$

Let $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$ be unit vectors in $\mathbb{R}^{d+1}$. Then for every $m = 1,\ldots,d$

$$\begin{aligned}\frac{\boldsymbol{o}_1^\top\partial\mathcal{V}_i(U)^{-1}\boldsymbol{o}_2}{\partial U_m} &= -\boldsymbol{o}_1^\top\mathcal{V}_i(U)^{-1}\left(\frac{\partial}{\partial U_m}\mathcal{V}_i(U)\right)\mathcal{V}_i(U)^{-1}\boldsymbol{o}_2 \\ &= -\boldsymbol{o}_1^\top\mathcal{V}_i(U)^{-1}\left(\sum_{j=1}^{n}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}^\top K'(Z_{ij}^\top D_U Z_{ij})Z_{ij,m}^2\right)\mathcal{V}_i(U)^{-1}\boldsymbol{o}_2.\end{aligned}$$

Lemma 7.1 and Assumption 5 yield

$$\left|\frac{\partial\boldsymbol{o}_1^\top\mathcal{V}_i(U)^{-1}\boldsymbol{o}_2}{\partial U_m}\right| \le \frac{2C_V^2}{(1-\alpha^*)|N_i(U)|^2}\sum_{j=1}^{n}\left|K'(Z_{ij}^\top D_U Z_{ij})\right|Z_{ij,m}^2.$$

Since $v_{ij,m}(U) = \left(1+|Z_{ij}|^2\right)^{1/2} \boldsymbol{e}_m^\top \mathcal{V}_i(U)^{-1}\boldsymbol{o}_2$ where $\boldsymbol{e}_m$ denotes the mth coordinate vector in $\mathbb{R}^{d+1}$ and $\boldsymbol{o}_2 = \left(1+|Z_{ij}|^2\right)^{-1/2}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}$, it follows for every pair i,j such that $Z_{ij}^\top D_U Z_{ij} \le 1$:

$$\begin{aligned}
\left|\frac{dv_{ij,m}(U)}{dU}\right| &\le \left(1+|Z_{ij}|^2\right)^{1/2}\left(\sum_{m'=1}^{d}\left|\frac{\partial \boldsymbol{e}_m^\top \mathcal{V}_i(U)^{-1}\boldsymbol{o}_2}{\partial U_{m'}}\right|^2\right)^{1/2}\\
&\le \frac{2^{3/2}C_V^2}{(1-\alpha^*)^{3/2}|N_i(U)|^2}\left[\sum_{m'=1}^{d}\left(\sum_{j=1}^{n}|K'(Z_{ij}^\top D_U Z_{ij})|\, Z_{ij,m'}^2\right)^2\right]^{1/2}\\
&\le \frac{2^{3/2}C_V^2}{(1-\alpha^*)^{3/2}|N_i(U)|^2}\sum_{m'=1}^{d}\sum_{j=1}^{n}|K'(Z_{ij}^\top D_U Z_{ij})|\, Z_{ij,m'}^2\\
&\le \frac{2^{3/2}C_V^2 C_{K'}}{(1-\alpha^*)^{3/2}|N_i(U)|^2}\sum_{j=1}^{n}|K'(Z_{ij}^\top D_U Z_{ij})|\,|Z_{ij}|^2 \le \frac{2^{3/2}C_V^2 C_{K'}}{(1-\alpha^*)^{5/2}N_i(U)}.
\end{aligned}$$

Since

$$\frac{dc_{j,lm}(U)}{dU} = \frac{1}{nh}\sum_{i=1}^{n} v_{ij,m}(U)\psi_{lm}(X_{i,m})\frac{dK_{ij}(U)}{dU} + \frac{1}{nh}\sum_{i=1}^{n}\frac{dv_{ij,m}(U)}{dU}K_{ij}(U)\psi_{lm}(X_{i,m}).$$

the use of (7.7) and Assumption 5 yields

$$\begin{aligned}
\left|\frac{dc_{j,lm}(U)}{dU}\right| &\le \frac{\sqrt{2}C_V\overline{\psi}_{lm}}{nh(1-\alpha^*)^{3/2}}\sum_{i=1}^{n}\frac{|K'(Z_{ij}^\top D_U Z_{ij})|}{N_i(U)} + \frac{2^{3/2}C_V^2 C_{K'}\overline{\psi}_{lm}}{nh(1-\alpha^*)^{5/2}}\sum_{i=1}^{n}\frac{K_{ij}(U)}{N_i(U)}\\
&\le \frac{\sqrt{2}C_V C_{K'}\overline{\psi}_{lm}}{nh(1-\alpha^*)^{3/2}} + \frac{2^{3/2}C_V^2 C_{K'}C_K\overline{\psi}_{lm}}{nh(1-\alpha^*)^{5/2}}
\end{aligned}$$

and assertion (iii) of the lemma follows. □

Since $\boldsymbol{E}|\zeta_{lm}(U)|^2 = \sigma^2\sum_{j=1}^{n} c_{j,lm}^2(U)$, condition (7.6) follows from Lemma 7.2, (i).

The following lemma is a minor modification of Lemma 8 of HJS.

Lemma 7.3. *Let r be a positive number and let Γ be a finite set. Let functions $a_{i,\gamma}(u)$ of $u \in \mathbb{R}^d$ obey the conditions*

$$\sup_{\gamma\in\Gamma}\sup_{|u-u^*|\le r}\left|\frac{d}{du}a_{i,\gamma}(u)\right| \le \kappa, \qquad i = 1,\dots,n. \tag{7.8}$$

If the ε_i's are independent $\mathcal{N}(0,\sigma^2)$-distributed random variables, then

$$\boldsymbol{P}\left(\sup_{\gamma\in\Gamma}\sup_{|u-u^*|\le r}\frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\{a_{i,\gamma}(u)-a_{i,\gamma}(u^*)\}\varepsilon_i\right| > \sigma\kappa\, r\, t\right) \le \frac{2}{n}$$

where $t = 2+\sqrt{2\log(n|\Gamma|)+d\log(4n)}$ and $|\Gamma|$ denotes the number of elements in Γ.

The claim (7.5) follows from Lemma 7.2, (ii) and (iii), by the preceding lemma applied with $a_{i,\gamma}(u) = \sqrt{n} c_{j,lm}(U)$, $\kappa = \frac{\kappa_\alpha}{h\sqrt{n}}$, $\Gamma = \{(m,l) : m = 1,\dots,d,\ l = 1,\dots,L\}$, and $r = |\boldsymbol{\alpha}|$. This completes the proof of the proposition.

Remark 7.1. In the proof of Proposition 7.1 we defined the random variables ξ_{lm} as $\xi_{lm} = \xi_{lm}(U^*)$. One can easily check that the result of the proposition continues to hold with ξ_{lm} replaced by $\xi_{lm}(U)$ for any $U \in \mathcal{U}_{\boldsymbol{\alpha}}$ and with the right hand-side of (7.2) and with the constant $C_{\alpha,n}$ multiplied by $(1-\alpha^*)^{-1}$. This fact is used in the proof Theorem 4.2.

7.3 Proof of Corollary 7.1

By Proposition 7.1

$$\sup_{U\in\mathcal{U}_\alpha} \max_{m=1,\dots,d} \left| P_{\rho,m}\left(\boldsymbol{E}\widehat{\beta}_m(U) - \beta^*_m \right) \right| \le \frac{C_g\, C_V^{1/2}}{1-\alpha^*}\, \rho^2 h$$

and on a random set of probability as least $1 - 2/n$

$$\left| P_{\rho,m}\left(\widehat{\beta}_m(U) - \boldsymbol{E}\widehat{\beta}_m(U) \right) - \frac{\xi_m}{h\sqrt{n}} \right| \le \frac{\sqrt{L}\sigma\, \overline{\psi}\, C_{\alpha,n}\, |\boldsymbol{\alpha}|}{h\sqrt{n}}, \qquad \forall m = 1,\dots,d,$$

where $\xi_m \in I\!R^L$, $m = 1,\dots,d$, are Gaussian random vectors with components $\xi_{l,m}$ from Proposition 7.1.

By Lemma 7 in HJS,

$$\boldsymbol{P}\left(|\xi_m| \ge z_n \sqrt{\boldsymbol{E}|\xi_m|^2} \right) \le 1/(nd).$$

In view of (7.2) $\boldsymbol{E}|\xi_m|^2 \le 2L\sigma^2 C_V^2 C_K$, and the corollary follows.

7.4 Proof of Proposition 7.2

The definition of τ_m implies

$$P^2_{\rho,m} v^*_m = (1 + \rho^{-2} v^*_m)^{-1} v^*_m = \tau^2_m \le \min\left\{\rho^2, v^*_m\right\}. \tag{7.9}$$

Lemma 7.4. *If $P_{\rho,m}|\widehat{\beta}_m - \beta^*_m| < \delta$, then $P^2_{\rho,m}|\widehat{v}_m - v^*_m| < \delta^2 + 2\delta\,\tau_m$.*

Proof. Define the vector $\widehat{u}_m \in I\!R^L$ (resp. u^*_m) whose elements are $P_{\rho,m}\widehat{\beta}_{lm}$ (resp. $P_{\rho,m}\beta^*_{lm}$). Clearly $P^2_{\rho,m}\widehat{v}_m(U) = |\widehat{u}_m(U)|^2$ and by (7.9) $P^2_{\rho,m} v^*_m = |u^*_m|^2 \le \tau^2_m$. It is easy to check that

$$\left| |\widehat{u}_m|^2 - |u^*_m|^2 \right| \le |\widehat{u}_m - u^*_m|^2 + 2|\widehat{u}_m - u^*_m| \cdot |u^*|, \tag{7.10}$$

and lemma follows. □

The proposition follows from Corollary 7.1 and Lemma 7.4.

7.5 Proof of Proposition 3.1

The proof of the first claim is a simplified version of the proof of Proposition 7.1: just set there $P_{\rho,m} = 1$, drop $\sup_U$, and repeat the proofs of (7.4) and (7.6). The factor $v^*_{(1)}$ in δ_1 comes from $\mathcal{R}^*$ in (3.1). Next, applying Lemma 7 of HJS one gets the claim (3.2). The claim (3.4) follows from (3.2) and Lemma 7.4 applied with $\rho = 1$.

7.6 Proof of Theorem 3.1

Let the numbers h_k and ρ_k be as in the algorithm description, $k = 1, \ldots, k_n$. Define successively the values δ_k and d-vectors $\boldsymbol{\alpha}_k$ with components $\alpha_{k,m}$ as follows: $\boldsymbol{\alpha}_1 = 0$, δ_1 as in (3.3), and for $k = 2, \ldots, k_n$

$$\begin{aligned} \delta_k &= \frac{C_g\, C_V^{1/2}}{(1-\alpha_k^*)}\, h_k\, \rho_k^2 + \frac{\sqrt{2L}\,\sigma\, C_V\, C_K^{1/2}\, z_n}{h_k\sqrt{n}} + \frac{\sqrt{L}\sigma\overline{\psi} C_{\alpha_k,n}\,|\boldsymbol{\alpha}_k|}{h_k\sqrt{n}}, \\ \alpha_{k,m} &= \rho_k^{-2}\left(2\delta_{k-1}\tau_{k,m} + \delta_{k-1}^2\right), \quad m = 1, \ldots, d \end{aligned} \tag{7.11}$$

with $\alpha_k^* = \max_{m=1,\ldots,m} \alpha_{k,m}$, $\tau_{k,m} = \rho_k\sqrt{v_m^*}\left(\rho_k^2 + v_m^*\right)^{-1/2} \le \min\{\rho_k, \sqrt{v_m^*}\}$, and with $\overline{\psi}$ defined in Proposition 7.1 and z_n in Corollary 7.1.

We will need the following two lemmas proofs of which require only minor modifications in the proofs of Lemmas 4 and 5 from HJS.

Lemma 7.5. *For n sufficiently large, the $\boldsymbol{\alpha}_k$'s satisfy $\max\limits_{k\le k_n} \alpha_k^* < 1/4$. In addition, for the last iteration k_n, it holds*

$$\mu_n := \frac{C_g\, C_V^{1/2}}{(1-\alpha_{k_n}^*)}\, h_{k_n}\, \rho_{k_n}^2 + \frac{\sqrt{L}\sigma\overline{\psi} C_{\alpha_{k_n},n}\,|\boldsymbol{\alpha}_{k_n}|}{h_{k_n}\sqrt{n}} \le C\left(\sigma^2 n^{-1} L \log n\right)^{2/3}$$

and $\delta_{k_n} \le \delta_n$, where δ_n is defined in (3.6) and C means a generic constant depending on d, M and the constants from Assumptions 1 through 5 only.

Proof. Note that $\alpha_{k,m} \le \delta_{k-1}^2/\rho_k^2$ for all $m \notin \mathcal{J}$ and $\alpha_{k,m} \le \delta_{k-1}^2/\rho_k^2 + 2\delta_{k-1}/\rho_k$ for $m \in \mathcal{J}$. The first assertion of the lemma easily follows from the fact that $h_k\rho_k$ decreases during iteration, cf. Lemma 4 of HJS.

Since the dimensionality of the nonlinear component is bounded by M, it follows

$$|\boldsymbol{\alpha}_k|^2 \le (d-M)\delta_{k-1}^4/\rho_k^4 + M\left(\delta_{k-1}^2/\rho_k^2 + 2\delta_{k-1}/\rho_k\right)^2.$$

Further, the inequality $|\boldsymbol{\alpha}_{k_n-1}| \le C_1$ with some constant C_1 depending on d and M only implies in view of $h_{k_n-1} \ge 1/a_h$ and $1 \le \rho_{k_n-1} \left(\sigma^2 n^{-1} L \log n\right)^{-1/3} \le 1/a_\rho$ that

$$\delta_{k_n-1} \le C \left(\sigma^2 n^{-1} L \log n\right)^{1/2}, \qquad |\boldsymbol{\alpha}_{k_n}| \le C \left(\sigma^2 n^{-1} L \log n\right)^{1/6}.$$

Substituting this bound in the formula for μ_n yields by $h_{k_n} \ge 1$ and $\rho_{k_n} = \left(\sigma^2 n^{-1} L \log n\right)^{1/3}$ that $\mu_n \le C \left(\sigma^2 n^{-1} L \log n\right)^{2/3}$ and therefore

$$\delta_{k_n} \le \sqrt{2}\, C_V\, C_K^{1/2} \left(\sigma^2 n^{-1} L z_n^2\right)^{1/2} + C \left(\sigma^2 n^{-1} L \log n\right)^{2/3}.$$

□

Lemma 7.6. *Let n be sufficiently large. There exist random sets $\mathcal{A}_1 \supseteq \ldots \supseteq \mathcal{A}_{k_n-1}$ such that $\boldsymbol{P}(\mathcal{A}_k) \ge 1 - \frac{3k}{n}$ and it holds on $\mathcal{A}_k$*

$$\max_{m=1,\ldots,d} \left| P_{\rho_{k+1},m} \left(\widehat{\beta}_m^{(k)} - \beta_m^*\right)\right| \le \delta_k, \qquad k = 1, \ldots, k_n - 1.$$

Proof. We proceed by induction in k. First by (3.2) there exists a random set $\mathcal{A}_1$ with $\boldsymbol{P}(\mathcal{A}_1) \ge 1 - 1/n$ such that $\max_{m=1,\ldots,d} |\widehat{\beta}_1 - \beta^*| \le \delta_1$ on $\mathcal{A}_1$. This obviously implies

$$\max_{m=1,\ldots,d} |P_{\rho_2,m}(\widehat{\beta}_1 - \beta^*)| \le \delta_1.$$

Suppose now that there is $\mathcal{A}_{k-1}$ such that $\boldsymbol{P}(\mathcal{A}_{k-1}) \ge 1 - \frac{3(k-1)}{n}$ and it holds on $\mathcal{A}_{k-1}$:

$$\max_{m=1,\ldots,d} \left| P_{\rho_k,m} \left(\widehat{\beta}_m^{(k-1)} - \beta_m^*\right)\right| \le \delta_{k-1}.$$

Then on $\mathcal{A}_{k-1}$ by Lemma 7.4 $P^2_{\rho_k,m} |\widehat{v}_m^{(k-1)} - v_m^*| < \delta_{k-1}^2 + 2\delta_{k-1}\tau_{k,m}$ simultaneously for all $m = 1, \ldots, d$, and denoting $U^{(k)}$ a d-vector with components $U_m^{(k)} = P^2_{\rho_k,m}(1 + \rho_k^{-2} \widehat{v}_m^{(k-1)})$, one gets $U^{(k)} \in \mathcal{U}_{\boldsymbol{\alpha}_k}$.

By Corollary 7.1 there exists another random set A_k with $\boldsymbol{P}(A_k) \ge 1 - 3/n$ such that on A_k it holds for every $U \in \mathcal{U}_{\boldsymbol{\alpha}_k}$

$$\max_{m=1,\ldots,d} |P_{\rho_k,m}(\widehat{\beta}_m(U) - \beta_m^*)| \le \delta_k,$$

so that, with $\mathcal{A}_k = \mathcal{A}_{k-1} \cap A_k$, we obtain $\boldsymbol{P}(\mathcal{A}_k) \ge 1 - 3k/n$ and it holds on $\mathcal{A}_k$

$$\max_{m=1,\ldots,d} |P_{\rho_k,m}(\widehat{\beta}_m^{(k)} - \beta_m^*)| \le \delta_k.$$

and, since for every m $P_{\rho_{k+1},m} \le P_{\rho_k,m}$, the assertion follows. □

Let now $\mathcal{A}_{k_n-1}$ be the random set with $\boldsymbol{P}(\mathcal{A}_{k_n-1}) \ge 1 - \frac{3k_n-3}{n}$ shown in Lemma 7.6 so that on this set

$$\max_{m=1,\ldots,d} |P_{\rho_{k_n},m}(\widehat{\beta}_m^{(k_n-1)} - \beta_m^*)| \le \delta_{k_n-1},$$

and for the corresponding d-vector $U^{(k_n)}$ with components $U_m^{(k_n)} = P^2_{\rho_{k_n},m}(1+\rho_{k_n}^{-2}\widehat{v}_m^{(k_n-1)})$, it holds $U^{(k_n)} \in \mathcal{U}_{\boldsymbol{\alpha}_{k_n}}$.

Let then ξ_m be the Gaussian L-vector with the components ξ_{lm} from Proposition 7.1 applied with $h = h_{k_n}$ and $\rho = \rho_{k_n}$. Due to this proposition, there exists a random set A_{k_n} with $\boldsymbol{P}(A_{k_n}) \geq 1 - 2/n$, so that on A_{k_n} it holds for all $U \in \mathcal{U}_{\boldsymbol{\alpha}_{k_n}}$:

$$\max_{m=1,\ldots,d} |P_{\rho_{k_n},m}(\widehat{\beta}_m(U) - \beta^*_m) - \frac{\xi_m}{h\sqrt{n}}| \leq \mu_n,$$

where μ_n is defined in Lemma 7.5. This yields for the set $\mathcal{A}_{k_n} = \mathcal{A}_{k_n-1} \cap A_n$ that $\boldsymbol{P}(\mathcal{A}_{k_n}) \geq 1 - \frac{3k_n-1}{n}$ and the final estimator $\widehat{\beta}_m = \widehat{\beta}_m^{(k_n)}$ satisfies on $\mathcal{A}_{k_n}$:

$$\max_{m=1,\ldots,d} \left| P_{\rho_{k_n}}(\widehat{\beta}_m - \beta^*) - n^{-1/2}\xi^*_m \right| \leq \mu_n$$

where $\xi^*_m = h^{-1}\xi_m$. In view of $h = h_{k_n} \geq 1$

$$\boldsymbol{E}|\xi^*_{lm}|^2 = h^{-2}\boldsymbol{E}|\xi_{lm}|^2 \leq 2\sigma^2 C_V^2 C_K$$

and the first two claims in (3.5) follow from Lemma 7.5. The last claim in (3.5) follow by applying Lemma 7 of HJS and Lemma 7.4. The first two inequalities in (3.7) follow from (3.5) by setting $P_{\rho,m} = 1$ and $\beta^*_m = 0$. The last one is proved similarly to Lemma 7.4.

7.7 Proof of Theorem 3.2

The proof can be done similarly to Spokoiny (2002) using the bound for the bias of estimation from the proof of Proposition 7.1. We omit the details to save the space.

7.8 Proof of Theorem 4.1

In view of Theorem 3.1 on the set A, it holds $\widehat{v}_m \leq \delta_n^2$ for all $m \notin \mathcal{J}$. Therefore, it suffices to show that on A, it holds $\widehat{v}_m > r^2\delta_n^2$ for every $m \in \mathcal{J}$. Next, by Theorem 3.1 again, for $m \in \mathcal{J}$

$$\widehat{v}_m > v^*_m - P_\rho^{-2}\left(\delta_n^2 + 2\delta_n P_\rho v^*_m\right) = v^*_m - \delta_n^2(1 + v^*_m\rho^{-2}) - 2\delta_n(1 + v^*_m\rho^{-2})^{1/2}v^*_m.$$

Define $s^2 = v^*_m/\delta_n^2$ and $u_n = \delta_n/\rho$. Then, on A,

$$\delta_n^{-2}\widehat{v}_m > s^2 - 1 - s^2u_n^2 - 2s(1 + s^2u_n^2)^{1/2} \geq s^2(1 - u_n^2 - 2u_n) - 1 - 2s.$$

It is straightforward to check that the right hand-side of this inequality as a function of s is greater than r^2 for all $s \geq s_r$. Therefore, on A, $\delta_n^{-2}\widehat{v}_m > r^2$ for $m \in \mathcal{J}$ as required.

7.9 Proof of Theorem 4.2

To simplify the exposition, we suppose that the resampling scheme of Section 4.1 utilizes the true variance σ^2 instead of the estimated variance $\widehat{\sigma}^2$. This assumption is easily justified by the result of Theorem 3.2 claiming root-n consistent estimation of σ^2 by $\widehat{\sigma}^2$.

The idea of the proof is to show that the variable $\widehat{v}_{(M+1)}$ and the similarly defined variable $\widetilde{v}_{(M+1)}$ for the resampling model have approximately the same distribution. Let A be the random set from Theorem 3.1 with $\boldsymbol{P}(A) \geq 1 - 3k_n/n$. It is obviously sufficient to show that

$$\boldsymbol{P}(\widehat{\mathcal{J}}_M \neq \mathcal{J} \mid A) \quad \leq \quad \alpha + 3/n.$$

We therefore suppose that the event A holds true. Then, under the assumptions of the theorem, the nonlinear component is correctly identified and all the bounds of Theorem 3.1 hold. Moreover, for every $m \notin \mathcal{J}$, the value $n\widehat{v}_m$ can be approximated by $|\xi_m^*|^2$, where the distribution of the vector ξ_m^* depend on the 'ideal' bandwidth $\boldsymbol{b}^* = \boldsymbol{b}^{*(k_n)}$, the kernel K, basis functions $\psi_{lm}(\cdot)$, and the design $X_1, \ldots, X_n$ only.

Next we consider the model we resample from. This artificial model has the same structure (i.e. the same linear and nonlinear components) and differs from the original one only by the parameters of the linear component (they are equal to zero in the resampling model) and by the nonlinear link function. More specifically, the estimators $\widehat{v}_m$ based on the original model are the "true" values for the resampling model and the last step bandwidth $\boldsymbol{b} = \boldsymbol{b}^{(k_n)}$ is the "ideal" bandwidth for the resampling model. Since the resampling model fulfills all the conditions that we impose on the original model, Theorem 3.1 (or Proposition 7.1 with $\boldsymbol{\alpha} = 0$ and $\boldsymbol{b} = \boldsymbol{b}^{(k_n)}$) continues to apply. This yields, in particular, that on a set $\widetilde{A}_M$ with $\boldsymbol{P}(\widetilde{A}_M) \geq 1 - 3/n$, the nonlinear component of the resampling model will be correctly identified. Moreover, due to Remark 7.1, every variable $n\widetilde{v}_m$ with $m \notin \mathcal{J}$ can be approximated by the squared norm of a Gaussian random vector with the same distribution as ξ_m^*. And thus, it is true for $n\widetilde{v}_{(M+1)}$. This yields, in particular, that the $(1-\alpha)$-quantile evaluated from the distribution of $n\widetilde{v}_{(M+1)}$ applies up to the approximation error to $n\widehat{v}_{(M+1)}$. It follows from Theorem 3.1 that the error of approximation of $n\widehat{v}_m$ by $|\xi_m^*|^2$ can be bounded by $n(\omega_n^2 + \omega_n\delta_n) \leq C' n^{-1/6}(\log n)^{5/6}$ for some constant C'. Therefore, at least for sufficiently large n, the approximation error is small and the assertion of the theorem follows.

7.10 Proof of Theorem 4.3

Let A be the random set described in Theorem 3.1 with $\boldsymbol{P}(A) \geq 1 - 3k_n/n$. In view of Theorem 4.1, it is sufficient to prove that $\boldsymbol{P}(\widehat{\mathcal{M}} \neq M \mid A) \leq \alpha + 3M/n$.

On A it holds $\widehat{v}_m \leq \delta_n^2$ for all $m \notin \mathcal{J}$ and $\widehat{v}_m > (r\delta_n)^2$ for all $m \in \mathcal{J}$ and $r = s_1$. Thus $\widehat{v}_{(\mathcal{M})} > (s_1\delta_n)^2$ for all $\mathcal{M} \leq M$ and $\widehat{v}_{(M+1)} \leq \delta_n^2$. For every $\mathcal{M} < M$, we resample from the model having precisely $\mathcal{M}$ nonlinear regressors with $\widehat{v}_m$ being the 'true' measure of nonlinearity for every $m \in \widehat{\mathcal{J}}_{\mathcal{M}}$.

Application of Propositions 7.1 and 7.2 with $\boldsymbol{\alpha} = 0$ to this artificial models and again Theorem 4.1 with $r = 1$ ensures that on a set $\widetilde{A}_{\mathcal{M}}$ with $\boldsymbol{P}(\widetilde{A}_{\mathcal{M}}) \geq 1 - 3/n$, every $\widetilde{v}_m$ for $m \notin \mathcal{J}_{\mathcal{M}}$ fulfills $\widetilde{v}_m \leq \delta_n^2$. Hence, $\widetilde{v}_{(\mathcal{M}+1)} \leq \delta_n^2$ on $\widetilde{A}_{\mathcal{M}}$ and the same holds for the $1-\alpha$ quantile of $\widetilde{v}_{(\mathcal{M}+1)}$ provided that $\alpha > 3/n$. Therefore, for every $\mathcal{M} < M$, the hypothesis $M \leq \mathcal{M}$ will be rejected on the intersection $A \cap \widetilde{A}_{\mathcal{M}}$. This yields

$$\boldsymbol{P}\left(\widehat{\mathcal{M}} < M \mid A\right) \leq 3(M-1)/n. \tag{7.12}$$

Next the definition of $\widehat{\mathcal{M}}$ implies the inclusion

$$\{\widehat{\mathcal{M}} > M\} \subseteq \{\widehat{v}_{(M+1)} > t_\alpha(M)\},$$

where $t_\alpha(M)$ is evaluated in the resampling procedure with $\mathcal{M} = M$. Applying now Theorem 4.2 we get, using also (7.12), the desired bound for $\boldsymbol{P}(\widehat{\mathcal{M}} \neq M)$, and the theorem follows.

References

[1] Bhattacharya, P.K. and Zhao, P.L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.*, **25**, 244–262.

[2] Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J. (1998). *Efficient and adaptive estimation for semiparametric models.* New York: Springer. xix, 560 p.

[3] Carroll, R.J., Fan, J., Gijbels, I., and Wand, M.P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.*, **92**, no. 438, 477–489.

[4] Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.*, **16**, 136–146.

[5] Chen, H. and Chen, K. W. (1991). Selection of the splined variables and convergence rates in a partial linear model. *Canadian J. Statist.*, **19**, 323–339.

[6] Cuzick, J. (1992a). Semiparametric additive regression. *J. Royal Statist. Soc., Series B*, **54**, 831–843.

[7] Cuzick, J. (1992b). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Ann. Statist.*, **20**, 1129–1136.

[8] Engle, R.F., Granger, C.W.J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Ass.*, **81**, 310–320.

[9] Eubank, R.L. and Spiegelman, C.H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Stat. Ass.*, **85**, 387–392.

[10] Eubank, R.L. and Hart, J.D. (1992). Testing goodness-of-fit in regression via order selection criteria. *Ann. Statist.*, **20**, no. 3, 1424–1425.

[11] Eubank, R.L., Kambour, E.L., Kim, J.T., Klipple, K., Reese, C.S. and Schimek, M. (1998). Estimation in partially linear models. *Computational Statistics Data Analysis*, **29**, 27–34.

[12] Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Ass.*, **91**, 674–688.

[13] Gao, J. T. (1995). Asymptotic theory for partly linear models. *Communications in Statistics, Theory Methods*, **24**, 1985–2010.

[14] Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

[15] Golubev, G. and Härdle, W. (2000). On the second order minimax estimation in a partly linear model. *Math. Methods Statist.*, **9**, 160–175.

[16] Hamilton, S.A. and Truong, Y.K. (1997). Local linear estimation in partly linear models. *J. Multivariate Analysis*, **60**, 1–19.

[17] Härdle, W. and Korostelev, A. (1996). Search of significant variables in noparametric additive regression. *Biometrika*, **83**, 541–549.

[18] Härdle, W., Liang, H., and Gao, J. (2000). *Partially linear models*. Physica Verlag and Springer Verlag, Germany.

[19] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **4**, 1926–1947.

[20] Härdle, W., Sperlich, S., and Spokoiny, V. (2001). Structural tests for additive regression. *J. Amer. Stat. Acc.*, **96**, no. 456, 1333–1347.

[21] Hart, J. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer Verlag, New York.

[22] Horowitz, J.L. and Spokoiny, V. (2001). An adaptive, rate-optimal test of a parametric mean regression model against a nonparametric alternative. *Econometrica*, **69**, 599–631.

[23] Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, **29**, no. 3, 595–623.

[24] Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.*, **29**, no. 6, 1537–1566.

[25] Heckman, N.E. (1986). Spline smoothing in partly linear models. *J. Royal Statist. Soc., Series B*, **48**, 244–248.

[26] Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Stat. Ass.*, **89**, no. 427, 1000–1005.

[27] Mammen, E. and van de Geer, S. (1997). Penalized estimation in partial linear models. *Ann. Statist.*, **25**, 1014–1035.

[28] Rice, J. (1986). Convergence rates for partially splined models. *Statistics Probability Letters*, **4**, 203–208.

[29] Robinson, P.M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56**, 931–954.

[30] Schick, A. (1996a). Weighted least squares estimates in partly linear regression models. *Statistics Probability Letters*, **27**, 281–287.

[31] Schick, A. (1996b). Root-n consistent estimation in partly linear regression models. *Statistics Probability Letters*, **28**, 353–358.

[32] Schimek, M. (2000). Estimation and inference in partially linear models with splines. *J. Statistical Planning Inference*, **91**, 525–540.

[33] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. R. Statist. Soc., Series B*, **50**, 413–436.

[34] Sperlich, S. (1998). *Additive modelling and testing model specification*. Aachen, Shaker.

[35] Spokoiny, V. (2001). Data driven testing the fit of linear models. *Math. Methods of Statistics*, **10**, no. 4, 465–497.

[36] Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. of Multivariate Analysis*, pp. 1–23 (doi:10.1006/jmva.2001.2023).

[37] Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.*, **25**, no. 2, 613–641.