Multivariate orthogonal series estimates for random design regression

Michael Kohler*

Department of Mathematics, University of Saarbrücken, Postfach 151150, D-66041 Saarbrücken, Germany, email: kohler@math.uni-sb.de

Abstract

In this paper a new multivariate regression estimate is introduced. It is based on ideas derived in the context of wavelet estimates and is constructed by hard thresholding of estimates of coefficients of a series expansion of the regression function. Multivariate functions constructed analogously to the classical Haar wavelets are used for the series expansion. These functions are orthogonal in $L_2(\mu_n)$, where μ_n denotes the empirical design measure. The construction can be considered as designing adapted Haar wavelets.

Bounds on the expected L_2 error of the estimate are presented, which imply that the estimate is able to adapt to local changes in the smoothness of the regression function and to the distribution of the design. This is also illustrated by simulations.

AMS classification: Primary 62G07; secondary 62G05.
Key words and phrases: hard thresholding, L_2 error, nonparametric regression, orthogonal series estimates, rate of convergence, regression estimation.

1 Introduction

1.1 Nonparametric regression. In *regression analysis* an $\mathbb{R}^d \times \mathbb{R}$-valued random vector (X, Y) with $\mathbf{E}Y^2 < \infty$ is considered and the dependency of Y on the value of X is of interest. More precisely, the goal is to find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(X)$ is a "good approximation" of Y. In the sequel we assume that the main aim of the analysis is

*Corresponding author. Tel: +49-681-302-2435, Fax: +49-681-302-6583

Running title: *Multivariate orthogonal series estimates*

minimization of the mean squared prediction error or L_2 *risk*

$$\mathbf{E}\{|f(X) - Y|^2\}. \tag{1}$$

In this case the optimal function is the so-called *regression function* $m : \mathbb{R}^d \to \mathbb{R}$, $m(x) = \mathbf{E}\{Y|X = x\}$. Indeed, let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary (measurable) function and denote the distribution of X by μ. Then

$$\begin{aligned} \mathbf{E}\{|f(X) - Y|^2\} &= \mathbf{E}\{((f(X) - m(X)) + (m(X) - Y))^2\} \\ &= \mathbf{E}\{|f(X) - m(X)|^2\} + \mathbf{E}\{|m(X) - Y|^2\} \\ &= \mathbf{E}\{|m(X) - Y|^2\} + \int |f(x) - m(x)|^2 \mu(dx). \end{aligned} \tag{2}$$

Here the second equation follows from

$$\mathbf{E}\{(f(X) - m(X)) \cdot (m(X) - Y)\} = \mathbf{E}\{(f(X) - m(X)) \cdot \mathbf{E}\{(m(X) - Y)|X\}\} = 0.$$

Since the integral on the right-hand side of (2) is always nonnegative, (2) implies that the regression function is the optimal predictor in view of minimization of the L_2 risk:

$$\mathbf{E}\{|m(X) - Y|^2\} = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbf{E}\{|f(X) - Y|^2\}. \tag{3}$$

In addition, any function f is a good predictor in the sense that its L_2 risk is close to the optimal value, if and only if the so-called L_2 *error*

$$\int |f(x) - m(x)|^2 \mu(dx) \tag{4}$$

is small. This motivates to measure the error caused by using a function f instead of the regression function by the L_2 error (4).

In applications, usually the distribution of (X, Y) (and hence also the regression function) is unknown. But often it is possible to observe a sample of the underlying distribution. This leads to the *regression estimation problem*. Here (X, Y), (X_1, Y_1), $\ldots$, (X_n, Y_n) are independent and identically distributed random vectors. The set of data

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

is given, and the goal is to construct an estimate

$$m_n(\cdot) = m_n(\cdot, \mathcal{D}_n) : \mathbb{R}^d \to \mathbb{R}$$

of the regression function such that the L_2 error

$$\int |m_n(x) - m(x)|^2 \mu(dx)$$

is small. For a detailed introduction to nonparametric regression we refer the reader to the monography Györfi et al. (2002).

1.2 Orthonormal series estimates. Orthonormal series regression estimates have been originally introduced in the context of regression estimation with fixed, equidistant design (see Donoho and Johnston (1994), Donoho et al. (1995), and the literature cited therein). In the sequel we motivate their use for random design regression problems.

Let $L_2(\mu)$ be the set of all square integrable functions $f : [0,1]^d \to \mathbb{R}$ with respect to μ. Let $\{f_j\}_{j\in\mathbb{N}}$ be a complete orthonormal system in $L_2(\mu)$, i.e., assume that

$$< f_j, f_k >_{L_2(\mu)} := \int f_j(x) f_k(x) \mu(dx) = \begin{cases} 1 & \text{if} \quad j = k, \\ 0 & \text{if} \quad j \neq k, \end{cases}$$

and that each function in $L_2(\mu)$ can be approximated arbitrarily exactly (with respect to $\| \cdot \|_{L_2(\mu)}$) by linear combinations of the $\{f_j\}_{j\in\mathbb{N}}$. Let m_n be an arbitrary estimate of the regression function m and assume $m, m_n \in L_2(\mu)$. Then analogously to the classical Fourier series expansion it can be shown that

$$m = \sum_{j=1}^{\infty} a_j \cdot f_j \quad \text{where } a_j =< m, f_j >_{L_2(\mu)} = \int m(x) \cdot f_j(x) \mu(dx)$$

and

$$m_n = \sum_{j=1}^{\infty} \hat{b}_j \cdot f_j \quad \text{where } \hat{b}_j =< m_n, f_j >_{L_2(\mu)} = \int m_n(x) \cdot f_j(x) \mu(dx).$$

Furthermore, the L_2 error of the estimate is equal to the sum of the squared distances between the coefficients of the two series expansions:

$$\int |m_n(x) - m(x)|^2 \mu(dx) = \int \left(\sum_{j=1}^{\infty} (\hat{b}_j - a_j) \cdot f_j(x) \right)^2 \mu(dx) = \sum_{j=1}^{\infty} (\hat{b}_j - a_j)^2. \tag{5}$$

This shows that it is important to construct regression estimates in such a way that the estimated coefficients $\hat{b}_j$ of the above series expansion of m are close to the actual coefficients a_j. And it motivates to consider so-called *orthonormal series regression estimates* defined by

$$m_n(\cdot) = \sum_{j \in J} \hat{a}_j \cdot f_j, \tag{6}$$

where $\{f_j\}_j$ is an orthonormal system in $L_2(\mu)$ (which we assume temporarily to be given), $\hat{a}_j = \hat{a}_j(\mathcal{D}_n)$ are estimates of the coefficients $a_j =< m, f_j >_{L_2(\mu)}$ based on the data $\mathcal{D}_n$ and J is a (usually finite) subset of $\mathbb{N}$.

A reasonable estimate of a_j is

$$\hat{a}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \cdot f_j(X_i), \tag{7}$$

which is an unbiased estimate of a_j:

$$\begin{aligned} \mathbf{E}\{\hat{a}_j\} &= \mathbf{E}\{\mathbf{E}\{\hat{a}_j|X_1,\ldots,X_n\}\} = \mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} m(X_i)\cdot f_j(X_i)\right\} \\ &= \int m(x)\cdot f_j(x)\,\mu(dx) = a_j. \end{aligned} \tag{8}$$

To motivate a good choice for J we consider for fixed $J \subseteq \mathbb{N}$ the expected L_2 error of the estimate m_n defined by (6) and (7). Using (5) and (8) we get

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mu(dx) = \mathbf{E}\left\{\sum_{j\in J}(\hat{a}_j - a_j)^2 + \sum_{j\notin J} a_j^2\right\} = \sum_{j\in J}\mathbf{Var}\{\hat{a}_j\} + \sum_{j\notin J} a_j^2.$$

If we assume for simplicity that $\mathbf{Var}\{\hat{a}_j\} = \frac{1}{n}\mathbf{Var}\{Y f_j(X)\}$ does not depend on j, then

$$\mathbf{E}\int |m_n(x) - m(x)|^2 \mu(dx) = const\cdot|J| + \sum_{j\notin J} a_j^2.$$

So, including an index j in the set J increases the expected L_2 error by the constant *const*, while not including it increases it by a_j^2. As a consequence, the expected L_2 error is minimal if we choose

$$J = J_{optimal} = \left\{j \in \mathbb{N} : |a_j|^2 > const\right\}.$$

Clearly, this choice is not possible in an application, because it depends on the unknown coefficients a_j. But what can be done in an application is to approximate it by

$$\hat{J} = \left\{j \in \mathbb{N} : |\hat{a}_j|^2 > const\right\}.$$

The heuristic behind this choice is that even if not all of the estimates $\hat{a}_j$ are accurate, they will hopefully at least have the same order of magnitude as the a_j's.

This leads to so-called *hard-thresholding* orthogonal series estimates

$$m_n(x) = \sum_{j=1}^{K} \eta_\delta(\hat{a}_j) \cdot f_j(x),$$

where

$$\eta_\delta(\hat{a}_j) = \begin{cases} \hat{a}_j & \text{if} \quad |\hat{a}_j| > \delta, \\ 0 & \text{if} \quad |\hat{a}_j| \leq \delta, \end{cases} \tag{9}$$

$\delta \in \mathbb{R}_+$ is a parameter of the estimate (so-called threshold), and K is usually chosen to be approximately equal to n.

Until now we have worked under the assumption that the orthonormal system $\{f_j\}_j$ in $L_2(\mu)$ is given. Clearly, this is not a reasonable assumption in most applications, because there the distribution of (X, Y), and hence also the distribution of X, is unknown. And even if the distribution of X is known, it is not obvious what a proper choice for the orthonormal system is.

There is one special situation, where it is easy to choose an orthonormal system in $L_2(\mu)$: If X is uniformly distributed on $[0,1]^d$, then one needs an orthonormal system in $L_2(\lambda)$, where λ is the Lebesgue measure on $[0,1]^d$, and as was shown, e.g., in Donoho and Johnstone (1994) and Donoho et al. (1995), in this case the use of wavelet systems leads to estimates which have many nice properties.

Motivated by the success of these estimates for this special case, it was suggested to use also for more general design measures orthonormal systems in $L_2(\lambda)$ and not in $L_2(\mu)$. But there are two problems in the above considerations if $\{f_j\}_j$ is an orthonormal system in $L_2(\lambda)$ and X is not uniformly distributed on $[0,1]^d$. Firstly, in this case the estimate (7) is no longer reasonable (in particular it is no longer unbiased). But as was shown, e.g., in Neuman and Spokoiny (1995), Hall and Turlach (1997), and Kovac and Silverman (2000), even then reasonable estimates of $a_j = \int m(x) f_j(x) dx$ can be constructed. But secondly, and more important, in addition relation (5) does no longer hold. It is this relation, which ensures that it makes sense to estimate the coefficients of a series expansion of the regression function. Obviously, to motivate this it is not necessary that the L_2 error is exactly equal to the sum of the squared differences between the coefficients and its estimates. It suffices, that the L_2 error is bounded from above and from below by some constant times the latter term. This in turn is satisfied if the distribution of X has a

density with respect to the Lebesgue measure, which is bounded away from zero and infinity on $[0,1]^d$. So under the last assumption the above considerations can be (and have already been) modified (see, e.g., Neuman and Spokoiny (1995) and Hall and Turlach (1997)). But in many multivariate applications some of the components of X are discrete so that X cannot have a density with respect to the Lebesgue measure. Therefore we want in the sequel to avoid to assume that a density of X exists.

1.3 Description of the main results. In this paper we use a different approach to apply the above ideas to regression estimation problems with general design measures. The basic idea is to estimate the distribution of X by the empirical distribution

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}} \quad (A \subseteq \mathbb{R}^d),$$

and to use an orthonormal system in $L_2(\mu_n)$. This orthonormal system is constructed by using ideas from the classical Haar wavelets.

Let $f_1, \ldots, f_K : \mathbb{R}^d \to \mathbb{R}$ be the piecewise constant orthonormal system in $L_2(\mu_n)$ defined in Section 2 below. Due to the orthonormality of the functions the best approximation (with respect to $\|\cdot\|_{L_2(\mu_n)}$) of the regression function by a linear combination of these functions is given by

$$\sum_{j=1}^{K} < m, f_j >_{L_2(\mu_n)} \cdot f_j. \tag{10}$$

We estimate

$$< m, f_j >_{L_2(\mu_n)} = \frac{1}{n}\sum_{i=1}^{n} m(X_i) f(X_i)$$

by

$$\hat{a}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i f(X_i), \tag{11}$$

and by hard-thresholding of the estimated coefficients we define the estimate

$$\tilde{m}_n = \sum_{j=1}^{K} \eta_\delta(\hat{a}_j) \cdot f_j. \tag{12}$$

Here the threshold $\delta > 0$ is a parameter of the estimate and η_δ is the hard-thresholding defined in (9).

The main theoretical result of this paper is Theorem 1 below, in which we derive an upper bound on the expected L_2 error of a truncated version m_n of the estimate $\tilde{m}_n$. In

this bound the estimate m_n is compared with an ideal piecewise constant least squares estimate, where the underlying partition is chosen in an optimal way for the distribution of (X, Y). Such an optimal choice of the partition is never computable in an application, but as is shown in Theorem 1 below the expected L_2 error of the estimate m_n is bounded from above by some logarithmic factor times a term which is approximately equal to the L_2 error of this ideal estimate. Since for this ideal estimate a partition can be chosen, which is especially fine in areas, where the regression function changes a lot or where the integration in the L_2 error gives large weight to the pointwise error, this theoretical result indicates that the estimate m_n is able to adapt to local changes in the smoothness of the regression function and to the distribution of the design. By applying the estimate to simulated data we show that this is (at least in the examples which we consider) indeed true.

In addition, we conclude from Theorem 1 that the estimate achieves (up to some logarithmic factor) the optimal rate of convergence for Lipschitz continuous regression function. Here the estimate is able to automatically adapt to the Lipschitz constant. For univariate X we can improve this result provided that X has a bounded density: In this case the estimate achives for regression functions, which have finitely many jump points and are otherwise Lipschitz continuous, again up to some logarithmic factor the optimal rate of convergence for Lipschitz continuous regression functions. Here the estimate is able to adapt to the Lipschitz constant and to the location of the jump points.

1.4 Discussion of related results. As described above there have been several attemps to generalize wavelet estimates from regression estimation with fixed equidistant design to random design regression (which is difficult because the random design will be in applications usually neither equidistant nor univariate). In most of them the orthonormal system is chosen as for fixed equidistant design and the way of estimating the coefficients has been adjusted to the design (see, e.g., Antoniadis, Gijbels and Grégoire (1997), Antoniadis, Grégoire and Vial (1997), Hall and Turlach (1997), Kovac and Silverman (2000) and Neuman and Spokoiny (1995)).

In this article we use the idea of adapting the wavelets to the random design. This idea was already used in Kohler (2000, 2003) and in Delouille, Franke and von Sachs (2001) in case of univariate X. In Kohler (2000, 2003) the adaptation of the wavelets to the random

design was done by constructing orthonormal systems consisting of piecewise polynomials of some fixed degree M. For $M = 0$ the resulting orthonormal system can be considered as design adapted Haar wavelets and is the same as the one used in Delouille, Franke and von Sachs (2001). In Kohler (2000, 2003) and in Delouille, Franke and von Sachs (2001) the position of the jump points of these Haar wavelets are chosen as quantiles of the empirical design measure. Unfortunately, it doesn‘t seem possible to generalize this idea to multivariate X. The reason is that for multivariate X a partition into tensor products of intervals does in general not satisfy that each set in it contains the same number of data points. Therefore we use in this article a different idea, namely to choose the jump points as in case of the classical Haar wavelets but to adapt the values of the functions to the empirical design measure. As in the univariate case we adapt our orthonormal system to the design points, however in contrast to the univariate case our wavelets are piecewise constant with respect to a cubic (and data-independent) partition. This allows us to define multivariate orthonormal systems and hence also multivariate estimates, while the approach in Kohler (2000, 2003) and in Delouille, Franke and von Sachs (2001) leads only to univariate estimates. In addition, in contrast to the articles cited just before, the approach used here allows us also to adapt the basis to the regression function, e.g., to choose in areas, where the regression function chances a lot, a "finer" basis than elsewhere (cf. Section 4). The multivariate Haar wavelets which we use can be considered as special cases of the unbalanced Haar wavelets introduced in Girardi and Sweldens (1997).

Our main result (Theorem 1 below) is similar to Theorem 2 in Kohler (2003). However, the choice of the partition of the ideal least squares estimate, which is compared to the orthogonal series estimate, is for univariate X in Theorem 1 below much more restricted than in Theorem 2 in Kohler (2003). As a consequence, the derived rate of convergence result for piecewise Lipschitz continuous regression function requires in this article the existence of a bounded density of the design measure with respect to the Lebesgue-Borel measure, while this is not necessary for the corresponding result in Kohler (2003). On the other hand, this time the result is also valid for multivariate X.

Comparing our results with the theoretical results in Delouille, Franke and von Sachs (2001) we see again that the estimate in Delouille, Franke and von Sachs (2001) works only for univariate X and that, in addition, the derived rate of convergence there, which is

similar to the result in Corollary 1 below, requires the existence of a density (with respect to the Lebesgue-Borel measure) of the design measure, which is not required in Corollary 1.

The proof of our main result is based on the fact that the estimate m_n minimizes a penalized empirical risk:

$$\frac{1}{n}\sum_{i=1}^{n}|\tilde{m}_n(X_i)-Y_i|^2+|\{j:|\hat{a}_j|>\delta\}|\cdot\delta^2$$

$$=\min_{J\subseteq\{1,\ldots,K\}}\left(\min_{a_j:j\in J}\frac{1}{n}\sum_{i=1}^{n}|\sum_{j\in J}a_jf_j(X_i)-Y_i|^2+|J|\cdot\delta^2\right). \tag{13}$$

This property of the estimates follows from Section 4 in Kohler (2003). In the theory of least squares estimates these kind of estimates are quite well understood (for results concerning estimates which minimize such a penalized empirical risk, see, e.g., Barron, Birgé and Massart (1999), Kohler (1998), Krzyżak and Linder (1998), van de Geer (2001), and the literature cited therein). In the proof we will use a result from van de Geer (2001), in which orthogonal series estimates are analyzed in a fixed design regression setting. This result will more or less directly imply a fixed design regression version of our main result.

The above connection between orthogonal series estimates and (penalized) least squares estimates was already used in Donoho (1997). There piecewise constant least squares estimates were analyzed by using results derived for orthogonal series estimates. As in Kohler (2000, 2003) we use this connection in this article in the opposite direction to analyze orthogonal series estimates by the aid of results for least squares estimates.

1.5 Notation. $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{R}_+$ are the sets of natural, real and nonnegative real numbers, resp. For $x\in\mathbb{R}$ we denote the smallest integer greater than or equal to x by $\lceil x\rceil$. $I_{\{x\in A\}}$ denotes the indicator function, $|A|$ the cardinality of a set A. The natural logarithm is denoted by $\log(\cdot)$, the logarithm with base two by $\log_2(\cdot)$, and the distribution of X is denoted by μ. The Euclidean norm of $x\in\mathbb{R}^d$ is denoted by $\|x\|$, the components of x are denoted by $x^{(1)},\ldots,x^{(d)}$.

1.6 Outline. An orthornormal system in $L_2(\mu_n)$ consisting of piecewise constant functions is defined in Section 2. The main theoretical results concerning rate of convergence of the estimates are described in Section 3 and proven in Section 5. Section 4 describes applications of the estimate to simulated data.

2 A multivariate orthormal system in $L_2(\mu_n)$

In this section we construct an orthonormal system in $L_2(\mu_n)$ consisting of piecewise constant functions, i.e., we construct piecewise constant functions $f_1, \ldots, f_K$ with the property

$$< f_j, f_k >_{L_2(\mu_n)} := \frac{1}{n} \sum_{i=1}^{n} f_j(X_i) \cdot f_k(X_i) = 0 \tag{14}$$

for $1 \leq j < k \leq K$ and

$$\|f_j\|^2_{L_2(\mu_n)} :=< f_j, f_j >_{L_2(\mu_n)} = \frac{1}{n} \sum_{i=1}^{n} f_j(X_i)^2 = 1 \tag{15}$$

for $1 \leq j \leq K$.

Clearly, the main difficulty here is to construct an orthogonal system in $L_2(\mu_n)$, i.e., to construct functions $f_1, \ldots, f_K$ with satisfy (14). From such functions an orthonormal system can be constructed by skipping all those functions with $\|f_j\|_{L_2(\mu_n)} = 0$ and by renormalizing the rest of them (i.e., by replacing each remaining function f by $f/\|f\|_{L_2(\mu_n)}$). Here functions with $\|f_j\|_{L_2(\mu_n)} = 0$ vanish on all x-values of the data points and hence do not contribute anything to the minimization of the empirical L_2 risk (defined as the first term on the left-hand side of (13)) of a linear combination of the orthogonal functions.

The construction of the orthogonal system is done analogously to the classical Haar-wavelets. This leads (up to some multiplicative constants) to the following orthogonal system in $L_2(\mu_n)$:

$$f_1(x) = \begin{cases} 1 & , x \in [0,1], \\ 0 & , \text{ else}, \end{cases}$$

$$f_2(x) = \begin{cases} \sum_{i=1}^n I_{\{X_i \in [1/2,1]\}} & , x \in [0,1/2), \\ -\sum_{i=1}^n I_{\{X_i \in [0,1/2)\}} & , x \in [1/2,1], \\ 0 & , \text{ else}, \end{cases}$$

$$f_3(x) = \begin{cases} \sum_{i=1}^n I_{\{X_i \in [1/4,1/2)\}} & , x \in [0,1/4), \\ -\sum_{i=1}^n I_{\{X_i \in [0,1/4)\}} & , x \in [1/4,1/2), \\ 0 & , \text{ else}, \end{cases}$$

$$f_4(x) = \begin{cases} \sum_{i=1}^n I_{\{X_i \in [3/4,1]\}} & , x \in [1/2,3/4), \\ -\sum_{i=1}^n I_{\{X_i \in [1/2,3/4)\}} & , x \in [3/4,1], \\ 0 & , \text{ else}, \end{cases}$$

and so on. Except f_1, all these functions are of the form

$$f(x) = \begin{cases} \sum_{i=1}^n I_{\{X_i \in B\}} & , x \in A, \\ -\sum_{i=1}^n I_{\{X_i \in A\}} & , x \in B, \\ 0 & , \text{ else,} \end{cases}$$

for some disjoint intervals A and B, and satisfy therefore

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n f(X_i) &= \frac{1}{n}\sum_{i=1}^n I_{\{X_i \in A\}} \cdot f(X_i) + \frac{1}{n}\sum_{i=1}^n I_{\{X_i \in B\}} f(X_i) + 0 \\ &= \frac{1}{n}\sum_{i=1}^n I_{\{X_i \in A\}} \cdot \sum_{i=1}^n I_{\{X_i \in B\}} + \frac{1}{n}\sum_{i=1}^n I_{\{X_i \in B\}} \cdot (-1) \cdot \sum_{i=1}^n I_{\{X_i \in A\}} \\ &= 0. \end{aligned}$$

As for the classical Haar wavelets, this together with the fact that for $j < k$ f_j is constant (maybe even constant zero) on the support of f_k, implies that f_1, f_2, ... are indeed orthogonal in $L_2(\mu_n)$.

In the sequel we use the same approach to construct a multivariate orthonormal system in $L_2(\mu_n)$. We start again with the indicator function of the cube $[0,1]^d$:

$$f_1(x) = \begin{cases} 1 & , \ x \in [0,1]^d, \\ 0 & , \text{ else,} \end{cases}$$

But instead of a subdivision of $[0,1]$ into two equidistant subintervals, we subdivide now $[0,1]^d$ into the 2^d equivolume cubes $B_1 \times B_2 \times \ldots \times B_d$, where B_i is either $[0,1/2)$ or $[1/2,1]$. As explained below this subdivision will give us additional $2^d - 1$ functions for the orthogonal system. After that we apply the same procedure recursively to each of the 2^d cubes constructed above, to get the next $2^d \cdot (2^d - 1)$ functions for the orthonormal system, and so on.

In the sequel we explain how one gets the $2^d - 1$ additional functions for the orthogonal system corresponding to a subdivision of a set $A_1 \times \ldots \times A_d$ (in the first step this set will be equal to $[0,1]^d$). We need the following notation: For an interval $A = [a,b]$ we define the two subintervals which one gets via equidistant subdivision of this interval by

$$A^L = [a, (a+b)/2) \quad \text{and} \quad A^R = [(a+b)/2, b].$$

Analogously we define for $A = [a,b)$

$$A^L = [a, (a+b)/2) \quad \text{and} \quad A^R = [(a+b)/2, b).$$

The functions which we construct are piecewise constant with respect to the partition

$$\left\{A_1^{M_1} \times \ldots \times A_d^{M_d} \ : \ M_i \in \{L, R\}\right\}$$

of $A_1 \times \ldots \times A_d$. Each function corresponds to one of the $2^d - 1$ sets

$$\begin{aligned}
&A_1 \times \ldots \times A_d, \quad A_1^L \times A_2 \times \ldots \times A_d, \quad A_1^R \times A_2 \times \ldots \times A_d,\\
&A_1^L \times A_2^L \times A_3 \times \ldots \times A_d, \quad A_1^L \times A_2^R \times A_3 \times \ldots \times A_d,\\
&A_1^R \times A_2^L \times A_3 \times \ldots \times A_d, \quad A_1^R \times A_2^R \times A_3 \times \ldots \times A_d,\\
&A_1^L \times A_2^L \times A_3^L \times A_4 \times \ldots \times A_d, \quad \ldots \quad , A_1^R \times \ldots \times A_{d-1}^R \times A_d.
\end{aligned}$$

Let $A_1^{M_1} \times \ldots \times A_{j-1}^{M_{j-1}} \times A_j \times \ldots \times A_d$ with $j \in \{1, \ldots, d\}$ and $M_1, \ldots, M_{j-1} \in \{L, R\}$ be one of these sets. We subdivide this set in the j-th component in two sets A and B, where

$$A = A_1^{M_1} \times \ldots \times A_{j-1}^{M_{j-1}} \times A_j^L \times \ldots \times A_d \tag{16}$$

and

$$B = A_1^{M_1} \times \ldots \times A_{j-1}^{M_{j-1}} \times A_j^R \times \ldots \times A_d. \tag{17}$$

Then the corresponding function is

$$f(x) = \begin{cases} \sum_{i=1}^n I_{\{X_i \in B\}} & , \ x \in A, \\ -\sum_{i=1}^n I_{\{X_i \in A\}} & , \ x \in B, \\ 0 & , \ \text{else.} \end{cases} \tag{18}$$

By construction, it satisfies

$$\frac{1}{n} \sum_{i=1}^n f(X_i) = 0. \tag{19}$$

Furthermore, if we choose two of those functions then one of them is constant on the support of the other, which together with (19) implies that they are orthogonal in $L_2(\mu_n)$. By the same argument we see that they are orthogonal to any function which is constant on $A_1 \times \ldots \times A_d$.

Let $f_1, \ldots, f_{2^d-1}$ be those functions (where $f_j = 0$ is possible). By construction, the linear span of

$$I_{\{x \in A_1 \times \ldots \times A_d\}}, f_1, \ldots, f_{2^d-1} \tag{20}$$

is a subset of the linear span of

$$\left\{ I_{\{x \in A_1^{M_1} \times \ldots \times A_d^{M_d}\}} \ : \ M_1, \ldots, M_d \in \{L, R\} \right\}. \tag{21}$$

Since the linear spans of $I_{\{x \in A\}}$ and $I_{\{x \in B\}}$, and of $I_{\{x \in A \cup B\}}$ and f, where A, B and f are defined in (16), (17) and (18), are equal (here we consider the functions as elements of $L_2(\mu_n)$ and identify two functions which have the same value at all x-components of the data), it follows furthermore that the two sets of functions are equal. So the linear span of the functions in (20) consists of all functions which are piecewise constant with respect to a partition constructed by subdivision of $A_1 \times \ldots \times A_d$ into 2^d equivolume cubes and which are zero outside from $A_1 \times \ldots \times A_d$.

To summarize, for arbitrary dimension d we construct the orthonormal system as follows: We start with $I_{\{x \in [0,1]^d\}}$. Then we construct as described above $2^d - 1$ additional functions corresponding to the subdivision of $[0,1]^d$ into 2^d equivolume cubes. For each of these cubes we construct additional $2^d - 1$ functions by subdividing it again. We recursively apply this procedure $k = \lceil \log_2(n)/d \rceil$ times, which gives as all together

$$1 + (2^d - 1) + 2^d(2^d - 1) + \ldots + (2^d)^{k-1}(2^d - 1) = (2^d)^k \approx n$$

orthogonal functions. Choosing k larger than above would lead to much more functions in the orthogonal system, which would imply that it is no longer possible to compute the orthonormal system in time $O(n \cdot \log(n))$.

To get an orthonormal system, we skip all those functions which vanish on all x-components of the data points and renormalize the rest of the functions such that each function has $L_2(\mu_n)$-norm one. This gives us an orthonormal system $\{f_j\}_{j=1,\ldots,K}$ (with $K \leq n$), which we will use for our orthogonal series estimate.

This orthogonal system can be used to represent special piecewise constant functions in an efficient way. These functions are piecewise constant with respect to partitions $\pi \in \cup_{k=1}^{n} \Pi_k$, where Π_k is recursively defined as follows: $\Pi_1 = \{\{[0,1]^d\}\}$ and Π_{k+1} is the set of all partitions which one obtains by choosing a partion of Π_k and by subdividing one of the sets of this partition into 2^d equivolume subsets. More precisely, Π_{k+1} consists of all partitions

$$\{\pi \setminus A_1 \times \ldots \times A_d\} \cup \left\{ A_1^{M_1} \times \ldots \times A_d^{M_d} : M_i \in \{L, R\} \right\}$$

where $\pi \in \Pi_k$, $A_1 \times \ldots \times A_d \in \pi$ and A_i are intervals of length greater than or equal to $2^{-\lceil \log_2(n)/d \rceil + 1}$.

For a partition π let $\mathcal{G}_c \circ \pi$ be the set of all piecewise constant functions with respect to that partition. As our next lemma shows, with the orthonormal system $\{f_j\}_{j=1,\ldots,K}$ one can represent in an efficient way functions from $\mathcal{G}_c \circ \pi$ for arbitrary partitions $\pi \in \cup_{k=1}^{n} \Pi_k$.

Lemma 1 *Let $\{f_j\}_{j=1,\ldots,K}$ be the orthonormal system (in $L_2(\mu_n)$) constructed above. Let $k \in \{1,\ldots,n\}$ and $\pi \in \Pi_k$ be arbitrary. Then there exist indices $j_1,\ldots,j_l \in \{1,\ldots,K\}$ such that*

$$span\{f_{j_1},\ldots,f_{j_l}\} = \mathcal{G}_c \circ \pi \quad in\ L_2(\mu_n) \quad and \quad l \leq |\pi|.$$

The proof of Lemma 1 will be given in Section 5.

3 Rate of convergence

In this section we present bounds on the expected L_2 error of the estimate. Throughout this section we will impose the following three regularity assumptions on the underlying distribution:

(A1) $X \in [0,1]^d$ *a.s.*,

(A2) $Y - m(X)$ is uniformly Sub-Gaussian, i.e.,

$$R^2 \mathbf{E}\left\{\exp((Y-m(X))^2/R^2) - 1 | X\right\} \leq \sigma_0^2 \quad a.s.$$

for some $R, \sigma_0 > 0$,

(A3) There exists a constant $L \in \mathbb{R}_+$ such that $|m(x)| \leq L$ for $x \in [0,1]^d$.

$(A1)$ requires that X takes on with probability one only values from some bounded set. By translating and rescaling of X we can assume w.l.o.g. that this bounded set is contained in $[0,1]^d$.

In $(A2)$ we impose a condition on the exponential moment of $Y - m(X)$. This condition is, e.g., satisfied if $Y - m(X)$ is bounded in absolute value by some constant $\beta > 0$ (take $R = \beta$ and $\sigma_0^2 = (e-1)\beta^2$), or if $Y - m(X)$ is independent of X and normally distributed with mean zero and variance σ^2 (take $R = 2\sigma$ and $\sigma_0^2 = 3\sigma^2$).

In $(A3)$ we assume that the regression function is bounded in absolute value by some known constant $L > 0$. If this is indeed true, then truncation of any estimate at $\pm L$ leads to an estimate with smaller L_2 error. We will denote this truncated estimate by $T_L \tilde{m}_n$, i.e., we will set

$$(T_L \tilde{m}_n)(x) = \begin{cases} L & , \tilde{m}_n(x) > L, \\ \tilde{m}_n(x) & , -L \leq \tilde{m}_n(x) \leq L, \\ -L & , \text{else.} \end{cases}$$

It is known that many least squares estimates need some kind of truncation to be consistent in random design regression (cf., Problem 10.3 in Györfi et al. (2002)). We do not know whether this is also the case for the estimate considered in this paper.

In the next theorem we compare the L_2 error of our estimate with the L_2 error of an arbitary estimate which fits a piecewise constant function to the data, where the underlying partition is chosen from the set $\cup_{k=1}^n \Pi_k$ of partitions defined at the end of Section 2.

Let m_n be an arbitrary estimate constructed by fitting via least squares a piecewise constant function, which is defined with respect to a given partition $\pi \in \cup_{k=1}^n \Pi_k$, to the data. Clearly, such an estimate cannot approximate the regression function better than the "best" piecewise constant function in $\mathcal{G}_c \circ \pi$, which induces an (approximation) error of at least

$$\inf_{f \in \mathcal{G}_c \circ \pi} \int |f(x) - m(x)|^2 \mu(dx).$$

Furthermore, estimation of the $|\pi|$ function values of the piecewise constant function induces an additional error ("variance") of order

$$\frac{|\pi|}{n}.$$

Now assume that one has an oracle available, which produces in dependency of the underlying distribution of (X, Y) an ideal partition for the above estimate. The expected L_2 error of the resulting estimate will be of order

$$\min_{k \in \{1,\ldots,n\}} \inf_{\pi \in \Pi_k} \left\{ \frac{|\pi|}{n} + \inf_{f \in \mathcal{G}_c \circ \pi} \int |f(x) - m(x)|^2 \mu(dx) \right\}.$$

Clearly, such an estimate will never be applicable in practice, because there we do not have an oracle which helps us to choose the underlying partition in an optimal way. But as our next theorem shows, a truncated version of the estimate $\tilde{m}_n$ has (up to some logarithmic factor) this (optimal) lower bound as an upper bound for the expected L_2 error.

Theorem 1 *Let $L > 0$ be arbitrary. Set $m_n(x) = T_L \tilde{m}_n(x)$, where $\tilde{m}_n$ is the estimate defined via (11) and (12) with $\delta = \sqrt{c_1 \log(n)/n}$ and $c_1 > 0$ sufficiently large. Then there exists a constant $c_2 > 0$ which depends only on R, σ_0 and L such that*

$$\begin{aligned}&\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \\ &\leq 4 \min_{k \in \{1,\ldots,n\}} \min_{\pi \in \Pi_k} \left\{ c_1 \cdot \log(n) \cdot \frac{|\pi|}{n} + \inf_{f \in \mathcal{G}_c \circ \pi} \int |f(x) - m(x)|^2 \mu(dx) \right\} + \frac{c_2}{n}\end{aligned}$$

for all distributions of (X, Y) which satisfy $(A1)$, $(A2)$ and $(A3)$.

The error bound above depends on the quality of the approximation of the regression function by piecewise constant functions. If we impose smoothness assumptions on m, we can control the approximation error.

Definition 1 *Let $0 < p \leq 1$ and $C \in \mathbb{R}_+$. A function $f : [0,1]^d \to \mathbb{R}$ is called (p, C)-smooth if*

$$|f(x) - f(z)| \leq C \cdot \|x - z\|^p$$

for all $x, z \in [0,1]^d$.

Corollary 1 *Let $0 < p \leq 1$, $C \in \mathbb{R}_+$ and $L > 0$ be arbitrary. Set $m_n(x) = T_L \tilde{m}_n(x)$, where $\tilde{m}_n$ is the estimate defined via (11) and (12) with $\delta = \sqrt{c_1 \log(n)/n}$ and $c_1 > 0$ sufficiently large. Assume that the distribution of (X, Y) satisfies $(A1)$, $(A2)$ and $(A3)$, and that the regression function is (p, C)-smooth. Then there exists a constant $c_3 > 0$ which depends only on d, c_1, R, σ_0 and L such that for n sufficiently large*

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq c_3 \cdot C^{2d/(2p+d)} \cdot \left(\frac{\log(n)}{n} \right)^{2p/(2p+d)}.$$

Proof. Let the partitions π_l be recursively defined as follows: $\pi_1 = \{[0,1]^d\}$ and π_{l+1} is obtained from π_l by subdividing each set in π_l into 2^d equivolume cubes. So π_l consists of $2^{d(l-1)}$ cubes of side length $2^{-(l-1)}$.

Set $l = 1 + \lceil \log_2(C^{2/(2p+d)}(n/\log(n))^{1/(2p+d)}) \rceil$. By approximating m on each set of π_l by the value of m at the center of this set, we can conclude from Theorem 1 and the (p, C)–smooth property of m

$$\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx)$$

$$
\begin{aligned}
&\leq 4c_1 \cdot \log(n) \cdot \frac{|\pi_l|}{n} + 4 \inf_{f \in \mathcal{G}_c \circ \pi_l} \int |f(x) - m(x)|^2 \mu(dx) + \frac{c_2}{n} \\
&\leq 4c_1 \cdot \log(n) \cdot \frac{2^{d(l-1)}}{n} + 4 \sup_{x,z \in [0,1]^d, \|x-z\| \leq 2\sqrt{d} 2^{-l}} |m(x) - m(z)|^2 + \frac{c_2}{n} \\
&\leq 4c_1 \cdot \log(n) \cdot \frac{2^{d(l-1)}}{n} + 4 \cdot C^2 2^{-2l \cdot p} \cdot 2^{2p} \cdot d^p + \frac{c_2}{n} \\
&\leq c_3 \cdot C^{2d/(2p+d)} \cdot \left(\frac{\log(n)}{n} \right)^{2p/(2p+d)}
\end{aligned}
$$

for n sufficiently large. □

Let $d = 1$ and let $f : [0,1] \to \mathbb{R}$ be piecewise constant with respect to a partition π consisting of finitely many intervals. Then we can find a function $g \in \mathcal{G}_c \circ \tilde{\pi}$ with $\tilde{\pi} \in \Pi_{\log_2(n) \cdot (|\pi|+1)}$ such that g is equal to f except on some set with small Lebesgue measure (see proof of Corollary 2 below). If we assume that this set has also small μ measure, then we can derive error bounds for regression functions which are piecewise smooth according to the following definition.

Definition 2 *Let $0 < p \leq 1$, $C \in \mathbb{R}_+$ and let π be a partition of $[0,1]$ consisting of finitely many intervals. A function $f : [0,1]^d \to \mathbb{R}$ is called piecewise (p, C)-smooth with respect to π, if*

$$
|f(x) - f(z)| \leq C \cdot \|x - z\|^p \quad (x, z \in A)
$$

for all $A \in \pi$.

Corollary 2 *Let $0 < p \leq 1$, $C \in \mathbb{R}_+$, $L > 0$ and let π be a partition of $[0,1]$ consisting of finitely many intervals. Set $m_n(x) = T_L \tilde{m}_n(x)$, where $\tilde{m}_n$ is the estimate defined via (11) and (12) with $\delta = \sqrt{c_1 \log(n)/n}$ and $c_1 > 0$ sufficiently large. Let $d = 1$. Assume that the distribution of (X, Y) satisfies $(A1)$, $(A2)$ and $(A3)$, and that the regression function is piecewise (p, C)-smooth with respect to π, and that X has a density with respect to the Lebesgue-Borel measure which is bounded on $[0,1]$. Then there exists a constant $c_4 > 0$ which depends only on c_1, R, σ_0 and L such that for n sufficiently large*

$$
\mathbf{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq c_4 \cdot C^{2/(2p+1)} \cdot \left(\frac{\log(n)}{n} \right)^{2p/(2p+1)}.
$$

Proof. Let the partition π_l be recursively defined as in the proof of Corollary 1, and set $l = 1 + \lceil \log_2(C^{2/(2p+1)}(n/\log(n))^{1/(2p+1)}) \rceil$. Construct a partition $\tilde{\pi}$ from π by replacing

recursively all intervalls $A \in \pi_l$, which contain an endpoint of an interval of π, by the two intervals A^L and A^R, unless all those intervals have length less than $1/n$. Let $A_1, \ldots, A_N$, with $N \leq |\pi| + 1$, be those intervals. Using on $[0,1] \setminus \cup_{j=1}^N A_j$ the same approximation of m by piecewise constant functions as in the proof of Corollary 1, we get

$$\begin{aligned}
&\inf_{f \in \mathcal{G}_c \circ \tilde{\pi}} \int |f(x) - m(x)|^2 \mu(dx) \\
&\leq \sup_{x,z \in [0,1] \setminus (\cup_{j=1}^N A_j), |x-z| \leq 2 \cdot 2^{-l}} |m(x) - m(z)|^2 + (4L)^2 \mu\left(\cup_{j=1}^N A_j\right) \\
&\leq C^2 2^{-2l \cdot p} \cdot 2^p + (4L)^2 (|\pi| + 1) \cdot \sup_{x \in [0,1]} |f(x)| \cdot \frac{1}{n},
\end{aligned}$$

where f is the density of X. Application of Theorem 1 yields the desired result. □

Remark 1. It follows from Stone (1982) that for (p, C)–smooth (and thus especially for piecewise (p, C)–smooth) regressions functions no estimate can achieve a minimax rate better than $C^{2d/(2p+d)} n^{-\frac{2p}{2p+d}}$ (see also Chapter 3 in Györfi et al. (2002)). The estimate in Corollary 1 achieves this rate up to the logarithmic factor $(\log(n))^{\frac{2p}{2p+d}}$, although its definition depends not on the smoothness (measured by (p, C)) of the regression function.

Remark 2. It follows from the proof of Corollary 2 that the result is also valid, if the number of discontinuities of the regression function increases with growing sample size at a rate not faster than $O((\frac{n}{\log(n)})^{\frac{1}{2p+1}} / \log(n))$.

Remark 3. We want to stress that in Theorem 1 and Corollary 1 there is no assumption on the distribution of X besides boundedness, especially it is not required that X has a density with respect to the Lebesgue–Borel measure.

Remark 4. By using a efficient implementation for computing

$$\sum_{i=1}^n I_{\{X_i \in A\}} \quad \text{and} \quad \sum_{i=1}^n I_{\{X_i \in A\}} \cdot Y_i$$

for all sets (16) and (17) used in the construction of the orthonormal system, the estimate can be computed in time $O(n \cdot \log(n))$ for a sample of size n. Hence it is applicable also to very large data sets.

Remark 5. The results above require that $c_1 > 0$ is chosen sufficiently large depending on the constants L, R and σ_0 from **(A2)** and **(A3)**. In any application c_1 has to be chosen such that it depends only on the given data, e.g. by splitting of the sample (cf. Section 4).

Remark 6. In view of having better approximation properties for smoother functions it would be nice to construct orthogonal systems in $L_2(\mu_n)$ consisting of smooth functions. As far as the author knows, it is an open research problem whether this is possible for general multivariate design measures.

4 Applications to simulated data

In our applications we choose the threshold in a data-dependent way by splitting of the sample. We split the sample of size n in a learning sample of size $n_l < n$ and a testing sample of size $n_t = n - n_l$. We use the learning sample to define for a fixed value δ of the threshold an estimate $\tilde{m}_{n_l,\delta}$, and compute the empirical L_2 risk of this estimate on the testing sample. Since the testing sample is independent of the learning sample, this gives us an unbiased estimate of the L_2 risk of $\tilde{m}_{n_l,\delta}$. Then we choose δ by minimizing this estimate with respect to δ. Our choice of n_l and n_t is mostly ad hoc, but motivated by theoretical considerations which show that splitting of the sample gives an estimate which has an L_2 error bounded by some constant times the optimal L_2 error (i.e., the L_2 error of the estimate which threshold chosen in an optimal way), plus some *log*-factor divided by the size of the testing sample (cf. Hamers and Kohler (2003)). This indicates that n_t might be much smaller than n provided n is large. In the sequel we use $n = 4000$ and $n_t = 1000$.

In order to compute the L_2 error of our estimates, we use MC integration, i.e., we approximate

$$\int |\tilde{m}_n(x) - m(x)|^2 \mu(dx) = \mathbf{E}\{|\tilde{m}_n(X) - m(X)|^2 | X, \mathcal{D}_n\}$$

by

$$\frac{1}{N} \sum_{j=1}^{N} |\tilde{m}_n(\tilde{X}_j) - m(\tilde{X}_j)|^2,$$

where the random variables $\tilde{X}_1, \tilde{X}_2, \ldots$ are i.i.d. with distribution μ and independent of $\mathcal{D}_n$. In the sequel we use $N = 2000$.

In our first example we define the distribution of (X, Y) by

$$Y = m(X^{(1)}, X^{(2)}) + \epsilon,$$

where X is uniformly distributed on $[-1,1]^2$, $m(x,z) = 4 - 4x^2 + 4z^3$, and ϵ is standard normally distributed and independent of X. We choose $n = 4000$, $n_l = 3000$ and $n_t = 1000$. In Figure 1 we plot the regression function. Figure 2 shows the estimate $\tilde{m}_n$. By

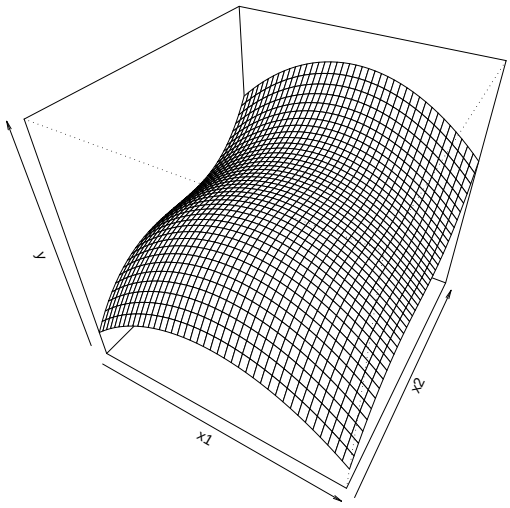

Figure 1: Regression function in first example.

MC integration with $N = 2000$ we get that the L_2 error of the estimate is approximately equal to 0.307.

The underlying partition used by the estimate $\tilde{m}_n$ in Figure 2 is finer at the border than in the center of $[-1,1]^2$. This shows that the estimate is able to adapt to the local behaviour of the regression function (which changes on the border of $[-1,1]^2$ more than in the center), and uses a especially fine partition in areas where the values of the regression function change a lot.

In our second example the regression function and Y are chosen as above, but X is with probability 0.4 uniformly distributed on $[0,1]^2$, and with probability 0.6 uniformly distributed on $[-1,1]^2 \setminus [0,1]^2$, so X gives $[0,1]^2$ twice as much probability than, e.g., $[-1,0]^2$.

Figure 3 shows the estimate for this data, again with $n = 4000$, $n_l = 3000$ and

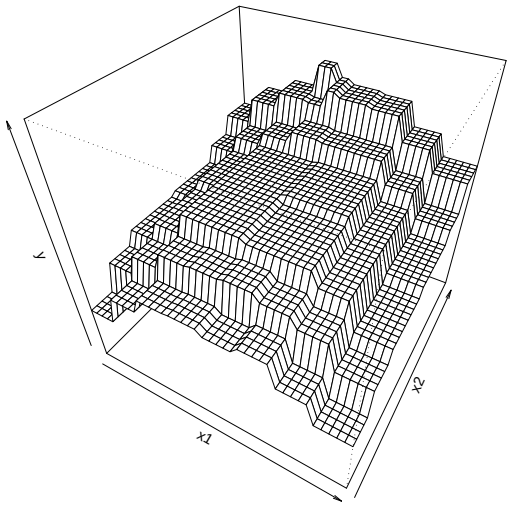

Figure 2: Estimate $\tilde{m}_n$ applied to data from the first example.

$n_t = 1000$. By MC integration with $N = 2000$ we get that the L_2 error of the estimate is approximately equal to 0.205. Figure 3 indicates that the estimate also adapts to the distribution of X in the sense that in areas with high μ-measure (which have especially large weight in the L_2 error) it tries to approximate the regression function especially well.

In our next two examples we choose

$$m(x, z) = \frac{10}{1 + 5x^2 + 5z^2}$$

and X and Y as in the first two examples. Again we use sample size $n = 4000$, $n_l = 3000$ and $n_t = 1000$. Figure 4 shows the regression function, in Figure 5 we see the estimate applied to the data with X uniformly distributed on $[-1, 1]^2$, and in Figure 6 we see the estimate for data with X chosen as in the second example above. Estimation of the L_2 error via MC integration gives 0.250 for the estimate in Figure 5 and 0.199 for the estimate in Figure 6.

Again we see in Figure 5, that in areas where the values of the regression function change a lot (e.g., away from the center and from the border of $[-1, 1]^2$) the estimate uses a finer partition than elsewhere. In addition, we see in Figure 6, that the approximation

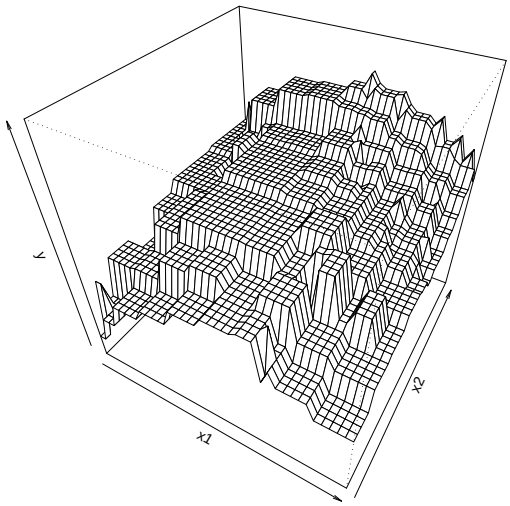

Figure 3: Estimate $\tilde{m}_n$ applied to data from the second example. Here the distribution of X gives $[0,1]^2$ and $[-1,1]^2 \setminus [0,1]^2$ probability 0.4 and 0.6 resp., so for the L_2 error it is important that the approximation on $[0,1]^2$ is better than on each of the sets $[-1,0]^2$, $[-1,0] \times [0,1]$ and $[0,1] \times [-1,0]$.

is on $[0,1]^2$ (where the measure μ has larger values) better than in the rest of $[-1,1]^2$.

In order to compare the estimates proposed in this paper with other nonparametric regression estimates we made a small simulation study analogously to the one in Beliakov and Kohler (2005). Here we define (X,Y) by

$$Y = m(X) + 0.2 \cdot \epsilon$$

for X uniformly distributed on $[-2,2]^d$ with $d \in \{1,2,3,4\}$, ϵ standard normally distributed and independent of X, and

$$m(x^{(1)}, \ldots, x^{(d)}) = \sum_{j=1}^{d} (-1)^{j+1} \cdot x^{(j)} \cdot \sin((x^{(j)})^2).$$

We compare our estimate with neural networks and regression trees (as implemented in R) by applying every one of these three estimates to 100 samples of size $n \in \{500, 3000\}$.

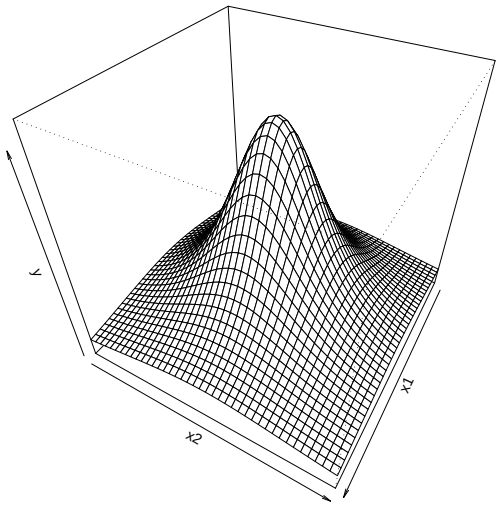

Figure 4: Regression function in the third example.

Table 1 below reports the mean and the standard deviation of Monte Carlo estimates of the corresponding L_2 errors of the estimates. Here the errors for neural networks and regression trees have been computed in Beliakov and Kohler (2005).

sample size	dimension	orthogonal series	neural networks	regression trees
500	1	0.018 (0.004)	**0.0019** (0.01)	0.05 (0.015)
500	2	0.350 (0.042)	**0.096** (0.05)	0.27 (0.01)
500	3	1.396 (0.101)	**0.53** (0.06)	0.88 (0.02)
500	4	1.962 (0.152)	**0.86** (0.50)	1.10 (0.01)
3000	1	0.005 (0.0008)	**0.0021** (0.02)	0.04 (0.01)
3000	2	**0.072** (0.006)	0.084 (0.06)	0.25 (0.01)
3000	3	0.670 (0.050)	**0.4**6 (0.05)	0.79 (0.01)
3000	4	1.757 (0.145)	**0.51** (0.4)	0.95 (0.01)

Table 1. Mean (and in brackets: standard deviation) of the L_2 error for the orthogonal series regression estimates, compared to L_2 error of Neural Networks and regression trees.

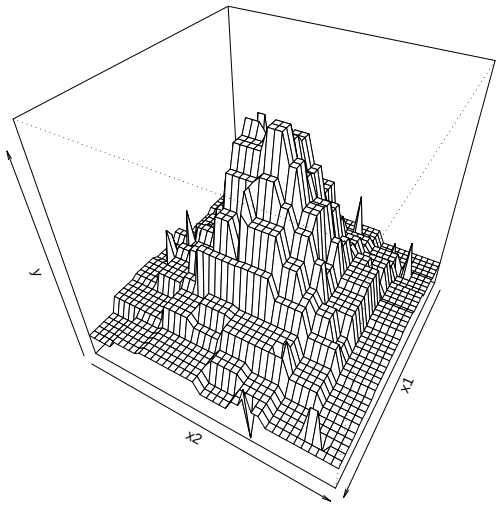

Figure 5: Estimate $\tilde{m}_n$ applied to the data from the third example.

Not surprisingly, Table 1 shows that our estimate behaves pourly if the sample size is small (i.e., for $n = 500$) or if the dimension is large (say, $d \geq 4$). The later point is due to the curse of dimensionality. However, for $n = 3000$ and $d \leq 3$ it behaves for the above distribution better than regression trees and comparable (for $d = 2$ even better) than neural networks. Here we consider for $d = 1$ also the standard deviation of the errors, which is for neural networks rather large.

From the above simulation one can expect that the newly proposed estimate is reasonable for large sample sizes and moderate dimensions.

5 Proofs

5.1 Proof of Lemma 1

Since we are interested only in the equality of the function spaces in $L_2(\mu_n)$, it suffices to proof the assertion for the orthogonal system constructed in Section 2, which we denote again by $\{f_j\}_{j=1,\ldots,K}$. We proceed by induction on k.

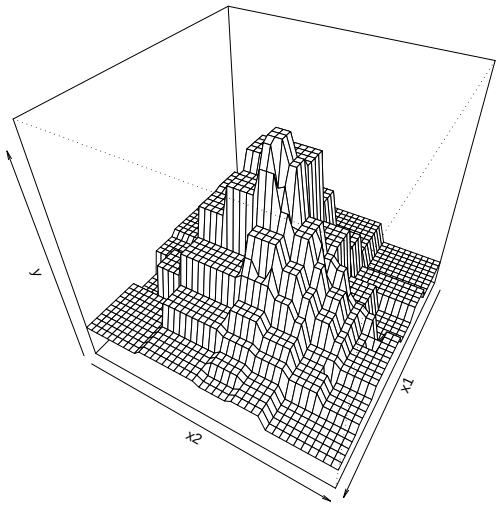

Figure 6: Estimate $\tilde{m}_n$ applied to data from the fourth example. Here the distribution of X gives $[0,1]^2$ and $[-1,1]^2 \setminus [0,1]^2$ probability 0.4 and 0.6 resp., so for the L_2 error it is important that the approximation on $[0,1]^2$ is better than on each of the sets $[-1,0]^2$, $[-1,0] \times [0,1]$ and $[0,1] \times [-1,0]$.

The assertion is trivial for $k = 1$, because in this case we have $\pi = \{[0,1]^d\}$ and $\mathcal{G}_c \circ \pi = span\{f_1\}$.

Let

$$\pi = \{\tilde{\pi} \setminus A_1 \times \ldots \times A_d\} \cup \{A_1^{M_1} \times \ldots \times A_d^{M_d} \,:\, M_i \in \{L, R\}\} \tag{22}$$

for some $\tilde{\pi} \in \Pi_k$ and $A_1 \times \ldots \times A_d \in \tilde{\pi}$ where the A_i's are intervals of length greater than or equal to $2^{-\lceil \log_2(n)/d \rceil + 1}$, and assume that the assertion holds for $\tilde{\pi}$. Then there exists $j_1, \ldots, j_l \in \{1, \ldots, K\}$ such that

$$span\{f_{j_1}, \ldots, f_{j_l}\} = \mathcal{G}_c \circ \tilde{\pi} \quad \text{and} \quad l \leq |\tilde{\pi}|.$$

By (22) we can conclude

$$span\left\{\{f_{j_1}, \ldots, f_{j_l}\} \cup \{I_{\{x \in A_1^{M_1} \times \ldots \times A_d^{M_d}\}} : M_i \in \{L, R\}\}\right\} = \mathcal{G}_c \circ \pi.$$

It follows from the construction of the orthogonal system in Section 2 (cf. proof of the equality of the linear spans of the functions in (20) and (21)) that there exists $k_1, \ldots, k_{2^d-1} \in \{1, \ldots, K\}$ such that

$$span\left\{I_{\{x \in A_1 \times \ldots \times A_d\}}, f_{k_1}, \ldots, f_{k_{2^d-1}}\right\} = span\left\{I_{\{x \in A_1^{M_1} \times \ldots \times A_d^{M_d}\}} : M_i \in \{L, R\}\right\}.$$

Now

$$I_{\{x \in A_1 \times \ldots \times A_d\}} \in \mathcal{G}_c \circ \tilde{\pi} = span\{f_{j_1}, \ldots, f_{j_l}\}$$

implies

$$span\{f_{j_1}, \ldots, f_{j_l}, f_{k_1}, \ldots, f_{k_{2^d-1}}\} = \mathcal{G}_c \circ \pi.$$

Because of

$$l + 2^d - 1 \leq |\tilde{\pi}| + 2^d - 1 = |\pi|,$$

the assertion follows. □

5.2 Proof of Theorem 1

Let $x_1, \ldots, , x_n \in \mathbb{R}^d$ and set $x_1^n = (x_1, \ldots, x_n)$. Define the distance $d_2(f, g)$ between $f, g : \mathbb{R}^d \to \mathbb{R}$ by

$$d_2(f, g) = \left(\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)|^2\right)^{\frac{1}{2}}.$$

Let $\mathcal{F}$ be a set of functions $f : \mathbb{R}^d \to \mathbb{R}$. An ϵ–cover of $\mathcal{F}$ (w.r.t. the distance d_2) is a set of functions $f_1, \ldots, f_k : \mathbb{R}^d \to \mathbb{R}$ with the property

$$\min_{1 \leq j \leq k} d_2(f, f_j) < \epsilon \quad \text{for all } f \in \mathcal{F}.$$

Let $\mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n)$ denote the size k of the smallest ϵ–cover of $\mathcal{F}$ w.r.t. the distance d_2, and set $\mathcal{N}_2(\epsilon, \mathcal{F}, x_1^n) = \infty$ if there doesn't exist any ϵ–cover of $\mathcal{F}$ of finite size.

In the proof of Theorem 1 we will need the following two auxiliary results.

Lemma 2 *Let $L \geq 1$, let $m : \mathbb{R}^d \to [-L, L]$ and let $\mathcal{F}$ be a class of functions $f : \mathbb{R}^d \to [-L, L]$. Let $0 < \epsilon < 1$ and $\alpha > 0$. Assume that*

$$\sqrt{n}\epsilon\sqrt{\alpha} \geq 1152L$$

and that, for all $x_1, \ldots, x_n \in \mathbb{R}^d$ *and all* $\delta \geq 2L^2\alpha$,

$$\frac{\sqrt{n}\epsilon\delta}{768\sqrt{2}L^2} \geq$$
$$\int_{\frac{\epsilon\delta}{128L^2}}^{\sqrt{\delta}} \left(\log \mathcal{N}_2 \left(\frac{u}{4L}, \left\{ f - m \ : \ f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} |f(x_i) - m(x_i)|^2 \leq \frac{\delta}{L^2} \right\}, x_1^n \right) \right)^{1/2} du. \quad (23)$$

Then

$$\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{\left|\mathbf{E}\{|f(X) - m(X)|^2\} - \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - m(X_i)|^2\right|}{\alpha + \mathbf{E}\{|f(X) - m(X)|^2\} + \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - m(X_i)|^2} > \epsilon \right\}$$
$$\leq 15 \exp\left(-\frac{n\alpha\epsilon^2}{512 \cdot 2304 L^2} \right).$$

Proof. See Lemma 5 in Kohler (2006). □

Lemma 3 *Let* $\mathcal{F}$ *be a set of functions* $f : \mathbb{R}^d \to \mathbb{R}$. *Assume that* $\mathcal{F}$ *is a linear vector space of dimension* D. *Then one has for arbitrary* $R > 0$, $u > 0$ *and* $x_1, \ldots, x_n \in \mathbb{R}^d$*:*

$$\mathcal{N}_2 \left(u, \left\{ f \in \mathcal{F} \quad : \quad \frac{1}{n}\sum_{i=1}^{n} |f(x_i)|^2 \leq R^2 \right\}, x_1^n \right) \leq \left(\frac{4R + u}{u} \right)^D.$$

Proof. See Corollary 2.6 in van de Geer (2000). □

Proof of Theorem 1. Set

$$\|f\|^2 = \int |f(x)|^2 \mu(dx),$$

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} |f(X_i)|^2$$

and

$$pen_n(\sum_{j=1}^{K} \alpha_j f_j) = c_1 \cdot \frac{\log(n) \cdot |\{j : \alpha_j \neq 0\}|}{n}.$$

We use the error decomposition

$$\int |m_n(x) - m(x)|^2 \mu(dx) = T_{1,n} + T_{2,n}, \quad (24)$$

where

$$T_{1,n} = \|m_n - m\|^2 - 2 \cdot \|m_n - m\|_n^2 - 2 \cdot pen_n(\tilde{m}_n)$$

and

$$T_{2,n} = 2 \cdot \left(\|m_n - m\|_n^2 + pen_n(\tilde{m}_n) \right).$$

In the first part of the proof we bound $\mathbf{E}T_{2,n}$. Because of $(A3)$ we have

$$|m_n(x) - m(x)| = |T_L \tilde{m}_n(x) - m(x)| \leq |\tilde{m}_n(x) - m(x)|$$

for all $x \in [0,1]^d$, which implies

$$T_{2,n} \leq 2 \cdot \left(\|\tilde{m}_n - m\|_n^2 + pen_n(\tilde{m}_n) \right).$$

By Lemma 3.8 in van de Geer (2001) and (13) we get for the latter term

$$\begin{aligned}
&\mathbf{E}\left\{ \|\tilde{m}_n - m\|_n^2 + pen_n(\tilde{m}_n) | X_1, \ldots, X_n \right\} \\
&\leq 2 \cdot \inf_{f = \sum_{j=1}^K \alpha_j f_j : \alpha_j \in \mathbb{R}} \left\{ c_1 \frac{\log(n) \cdot |\{j : \alpha_j \neq 0\}|}{n} + \frac{1}{n} \sum_{i=1}^n |f(X_i) - m(X_i)|^2 \right\} + \frac{c_5}{n}.
\end{aligned}$$

Application of Lemma 1 yields that the right-hand side above is bounded by

$$2 \min_{k \in \{1,\ldots,n\}} \inf_{\pi \in \Pi_k} \left\{ c_1 \frac{\log(n) \cdot |\pi|}{n} + \inf_{f \in \mathcal{G}_c \circ \pi} \frac{1}{n} \sum_{i=1}^n |f(X_i) - m(X_i)|^2 \right\} + \frac{c_5}{n}.$$

Summarizing the above results we get

$$\begin{aligned}
&\mathbf{E}\{T_{2,n}\} \\
&\leq 2 \cdot \mathbf{E}\left\{ \mathbf{E}\left\{ \|\tilde{m}_n - m\|_n^2 + pen_n(\tilde{m}_n) | X_1, \ldots, X_n \right\} \right\} \\
&\leq 4 \cdot \mathbf{E}\left\{ \min_{k \in \{1,\ldots,n\}} \inf_{\pi \in \Pi_k} \left\{ c_1 \frac{\log(n) \cdot |\pi|}{n} + \inf_{f \in \mathcal{G}_c \circ \pi} \frac{1}{n} \sum_{i=1}^n |f(X_i) - m(X_i)|^2 \right\} \right\} + \frac{c_5}{n} \\
&\leq 4 \cdot \min_{k \in \{1,\ldots,n\}} \inf_{\pi \in \Pi_k} \left\{ c_1 \frac{\log(n) \cdot |\pi|}{n} + \inf_{f \in \mathcal{G}_c \circ \pi} \int |f(x) - m(x)|^2 \mu(dx) \right\} + \frac{c_5}{n}. \qquad (25)
\end{aligned}$$

In the second part of the proof we bound $\mathbf{E}T_{1,n}$. By construction of the orthogonal system in Section 2, each function f_j is piecewise constant with respect to a partition of $[0,1]^d$ into three sets, where the first two sets are of the form (16) and (17) and f_j vanishes on the third set. Here the sets (16) and (17) do not depend on the data. In the construction of the orthonormal system there occur at most

$$2 + 2 \cdot (2^d - 1) + 2 \cdot 2^d (2^d - 1) + \ldots + 2 \cdot (2^d)^{k-1} (2^d - 1) = 2 \cdot (2^d)^k \leq 2^{d+1} n$$

different sets of this form. If we take two pairs of sets of the form (16) and (17), then they are either disjoint, or the sets of one of the pairs are contained in one set of the other pair. Hence for $j_1, \ldots, j_l \in \{1, \ldots, K\}$ arbitrary we have

$$span\{f_{j_1}, \ldots, f_{j_l}\} \subseteq \mathcal{F}_{2l+1},$$

where $\mathcal{F}_{2l+1}$ is the set of all functions which are piecewise constant with respect to a partition of $[0,1]^d$ consisting of $2l$ sets, which are constructed by choosing $2l$ sets of the form (16) and (17) and by intersecting each of these sets with the complements of all those of the $2l-1$ remaining sets which are contained in this set, and one additional set, on which the functions in $\mathcal{F}_{2l+1}$ vanish. Because of

$$m_n(x) = T_L\left(\sum_{j=1}^{K} \eta_\delta(\hat{a}_j) \cdot f_j(x)\right)$$

we can conclude

$$m_n(\cdot) \in \mathcal{F}_{2\cdot|\{j:\eta_\delta(\hat{a}_j)\neq 0\}|+1} \quad \text{and} \quad \|m_n\|_\infty \leq L.$$

Using this we get for $t > 0$ arbitrary:

$$\begin{aligned}
&\mathbf{P}\{T_{1,n} > t\} \\
&\leq \mathbf{P}\left\{\frac{\|m_n - m\|^2 - \|m_n - m\|_n^2}{t + 2\cdot pen_n(\tilde{m}_n) + \|m_n - m\|^2} > \frac{1}{2}\right\} \\
&\leq \sum_{k=1}^{2^d n} \mathbf{P}\left\{\exists f \in \mathcal{F}_{2k+1} : \|f\|_\infty \leq L \quad \text{and} \quad \frac{\|f - m\|^2 - \|f - m\|_n^2}{t + 2c_1 \cdot \frac{\log(n)\cdot k}{n} + \|f - m\|^2} > \frac{1}{2}\right\}. \quad (26)
\end{aligned}$$

To bound the above probabilities, we use Lemma 2. There are at most

$$\binom{2^{d+1}n}{2k} \leq (2^{d+1}n)^{2k}$$

possibilities to choose the $2k$ sets of the form (16) and (17) used in the definition of $\mathcal{F}_{2k+1}$. Therefore

$$\{f - m \,:\, f \in \mathcal{F}_{2k+1}, \|f\|_\infty \leq L\} \subseteq \{f + \alpha\cdot m \,:\, \alpha \in \mathbb{R}, f \in \mathcal{F}_{2k+1}, \|f + \alpha\cdot m\|_\infty \leq 2L\}$$

is a subset of a union of at most $(2^{d+1}n)^{2k}$ linear vector spaces of dimension $2k+1$. Using this together with Lemma 3 we get for arbitrary $u > c_6/n$ and arbitrary $x_1^n \subseteq (\mathbb{R}^d)^n$

$$\begin{aligned}
&\mathcal{N}_2\left(\frac{u}{4L}, \{f - m \,:\, f \in \mathcal{F}_{2k+1}, \|f\|_\infty \leq L\}, x_1^n\right) \\
&\leq (2^{d+1}n)^{2k} \cdot \left(\frac{4\cdot(2L) + u/(4L)}{u/(4L)}\right)^{2k+1} \\
&\leq (c_7 \cdot n)^{4k+1}.
\end{aligned}$$

Hence for $\delta \geq c_8/n$ (23) follows from

$$\frac{\sqrt{n}\delta/2}{768\sqrt{2}L^2} \geq \sqrt{\delta}\cdot((4k+1)\cdot\log(c_9\cdot n))^{1/2}.$$

The last inequality is in turn implied by

$$\delta \geq c_{10} \cdot \frac{\log(n) \cdot k}{n}.$$

Application of Lemma 2 with $\alpha = t + 2 \cdot c_1 \cdot \frac{\log(n) \cdot k}{n}$ yields for c_1 sufficiently large (i.e., for $2c_1 \geq c_{10}$)

$$\begin{aligned} \mathbf{P}\{T_{1,n} > t\} &\leq \sum_{k=1}^{2^d n} 15 \cdot \exp\left(-\frac{n/4}{512 \cdot 2304 \cdot L^2} \cdot \left(t + 2c_1 \cdot \frac{\log(n) \cdot k}{n}\right)\right) \\ &\leq c_{11} \cdot \exp\left(-\frac{n \cdot t}{c_{11}}\right). \end{aligned}$$

From this we get

$$\mathbf{E}\{T_{1,n}\} \leq \int_0^\infty \mathbf{P}\{T_{1,n} > t\}\, dt \leq \frac{c_{11}^2}{n}. \tag{27}$$

The assertion follows from (24), (25) and (27). □

Acknowledgments

The author wishes to thank two anonymous referees for many very helpful comments.

References

[1] Antoniadis, A., Gijbels, I., Grégoire, G., 1997. Model selection using wavelet decomposition and applications. *Biometrika* **84**, 751–763.

[2] Antoniadis, A., Grégoire, G., Vial, P., 1997. Random design wavelet curve smoothing. *Statistics and Prob. Letters* **35**, 225-232.

[3] Barron, A. R., Birgé, L., Massart, P., 1999. Risk bounds for model selection via penalization. *Probability Theory and Related Fields* **113**, 301-413.

[4] Beliakov, G. and Kohler, M. (2005). Estimation of regression functions by Lipschitz continuous functions. Submitted for publication.

[5] Delouille, V., Franke, J., von Sachs, R., 2001. Nonparametric stochastic regression with design–adapted wavelets. *Sankhya Series A* **63**, 328–366.

[6] Donoho, D., 1997. CART and best-ortho-basis: a connection. *Ann. Statist.* **25**, 1870–1911.

[7] Donoho, D., Johnstone, I.M., 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

[8] Donoho, D., Johnstone, I.M., Kerkyacharian, G., Picard, D., 1995. Wavelet Shrinkage: Asymptopia? *J. R. Statist. Soc. B.* **57**, 301–369.

[9] Girardi, M., Sweldens, W., 1997. A new class of unbalanced Haar wavelets that form a unconditional basis for L_p on general measure spaces. *J. Fourier Anal. Appl.* **3**, 457–474.

[10] Györfi, L., Kohler, M., Krzyżyak, A., Walk,H.. *A distribution–free theory of nonparametric regression*. Springer–Verlag, New York, 2002.

[11] Hall, P., Turlach, B. A., 1997. Interpolation methods for nonlinear wavelet regression with irregular spaced design. *Ann. Statist.* **25**, 1912–1925.

[12] Hamers, M., Kohler, M., 2003. A bound on the expected maximal deviation of averages from their means. *Statistics and Probability Letters* **62**, 137-144.

[13] Kohler, M., 1998. Nonparametric Regression Function Estimation Using Interaction Least Squares Splines and Complexity Regularization. *Metrika* **47**, 147-163.

[14] Kohler, M., 2000. *Analyse von nichtparametrischen Regressionsschätzern unter minimalen Voraussetzungen*. Habilitation thesis. Shaker Verlag, Aachen.

[15] Kohler, M., 2003. Nonlinear orthogonal series estimates for random design regression. *Journal of Statistical Planning and Inference* **115**, 491-520.

[16] Kohler, M., 2006. Nonparametric regression with additional measurements errors in the dependent variable. To appear in *Journal of Statistical Planning and Inference*, 2006.

[17] Kovac, A., Silverman, B.W., 2000. Extending the scope of wavelet regression methods by coefficient–dependent thresholding. *Journal of the American Statistical Association* **95**, 172–183.

[18] Krzyżak, A., Linder, T., 1998. Radial basis function networks and complexity regularization in function learning. *IEEE Transaction on Neural Networks* **9**, 247–256.

[19] Neumann, M.H., Spokoiny, V.G., 1995. On the efficiency of wavelet estimators under arbitrary error distributions. *Mathematical Methods of Statistics* **4**, 137–166.

[20] Stone, C. J., 1982. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

[21] van de Geer, S., 2000. *Empirical Processes in M–estimation.* Cambridge University Press.

[22] van de Geer, S., 2001. Least squares estimation with complexity penalties. *Mathematical Methods of Statistics* **10**, 355-374.