Weierstraß–Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Multivariate wavelet thresholding: a remedy against the curse of dimensionality?

Michael H. Neumann

submitted: 7th May 1996

Weierstrass Institute
for Applied Analysis
and Stochastics
Mohrenstraße 39
D – 10117 Berlin
Germany

Preprint No. 239
Berlin 1996

1991 Mathematics Subject Classification. Primary 62G07; Secondary 62G20.
Key words and phrases. Nonparametric curve estimation, multivariate wavelet estimators, nonlinear thresholding, curse of dimensionality, anisotropic wavelet basis, anisotropic smoothness classes, smoothness classes with dominating mixed derivatives, optimal rate of convergence.

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

Fax: + 49 30 2044975
e–mail (X.400): c=de;a=d400-gw;p=WIAS-BERLIN;s=preprint
e–mail (Internet): preprint@wias-berlin.de

ABSTRACT. It is well-known that multivariate curve estimation suffers from the "curse of dimensionality". However, reasonable estimators are possible, even in several dimensions, under appropriate restrictions on the complexity of the curve. In the present paper we explore how much appropriate wavelet estimators can exploit typical restrictions on the curve, which require a local adaptation to different degrees of smoothness in the different directions. It turns out that the application of a anisotropic multivariate basis, which has in contrast to the conventional multivariate resolution scheme a multidimensional scale parameter, is essential. Some simulations indicate the possible gains by this new method over thresholded estimators based on the multiresolution basis with a one-dimensional scale index.

1. INTRODUCTION

Multivariate curve estimation is often considered with some scepticism, because it is associated with the term of the "curse of dimensionality". This notion reflects the fact that nonparametric statistical methods lose much of their power if the dimension d is large. In the presence of r bounded derivatives, the usual optimal rate of convergence in regression or density estimation is $n^{-2r/(2r+d)}$, where n denotes the number of observations. To get the same rate as in the one-dimensional case, one has to assume a smoothness of order rd rather than r. This phenomenon can also be explained by the sparsity of data in high dimensions. If we have a uniformly distributed sample over the hypercube $[-1,1]^d$, then we will find only a fraction of about 2^{-d} of the data in the hypercube $[0,1]^d$.

Nevertheless, there is sometimes some hope for a successful statistical analysis in higher dimensions. Often the true complexity of a multivariate curve is much lower than it could be expected from a statement that the curve is a member of a certain Sobolev class $W_p^r(\mathbb{R}^d)$ with degree of smoothness r. Scott (1992, Chapter 7) claims: "Multivariate data in $\mathbb{R}^d$ are almost never d-dimensional. That is, the *underlying structure* of data in $\mathbb{R}^d$ is almost always of dimension lower than d." Even if this statement applies often not in this pure form, one has sometimes the situation that the variability in some of the directions is smaller than that described by a conservative multivariate smoothness assumption. This phenomenon can be adequately modelled by anisotropic smoothness classes, which therefore provide a good point of departure for rigorous mathematical considerations in this context. According to such an assumption, an appropriate smoothing method has to apply different degrees of smoothing in the various directions.

Another, even more restrictive, remedy to problems with high dimensionality consists in imposing additional *structural* assumptions which restrict the complexity of the curve. Well-established extreme cases in this direction are additive models and single index models. It is known that one can estimate in both cases the curve with a rate corresponding to the one-dimensional case; see, e.g., Stone (1985) and Härdle, Hall and Ichimura (1993). Compared to a high-dimensional nonparametric estimate, such additive functions are sometimes easier to interpret. It is clear that such a strong structural assumption is almost always inadequate, and one actually estimates some

kind of projection of the true function on the lower-dimensional functional class. Stone (1985) derived his results in this general setting of a possibly inadequate model. From the pure estimation point of view, this approach has an obvious drawback. Except for the rather rare cases that such structural assumptions are actually *exactly* fulfilled, such estimators are even not consistent as the sample size n tends to infinity. Hence, there is some motivation for a more flexible approach, which provides an effective dimension reduction if appropriate, but which leads at least to a consistent estimate in the general case.

Since the seminal papers by Donoho and Johnstone (1992) and Donoho, Johnstone, Kerkyacharian and Picard (1995) nonlinear wavelet estimators have developed to a widely accepted alternative to traditional methods like kernel or spline estimators. In particular, they are known to be able to successfully deal with spatially varying smoothness properties, which are summarized under the notion of "inhomogeneous smoothness". Assume we measure the loss in L_2. Inhomogeneous smoothness is then often modelled by Besov constraints, that is the unknown curve is assumed to lie in a Besov class $B^m_{p,q}(K)$ with $p < 2$. It is well-known that higher-dimensional wavelet bases can be obtained by taking tensor products of appropriately combined functions from one-dimensional bases. In almost all statistical papers the authors used an isotropic multiresolution construction, where one-dimensional basis functions coming from the same resolution scale are combined with each other. However, it was shown in Neumann and von Sachs (1995) for the special case of two-dimensional anisotropic Sobolev classes that this basis does not provide an optimal data compression if different degrees of smoothness are present in the two directions. Accordingly, the commonly used coordinatewise thresholding approach does not provide the optimal rate of convergence in such a case. Neumann and von Sachs (1995) proposed an alternative construction of a higher-dimensional basis, which involves tensor products of one-dimensional basis functions from different resolution scales, too. It was shown in the abovementioned special case that a thresholded wavelet estimator based on this basis can really adapt to different degrees of smoothness in different directions and can attain the optimal rate of convergence. In Section 2 we extend these results to higher dimensions and to Besov constraints, which admit also fractional degrees of smoothness.

In Section 3 we study another situation, which more implicitly requires directional adaptivity. We seek an as large as possible functional classes, where our directionally adaptive estimation method still attains a rate close to the one-dimensional case. These classes have dominating mixed smoothness properties and are considerably larger than classes like $W_p^{rd}(\mathbb{R}^d)$, for example, and they involve somewhat like a restriction to functions with a lowerdimensional structure. Additive or multiplicative models are contained there as special cases, however the estimation method is more flexible than usual methods for such models. Since it is not explicitly based on this structural assumption, it delivers an asymptotically consistent estimate even if the true curve cannot be decomposed into additive or multiplicative components.

The multivariate estimation scheme considered in this article seems to be reasonable on general grounds and it could have been found also without the motivation by

anisotropic smoothness classes. Once the reasonability of this estimator is accepted, one could also raise the opposite question: What is the class of problems that our anisotropic wavelet basis is the solution to? The present paper provides at least a partial answer to this question by showing that certain anisotropic smoothness priors (and the case considered in Section 3 can also be interpreted in this sense) require a multivariate wavelet basis with mixed resolution scales rather than the commonly used multivariate basis with a one-dimensional scale parameter. In this sense, the present article contributes also to a better understanding of the estimation method. Following a recent trend, the theoretical derivations in the Sections 2 and 3 are made for the technically simplest model, signal plus Gaussian white noise. In Section 4 we transfer the results to actually interesting settings from the statistical point of view, density estimation and regression. The results of some simulations are reported in Section 5. The proofs are contained in Section 6.

2. Wavelet thresholding in anisotropic Besov classes

To keep the technical part as simple as possible, we assume that we have function-valued observations $Y(\underline{x})$, $\underline{x} = (x_1, \dots, x_d)' \in [0,1]^d$, according to the Gaussian white noise model

$$Y(\underline{x}) = \int_0^{x_1} \cdots \int_0^{x_d} f(z_1, \dots, z_d)\, dz_1 \cdots dz_d + \epsilon W(\underline{x}). \tag{2.1}$$

Here W is a Brownian sheet (cf., e.g., Walsh (1986)) and $\epsilon > 0$ is the noise level. We will consider a small-noise asymptotics, that is $\epsilon \to 0$, which mimics the situation of large-sample asymptotics in nonparametric regression or density estimation. The link between the asymptotics in model (2.1) and the usual asymptotics for regression and density estimation will be established by setting $\epsilon = n^{-1/2}$, where n denotes the sample size.

To investigate how well our estimation method adapts to varying smoothness properties in different directions, we assume that f lies in an anisotropic Besov class. We restrict ourselves to this *global* smoothness class mainly for technical convenience. This is sufficient for our particular purpose to investigate the capability of the estimator to adapt to different degrees of smoothness in different directions. Since wavelet thresholding is a spatially adaptive procedure in that it automatically chooses a reasonable degree of smoothing according to the *local* smoothness properties of the function, one could expect a favourable behaviour of our estimator in the case of spatially varying anisotropic smoothness properties of f, too.

Following Besov, Il'in and Nikol'skii (1979), we introduce now smoothness classes in anisotropic Besov spaces. Denote by $e_i = (0, \dots, 0, 1, 0, \dots, 0)'$ the ith unit vector. We define the finite difference of the function f in direction of x_i as

$$\Delta_{i,h} f(\underline{x}) = f(\underline{x} + h e_i) - f(\underline{x}).$$

By induction we get the kth difference in direction of x_i as

$$\Delta_{i,h}^k f(\underline{x}) = \Delta_{i,h} \Delta_{i,h}^{k-1} f(\underline{x}) = \sum_{l=0}^{k} (-1)^{l+k} \binom{k}{l} f(\underline{x} + l h e_i).$$

Fix any integer $k_i > r_i$. Similar to the one-dimensional case we define the Besov norm in direction of x_i as

$$\|f\|_{b^{r_i}_{i,p_i,q}} = \left(\int_{-1}^{1} |h|^{-r_i q-1} \|\Delta^{k_i}_{i,h} f\|^q_{L_{p_i}(g_{i,k})}\, dh\right)^{1/q}$$

for $q < \infty$, and

$$\|f\|_{b^{r_i}_{i,p_i,\infty}} = \sup_{|h|\le 1} \left\{|h|^{-r_i} \|\Delta^{k_i}_{i,h} f\|_{L_{p_i}(g_{i,k})}\right\},$$

where $g_{i,h} = \underbrace{[0,1]\times\ldots\times[0,1]}_{i-1} \times [0\vee k_i h, 1\wedge(1-k_i h)] \times \underbrace{[0,1]\times\ldots\times[0,1]}_{d-i}$. Note that $\|.\|_{b^{r_i}_{i,p_i,q}}$ measures only smoothness of f in direction of x_i. Setting $\underline{r} = (r_1,\ldots,r_d)'$ and $\underline{p} = (p_1,\ldots,p_d)'$ we define

$$B^{\underline{r}}_{\underline{p},q}(K) = \left\{ f \,\middle|\, \sum_{i=1}^{d} \left[\|f\|_{L_{p_i}([0,1]^d)} + \|f\|_{b^{r_i}_{i,p_i,q}}\right] \le K \right\}.$$

Assume we have an orthonormal basis of compactly supported wavelets of $L_2[0,1]$, $\{\phi_{l,k}\}_k \cup \{\psi_{j,k}\}_{j\ge l,k}$. Such bases are given by Meyer (1991) and Cohen, Daubechies and Vial (1993).
Let V_j be the subspace of $L_2[0,1]$, which is generated by $\{\phi_{j,k}\}_k$. It is known that

$$L_2([0,1]^d) = \overline{\bigcup_{j=l}^{\infty} V_j \otimes \cdots \otimes V_j},$$

which shows the possibility to build a basis of $L_2([0,1]^d)$ from tensor products of functions from a one-dimensional basis $\{\phi_{lk}\}_k \cup \{\psi_{jk}\}_{j\ge l;k}$.
Setting $W_{l-1} := V_l$ we obtain the decomposition

$$\begin{aligned} V^d_{j^*} &= V_{j^*} \otimes \cdots \otimes V_{j^*} \\ &= (V_l \oplus W_l \oplus \cdots \oplus W_{j^*-1}) \otimes \cdots \otimes (V_l \oplus W_l \oplus \cdots \oplus W_{j^*-1}) \\ &= \bigoplus_{j_1,\ldots,j_d=l-1}^{j^*-1} W_{j_1} \otimes \cdots \otimes W_{j_d}. \end{aligned} \tag{2.2}$$

Accordingly, we obtain a basis $\mathcal{B}$ of $L_2([0,1]^d)$ as

$$\mathcal{B} = \bigcup_{j_1,\ldots,j_d=l-1}^{\infty} \{\psi_{j_1,k_1}(x_1)\cdots\psi_{j_d,k_d}(x_d)\}_{k_1,\ldots,k_d}, \tag{2.3}$$

where $\psi_{l-1,k} := \phi_{l,k}$. This construction provides a multidimensional basis, where the resolution scales $j_1,\ldots,j_d$ are completely mixed.

To introduce another construction of a higher-dimensional basis, we set $V_j^{(0)} := V_j$, $V_j^{(1)} := W_j$, and $\phi_{j,k}^{(0)} := \phi_{j,k}$, $\phi_{j,k}^{(1)} := \psi_{j,k}$. Now we can write $V_{j^*}^d$ as

$$V_{j^*}^d = \left(V_l^{(0)} \otimes \cdots \otimes V_l^{(0)}\right) \oplus \\ \oplus \bigoplus_{j \geq l} \bigoplus_{(i_1,\ldots,i_d) \in \{0,1\}^d \setminus \{(0,\ldots,0)\}} \left(V_j^{(i_1)} \otimes \cdots \otimes V_j^{(i_d)}\right), \tag{2.4}$$

which corresponds to the following basis $\overline{\mathcal{B}}$ of $L_2([0,1]^d)$:

$$\overline{\mathcal{B}} = \left\{\phi_{l,k_1}^{(0)}(x_1) \cdots \phi_{l,k_d}^{(0)}(x_d)\right\}_{k_1,\ldots,k_d} \cup \\ \cup \bigcup_{j \geq l} \bigcup_{(i_1,\ldots,i_d) \in \{0,1\}^d \setminus \{(0,\ldots,0)\}} \left\{\phi_{j,k_1}^{(i_1)}(x_1) \cdots \phi_{j,k_d}^{(i_d)}(x_d)\right\}_{k_1,\ldots,k_d}. \tag{2.5}$$

The latter basis $\overline{\mathcal{B}}$ provides a d-dimensional multiresolution analysis. On first sight it seems to be more appealing than $\mathcal{B}$ and it is almost exclusively used in statistics; see, e.g., Delyon and Juditsky (1993), Tribouley (1995), and von Sachs and Schneider (1994). Appropriate wavelet estimators based on $\overline{\mathcal{B}}$ can attain minimax rates of convergence in isotropic smoothness classes, which justifies its use in statistics.
However, it was shown in Neumann and von Sachs (1995) in the two-dimensional case that $\overline{\mathcal{B}}$ is not really able to adapt to different degrees of smoothness in different directions. Expressed in terms of the kernel–estimator language, a projection estimator using basis functions from $\overline{\mathcal{B}}$ cannot mimic a multivariate kernel estimator based on a product kernel with different (directional) bandwidths $h_1, \ldots, h_d$. In contrast, we will show that estimators based on $\mathcal{B}$ can attain minimax rates of convergence in anisotropic smoothness classes. Furthermore, the superiority of $\mathcal{B}$ extends beyond the rigorous, but sometimes quite pessimistic minimax approach. The use of such a multiscale method seems to be important in many estimation problems, whenever – globally or locally – different degrees of smoothness are present. An alternative method of adapting to different degrees of smoothness in different directions was developed by Donoho (1995) in the framework of anisotropic Hölder classes. He proposed a CART-like recursive scheme to obtain adequate degrees of smoothing in each direction.

2.1. A lower bound to the rate of convergence. To set a benchmark for the estimation scheme to be developed, we establish a lower bound to the rate at which the risk can decrease in anisotropic Besov classes. Since we are only interested in the optimal *rate*, we can use an easily implemented approach developed in Bretagnolle and Huber (1979).
To study the complexity of the functional class $B_{\underline{p},q}^{\underline{r}}(K)$, we take any function μ, which is Hölder continuous of order $\max\{r_1, \ldots, r_d\}$, supported on $[0,1)$, and satisfies $\|\mu\|_{L_2} = 1$. Let, for some positive C_0 to be precised in the proof of Lemma 2.1, j_i be chosen such that

$$2^{j_i} \leq C_0 \epsilon^{-(2/r_i)/(1/r_1+\ldots+1/r_d+2)} < 2^{j_i+1}.$$

Define

$$\mu_{k_1,\dots,k_d}(\underline{x}) = 2^{(j_1+\dots+j_d)/2}\mu(2^{j_1}x_1 - k_1)\cdots\mu(2^{j_d}x_d - k_d).$$

It is easy to see that

$$\|\mu_{k_1,\dots,k_d}\|_{L_2} = 1 \tag{2.6}$$

and

$$supp(\mu_{k_1,\dots,k_d}) \cap supp(\mu_{k_1',\dots,k_d'}) = \emptyset, \quad \text{if } (k_1,\dots,k_d) \neq (k_1',\dots,k_d'). \tag{2.7}$$

Let $D = D(\epsilon) = 2^{j_1+\dots+j_d} \asymp (\epsilon^2)^{-(1/r_1+\dots+1/r_d)/(1/r_1+\dots+1/r_d+2)}$. Now we define a class of functions, parametrized by the D-dimensional parameter $\underline{\theta} = (\theta_{k_1,\dots,k_d})_{0\le k_i\le 2^{j_i}-1}$, by

$$\mu_{\underline{\theta}}(\underline{x}) = \sum_{k_1=1}^{2^{j_1}} \cdots \sum_{k_d=1}^{2^{j_d}} \theta_{k_1,\dots,k_d}\mu_{k_1,\dots,k_d}(\underline{x}). \tag{2.8}$$

It is not difficult to prove the following lemma.

Lemma 2.1. *If C_0 is chosen small enough, then*

$$\max_{\underline{\theta}\in\{0,\epsilon\}^D} \left\{ \|\mu_{\underline{\theta}}\|_{B^{\underline{r}}_{\underline{p},q}} \right\} \le K.$$

Using (2.6), (2.7), and Lemma 2.1, we obtain by the method introduced in Bretagnolle and Huber (1979) a lower bound to the rate of convergence in $B^{\underline{r}}_{\underline{p},q}(K)$.

Theorem 2.1. *It holds that*

$$\inf_{\widehat{f}_\epsilon} \sup_{f\in B^{\underline{r}}_{\underline{p},q}(K)} \left\{ E\|\widehat{f}_\epsilon - f\|^2_{L_2} \right\} \ge C\epsilon^{2\vartheta(r_1,\dots,r_d)},$$

where

$$\vartheta(r_1,\dots,r_d) = 2\widetilde{r}/(2\widetilde{r}+d), \quad \widetilde{r} = \left[\frac{1}{d}\left(\frac{1}{r_1}+\dots+\frac{1}{r_d}\right)\right]^{-1}.$$

2.2. Optimal wavelet thresholding in anisotropic Besov classes. In this subsection we develop thresholding schemes based on the anisotropic basis $\mathcal{B}$, which provide the optimal or a near–optimal rate of convergence in anisotropic Besov spaces. First, we show that the rate given in Theorem 2.1 is actually attainable by certain wavelet estimators. It will turn out, that this method depends on the unknown smoothness parameters $r_1,\dots,r_d$. Hence, an additional adaptation step would be necessary to obtain a fully adaptive method. Alternatively, one can use a universal estimation method, as proposed in a series of papers by Donoho and Johnstone, also contained in Donoho *et al.* (1995).

As a starting point we take a one–dimensional boundary–adjusted wavelet basis of $L_2([0,1])$, e.g., those of Meyer (1991) or Cohen, Daubechies and Vial (1993). We assume that

(A1) (i) $\int \phi(t)\,dt = 1$,
(ii) $\int \psi(t)t^k\,dt = 0 \quad \text{for} \quad 0 \le k \le \max\{r_1,\dots,r_d\} - 1.$

(As mentioned in Delyon and Juditsky (1993, Section 5.2), we do not need the frequently assumed smoothness of the wavelet itself for the particular purpose of obtaining certain rates of convergence.)
For the sake of notational convenience we write $\psi_{l-1,k} = \phi_{l,k}$. As explained above, we get a d–dimensional orthonormal basis by setting

$$\psi_{j_1,\dots,j_d,k_1,\dots,k_d}(\underline{x}) = \psi_{j_1,k_1}(x_1)\cdots\psi_{j_d,k_d}(x_d). \tag{2.9}$$

To simplify notation, we use the multiindex I for $(j_1,\dots,j_d,k_1,\dots,k_d)'$, whenever possible. The true wavelet coefficients are defined as

$$\theta_I = \int_{[0,1]^d} \psi_I(\underline{x}) f(\underline{x})\,d\underline{x}. \tag{2.10}$$

Having observations according to model (2.1), we obtain empirical versions of these coefficients as

$$\tilde{\theta}_I = \int_{[0,1]^d} \psi_I(\underline{x})\,dY(\underline{x}) = \theta_I + \epsilon\xi_I, \tag{2.11}$$

where $\xi_I \sim N(0,1)$ are i.i.d.
Now we proceed in the usual way. An appropriate smoothing is obtained by nonlinear thresholding of the empirical coefficients, which includes a truncation of the infinite wavelet series as a special case. Finally, we obtain an estimate of f by applying the inverse wavelet transform to the thresholded empirical coefficients.
Two commonly used rules to treat the coefficients are

1) hard thresholding

$$\delta^{(h)}(\tilde{\theta}_I,\lambda) = \tilde{\theta}_I I\left(|\tilde{\theta}_I| \ge \lambda\right)$$

and

2) soft thresholding

$$\delta^{(s)}(\tilde{\theta}_I,\lambda) = \left(|\tilde{\theta}_I| - \lambda\right)_+ sgn(\tilde{\theta}_I).$$

In the following we denote by $\delta^{(\cdot)}$ either $\delta^{(h)}$ or $\delta^{(s)}$.
As a basis for our particular choice of the threshold values we take an upper estimate of the risk of $\delta^{(\cdot)}(\tilde{\theta}_I,\lambda)$ as an estimate of θ_I. By Lemma 1 of Donoho and Johnstone (1994a) we can prove that the relation

$$E\left(\delta^{(\cdot)}(\tilde{\theta}_I,\lambda) - \theta_I\right)^2 \le C\left(\epsilon^2\left(\frac{\lambda}{\epsilon}+1\right)\varphi(\frac{\lambda}{\epsilon}) + \min\{\lambda^2,\theta_I^2\}\right) \tag{2.12}$$

holds uniformly in $\lambda \geq 0$ and $\theta_I \in \mathbb{R}$, where φ denotes the standard normal density. Accordingly, we get by

$$\Omega_\epsilon((\lambda_I), \Theta) := \sup_{(\theta_I)\in\Theta} \left\{ \sum_I \left(\epsilon^2 \left(\frac{\lambda_I}{\epsilon} + 1 \right) \varphi(\frac{\lambda_I}{\epsilon}) + \min\{\lambda_I^2, \theta_I^2\} \right) \right\} \tag{2.13}$$

an upper rate bound for the estimator

$$\hat{f} = \sum_i \delta^{(\cdot)}(\tilde{\theta}_I, \lambda_I)\psi_I,$$

which is uniform in the functional class $\{f = \sum_I \theta_I \psi_i \mid (\theta_I) \in \Theta\}$.
A closely related quantity,

$$\Omega_\epsilon(\Theta) = \inf_{(\lambda_I)} \Omega_\epsilon((\lambda_I), \Theta) \tag{2.14}$$

was used in Neumann and von Sachs (1995) as a chracterization of the difficulty of estimation in the functional class given by Θ. A different quantity,

$$\tilde{\Omega}_\epsilon(\Theta) = \sup_{(\theta_I)\in\Theta} \left\{ \sum_I \min\{\epsilon^2, \theta_I^2\} \right\},$$

has been considered in Donoho and Johnstone (1994) to establish the link between optimal estimation and approximation theory. There it was shown that $\tilde{\Omega}_\epsilon(\Theta)$ can be attained by the risk of an appropriately thresholded wavelet estimator within some logarithmic factor, $(\log \epsilon^{-1})^\rho$, $\rho > 0$. We modify $\tilde{\Omega}_\epsilon(\Theta)$ by $\Omega_\epsilon((\lambda_I), \Theta)$ in order to remove the logarithmic factor, which does not occur in the lower bound given in Theorem 2.1. This factor appeared in Donoho and Johnstone (1994) because $\tilde{\Omega}_\epsilon$ does not appropriately capture the additional difficulty due to sparsity of the signal; and hence ϵ had to be replaced by $\epsilon\sqrt{\log \epsilon^{-1}}$. In contrast, Ω_ϵ penalizes sparsity of the signal, which arises due to ignorance of the significant coefficients in a large set of potentially important ones, by the additional terms $(\lambda_I/\epsilon + 1)\varphi(\lambda_I/\epsilon)$. They arise from upper estimates of tail probabilities of Gaussian random variables.
Now we intend to show how the lower risk bound given in Theorem 2.1 can be attained by a particular estimator. This will be a thresholded wavelet estimator, where the choice of the thresholds is motivated by the upper bound given by (2.13).
Let $j_1^*, \ldots, j_d^*$ be chosen in such a way that

$$2^{j_i^* r_i} \asymp \epsilon^{-2/(1/r_1 + \ldots + 1/r_d + 2)}. \tag{2.15}$$

In "homogeneous smoothness classes", that is in the case of $p_i \geq 2$ for $i = 1, \ldots, d$, we would attain the optimal rate of convergence by the linear projection estimator on the linear space $V_{j_1^*} \otimes \cdots \otimes V_{j_d^*}$; see also the next lemma for an upper estimate of the error due to truncation. In the more difficult case of "inhomogeneous smoothness classes", that is if $p_i < 2$ for any i, we have to employ a more refined method.
We define the following thresholds:

$$\lambda_I^{opt} = \epsilon\kappa \sqrt{\max_{1\leq i\leq d} \{(j_i - j_i^*)_+ r_i\}(1/r_1 + \ldots + 1/r_d)}, \tag{2.16}$$

where κ is any constant satisfying

$$\kappa > \sqrt{2\log(2)}. \tag{2.17}$$

These particular choices of the λ_I's are similar to those in Delyon and Juditsky (1993), which has been proposed for isotropic smoothness classes. We consider the estimator

$$\widehat{f}_\epsilon^{opt}(\underline{x}) = \sum_I \delta^{(.)}(\tilde{\theta}_I, \lambda_I^{opt})\psi_I(\underline{x}). \tag{2.18}$$

The following theorem establishes the desired result for the rate of convergence.

Theorem 2.2. *Assume (A1) and*

$$p_i > (1 - p_i/2)(1/r_1 + \ldots + 1/r_d) \quad \textit{for all } i = 1, \ldots, d.$$

Then

$$\sup_{f \in B^{\underline{r}}_{\underline{p},q}(K)} \left\{ E\|\widehat{f}_\epsilon^{opt} - f\|^2_{L_2} \right\} = O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right).$$

Note that the above thresholding scheme depends on the unknown parameters $r_1, \ldots, r_d$. Hence, its practical implementation would require an additional adaptation step. There exists a wide variety of possible approaches to achieve this in many statistical models of interest. However, there seems to be no universal recipe for all purposes. To avoid these difficulties one could use an alternative approach propagated in a series of papers by Donoho and Johnstone, also contained in Donoho *et al.* (1995). It consists of truncating the infinite wavelet expansion of f at a sufficiently high resolution scale and then treating the remaining empirical coefficients by some universal thresholding rule. First, consider the error incurred by truncation at a given level.

Lemma 2.2. *Assume (A1). Let* $\tilde{V}_J = \bigoplus_{j_1+\ldots+j_d=J} (V_{j_1} \otimes \cdots \otimes V_{j_d})$. *Then*

$$\sup_{f \in B^{\underline{r}}_{\underline{p},q}(K)} \left\{ \|f - Proj_{\tilde{V}_{J^*}} f\|^2_{L_2} \right\} = O\left(2^{-J^*\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)}\right),$$

where

$$\gamma(r_1, \ldots, r_d, p_1, \ldots, p_d) = \{2 + [(1 - 2/\tilde{p}_1)/r_1 + \cdots + (1 - 2/\tilde{p}_d)/r_d]\} / (1/r_1 + \cdots + 1/r_d)$$

and $\tilde{p}_i = \min\{p_i, 2\}$.

Provided that $\gamma(r_1, \ldots, r_d, p_1, \ldots, p_d) > 0$, this lemma basically means that an approximation rate of ϵ^ρ $(\rho < \infty)$ can be attained by an appropriate set of basis functions which has algebraic cardinality, say $\epsilon^{-\nu(\rho)}$ for some $\nu(\rho) < \infty$.
Define $\mathcal{I}_\epsilon = \{I \mid j_1 + \ldots + j_d \leq J_\epsilon^*\}$, where $2^{J_\epsilon^*} = O(\epsilon^{-\nu})$ for any $\nu < \infty$. We consider the estimator

$$\widehat{f}_\epsilon^{univ}(\underline{x}) = \sum_{I \in \mathcal{I}_\epsilon} \delta^{(.)}(\tilde{\theta}_I, \lambda_I^{univ})\psi_I(\underline{x}), \tag{2.19}$$

where

$$\lambda_I^{univ} = \epsilon\sqrt{2\log(\#\mathcal{I}_\epsilon)}. \tag{2.20}$$

This estimator $\widehat{f}_\epsilon^{univ}$ is much less dependent than $\widehat{f}_\epsilon^{opt}$ on prior assumptions about the smoothness of f. In practice, one should take some reasonably large ν in order to keep the truncation bias small in a wide range of smoothness classes. In view of results of Donoho *et al.* (1995), it is not surprising at all that $\widehat{f}_\epsilon^{univ}$ attains the optimal rate of convergence within some logarithmic factor. For reader's convenience we formally establish this in the following theorem.

Theorem 2.3. *Assume (A1) and*

$$p_i > (1 - p_i/2)(1/r_1 + \ldots + 1/r_d) \quad \textit{for all } i = 1, \ldots, d.$$

Then

$$\sup_{f \in B^r_{\underline{p},q}(K)} \left\{ E\|\widehat{f}_\epsilon^{univ} - f\|_{L_2}^2 \right\} = O\left((\epsilon^2 \log(\epsilon^{-1}))^{\vartheta(r_1,\ldots,r_d)} \right) + O\left(2^{-J_\epsilon^* \gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)} \right).$$

If $\gamma(r_1, \ldots, r_d, p_1, \ldots, p_d) > 0$, the value of J_ϵ^* can be chosen so large, that the upper bound given in Theorem 2.3 is dominated by the first term on the right-hand side. Hence, we obtain the optimal rate of convergence within some logarithmic factor.

Remark 1. (The corresponding kernel estimator)
As already mentioned, we can attain the optimal rate of convergence by a projection estimator on the space $V_{j^*} \otimes \cdots \otimes V_{j^*}$ in the class $B^r_{\underline{p},q}(K)$, if $p_i \geq 2$ for all $i = 1, \ldots, d$. Alternatively, we can also use a multivariate kernel estimator with a product kernel $K(\underline{x}) = K_1(x_1) \cdots K_1(x_d)$, where K_1 is a boundary corrected kernel satisfying $\int K_1(x) x^k \, dx = \delta_{0k}$ $0 \leq k \leq \max\{r_1, \ldots, r_d\} - 1$. Choosing a product bandwidth $\underline{h} = (h_1, \ldots, h_d)$ with $h_i \asymp \epsilon^{(2/r_i)/(1/r_1 + \ldots + 1/r_d + 2)}$, we obtain the optimal rate of convergence.

3. A MULTIVARIATE FUNCTIONAL CLASS, WHICH ADMITS RATES OF CONVERGENCE CLOSE TO THE ONE-DIMENSIONAL CASE

In this section we proceed with the investigation of what wavelet methods can offer for multivariate estimation problems. Although again nonlinear thresholding in the anisotropic wavelet basis is used, the object under consideration is quite different from that considered in the previous section: There we studied the ability of our estimator to adapt to different degrees of smoothness in different directions, which were modeled by anisotropic Besov classes. The "effective dimension" of such a class in $[0,1]^d$ is d, and therefore at least some of the directional smoothness parameters r_i must be sufficiently large to make a successful estimation in several dimensions possible. Here we consider the opposite situation, where the effective dimension of our multivariate functional class in $[0,1]^d$ is still one, or at least very close to one.

Some motivation for the definition of the particular functional classes considered here comes from additive models, which are known to allow rates of convergence corresponding to the one-dimensional case. As we will see below, the approximate preservation of the one-dimensional rate goes considerably beyond the case of such semiparametric models. Having in mind that nonlinear thresholding in the anisotropic basis adapts locally to the presence of a different complexity in the various directions, we seek an as large as possible class of functions that does not suffer from the curse of dimensionality. It turns out that appropriate functional classes are those with dominating mixed derivatives; see, e. g., Schmeißer and Triebel (1987, Chapter 2).
For the sake of simplicity we first restrict our considerations to the case of L_2-Sobolev constraints, although other definitions of smoothness like Besov constraints would also be possible. Let, for some fixed K,

$$\mathcal{F}_r^{(d)}(K) = \left\{ f \;\middle|\; \sum_{0 \le r_1, \dots, r_d \le r} \left\| \frac{\partial^{r_1+\dots+r_d}}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}} f \right\|_{L_2} \le K \right\}. \tag{3.1}$$

In contrast to usually considered isotropic smoothness classes like the d-dimensional Sobolev class

$$\mathcal{F}_s(K) = \left\{ f \;\middle|\; \sum_{0 \le r_1+\dots+r_d \le s} \left\| \frac{\partial^{r_1+\dots+r_d}}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}} f \right\|_{L_2} \le K \right\},$$

the mixed derivatives play the dominant part in (3.1). Whereas we need a degree of smoothness of $s = rd$ in $\mathcal{F}_s(K)$ to get the rate $\epsilon^{4r/(2r+1)}$ for the minimax risk, we need only r partial derivatives in each direction in (3.1) to attain this rate up to a logarithmic factor.
The class $\mathcal{F}_r^{(d)}(K)$ contains additive models like, e. g.,

$$f(\underline{x}) = \sum_{i=1}^{d} f_i(x_i) + \sum_{i,j=1}^{d} f_{ij}(x_i, x_j), \tag{3.2}$$

if $f_i \in \mathcal{F}_r^{(1)}(K')$ and $f_{ij} \in \mathcal{F}_r^{(2)}(K')$, or a multiplicative model like

$$f(\underline{x}) = \prod_{i=1}^{d} f_i(x_i), \tag{3.3}$$

if $f_i \in \mathcal{F}_r^{(1)}(K')$, for appropriate K', as special cases. However, it is considerably larger than such semiparametric classes of functions in that it is a truely nonparametric functional class. The restriction of the complexity is attained by an appropriate smoothness assumption instead of rigorous structural assumptions as in (3.2) and (3.3).
As a benchmark for the estimation method to be considered, we derive first a lower bound to the minimax risk in $\mathcal{F}_r^{(d)}(K)$. Recall that ψ_I are the tensor product wavelets defined by (2.9), and $\theta_I = \int \psi_I(\underline{x}) f(\underline{x})\, d\underline{x}$ denotes the corresponding wavelet coefficient. For the one-dimensional scaling function ϕ and the wavelet ψ we assume that

(A2) (i) $\int \phi(t)\,dt = 1$,
(ii) $\int \psi(t)t^k\,dt = 0 \quad \text{for} \quad 0 \le k \le r$.

It will be shown below that membership in $\mathcal{F}_r^{(d)}(K)$ implies a constraint on the wavelet coefficients of the type

$$\sum_{j_1,\dots,j_d} 2^{2(j_1+\dots+j_d)r} \sum_{k_1,\dots,k_d} |\theta_I|^2 \le K'. \tag{3.4}$$

We again intend to apply the hypercube method to derive a lower risk bound. To get a sharp bound, we have to find the hardest cubical subproblem. To achieve this, we consider the level-wise contributions to the total risk by any hypothetical minimax estimator. At coarse scales, that is for $J = j_1 + \dots + j_d$ small, the coefficients θ_I are allowed to be quite large. Accordingly, the linear estimates $\tilde{\theta}_I$ are minimax and their level-wise contributions to the total risk are of order $\epsilon^2 \#\{I \mid j_1 + \dots + j_d = J\} \asymp \epsilon^2 2^J J^{d-1}$. At finer scales, the smoothness constraint of

$$\sum_{j_1+\dots+j_d=J} \sum_{k_1,\dots,k_d} |\theta_I|^2 \le K' 2^{-2(j_1+\dots+j_d)r}$$

becomes dominating, and not all coefficients are allowed to be in absolute value as large as the noise level ϵ at the same time. Despite the rapidly increasing number of coefficients at each level J as $J \to \infty$, the level-wise contribution of optimal estimators to the total risk will decrease.

In accordance with this heuristics, a sharp lower bound to the minimax rate of convergence will be generated by the problem of estimating the wavelet coefficients at a level which is at the border between the "dense case" and the "sparse case". Roughly speaking, the dense case corresponds to levels $\{(j_1,\dots,j_d) \mid j_1 + \dots + j_d = J\}$, where all coefficients can simultaneously attain the value ϵ, whereas the sparse case corresponds to levels at which only a fraction of these coefficients can be equal to ϵ at the same time. Correspondingly, the hardest level J_ϵ satisfies the relation

$$\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1} \asymp 2^{-2J_\epsilon r}, \tag{3.5}$$

which leads to

$$2^{J_\epsilon} \asymp \left(\epsilon^2 [\log(\epsilon^{-1})]^{d-1}\right)^{-1/(2r+1)}. \tag{3.6}$$

Let μ be any r times continuously differentiable wavelet supported on $[0,1]$ (In contrast to the case in Subsection 2.1 we need orthogonality of $\mu(2^j x - k)$ and $\mu(2^{j'} x - k')$ if $(j,k) \ne (j',k')$.). Using the multiindex $I = (j_1,\dots,j_d,k_1,\dots,k_d)$ we define

$$\mu_I(\underline{x}) = 2^{(j_1+\dots+j_d)/2} \mu(2^{j_1}x_1 - k_1)\cdots\mu(2^{j_d}x_d - k_d).$$

Define the following class of functions, parametrized by the multidimensional parameter $\theta = (\theta_I)_{I:\ j_1+\dots+j_d=J_\epsilon}$:

$$\mu_\theta(\underline{x}) = \sum_{j_1+\dots+j_d=J_\epsilon} \sum_{k_1,\dots,k_d} \theta_I \mu_I(\underline{x}).$$

The following lemma characterizes the complexity of the functional class $\mathcal{F}_\tau^{(d)}(K)$ via the dimensionality of θ.

Lemma 3.1. *Let J_ϵ be chosen according to (3.6). If C_0 is small enough, then*

$$\{\mu_\theta \mid \quad \theta_I \in \{0, C_0\epsilon\} \quad \text{for all} \quad I: \ j_1 + \ldots + j_d = J_\epsilon\} \subseteq \mathcal{F}_\tau^{(d)}(K).$$

Since the μ_I's are orthogonal, we immediately obtain by the hypercube method the following bound to the minimax rate of convergence in $\mathcal{F}_\tau^{(d)}(K)$.

Theorem 3.1. *It holds that*

$$\inf_{\widehat{f}_\epsilon} \sup_{f \in \mathcal{F}_\tau^{(d)}(K)} \left\{ E\|\widehat{f}_\epsilon - f\|_{L_2}^2 \right\} \geq C \left(\epsilon^2 [\log(\epsilon^{-1})]^{d-1} \right)^{2\tau/(2\tau+1)}.$$

Now we formulate an upper bound to the complexity of the functional class $\mathcal{F}_\tau^{(d)}(K)$ by an appropriate restriction on the wavelet coefficients.

Lemma 3.2. *Assume (A2). Then, for appropriate K',*

$$\mathcal{F}_\tau^{(d)}(K) \subseteq \left\{ f = \sum_I \theta_I \psi_I \;\middle|\; \sum_{j_1,\ldots,j_d} 2^{2(j_1+\ldots+j_d)\tau} \sum_{k_1,\ldots,k_d} |\theta_I|^2 \leq K' \right\}.$$

Note that the norm applied to the coefficients θ_I in Lemma 3.2 is of L_2-type. Therefore it is not surprising that even a simple projection estimator attains the minimax rate of convergence in $\mathcal{F}_\tau^{(d)}(K)$.

Theorem 3.2. *Assume (A2). Let J_ϵ be defined as in (3.6) and let*

$$\widehat{f}_\epsilon^P(\underline{x}) = \sum_{j_1+\ldots+j_d \leq J_\epsilon} \sum_{k_1,\ldots,k_d} \tilde{\theta}_I \psi_I(\underline{x}).$$

Then

$$\sup_{f \in \mathcal{F}_\tau^{(d)}(K)} \left\{ E\|\widehat{f}_\epsilon^P - f\|_{L_2}^2 \right\} = O\left(\left(\epsilon^2 [\log(\epsilon^{-1})]^{d-1} \right)^{2\tau/(2\tau+1)} \right).$$

Remark 2. In contrast to the case of anisotropic Besov classes considered in the previous section, the construction of an appropriate kernel estimator is not obvious at all. Note that the wavelet estimator $\widehat{f}_\epsilon^P$ projects the observations on the space $\bigoplus_{j_1+\ldots+j_d=J_\epsilon} V_{j_1} \otimes \cdots \otimes V_{j_d}$. Since the spaces $V_{j_1} \otimes \cdots \otimes V_{j_d}$ and $V_{j_1'} \otimes \cdots \otimes V_{j_d'}$ are not orthogonal for $(j_1, \ldots, j_d) \neq (j_1', \ldots, j_d')$, one has to devise a quite involved kernel-based projection scheme, which is then able to provide the optimal rate of

convergence.

Remark 3. Note that an assumption of different degrees of smoothness in different directions like

$$\sum_{0\le r_i\le R_i} \left\| \frac{\partial^{r_1+\ldots+r_d}}{\partial x_1^{r_1}\cdots\partial x_d^{r_d}} f \right\|_{L_2} \le K$$

does not lead to an essential change in the rate of convergence. Here the worst case described by $r = \min\{R_i\}$ drives essentially the rate of convergence, which is again not better than $\epsilon^{4r/(2r+1)}$. More exactly, the minimax rate of convergence is then $(\epsilon^2[\log(\epsilon^{-1})]^{D-1})^{2r/(2r+1)}$, where $D = \#\{r_i \mid \ r_i = \min\{r_j\}\}$ is the multiplicity of the worst direction.

Note that the optimal projection estimator $\widehat{f}_\epsilon^P$ depends, via J_ϵ, on the smoothness parameter r. To get a simple, fully adaptive method, we can again apply certain universal thresholds. Let $\widehat{f}_\epsilon^{univ}$ be defined as in (2.19) and (2.20).

Theorem 3.3. *Assume (A2). Then*

$$\sup_{f\in\mathcal{F}_r^{(d)}(K)} \left\{E\|\widehat{f}_\epsilon^{univ} - f\|_{L_2}^2\right\} = O\left(\left(\epsilon^2[\log(\epsilon^{-1})]^d\right)^{2r/(2r+1)}\right) + O\left(2^{-2J_\epsilon^* r}\right).$$

Note that the universally thresholded estimator misses the optimal rate of convergence, which is attained by the projection estimator considered in Theorem 3.2, by some logarithmic factor. This is because the universal estimator does not achieve the optimal tradeoff between squared bias and variance. The same effect is well-known for conventional smoothness classes; see, e.g., Donoho *et al.* (1995). As shown in Donoho and Johnstone (1992) for univariate Besov classes, the necessity for nonlinear estimators occurs in functional classes which allow more spatial inhomogeneity than L_2-classes. To show that appropriate thresholding works also in our framework of multivariate smoothness classes with dominating mixed derivatives, we consider now a slightly larger functional class, which allows a more inhomogeneous distribution of the smoothness over $[0,1]^d$. We define this class in analogy to the Besov space $B_{1,\infty}^r$, which is the largest one in the scale of spaces $B_{p,q}^r$ with degree of smoothness r and $1 \le p, q \le \infty$.
According to the inequality

$$(\#\mathcal{I})^{-1}\sum_{I\in\mathcal{I}}|\theta_I| \le \left((\#\mathcal{I})^{-1}\sum_{I\in\mathcal{I}}|\theta_I|^2\right)^{1/2},$$

we define the following functional class:

$$\mathcal{F}_{r,1,\infty}^{(d)}(K) = \left\{ f = \sum_I \theta_I\psi_I \;\middle|\; \sup_J \left\{ 2^{J(r-1/2)} J^{-(d-1)/2} \sum_{j_1+\ldots+j_d=J} \sum_{k_i,\ldots,k_d} |\theta_I| \right\} \le K \right\}. \quad (3.7)$$

By Lemma 3.2, we can easily see that $\mathcal{F}_r^{(d)}(K) \subseteq \mathcal{F}_{r,1,\infty}^{(d)}(K')$ holds for an appropriate K'. Moreover, these classes are considerably larger than $\mathcal{F}_r^{(d)}(K)$, since they contain, for example, one-dimensional functions $f(\underline{x}) = f_1(x_1)$ from the spatially inhomogeneous smoothness class $B_{1,\infty}^r(K')$. Since linear estimators are, even in this simple special case of $f(\underline{x}) = f_1(x_1)$, $f_1 \in B_{1,\infty}^r(K')$, restricted to a rate of convergence of $\epsilon^{4\widetilde{r}/(2\widetilde{r}+1)}$, $\widetilde{r} = r - 1/2$, we can only hope to get the desired rate of $(\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}$ by an appropriate nonlinear method.
Let J_ϵ be defined as in (3.6). We define the thresholds

$$\lambda_I^* = \begin{cases} 0, & \text{if } j_1 + \ldots + j_d \le J_\epsilon \\ \epsilon\kappa\sqrt{(j_1 + \ldots + j_d) - J_\epsilon}, & \text{if } j_1 + \ldots + j_d > J_\epsilon \end{cases}, \tag{3.8}$$

where κ is again any constant larger than $\sqrt{2\log 2}$. Further, let

$$\widehat{f}_\epsilon^*(\underline{x}) = \sum_I \delta^{(\cdot)}(\widetilde{\theta}_I, \lambda_I^*)\psi_I(\underline{x}). \tag{3.9}$$

The following theorem shows that $\widehat{f}_\epsilon^*$ is optimal in the class $\mathcal{F}_{r,1,\infty}^{(d)}(K)$.

Theorem 3.4. *Assume (A2). Then*

$$\sup_{f \in \mathcal{F}_{r,1,\infty}^{(d)}(K)} \left\{ E\|\widehat{f}_\epsilon^* - f\|_{L_2}^2 \right\} = O\left(\left(\epsilon^2[\log(\epsilon^{-1})]^{d-1} \right)^{2r/(2r+1)} \right).$$

4. Application to nonparametric regression and density estimation

In this section we intend to indicate how far the theoretical results from the previous sections are relevant for more realistic models like nonparametric regression and density estimation. Under reasonable assumptions and by setting $\epsilon \asymp n^{-1/2}$, the lower bounds from the previous sections can be transferred both to non-Gaussian regression and density estimation. (Note that the hypercube approach by Bretagnolle and Huber (1979) was just developed in the density estimation setting.)
Assume that d-dimensional independent observations Y_i, $i = 1, \ldots, n$, according to a density f are available. For simplicity of this discussion assume that we intend to estimate f only on some rectangular domain, say $[0,1]^d$, where f is bounded away from zero. Empirical wavelet coefficients are easily defined as

$$\widetilde{\theta}_I = n^{-1} \sum_{i=1}^n \psi_I(Y_i).$$

These coefficients are unbiased estimators for θ_I and have a variance $\sigma_I^2 = n^{-1}\left(\int \psi_I^2(\underline{x}) f(\underline{x})\, d\underline{x} - \theta_I^2\right)$. Using Lemma 3.1 and Lemma 2.3 in Saulis and Statulevicius (1991) we obtain that

$$P\left(\pm\frac{\widetilde{\theta}_I - \theta_I}{\sigma_I} \ge x \right) = (1 - \Phi(x))(1 + o(1)) \tag{4.1}$$

holds uniformly in $x = o((n2^{-(j_1+\ldots+j_d)})^{1/6})$. Let

$$\xi_I \sim N(\theta_I, \sigma_I^2). \tag{4.2}$$

Essentially by integration by parts, we can derive that

$$E\left(\delta^{(\cdot)}(\tilde{\theta}_I, \lambda) - \theta_I\right)^2 = E\left(\delta^{(\cdot)}(\xi_I, \lambda) - \theta_I\right)^2 (1 + o(1)) + O(n^{-\lambda}) \tag{4.3}$$

holds in a uniform manner in $\{I \mid \ 2^{j_1+\ldots+j_d} \leq n^\gamma\}$, for any fixed $\gamma < 1$ and arbitrary $\lambda < \infty$; see, e. g., Neumann (1994). This means that we can apply just the same estimation techniques which were developed for the Gaussian white noise model (2.1).

In view of the heteroscedastic structure of the approximating model (4.2), we think that a slight modification of the thresholding schemes from the previous sections is advisable. For example, the universal threshold $\lambda_\epsilon^{univ} = \epsilon\sqrt{2\log(\#\mathcal{I}_\epsilon)}$ defined in Section 2 should be replaced by thresholds

$$\lambda_I^{univ} = \widehat{\sigma}_I\sqrt{2\log(\#\mathcal{I}_n)}, \tag{4.4}$$

where $\mathcal{I}_n$ denotes the set of indices associated to coefficients that are thresholded, and

$$\widehat{\sigma}_I^2 = n^{-1}\left(n^{-1}\sum_i \psi_I^2(Y_i) - \tilde{\theta}_I^2\right) \tag{4.5}$$

is a consistent estimate of σ_I^2.

As long as $\mathcal{I}_n$ contains only indices I with $2^{j_1+\ldots+j_d} \leq n^\gamma$, for some $\gamma < 1$, asymptotic normality (4.1) and the risk equivalence (4.2) are valid, and we can expect analogous asymptotic results to hold as in the Gaussian case.

A similar connection to the Gaussian white noise model can be established for multivariate nonparametric regression. Assume we have n independent observations $(\underline{X}_1, Y_1), \ldots, (\underline{X}_n, Y_n)$, where $\underline{X}_i$ is distributed according to a d-dimensional density p. For $f(\underline{x}) = E(Y_i \mid \ \underline{X}_i = \underline{x})$ we obtain the usual nonparametric regression model

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.6}$$

where $E(\varepsilon_i \mid \ \underline{X}_i = \underline{x}) = 0$.

In contrast to the density case, we have to take care of the bias when we construct empirical wavelet coefficients. An obvious possibility is, to compute first a multivariate local polynomial estimator of f with some small bandwidth h_n and then to insert this estimate into formula (2.10).

To be more specific, assume that

(A3) (i) the marginal density p of X_i is bounded away from zero on $[0,1]^d$,
(ii) f is r-times continuously differentiable on $[0,1]^d$, where $r > d/2$,
(iii) for all $M < \infty$, there exist constants $C_M < \infty$, such that

$$\sup_{\underline{x}\in[0,1]^d} E\left(|\varepsilon_i|^M \mid \ \underline{X}_i = \underline{x}\right) \leq C_M.$$

Let $\tilde{f}(\underline{x})$ be a multivariate local polynomial estimator of order r with some bandwidth h_n; see, e. g., Ruppert and Wand (1994). Conditioned on $\underline{X}_1, \ldots, \underline{X}_n$, $\tilde{f}$ can be written in the form

$$\tilde{f}(\underline{x}) = \sum_i w_i(\underline{x}) Y_i.$$

For the bandwidth h_n we assume that, for any $\delta > 0$,

$$h^{2r} = O\left(n^{-1-\delta}\right), \quad h^{-d} = O\left(n^{1-\delta}\right). \tag{4.7}$$

It may be shown that, for C large enough, the relation

$$P\left(\sup_{\underline{x}\in[0,1]^d}\left\{\left|\sum_i w_i(\underline{x}) f(\underline{X}_i) - f(\underline{x})\right|\right\} > C h_n^r\right) = O\left(n^{-\lambda}\right)$$

is satisfied, that is, the maximal bias of $\tilde{f}(\underline{x})$ over $[0,1]^d$ is of order h_n^r with a probability exceeding $1 - O(n^{-\lambda})$. If we set $\tilde{\theta}_I = \int \psi_I(\underline{x}) \tilde{f}(\underline{x})\, d\underline{x}$, then

$$E\left(\tilde{\theta}_I \mid \underline{X}_1, \ldots, \underline{X}_n\right) - \theta_I = O(h_n^r)$$

is also satisfied with the above probability. Using Theorem 2 of Amosova (1972) we can prove that, conditioned on $\underline{X}_1, \ldots, \underline{X}_n$, the asymptotic relation

$$P\left(\pm\frac{\tilde{\theta}_I - \theta_I}{\sigma_I} \geq x\right) = (1 - \Phi(x))(1 + o(1)) + O(n^{-\lambda}), \tag{4.8}$$

where $\sigma_I^2 = var\left(\tilde{\theta}_I \mid \underline{X}_1, \ldots, \underline{X}_n\right)$, is satisfied for all I with $2^{j_1+\ldots+j_d} \leq n^\gamma$, $\gamma < 1$, with overwhelming probability. Hence, it is not difficult to transfer the methods from Sections 2 and 3 to multivariate nonparametric regression.

5. Simulations

In this section we briefly report on results of a simulation study, which was carried out to check how far the asymptotic results are relevant for moderate sample sizes. We used as a convenient programming environment the *XploRe* system, which has been developed by W. Härdle and coworkers, and runs on personal computers. A description of this is contained in Härdle, Klinke and Turlach (1995).
In accordance to the main theme of this paper, we considered a bivariate function $f(x_1, x_2) = 2\sin^2(x_1)$, which has an effective dimension one. This function is visualized in Figure 1, on the grid $G = \{((i - 1/2)/16, (j - 1/2)/16),\ i, j = 1, \ldots, 16\}$ with 256 grid points.

[Please insert Figure 1 about here]

First we compared the anisotropic wavelet basis with the usual (isotropic) multiresolution basis with regard to their ability to compress the signal f. Good compressibility means that most of the power of the signal is packet in an as small as possible number of coefficients. To compute the wavelet coefficients of the two-dimensional bases we used the one-dimensional fast wavelet transform (as well as its inverse for

backtransformation) as the main building block, which provides a quite efficient algorithm. Figure 2 shows magnitudes of the wavelet coefficients, $|\tilde{\theta}_I|$, of the two bases on a logarithmic scale. We omitted the "father × father-coefficient", $\theta_{0,0;1,1}$, in both cases and displayed the 50 largest coefficients in decreasing order. The solid line corresponds to coefficients of the anisotropic basis, whereas the dotted line refers to coefficients of the multiresolution basis.

[Please insert Figure 2 about here]

This picture underlines the superior ability of the anisotropic basis to compress signals like f, which have an effective dimension lower than the nominal one. Most of the power of the signal is packed in a small set of coefficients, whereas the multiresolution basis needs more functions to provide a comparable approximation to the function f. Asymptotic theory prescribes that this will have direct consequences to the performance of appropriately thresholded estimators in statistical models.
We added Gaussian white noise, which leads to the nonparametric regression model

$$Y_{ij} = f(x_i, x_j) + \varepsilon_{ij}, \quad i,j = 1,\dots,16,$$

where $x_i = (i - 1/2)/16$ and $\varepsilon_{ij} \sim N(0,\sigma^2)$ are independent. We chose $\sigma = 0.1$, which corresponds roughly to an amount of noise usually assumed in nonparametric regression, and $\sigma = 0.05$. Figures 3a and b show the true function f (solid line) and one set of observations Y_{ij} according to $\sigma = 0.1$ and $\sigma = 0.05$, respectively.

[Please insert Figures 3a and b about here]

After calculating 256 empirical coefficients in both cases, we applied thresholding at the universal thresholds $\lambda_I^{univ} = \sigma\sqrt{2\log(255)}$ to all coefficients $\tilde{\theta}_I$, $I \neq (0,0;1,1)$. We restricted our considerations to hard thresholding, that is $\delta^{(\cdot)}(\tilde{\theta}_I,\lambda) = \tilde{\theta}_I\, I(|\tilde{\theta}_I| \geq \lambda)$, since we know from extensive simulations in Marron *et al.* (1995) that hard thresholding is often better than soft and almost never essentially worse. This superiority may be quite clear in cases of strong inhomogeneity in the size of the coefficients θ_I.
Figures 4a and b show realizations of hard thresholded estimators based on the anisotropic wavelet basis and the multiresolution basis, respectively, in the case of $\sigma = 0.1$.

[Please insert Figures 4a and b about here]

The anisotropic estimator provides quite a good approximation to the true function f, whereas the isotropic one even does not capture the variability of f in x_1-direction. The numbers of active coefficients including $\tilde{\theta}_{0,0;1,1}$ are 3 and 1, and the L_2-losses are 0.174 and 0.500, respectively.
Figures 5a and b show realizations of the same estimators in the case $\sigma = 0.05$.

[Please insert Figures 5a and b about here]

The anisotropic estimator approximates f almost perfectly, whereas the isotropic one achieves a similar degree of approximation as the anisotropic estimator in the case

$\sigma = 0.1$. The numbers of active coefficients are 7 and 5, and the L_2-losses are 0.0003 and 0.174, respectively.
We studied also some other examples for f and σ as well as a modified threshold choice $\lambda_I^{mod} = \sigma\sqrt{2\log(n_{j_1+j_2}(\mathcal{B}))}$, where $n_J(\mathcal{B})$ denotes the number of wavelets in a certain basis $\mathcal{B}$ associated to levels (j_1, j_2) with $j_1 + j_2 = J$. This particular choice of the thresholds is somewhat less conservative than the above thresholds λ_I^{univ}. This alternative led sometimes to the inclusion of some more coefficients, which resulted in slightly improved estimators, but sometimes both thresholding schemes gave identical results. Concerning different choices for f and σ, the anisotropic estimators were never worse than the isotropic ones, and sometimes dramatically better.
We restricted our study to bivariate functions, mainly for the sake of a convenient visual presentation of the results. The whole program, including the use of the one-dimensional fast wavelet transform as the main building block of the implementation, can be carried out in higher dimensions, also. An appropriately thresholded estimator based on the anisotropic basis will be able to adapt to a lowerdimensional structure of a higherdimensional function, whereas the structure of the isotropic basis prevents corresponding estimators from exploiting an effective dimension lower than the nominal one. We think that this effect will become even more drastic when the difference between the full dimension and the effective dimension is larger than one.

6. Proofs

Proof of Lemma 2.1. It is easy to see that

$$\left|\frac{\partial^{[r_i]}}{\partial x_i^{[r_i]}}\mu_{k_1,\dots,k_d}(\underline{x}+he_i) - \frac{\partial^{[r_i]}}{\partial x_i^{[r_i]}}\mu_{k_1,\dots,k_d}(\underline{x})\right| \le C2^{(j_1+\dots+j_d)/2}2^{j_i r_i}h^{r_i-[r_i]} \asymp \epsilon^{-1}h^{r_i-[r_i]},$$

where the constant C can be made arbitrarily small by an appropriately small choice of C_0. Hence, $\{\mu_\theta,\ \ \theta \in \{0,\epsilon\}^D\} \subseteq H^{\underline{r}}(C)$. Since the Hölder class $H^{\underline{r}}(C)$ is embedded in $B^{\underline{r}}_{\underline{p},q}(K)$ for C small enough, we obtain the assertion. □

Proof of Theorem 2.2. By (2.12), we only have to study the decay of the functional $\Omega_\epsilon((\lambda_I^{opt}),\Theta)$ given by (2.13) as $\epsilon \to 0$, where $\Theta = \{(\theta_I) \mid \sum_I \theta_I\psi_I \in B^{\underline{r}}_{\underline{p},q}(K)\}$.
We proceed from the decomposition

$$\begin{aligned}
&\Omega_\epsilon\left((\lambda_I^{opt}),\Theta\right)\\
&\le \sum_{j_1\le j_1^*,\dots,j_d\le j_d^*}\ \sum_{k_1,\dots,k_d} \epsilon^2\\
&\quad + \sum_{i=1}^{d}\sum_{j_i=j_i^*+1}^{\infty}\ \sum_{(j_1,\dots,j_{i-1},j_{i+1},\dots,j_d):\ (j_k-j_k^*)r_k\le(j_i-j_i^*)r_i}\ \sum_{k_1,\dots,k_d} \epsilon^2\left(\frac{\lambda_I^{opt}}{\epsilon}+1\right)\varphi\left(\frac{\lambda_I^{opt}}{\epsilon}\right)\\
&\quad + \sum_{i=1}^{d}\sum_{j_i=j_i^*+1}^{\infty}\ \sum_{(j_1,\dots,j_{i-1},j_{i+1},\dots,j_d):\ j_k r_k\le j_i r_i}\ \sum_{k_1,\dots,k_d} \min\left\{(\lambda_I^{opt})^2,\theta_I^2\right\}\\
&= S_1 + S_2 + S_3. \qquad (6.1)
\end{aligned}$$

By (2.15), we have that

$$S_1 = O\left(\epsilon^2 2^{j_1^*+\ldots+j_d^*}\right) = O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right). \tag{6.2}$$

Fix for a moment i and $j_i > j_i^*$. Note that

$$\begin{aligned}\#\{I \mid \ (j_k - j_k^*)r_k \le (j_i - j_i^*)r_i\} &= O\left(2^{j_i r_i(1/r_1+\ldots+1/r_d)}\right) \\ &= O\left(2^{j_1^*+\ldots+j_d^* 2^{(j_i-j_i^*)r_i(1/r_1+\ldots+1/r_d)}}\right). \end{aligned} \tag{6.3}$$

This implies that

$$\begin{aligned}&\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ (j_k-j_k^*)r_k\le(j_i-j_i^*)r_i} \ \sum_{k_1,\ldots,k_d} \epsilon^2\left(\frac{\lambda_I^{opt}}{\epsilon}+1\right)\varphi\left(\frac{\lambda_I^{opt}}{\epsilon}\right) \\ &= O\left(\epsilon^2 2^{j_1^*+\ldots+j_d^*}\right) O\left(2^{(j_i-j_i^*)r_i(1/r_1+\ldots+1/r_d)}\sqrt{j_i-j_i^*}\exp(-\frac{\kappa^2(j_i-j_i^*)r_i(1/r_1+\ldots+1/r_d)}{2})\right) \\ &= O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right) O\left(\exp\left((j_i-j_i^*)r_i(1/r_1+\ldots+1/r_d)\left[\log(2)-\kappa^2/2\right]\right)\sqrt{j_i-j_i^*}\right).\end{aligned}$$

Since $[\log(2) - \kappa^2/2] < 0$, we obtain that

$$S_2 = O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right). \tag{6.4}$$

Let $\widetilde{p}_i = \min\{p_i, 2\}$. By $\|f\|_{b^{r_i}_{i,p_i,q}} \le K$ we obtain by straightforward calculations that

$$\sup_{(\theta_I)\in\Theta}\left\{\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ j_k r_k\le j_i r_i} \ \sum_{k_1,\ldots,k_d} |\theta_I|^{\widetilde{p}_i}\right\} \le C 2^{-j_i r_i \widetilde{p}_i} 2^{j_i r_i(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)}. \tag{6.5}$$

This implies

$$\begin{aligned}&\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ j_k r_k\le j_i r_i} \ \sum_{k_1,\ldots,k_d} \min\{(\lambda_I^{opt})^2, \theta_I^2\} \\ &\le (\lambda_I^{opt})^{2-\widetilde{p}_i} \sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ j_k r_k\le j_i r_i} \ \sum_{k_1,\ldots,k_d} |\theta_I|^{\widetilde{p}_i} \\ &= O\left((\lambda_I^{opt})^{2-\widetilde{p}_i} 2^{-j_i r_i \widetilde{p}_i} 2^{j_i r_i(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)}\right) \\ &= O\left(\epsilon^{2-\widetilde{p}_i}(j_i-j_i^*)^{1-\widetilde{p}_i/2} 2^{-j_i r_i \widetilde{p}_i} 2^{j_i r_i(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)}\right) \\ &= O\left(\epsilon^2 2^{j_1^*+\ldots+j_d^*}\right) O\left(\epsilon^{-\widetilde{p}_i} 2^{-j_i^* r_i \widetilde{p}_i} 2^{-(j_1^*+\ldots+j_d^*)\widetilde{p}_i/2}\right) * \\ &\quad * O\left(2^{-(j_i-j_i^*)r_i\widetilde{p}_i} 2^{j_i r_i(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)} 2^{(j_1^*+\ldots+j_d^*)(\widetilde{p}_i/2-1)}(j_i-j_i^*)^{1-\widetilde{p}_i/2}\right)\end{aligned} \tag{6.6}$$

First we can easily see that $\epsilon^{-\widetilde{p}_i} 2^{-j_i^* r_i \widetilde{p}_i} 2^{-(j_1^*+\ldots+j_d^*)\widetilde{p}_i/2} = O(1)$. Because of $2^{j_1^*+\ldots+j_d^*} \asymp 2^{j_i^* r_i(1/r_1+\ldots+1/r_d)}$ and by $(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d) - \widetilde{p}_i < 0$ we obtain

that

$$
\begin{aligned}
S_3 &= \sum_{i=1}^{d} \sum_{j_i=j_i^*+1}^{\infty} O\left(\epsilon^2 2^{j_1^*+\ldots+j_d^*} 2^{-(j_i-j_i^*)r_i\widetilde{p}_i} 2^{j_i r_i(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)} 2^{(j_1^*+\ldots+j_d^*)(\widetilde{p}_i/2-1)} (j_i-j_i^*)^{1-\widetilde{p}_i/2}\right) \\
&= O\left(\epsilon^2 2^{j_1^*+\ldots+j_d^*}\right) \sum_{i=1}^{d} \sum_{j_i=j_i^*+1}^{\infty} O\left(2^{(j_i-j_i^*)r_i[(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)-\widetilde{p}_i]} (j_i-j_i^*)^{1-\widetilde{p}_i/2}\right) \\
&= O\left(\epsilon^2 2^{j_1^*+\ldots+j_d^*}\right). \qquad (6.7)
\end{aligned}
$$

Collecting the estimates in (6.1), (6.2), (6.4), and (6.7) we obtain the assertion. □

Proof of Lemma 2.2. Fix for a moment j_i. Then we obtain, analogously to (6.5), that

$$
\begin{aligned}
&\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ j_1+\ldots+j_d=J} \sum_{k_1,\ldots,k_d} |\theta_I|^2 \\
&\le \left(\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ j_1+\ldots+j_d=J} \sum_{k_1,\ldots,k_d} |\theta_I|^{\widetilde{p}_i} \right)^{2/\widetilde{p}_i} \\
&= O\left(2^{-2j_i r_i+(j_1+\ldots+j_d)(2/\widetilde{p}_i-1)}\right). \qquad (6.8)
\end{aligned}
$$

Note that $j_i r_i - (j_1+\ldots+j_d)(1/\widetilde{p}_i - 1/2) \ge j_k r_k - (j_1+\ldots+j_d)(1/\widetilde{p}_k - 1/2)$ for all $k = 1,\ldots,d$ implies the relation

$$
2j_i r_i - (j_1+\ldots+j_d)(2/\widetilde{p}_i - 1) \ge (j_1+\ldots+j_d)\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d). \qquad (6.9)
$$

Using (6.8) and (6.9) we obtain that

$$
\begin{aligned}
&\|f - Proj_{\widetilde{V}_{J^*}} f\|_{L_2}^2 \\
&\le \sum_{J=J^*+1}^{\infty} \sum_{i=1}^{d} \sum_{\substack{(j_1,\ldots,j_d):\ j_1+\ldots+j_d=J, \\ j_i r_i - J(1/\widetilde{p}_i-1/2) \ge j_k r_k - J(1/\widetilde{p}_k-1/2)}} \sum_{k_1,\ldots,k_d} \theta_I^2 \\
&= \sum_{J=J^*+1}^{\infty} O\left(2^{-J\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)}\right) = O\left(2^{-J^*\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)}\right).
\end{aligned}
$$

□

Proof of Theorem 2.3. Setting formally $\lambda_I^{univ} = \infty$ if $I \notin \mathcal{I}_\epsilon$, we have again that

$$
\sup_{f \in B^r_{\underline{p},q}(K)} \left\{ E\|\widehat{f}_\epsilon^{univ} - f\|_{L_2}^2 \right\} \le C\Omega((\lambda_I^{univ}), \Theta), \qquad (6.10)
$$

where $\Theta = \{(\theta_I) \mid \sum_I \theta_I \psi_I \in B^r_{\underline{p},q}(K)\}$.

Let j_i^* be chosen such that $2^{j_i^* r_i} \asymp (\epsilon^2 \log(\epsilon^{-1}))^{-1/(1/r_1+\ldots+1/r_d+2)}$. We split up

$$\begin{aligned}
&\Omega\left((\lambda_I^{univ}), \Theta\right) \\
&\quad \leq \sum_{I \in \mathcal{I}_\epsilon} \epsilon^2 \left(\frac{\lambda_I^{univ}}{\epsilon} + 1\right) \varphi\left(\frac{\lambda_I^{univ}}{\epsilon}\right) \\
&\qquad + \sum_{j_i \leq j_i^*} \sum_{k_1,\ldots,k_d} \min\left\{(\lambda_I^{univ})^2, \theta_I^2\right\} \\
&\qquad + \sum_{j_i > j_i^*} \sum_{k_1,\ldots,k_d} \min\left\{(\lambda_I^{univ})^2, \theta_I^2\right\} \\
&\qquad + \sum_{I \notin \mathcal{I}_\epsilon} \theta_I^2. \qquad (6.11)
\end{aligned}$$

The first term on the right-hand side is obviously of order $\epsilon^2 \log(\epsilon^{-1})$.
The second term can be majorized by $C\epsilon^2 \log(\epsilon^{-1}) 2^{j_1^*+\ldots+j_d^*}$, which is $O((\epsilon^2 \log(\epsilon^{-1}))^{\vartheta(r_1,\ldots,r_d)}$
The third term can be estimated by

$$\begin{aligned}
&\sum_{i=1}^{d} (\lambda_I^{univ})^{2-\widetilde{p}_i} \sum_{j_i > j_i^*} \sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ j_k r_k \leq j_i r_i} \sum_{k_1,\ldots,k_d} |\theta_I|^{\widetilde{p}_i} \\
&\quad = \sum_i O\left((\epsilon^2 \log(\epsilon^{-1}))^{1-\widetilde{p}_i/2}\right) \sum_{j_i > j_i^*} 2^{j_i r_i [(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)-\widetilde{p}_i]} \\
&\quad = \sum_i O\left((\epsilon^2 \log(\epsilon^{-1}))^{1-\widetilde{p}_i/2} 2^{j_i^* r_i [(1-\widetilde{p}_i/2)(1/r_1+\ldots+1/r_d)-\widetilde{p}_i]}\right) \\
&\quad = O\left((\epsilon^2 \log(\epsilon^{-1}))^{\vartheta(r_1,\ldots,r_d)}\right).
\end{aligned}$$

Finally, by Lemma 2.2, the fourth term is of order $2^{-J_\epsilon^* \gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)}$, which completes the proof. □

Proof of Lemma 3.1. Let $0 \leq r_1, \ldots, r_d \leq r$. Then, for $j_1 + \ldots + j_d = J_\epsilon$,

$$\|\mu_I^{(r_1,\ldots,r_d)}\|_{L_2}^2 = O\left(2^{2j_1 r_1 + \ldots + 2j_d r_d}\right) = O\left(2^{2J_\epsilon r}\right).$$

Let $\mathcal{I} = \{I \mid \ j_1 + \ldots + j_d = J_\epsilon\}$. Further, since $\int \mu_I^{(r_1,\ldots,r_d)}(\underline{x})\, d\underline{x} = 0$, we have that

$$\sum_{J \in \mathcal{I}} \sup_{I \in \mathcal{I}} \left\{|(\mu_I^{(r_1,\ldots,r_d)}, \mu_{I+J}^{(r_1,\ldots,r_d)})|\right\} = O\left(2^{2J_\epsilon r}\right).$$

We have by (3.5) that

$$\begin{aligned}
&\left\| \sum_{I\in\mathcal{I}} \theta_I \mu_I^{(r_1,\dots,r_d)} \right\|_{L_2}^2 \\
&\leq \sum \theta_I^2 \|\mu_I^{(r_1,\dots,r_d)}\|_{L_2}^2 \\
&\quad + \sum_{I,J} \theta_I \theta_J \left| (\mu_I^{(r_1,\dots,r_d)}, \mu_J^{(r_1,\dots,r_d)}) \right| \\
&= O\left(\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1} 2^{2J_\epsilon r}\right) \\
&\quad + \sum_J \sup_{I\in\mathcal{I}} \left\{ |(\mu_I^{(r_1,\dots,r_d)}, \mu_{I+J}^{(r_1,\dots,r_d)})| \right\} \sum_I \theta_I \theta_{I+J} \\
&= O\left(\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1} 2^{2J_\epsilon r}\right) = O(1),
\end{aligned}$$

which proves the assertion. □

Proof of Lemma 3.2. Let $\mu_I(\underline{x}) = 2^{(j_1+\dots+j_d)r} \psi_{j_1,k_1}^{(-r)}(x_1) \cdots \psi_{j_d,k_d}^{(-r)}(x_d)$. Then, by integration by parts,

$$2^{(j_1+\dots+j_d)r} \theta_I = (-1)^{rd} \int f^{(r,\dots,r)}(\underline{x}) \mu_I(\underline{x})\, d\underline{x}.$$

Expanding $f^{(r,\dots,r)}$ in a homogeneous wavelet series, $f^{(r,\dots,r)} = \sum_{J\in\mathbb{Z}^{2d}} \gamma_J \psi_J$, we obtain

$$\sum \gamma_J^2 \leq C.$$

We restrict our considerations first to $\mathcal{I} = \{(j_1,\dots,j_d,k_1,\dots,k_d) \mid \ j_i \geq l \ \text{ for all } i\}$, that is to coefficients associated to products of wavelets ψ_{j_i,k_i} rather than scaling functions ϕ_{k_i}. The near-orthogonality of the system $\{\mu_I\}_{I\in\mathcal{I}}$ is characterized by the fact that

$$\sum_{J\in\mathbb{Z}^{2d}} \sup_{I\in\mathcal{I}} \{|(\psi_{J+I}, \mu_I)|\} \leq \infty.$$

This implies that

$$\begin{aligned}
&\sum_{(j_1,\dots,j_d):\ j_i\geq l} 2^{2(j_1+\dots+j_d)r} \sum_{k_1,\dots,k_d} \theta_I^2 \\
&= \sum_{I\in\mathcal{I}} \left| \int f^{(r,\dots,r)}(\underline{x}) \mu_I(\underline{x})\, d\underline{x} \right|^2 \\
&\leq \sum_{I\in\mathcal{I}} \sum_{J_1,J_2\in\mathbb{Z}^{2d}} |\gamma_{J_1+I}\gamma_{J_2+I}(\psi_{J_1+I}, \mu_I)(\psi_{J_2+I}, \mu_I)| \\
&\leq \left(\sum_{J\in\mathbb{Z}^{2d}} \sup_{I\in\mathcal{I}} \{|(\psi_{J+I}, \mu_I)|\} \right)^2 \sum_{I\in\mathcal{I}} \gamma_I^2 \leq C.
\end{aligned}$$

The remaining wavelet coefficients θ_I with $j_i = l-1$ for at least one i can be treated by similar considerations, which completes the proof. □

Proof of Theorem 3.2. By Lemma 3.2, (3.5), and (3.6) we obtain that

$$\begin{aligned} & E\|\widehat{f}_\epsilon^P - f\|_{L_2}^2 \\ &= \epsilon^2 \#\{I \mid \quad j_1 + \ldots + j_d \le J_\epsilon\} + \sum_{I:\ j_1+\ldots+j_d > J_\epsilon} \theta_I^2 \\ &= O\left(\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1}\right) + O\left(2^{-2J_\epsilon r}\right) = O\left((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}\right). \end{aligned}$$

□

Proof of Theorem 3.3. By (2.12), our coordinatewise upper risk estimate will be essentially driven by the functional $\min\{(\lambda_I^{univ})^2, \theta_I^2\}$. According to the heuristics leading to the balance relation (3.5), we choose J_ϵ as the presumably hardest level such that

$$\epsilon^2 \log(\epsilon^{-1}) 2^{J_\epsilon} J_\epsilon^{d-1} \asymp 2^{-2J_\epsilon r}$$

is satisfied. Since $J_\epsilon = O(\log(\epsilon^{-1}))$, we conclude from (2.12) that

$$\begin{aligned} & E\|\widehat{f}_\epsilon^{univ} - f\|_{L_2}^2 \\ &\le \epsilon^2 \#\mathcal{I}_\epsilon \left(\frac{\lambda_I^{univ}}{\epsilon} + 1\right) \varphi\left(\frac{\lambda_I^{univ}}{\epsilon}\right) \\ &\quad + (\lambda_I^{univ})^2 \#\{I \mid \quad j_1 + \ldots + j_d \le J_\epsilon\} \\ &\quad + \sum_{j_1+\ldots+j_d > \min\{J_\epsilon, J_\epsilon^*\}} \sum_{k_1,\ldots,k_d} \theta_I^2 \\ &= O\left(\epsilon^2 \sqrt{\log(\epsilon^{-1})} + \epsilon^2 \log(\epsilon^{-1}) 2^{J_\epsilon} J_\epsilon^{d-1} + 2^{-2(J_\epsilon \wedge J_\epsilon^*)r}\right) \\ &= O\left(\left(\epsilon^2[\log(\epsilon^{-1})]^d\right)^{2r/(2r+1)}\right) + O\left(2^{-2J_\epsilon^* r}\right). \end{aligned}$$

□

Proof of Theorem 3.4. By (2.12) and (3.7), the proof of the theorem is reduced to estimating $\Omega_\epsilon((\lambda_I^*), \Theta)$, where
$\Theta = \left\{(\theta_I) \middle|\ \sup_J \left\{2^{J(r-1/2)} J^{-(d-1)/2} \sum_{j_1+\ldots+j_d=J} \sum_{k_1,\ldots,k_d} |\theta_I|\right\} \le K\right\}$. We have

$$\begin{aligned} & \Omega_\epsilon\left((\lambda_I^*), \Theta\right) \\ &\le \sum_{j_1+\ldots+j_d \le J_\epsilon} \sum_{k_1,\ldots,k_d} \epsilon^2 \\ &\quad + \sum_{J=J_\epsilon+1}^{\infty} \sum_{j_1+\ldots+j_d=J} \sum_{k_1,\ldots,k_d} \epsilon^2 \left(\frac{\lambda_I^*}{\epsilon} + 1\right) \varphi\left(\frac{\lambda_I^*}{\epsilon}\right) \\ &\quad + \sum_{J=J_\epsilon+1}^{\infty} \sum_{j_1+\ldots+j_d=J} \sum_{k_1,\ldots,k_d} \min\left\{(\lambda_I^*)^2, \theta_I^2\right\} \\ &= T_1 + T_2 + T_3. \end{aligned} \tag{6.12}$$

From (3.5) and (3.6) we see that

$$T_1 = O\left(\epsilon^2 2^{J_\epsilon} [\log(\epsilon^{-1})]^{d-1}\right) = O\left((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}\right). \tag{6.13}$$

Since

$$\sum_{J>J_\epsilon} 2^{J-J_\epsilon}(J/J_\epsilon)^{d-1}\left(\frac{\lambda_I^*}{\epsilon}+1\right)\varphi\left(\frac{\lambda_I^*}{\epsilon}\right)$$
$$= \sum_{J>J_\epsilon} O\left(\exp\left((J-J_\epsilon)[\log(2)-\kappa^2/2]\right)(J/J_\epsilon)^{d-1}\sqrt{J-J_\epsilon}\right) = O(1),$$

we get

$$T_2 = O\left(\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1}\right)\sum_{J>J_\epsilon} O\left(2^{J-J_\epsilon}(J/J_\epsilon)^{d-1}\left(\frac{\lambda_I^*}{\epsilon}+1\right)\varphi\left(\frac{\lambda_I^*}{\epsilon}\right)\right)$$
$$= O\left(\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1}\right) = O\left((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}\right). \tag{6.14}$$

Finally, we have

$$\sum_{j_1+\ldots+j_d=J}\sum_{k_1,\ldots,k_d}\min\left\{(\lambda_I^*)^2,\theta_I^2\right\}$$
$$\leq \lambda_I^*\sum_{j_1+\ldots+j_d=J}\sum_{k_1,\ldots,k_d}|\theta_I|$$
$$= O\left(\epsilon\sqrt{J-J_\epsilon}2^{-J(r-1/2)}J^{(d-1)/2}\right)$$
$$= O\left(\epsilon 2^{-J_\epsilon(r-1/2)}J_\epsilon^{(d-1)/2}\right)O\left(2^{-(J-J_\epsilon)(r-1/2)}\sqrt{J-J_\epsilon}(J/J_\epsilon)^{(d-1)/2}\right),$$

which implies, by $\epsilon 2^{-J_\epsilon(r-1/2)}J_\epsilon^{(d-1)/2} \asymp 2^{-2J_\epsilon r} = O((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)})$, that

$$T_3 = O\left((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}\right). \tag{6.15}$$

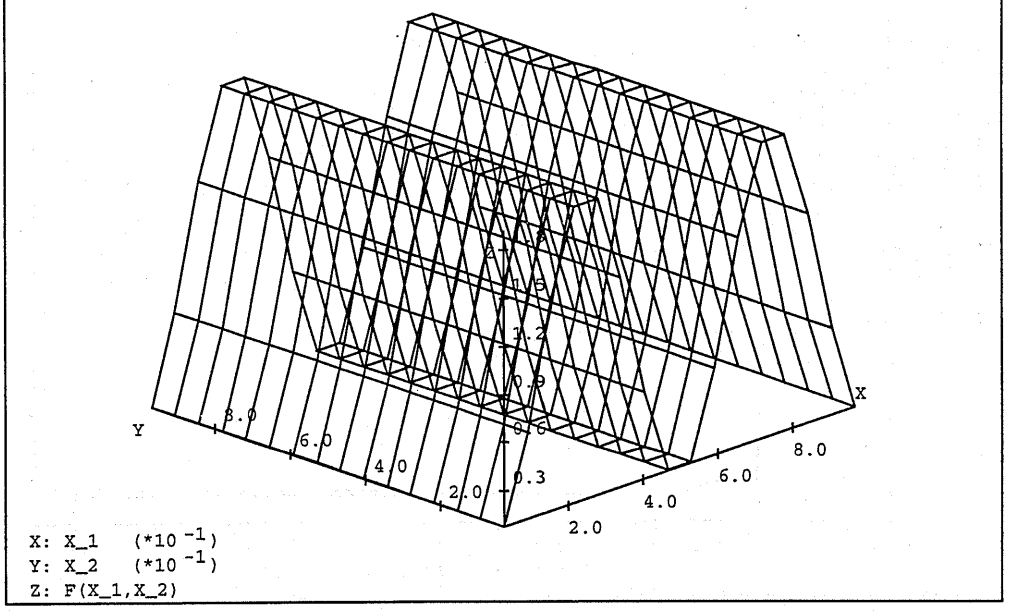
□

References

1. Amosova, N. N. (1972). On limit theorems for probabilities of moderate deviations. *Vestnik Leningrad. Univ.* **13**, 5–14. (in Russian)
2. Besov, O. V., Il'in, V. P. and Nikol'skii, S. M. (1979). *Integral Representations of Functions and Imbedding Theorems II.* Wiley, New York.
3. Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque mimimax. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **47**, 119–137.
4. Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transform. *Appl. Comp. Harmonic Anal.* **1**, 54–81.
5. Delyon, B. and Juditsky, A. (1993). Wavelet estimators, global error measures: revisited. Technical Report No. 782, Irisa, France.
6. Donoho, D. L. (1995). CART and Best-Ortho-Basis: a connection. Technical Report, Department of Statistics, Stanford University.
7. Donoho, D. L. and Johnstone, I. M. (1992). Minimax estimation via wavelet shrinkage. Technical Report No. 402, Department of Statistics, Stanford University, *Ann. Statist.*, to appear.
8. Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.
9. Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion) *J. R. Statist. Soc., Ser. B* **57**, 301–369.

10. Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single–index models. *Ann. Statist.* **21**, 157–178.
11. Härdle, W., Klinke, S. and Turlach, B. A. (1995). XploRe: an Interactive Statistical Computing Environment. Springer, New York.
12. Marron, J. S., Adak, S., Johnstone, I. M., Neumann, M. H. and Patil, P. (1995). Exact risk analysis of wavelet regression, Research Report No. SRR 035-95, Centre for Mathematics and its Applications, Australian National University, Canberra, Australia.
13. Meyer, Y. (1991). Ondelettes sur l'intervalle. *Revista Mathematica Ibero-Americana* **7** (2), 115–133.
14. Neumann, M. H. (1994). Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *J. Time Ser. Anal.*, to appear.
15. Neumann, M. H. and von Sachs, R. (1995). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra, *Ann. Statist.*, to appear.
16. Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.
17. von Sachs, R. and Schneider, K. (1994). Wavelet smoothing of evolutionary spectra by non-linear thresholding. *Appl. Comp. Harmonic Anal.*, to appear.
18. Saulis, L. and Statulevicius, V. A. (1991). *Limit Theorems for Large Deviations.* Kluwer, Dordrecht.
19. Schmeißer, H.-J. and Triebel, H. (1987). *Topics in Fourier Analysis and Function Spaces.* Geest & Portig, Leipzig.
20. Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualisation*, Wiley, New York.
21. Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689–705.
22. Tribouley, K. (1995). Practical estimation of multivariate densities using wavelet methods. *Statistica Neerlandica* **49**, 41–62.
23. Walsh, J. B. (1986). Martingales with a multidimensional parameter and stochastic integrals in the plane. In *Lectures in Probability and Statistics* (A. Dold and B. Eckmann, ed.) *Lecture Notes in Math.* **1215**, 329-491. Springer, Berlin.

FIG. 1. FUNCTION F



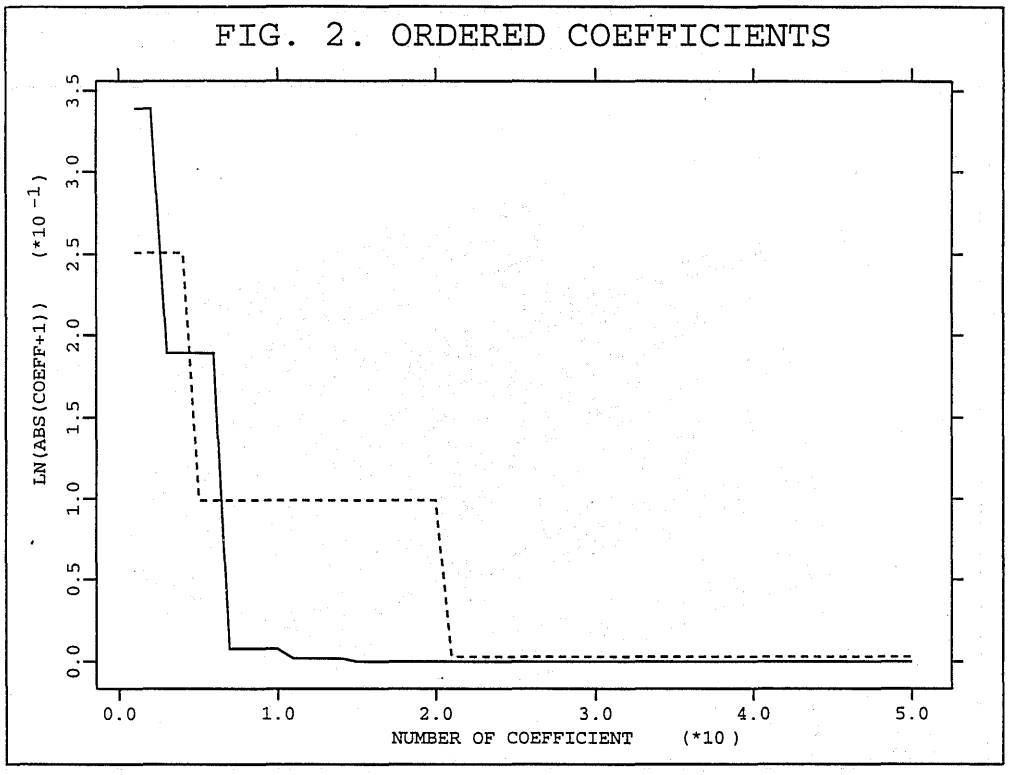


FIG. 2. ORDERED COEFFICIENTS
LN(ABS(COEFF+1)) (*10 -1)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
1.0
2.0
3.0
4.0
5.0
NUMBER OF COEFFICIENT (*10)

FIG. 3a. F VS. F+NOISE

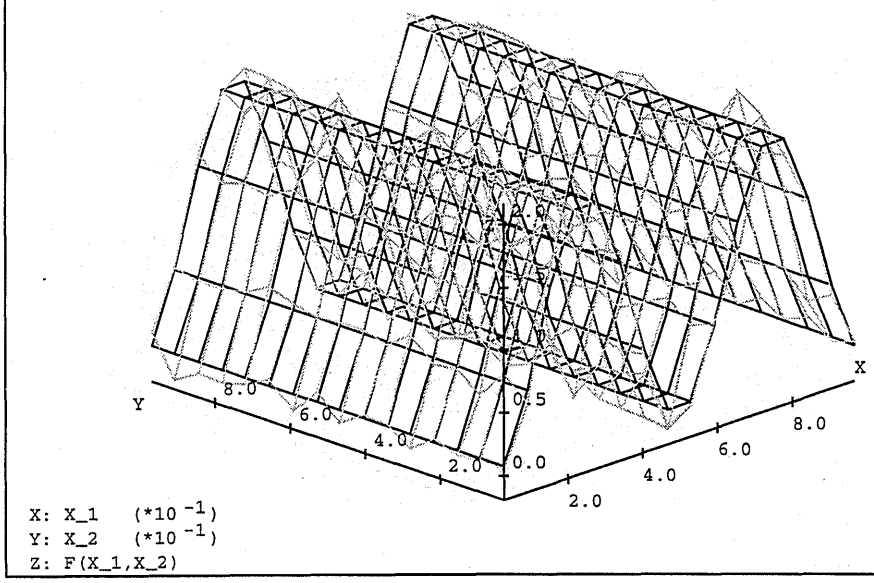


FIG. 3b. F VS. F+NOISE

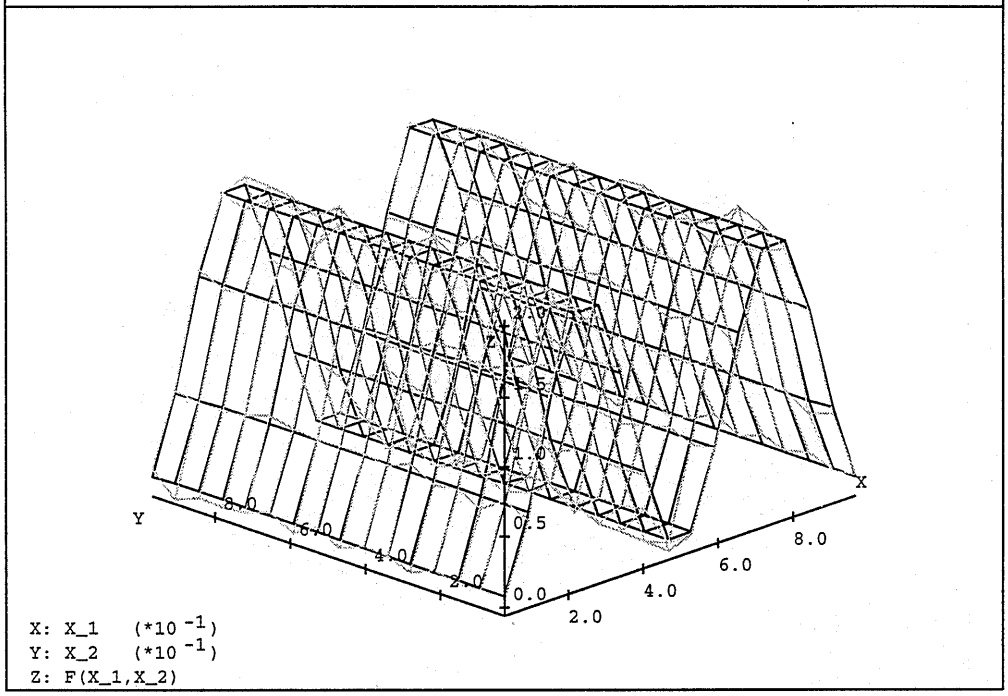
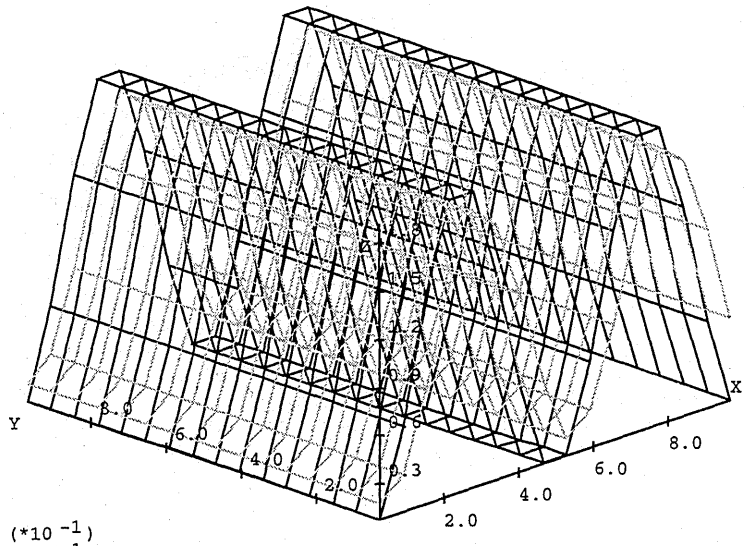


FIG. 4a. F VS. ANISOTROPIC EST.



X: X_1 (*10 $^{-1}$)
Y: X_2 (*10 $^{-1}$)
Z: F(X_1,X_2)

FIG. 4b. F VS. MULTIRES EST.

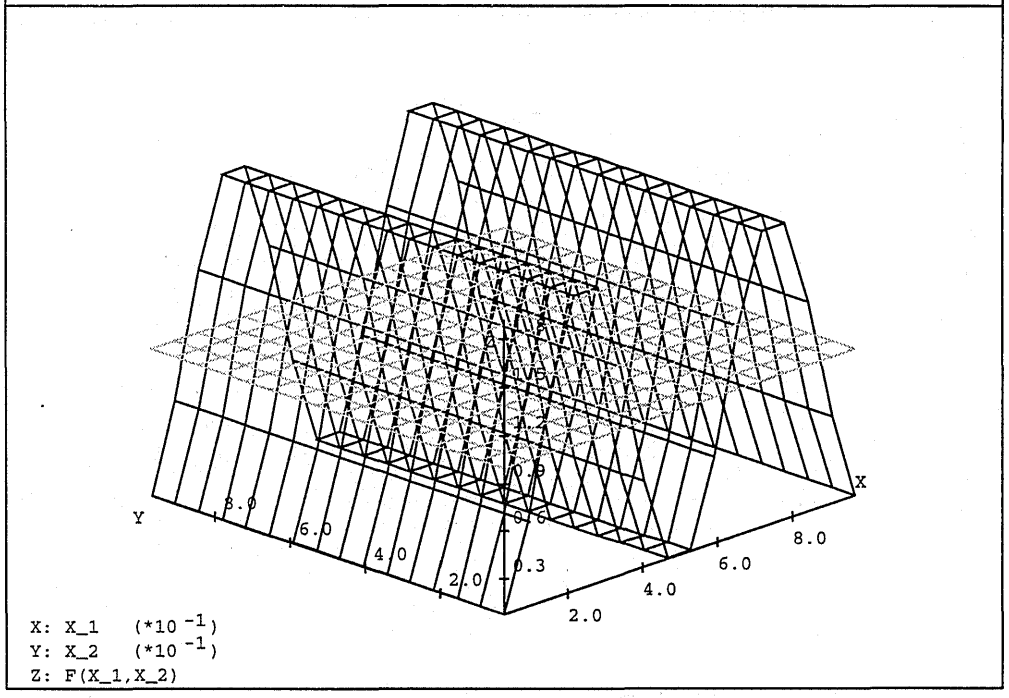


FIG. 5a. F VS. ANISOTROPIC EST.

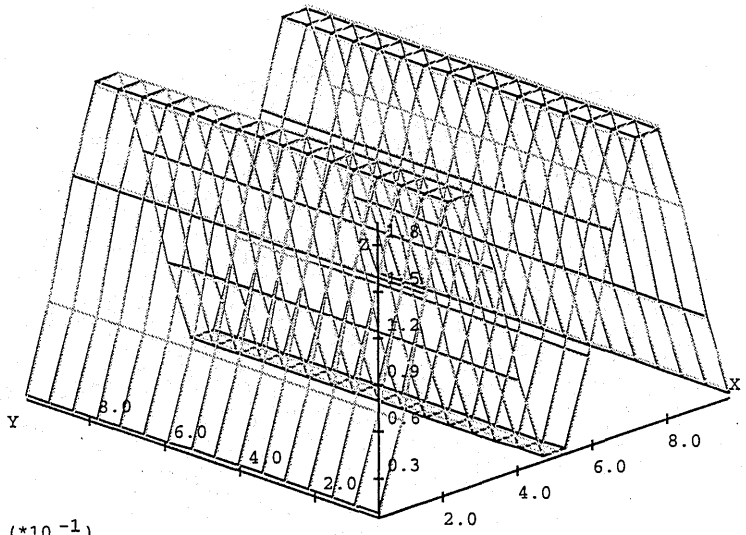
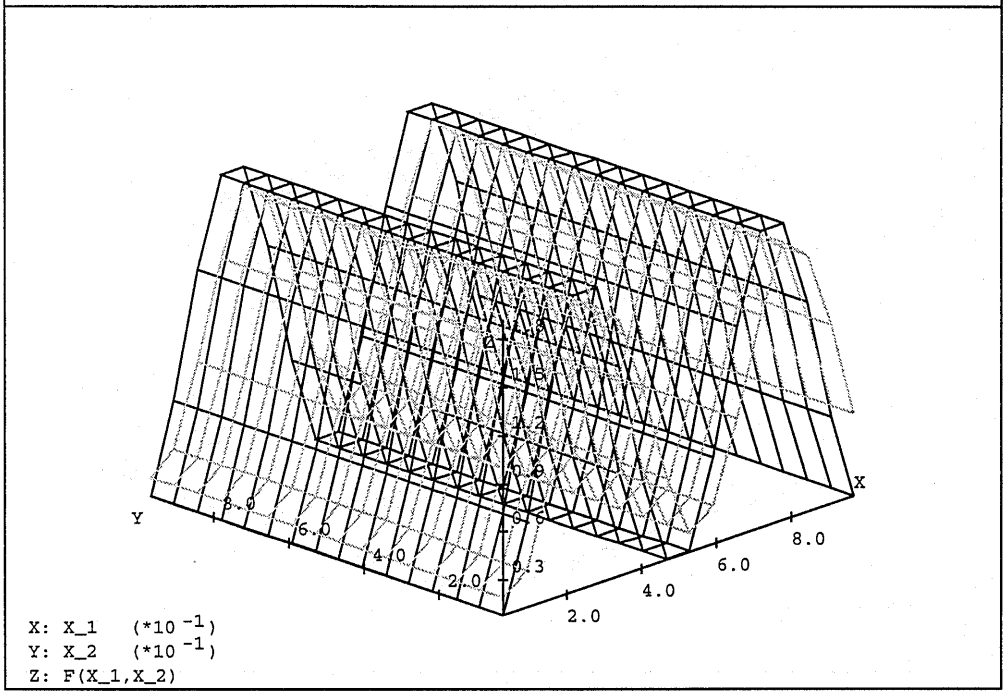


FIG. 5b. F VS. MULTIRES EST.



Recent publications of the Weierstraß–Institut für Angewandte Analysis und Stochastik

Preprints 1995

210. Günter Albinus: A thermodynamically motivated formulation of the energy model of semiconductor devices.

211. Dmitry Ioffe: Extremality of the disordered state for the Ising model on general trees.

212. Stefan Seelecke: Equilibrium thermodynamics of pseudoelasticity and quasiplasticity.

Preprints 1996

213. Björn Sandstede: Stability of N–fronts bifurcating from a twisted heteroclinic loop and an application to the FitzHugh–Nagumo equation.

214. Jürgen Sprekels, Songmu Zheng, Peicheng Zhu: Asymptotic behavior of the solutions to a Landau–Ginzburg system with viscosity for martensitic phase transitions in shape memory alloys.

215. Yuri I. Ingster: On some problems of hypothesis testing leading to infinitely divisible distributions.

216. Grigori N. Milstein: Evaluation of moment Lyapunov exponents for second order linear autonomous SDE.

217. Hans Günter Bothe: Shift spaces and attractors in non invertible horse shoes.

218. Gianfranco Chiocchia, Siegfried Prößdorf, Daniela Tordella: The lifting line equation for a curved wing in oscillatory motion.

219. Pavel Krejčí, Jürgen Sprekels: On a system of nonlinear PDE's with temperature–dependent hysteresis in one–dimensional thermoplasticity.

220. Boris N. Khoromskij, Siegfried Prößdorf: Fast computations with the harmonic Poincaré–Steklov operators on nested refined meshes.

221. Anton Bovier, Véronique Gayrard: Distribution of overlap profiles in the one–dimensional Kac–Hopfield model.

222. Jürgen Sprekels, Dan Tiba: A duality–type method for the design of beams.

223. Wolfgang Dahmen, Bernd Kleemann, Siegfried Prößdorf, Reinhold Schneider: Multiscale methods for the solution of the Helmholtz and Laplace equation.

224. Herbert Gajewski, Annegret Glitzky, Jens Griepentrog, Rolf Hünlich, Hans–Christoph Kaiser, Joachim Rehberg, Holger Stephan, Wilfried Röpke, Hans Wenzel: Modellierung und Simulation von Bauelementen der Nano– und Optoelektronik.

225. Andreas Rathsfeld: A wavelet algorithm for the boundary element solution of a geodetic boundary value problem.

226. Sergej Rjasanow, Wolfgang Wagner: Numerical study of a stochastic weighted particle method for a model kinetic equation.

227. Alexander A. Gushchin: On an information–type inequality for the Hellinger process.

228. Dietmar Horn: Entwicklung einer Schnittstelle für einen DAE–Solver in der chemischen Verfahrenstechnik.

229. Oleg V. Lepski, Vladimir G. Spokoiny: Optimal pointwise adaptive methods in nonparametric estimation.

230. Bernd Kleemann, Andreas Rathsfeld, Reinhold Schneider: Multiscale methods for boundary integral equations and their application to boundary value problems in scattering theory and geodesy.

231. Jürgen Borchardt, Ludger Bruell, Friedrich Grund, Dietmar Horn, Frank Hubbuch, Tino Michael, Horst Sandmann, Robert Zeller: Numerische Lösung großer strukturierter DAE–Systeme der chemischen Prozeßsimulation.

232. Herbert Gajewski, Klaus Zacharias: Global behaviour of a reaction–diffusion system modelling chemotaxis.

233. Fréderic Guyard, Reiner Lauterbach: Forced symmetry breaking perturbations for periodic solutions.

234. Vladimir G. Spokoiny: Adaptive and spatially adaptive testing of a nonparametric hypothesis.

235. Georg Hebermehl, Rainer Schlundt, Horst Zscheile, Wolfgang Heinrich: Simulation of monolithic microwave integrated circuits.

236. Georg Hebermehl, Rainer Schlundt, Horst Zscheile, Wolfgang Heinrich: Improved numerical solutions for the simulation of monolithic microwave integrated circuits.

237. Pavel Krejčí, Jürgen Sprekels: Global solutions to a coupled parabolic–hyperbolic system with hysteresis in 1–d magnetoelasticity.

238. Georg Hebermehl, Friedrich–Karl Hübner: Portabilität und Adaption von Software der linearen Algebra für Distributed Memory Systeme.