Statistica Sinica **10**(2000), 399-431

MULTIVARIATE WAVELET THRESHOLDING IN ANISOTROPIC FUNCTION SPACES

Michael H. Neumann

Humboldt-Universität zu Berlin

Abstract: It is well known that multivariate curve estimation under standard (isotropic) smoothness conditions suffers from the "curse of dimensionality". This is reflected by rates of convergence that deteriorate seriously in standard asymptotic settings. Better rates of convergence than those corresponding to isotropic smoothness priors are possible if the curve to be estimated has different smoothness properties in different directions and the estimation scheme is capable of making use of a lower complexity in some of the directions. We consider typical cases of anisotropic smoothness classes and explore how appropriate wavelet estimators can exploit such restrictions on the curve that require an adaptation to different smoothness properties in different directions. It turns out that nonlinear thresholding with an anisotropic multivariate wavelet basis leads to optimal rates of convergence under smoothness priors of anisotropic type. We derive asymptotic results in the model "signal plus Gaussian white noise", where a decreasing noise level mimics the standard asymptotics with increasing sample size.

Key words and phrases: Anisotropic smoothness classes, anisotropic wavelet basis, multivariate wavelet estimators, nonlinear thresholding, nonparametric curve estimation, optimal rate of convergence, smoothness classes with dominating mixed derivatives.

1. Introduction

Multivariate curve estimation is often considered with some scepticism, because it is associated with 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 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.

On the other hand, there is sometimes some hope for a successful statistical analysis in higher dimensions. Often the true complexity of a multivariate curve is considerably lower than could be expected from a statement that the curve is a member of a certain (isotropic) Sobolev class $W_p^r(\mathbb{R}^d)$ with degree of smoothness r. Scott (1992, Chapter 2) 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".

In view of this, the prior assumption that the function to be estimated lies in some isotropic smoothness class is sometimes too pessimistic. As a consequence, corresponding estimators are too rough and cannot really make full use of partial simplicities. We develop an estimator to take advantage of them. It turns out that the proposed estimation scheme is appropriate for different degrees of anisotropy of the function. First, we study the case of "weak anisotropy", that is, the function to be estimated possesses different degrees of smoothness along the different coordinate axes. In practice, this may happen if the coordinates have a different meaning such as, for example, time and frequency in the particular problem of estimating a time-varying spectral density. Second, some stronger kind of anisotropy is given when the "effective dimension" of the function is strictly less than the nominal dimension of the space. Obvious examples are multivariate functions that are composed of univariate functions or ridge functions. It is well known that such additive or single-index models allow rates of convergence that correspond to those in the one-dimensional case; see, for example, Stone (1985) and Härdle, Hall and Ichimura (1993). However, our considerations have to go beyond the case of *structural* models. Except for the rare (and in practice unlikely) cases that these assumptions are actually *exactly* fulfilled, such estimators are not even 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 (1998) (a first version of the paper dates back to the early 90's) and Donoho, Johnstone, Kerkyacharian and Picard (1995), nonlinear wavelet estimators have developed to a widely accepted alternative to such traditional methods as 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 have used an isotropic multiresolution construction, where one-dimensional basis functions from the same resolution scale are combined with each other. However, it was shown in Neumann and von Sachs (1997) 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 coordinate-wise thresholding approach does not provide the optimal rate of convergence. Neumann and von Sachs (1997) argued in favor of an alternative construction of a higher-dimensional basis, which involves tensor products of one-dimensional basis functions from different resolution scales. 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.

In Section 2 we extend these results to higher dimensions and to Besov constraints, which also admit fractional degrees of smoothness. To this end, we have to transfer the Besov conditions given in the function space to an appropriate condition on the wavelet coefficients. Since such a result is of potential independent interest we provide a separate lemma in the Appendix. It is shown that nonlinear thresholding in conjunction with the anisotropic multivariate wavelet basis leads to optimal rates of convergence in anisotropic Besov classes. In Section 3 we study another situation, which more implicitly requires directional adaptivity. We seek the largest possible function classes where our directionally adaptive estimation method still attains a rate close to that in the one-dimensional case. These classes have dominating mixed smoothness properties and are considerably larger than classes such as $W_p^{rd}(\mathbb{R}^d)$, since at most r partial derivatives are required in each direction. This could be interpreted as a restriction to functions with a lower-dimensional structure. Additive or multiplicative models are contained there as special cases but our estimation method is more flexible than commonly used special-purpose methods for such models. Since it is not explicitly based on this structural assumption, it delivers a consistent estimate even if the true curve cannot be decomposed into additively or multiplicatively connected univariate 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 of anisotropic smoothness classes. Once the estimator is accepted, one could also raise the opposite question: what is the class of problems for which the anisotropic wavelet basis is appropriate? 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 frequently 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 Sections 2 and 3 are made for the technically simplest model, signal plus Gaussian white noise.

All results are of an asymptotic nature, that is, it is assumed that the level of noise tends to zero, which mimics the asymptotics with increasing sample size in nonparametric regression and density estimation. Section 4 contains a brief discussion of how these results can be transferred to such statistically relevant settings. The proofs are contained in Section 6 while two additional technical lemmas are contained in the Appendix.

2. Wavelet Thresholding in Anisotropic Besov Classes

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

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

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

Following Besov, Il'in and Nikol'skii (1979b), we define smoothness classes in anisotropic Besov spaces. Denote by $e_i = (\delta_{i1}, \ldots, \delta_{id})'$ ($\delta_{ij} = I(i = j)$) the ith unit vector. We define the first difference of the function f in the direction of x_i as

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

and the second difference as

$$\Delta_{i,h}^2 f(\underline{x}) \;=\; \Delta_{i,h}\left(\Delta_{i,h} f(\underline{x})\right) \;=\; f(\underline{x} + 2h e_i) \;-\; 2 f(\underline{x} + h e_i) \;+\; f(\underline{x}).$$

For $i = 1, \ldots, d$, let $s_i = \lfloor r_i \rfloor$ be the largest integer strictly less than r_i. As in the one-dimensional case, we define the Besov norm in the direction of x_i as

$$\|f\|_{b^{r_i}_{i,p_i,q}} \;=\; \left(\int_0^1 |h|^{(s_i - r_i)q - 1} \left\| \Delta_{i,h}^2 \left(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f \right) \right\|^q_{L_{p_i}(g_{i,h})} dh \right)^{1/q}$$

for $q < \infty$, and

$$\|f\|_{b^{r_i}_{i,p_i,\infty}} \;=\; \sup_{0 \le h \le 1} \left\{ |h|^{s_i - r_i} \left\| \Delta_{i,h}^2 \left(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f \right) \right\|_{L_{p_i}(g_{i,h})} \right\},$$

where $g_{i,h} = (0,1)^{i-1} \times (0, 0 \vee (1-2h)) \times (0,1)^{d-i}$. Note that $\|.\|_{b^{r_i}_{i,p_i,q}}$ measures only smoothness of f in the direction of x_i. Setting $\underline{r} = (r_1, \ldots, r_d)'$ and $\underline{p} = (p_0, \ldots, p_d)'$, we define an anisotropic Besov class as

$$B^{\underline{r}}_{\underline{p},q}(K) \;=\; \left\{ f \;\middle|\; \|f\|_{B^{\underline{r}}_{\underline{p},q}} \;\leq\; K \right\},$$

where

$$\|f\|_{B^{\underline{r}}_{\underline{p},q}} \;=\; \|f\|_{L_{p_0}((0,1)^d)} \;+\; \sum_{i=1}^{d} \|f\|_{b^{r_i}_{i,p_i,q}}.$$

Assume we have a scaling function ϕ and a so-called wavelet ψ such that $\{2^{l/2}\phi(2^l \cdot - k)\}_{k\in\mathbb{Z}} \cup \{2^{j/2}\psi(2^j \cdot - k)\}_{j\geq l; k\in\mathbb{Z}}$ forms an orthonormal basis of $L_2(\mathbb{R})$. The construction of such functions ϕ and ψ, which are compactly supported, is described in Daubechies (1988). It is well known that the boundary-corrected Meyer wavelets (Meyer (1991)) or those developed by Cohen, Daubechies and Vial (1993) form orthonormal bases of $L_2[0,1]$. In both approaches Daubechies' wavelets are used to construct this basis, essentially by truncation of the above functions to the interval [0,1] and a subsequent orthonormalization step. Throughout this paper either of these bases can be used. It is denoted by $\{\phi_{l,k}\}_{k\in I_l^0} \cup \{\psi_{j,k}\}_{j\geq l; k\in I_j}$, where $\phi_{l,k}(x) = 2^{l/2}\phi(2^l x - k)$ and $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, and with certain modifications of those functions that have a support beyond the interval $[0,1]$. It is known that $\#I_j = 2^j$, and that $\#I_l^0 = 2^l$ for the Cohen, Daubechies, Vial bases, whereas for the Meyer bases, $\#I_l^0 = 2^l + N$ for some integer N depending on the regularity of the wavelet basis.

Daubechies (1992, Section 10.1) describes two possibilities for constructing multivariate wavelet bases from a given univariate basis. Let V_j be the subspace of $L_2(0,1)$ which is generated by $\{\phi_{j,k}\}_k$. It is known that

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

For $j \geq l$, denote by W_j the linear span of $\{\psi_{j,k}\}_k$. Setting $W_{l-1} := V_l$ we obtain the decomposition

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

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, for which 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

$$\begin{aligned} V_{j^*}^d = &\left(V_l^{(0)} \otimes \cdots \otimes V_l^{(0)}\right) \oplus \\ &\oplus \bigoplus_{l\le j\le j^*-1} \bigoplus_{(i_1,\ldots,i_d)\in\{0,1\}^d\setminus\{(0,\ldots,0)\}} \left(V_j^{(i_1)} \otimes \cdots \otimes V_j^{(i_d)}\right). \end{aligned} \tag{2.4}$$

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

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

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 theoretical work in statistics; see, e.g. Tribouley (1995), Delyon and Juditsky (1996), and von Sachs and Schneider (1996). 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 (1997) 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 (1997) in the framework of anisotropic Hölder classes. This author 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). An analogous lower bound was obtained by Nussbaum (1982) for smoothness classes with an L_p-restriction on some Hölder modulus of continuity.

To study the complexity of the functional class $B^{\underline{r}}_{\underline{p},q}(K)$, we take any function $\mu \in B^{r_1}_{p_1,q} \cap \cdots \cap B^{r_d}_{p_d,q}$ which is supported on $[0,1)$ and satisfies $\|\mu\|_{L_2} = 1$. Furthermore, μ and all its derivatives up to the order $\max\{s_1,\ldots,s_d\}$, where s_i is the largest integer strictly less than r_i, are assumed to be bounded. Let, for some positive C_0 whose choice is made precise in the proof of Lemma 2.1, j_i be chosen such that

$$2^{j_i} \le C_0 \epsilon^{-(2/r_i)/(1/r_1+\cdots+1/r_d+2)} < 2^{j_i+1}. \tag{2.6}$$

Define

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

It is easy to see that

$$\|\mu_{k_1,\ldots,k_d}\|_{L_2} \;=\; 1 \tag{2.7}$$

and

$$\operatorname{supp}(\mu_{k_1,\ldots,k_d}) \cap \operatorname{supp}(\mu_{k'_1,\ldots,k'_d}) = \emptyset \qquad \text{if} \quad (k_1,\ldots,k_d) \ne (k'_1,\ldots,k'_d). \tag{2.8}$$

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

$$\mu_{\underline{\theta}}(\underline{x}) \;=\; \sum_{k_1=0}^{2^{j_1}-1}\cdots\sum_{k_d=0}^{2^{j_d}-1}\theta_{k_1,\ldots,k_d}\mu_{k_1,\ldots,k_d}(\underline{x}). \tag{2.9}$$

The following lemma characterizes the complexity of the class $B^{\underline{r}}_{\underline{p},q}(K)$.

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 (2.7), (2.8), and Lemma 2.1 we obtain, by the method introduced in Bretagnolle and Huber (1979), a lower bound to the rate of convergence in $B^r_{\underline{p},q}(K)$.

Theorem 2.1. *It holds that*

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

where

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

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 classes. First, we show that the rate given in Theorem 2.1 is actually attainable by certain wavelet estimators. It turns out that this method depends on the unknown smoothness parameters $r_1,\ldots,r_d$. 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, Johnstone, Kerkyacharian and Picard (1995).

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

(A1) (i) $\int \phi(t)\,dt = 1$,
(ii) $\int \psi(t)t^k\,dt = 0$, for $0 \leq k \leq \max\{\lfloor r_1\rfloor,\ldots,\lfloor r_d\rfloor\}$.

(As mentioned in Delyon and Juditsky (1996, 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 obtain a d–dimensional orthonormal basis by setting

$$\psi_{j_1,\ldots,j_d;k_1,\ldots,k_d}(\underline{x}) = \psi_{j_1,k_1}(x_1)\cdots\psi_{j_d,k_d}(x_d). \tag{2.10}$$

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

$$\theta_I = \int_{(0,1)^d} \psi_I(\underline{x}) f(\underline{x})\,d\underline{x}. \tag{2.11}$$

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

$$\widetilde{\theta}_I \;=\; \int_{(0,1)^d} \psi_I(\underline{x})\, dY(\underline{x}) \;=\; \theta_I \;+\; \epsilon\xi_I, \tag{2.12}$$

where $\xi_I \sim N(0,1)$ are i.i.d.

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

$$\delta^{(h)}(\widetilde{\theta}_I, \lambda) \;=\; \widetilde{\theta}_I I\left(|\widetilde{\theta}_I| \geq \lambda\right);$$

(2) soft thresholding

$$\delta^{(s)}(\widetilde{\theta}_I, \lambda) \;=\; \Big(|\widetilde{\theta}_I| - \lambda\Big)_+ \operatorname{sgn}(\widetilde{\theta}_I).$$

In the following we denote either $\delta^{(h)}$ or $\delta^{(s)}$ by $\delta^{(.)}$.

As a basis for our particular choice of the threshold values we take an upper estimate of the risk of $\delta^{(.)}(\widetilde{\theta}_I, \lambda)$ as an estimate of θ_I. It follows from Lemma 1 of Donoho and Johnstone (1994) that the relation

$$E\left(\delta^{(.)}(\widetilde{\theta}_I, \lambda) \;-\; \theta_I\right)^2 \;\leq\; C\Big(\epsilon^2\Big(\frac{\lambda}{\epsilon} + 1\Big)\varphi(\frac{\lambda}{\epsilon}) \;+\; \min\{\lambda^2, \theta_I^2\}\Big) \tag{2.13}$$

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

$$\Omega_\epsilon\left((\lambda_I), \Theta\right) := \sup_{(\theta_I)\in\Theta} \Big\{ \sum_I \Big(\epsilon^2\Big(\frac{\lambda_I}{\epsilon} + 1\Big)\varphi(\frac{\lambda_I}{\epsilon}) \;+\; \min\{\lambda_I^2, \theta_I^2\}\Big)\Big\} \tag{2.14}$$

an upper rate bound for the L_2 risk of the estimator

$$\widehat{f} \;=\; \sum_I \delta^{(.)}(\widetilde{\theta}_I, \lambda_I)\psi_I,$$

uniform in the function class $\{f = \sum_I \theta_I\psi_I \mid (\theta_I) \in \Theta\}$.

A closely related quantity,

$$\Omega_\epsilon(\Theta) \;=\; \inf_{(\lambda_I)} \left\{\Omega_\epsilon\left((\lambda_I), \Theta\right)\right\}, \tag{2.15}$$

was used in Neumann and von Sachs (1997) as a characterization of the difficulty of estimation in the functional class given by Θ. A different quantity,

$$\widetilde{\Omega}_\epsilon(\Theta) \;=\; \sup_{(\theta_I)\in\Theta} \Big\{ \sum_I \min\{\epsilon^2, \theta_I^2\}\Big\},$$

has been considered in Donoho and Johnstone (1994) to establish the link between optimal statistical estimation and approximation theory. There it was shown that $\widetilde{\Omega}_\epsilon(\Theta)$ can be attained by the risk of an appropriately thresholded wavelet estimator up to some logarithmic factor, $(\log(\epsilon^{-1}))^\rho$, $\rho > 0$. We modify $\widetilde{\Omega}_\epsilon(\Theta)$ 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 $\widetilde{\Omega}_\epsilon$ 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 (2.14).

Let $j_1^*, \ldots, j_d^*$ be chosen in such a way that

$$2^{j_i^*} \asymp \epsilon^{-(2/r_i)/(1/r_1+\cdots+1/r_d+2)}. \tag{2.16}$$

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 onto the subspace $V_{j_1^*} \otimes \cdots \otimes V_{j_d^*}$; see also Lemma 2.2 below 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 at least one i, we have to employ a more refined method.

We define the following thresholds:

$$\lambda_I^{opt} = \epsilon\kappa\sqrt{\max_{1\leq i\leq d}\{(j_i - j_i^*)_+ r_i\}(1/r_1 + \cdots + 1/r_d)}, \tag{2.17}$$

where κ is any constant satisfying

$$\kappa > \sqrt{2\log(2)}. \tag{2.18}$$

These particular choices of the λ_I^{opt} are similar to those in Delyon and Juditsky (1996), proposed for isotropic smoothness classes. We consider the estimator

$$\widehat{f}_\epsilon^{opt}(\underline{x}) = \sum_I \delta^{(\cdot)}(\widetilde{\theta}_I, \lambda_I^{opt})\psi_I(\underline{x}). \tag{2.19}$$

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

Theorem 2.2. *Assume* (A1) *and*

$$\widetilde{p}_i > (1-\widetilde{p}_i/2)(1/r_1+\cdots+1/r_d), \qquad \textit{for all} \quad i=1,\ldots,d,$$

where $\widetilde{p}_i = \min\{p_i, 2\}$. *Then*

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

Notice 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, Johnstone, Kerkyacharian and Picard (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. To investigate this approach, we consider first the error incurred by truncation at a given level.

Lemma 2.2. *Assume* (A1) *and, for* $\widetilde{p}_i = \min\{p_i, 2\}$,

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

Define $\widetilde{V}_J = \bigoplus_{j_1+\cdots+j_d=J} (V_{j_1}\otimes\cdots\otimes V_{j_d})$. *Then*

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

where

$$\begin{aligned}&\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)\\ &= \{2 + [(1-2/\widetilde{p}_1)/r_1+\cdots+(1-2/\widetilde{p}_d)/r_d]\} / (1/r_1+\cdots+1/r_d).\end{aligned}$$

Provided $\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d) > 0$, 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 $\mathcal{I}_\epsilon = \{I \mid \ j_1+\cdots+j_d \le J_\epsilon^*\}$, where $2^{J_\epsilon^*} = O(\epsilon^{-\nu})$ for some $\nu<\infty$. We consider the estimator

$$\widehat{f}_\epsilon^{univ}(\underline{x}) = \sum_{I\in\mathcal{I}_\epsilon} \delta^{(.)}(\widetilde{\theta}_I, \lambda_\epsilon^{univ})\psi_I(\underline{x}), \tag{2.20}$$

where

$$\lambda_\epsilon^{univ} \;=\; \epsilon\sqrt{2\log(\#\mathcal{I}_\epsilon)}. \tag{2.21}$$

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, Johnstone, Kerkyacharian and Picard (1995), it is not surprising at all that $\widehat{f}_\epsilon^{univ}$ attains the optimal rate of convergence up to some logarithmic factor. For the reader's convenience we formally establish this in the following theorem.

Theorem 2.3. *Assume* (A1) *and, for* $\widetilde{p}_i = \min\{p_i, 2\}$,

$$\widetilde{p}_i \;>\; (1-\widetilde{p}_i/2)(1/r_1+\cdots+1/r_d), \qquad \textit{for all} \quad i=1,\ldots,d.$$

Then

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

Since, under the above condition, $\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d) > 0$ holds, 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 in the class $B^{\underline{r}}_{\underline{p},q}(K)$ by a projection estimator onto the space $V_{j^*}\otimes\cdots\otimes V_{j^*}$, 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 $\int K_1(x)x^k\,dx = \delta_{0k}$, for $0 \leq k \leq \max\{\lfloor r_1\rfloor,\ldots,\lfloor r_d\rfloor\}$. Choosing a vector of directional bandwidths, $\underline{h} = (h_1,\ldots,h_d)$, with $h_i \asymp \epsilon^{(2/r_i)/(1/r_1+\cdots+1/r_d+2)}$ we obtain the optimal rate of convergence.

Remark 2. So far, our considerations were restricted to anisotropic Besov classes that impose some *global* smoothness condition. This is sufficient for our particular purpose of investigating 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 favorable behavior of our estimator in the case of spatially varying anisotropic smoothness properties of f as well.

3. Wavelet Estimation in Smoothness Classes with Dominating Mixed Derivatives

In this section we proceed with the investigation of what wavelet methods can offer for multivariate estimation problems. Although nonlinear thresholding in the anisotropic wavelet basis is used again, 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 function class in $(0,1)^d$ is still one, or at least very close to one.

Some motivation for the definition of the particular function classes considered here arises 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. Bearing in mind that nonlinear thresholding in the anisotropic basis adapts locally to the presence of a different complexity in different directions, we seek an as large as possible class of functions that allows a rate of convergence comparable to the one-dimensional case. It turns out that appropriate function 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 such as Besov constraints would also be possible. Let, for some integer r and $K < \infty$,

$$\mathcal{F}_r^{(d)}(K) = \Big\{ f \Big| \quad \sum_{0 \le r_1,\dots,r_d \le r} \Big\| \frac{\partial^{r_1+\cdots+r_d}}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}} f \Big\|_{L_2((0,1)^d)} \le K \Big\}. \tag{3.1}$$

In contrast to usually considered isotropic smoothness classes such as the d-dimensional Sobolev class,

$$\mathcal{F}_s(K) = \Big\{ f \Big| \quad \sum_{0 \le r_1+\cdots+r_d \le s} \Big\| \frac{\partial^{r_1+\cdots+r_d}}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}} f \Big\|_{L_2((0,1)^d)} \le K \Big\},$$

the mixed derivatives play the dominant role 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 such as

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

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

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

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 truly nonparametric function 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 (2.10), 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) $\int \psi(t)t^k dt = 0$, for $0 \le k \le r$.

It will be shown in Lemma 3.2 below that membership in $\mathcal{F}_r^{(d)}(K)$ implies a constraint on the wavelet coefficients of the type

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

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

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

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 heuristic, a sharp lower bound to the minimax rate of convergence will be generated by the subproblem 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 $\{(j_1, \dots, j_d) \mid j_1 + \cdots + j_d = J\}$, 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. According to (3.4) and Lemma 3.1 below, the hardest level J_ϵ satisfies the relation

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

This yields $J_\epsilon \asymp \log(\epsilon^{-1})$ and, therefore,

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

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

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

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

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

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

Lemma 3.1. *Let J_ϵ be chosen according to* (3.5). *If C_0 is small enough, then*

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

Since the μ_I's are orthogonal, we immediately obtain by the hypercube method of Bretagnolle and Huber (1979) that the minimax rate of convergence in $\mathcal{F}_r^{(d)}(K)$ can be bounded from below by $\epsilon^2 \#\{I:\ j_1 + \cdots + j_d = J_\epsilon\}$, which leads to the following theorem.

Theorem 3.1. *It holds that*

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

This rate of convergence could of course be expected, perhaps except for the logarithmic factor. This factor is also present in approximation-theoretic results for smoothness classes with dominating mixed derivatives; see Wahba (1990, pp.145-146) and Barron (1993, Section II).

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

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

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

A similar result for Besov-type spaces with dominating mixed derivatives and coefficients of a function in a tensor-product basis constructed from Franklin functions or Schauder functions can be found in Kamont (1994, Theorem A.1). 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 $\mathcal{F}_r^{(d)}(K)$.

Theorem 3.2. *Assume* (A2). *Let* J_ϵ *be defined according to the balance relation* (3.5) *and let*

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

Then

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

Remark 3. 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. Notice that the wavelet estimator $\widehat{f}_\epsilon^P$ projects the observations onto the space $\bigoplus_{j_1+\cdots+j_d=J_\epsilon} V_{j_1} \otimes \cdots \otimes V_{j_d}$. 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 4. An assumption of different degrees of smoothness in different directions such as

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

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 $(\epsilon^2[\log(\epsilon^{-1})]^{\widetilde{d}-1})^{2r/(2r+1)}$, where $\widetilde{d} = \#\{R_i \mid \quad R_i = \min\{R_j\}\}$ 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 (2.20) and (2.21).

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, Johnstone, Kerkyacharian and Picard (1995).

As shown in Donoho and Johnstone (1998) for univariate Besov classes, the necessity for nonlinear estimators occurs in function 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 function 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

$$(\#\mathcal{I})^{-1/2} \sum_{I\in\mathcal{I}} |\theta_I| \;\le\; \left(\sum_{I\in\mathcal{I}} |\theta_I|^2\right)^{1/2}, \tag{3.7}$$

we define the following function class:

$$\mathcal{F}_{r,1,\infty}^{(d)}(K) = \Big\{f = \sum_I \theta_I\psi_I \Big| \sup_J \{2^{J(r-1/2)} J^{-(d-1)/2} \sum_{j_1+\cdots+j_d=J} \sum_{k_1,\ldots,k_d} |\theta_I|\} \le K\Big\}. \tag{3.8}$$

By (3.7) and 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 according to (3.5). We define the thresholds

$$\lambda_I^* = \begin{cases} 0 & \text{if } j_1 + \cdots + j_d \le J_\epsilon, \\ \epsilon\kappa\sqrt{(j_1 + \cdots + j_d) - J_\epsilon} & \text{if } j_1 + \cdots + j_d > J_\epsilon, \end{cases} \tag{3.9}$$

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

$$\widehat{f}_\epsilon^*(\underline{x}) = \sum_I \delta^{(\cdot)}(\widetilde{\theta}_I, \lambda_I^*)\psi_I(\underline{x}). \tag{3.10}$$

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*

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

Remark 5. A different scale of multivariate smoothness classes which allow dimension-independent rates for data compression and, therefore, also dimension-independent rates in statistical estimation is considered by Barron (1993). Barron's classes are defined via an integrability condition on the Fourier transform and allow rates of convergence in the context of neural networks known from the one-dimensional case. As pointed out in Section 2 in Barron (1993), the smoothness conditions in such classes cannot be equivalently formulated in terms of restrictions on the derivatives. Hence, our study complements the work of Barron (1993) in a more traditional framework.

4. Discussion

4.1. More realistic statistical models

We used the model "signal plus Gaussian white noise" since it contains all essential features of problems in nonparametric curve estimation and since we do not want to hide the main messages of this article behind technicalities that unavoidably occur with more realistic statistical models. It is well known that analogous results are usually attainable for models in which statisticians are really interested. Under reasonable assumptions and by setting $\epsilon \asymp n^{-1/2}$, the lower bounds from the previous sections can be transferred both to nonparametric regression (even with non-Gaussian errors) and density estimation. (Note that the hypercube approach of Bretagnolle and Huber (1979) was developed in the density estimation setting.)

On the other hand, convergence rates derived in the Gaussian white noise model are also attainable in usually considered settings of nonparametric curve estimation. It turns out that in many of those models appropriate versions of empirical wavelet coefficients are asymptotically normally distributed. As mentioned in Neumann (1995), asymptotic normality in terms of probabilities of large deviations leads to the equivalence, on the level of risks, to the Gaussian case. This is described in more detail in Neumann and von Sachs (1995), while Neumann and Spokoiny (1995), Neumann (1996), Neumann and von Sachs (1997) and Dahlhaus, Neumann and von Sachs (1998) provide rigorous derivations in the particular cases of nonparametric regression, spectral density estimation for stationary and nonstationary processes, and nonparametric estimation of time-varying autoregression parameters, respectively.

4.2. What can wavelets offer for multivariate curve estimation?

It is well known that nonparametric curve estimation in conventional (isotropic) smoothness classes suffers from the curse of dimensionality. The minimax rate of convergence in Sobolev classes $W_p^r(\mathbb{R}^d)$ is known to be $\epsilon^{4r/(2r+d)}$ in the Gaussian white noise model or, analogously, $n^{-2r/(2r+d)}$ in nonparametric regression and density estimation. These rates deteriorate seriously when d grows.

On the other hand, assessing the difficulty of estimation by the membership to such a large function class is sometimes rather pessimistic. Under certain circumstances, when the true complexity of the target function f is lower than those prescribed by $W_p^{r(f)}(\mathbb{R}^d)$, where $r(f)$ is the maximal value such that f is contained in this class, one can hope for better rates of convergence. Depending on particular prior assumptions such as additivity for example, there exist specific estimators that enjoy better rates of convergence. A typical example is the integration estimator proposed by Tjøstheim and Auestad (1994) and Linton and Nielsen (1995) in nonparametric additive models. However, although such specific procedures are quite successful in cases they are designed for, they can completely fail if the true function does not obey the presumed structural assumption. Our anisotropic wavelet estimators attain optimal or near-optimal rates of convergence in cases where directional adaptation improves the rate corresponding to the respective isotropic smoothness class. However, they are more flexible than some of the specific methods since they are even consistent without extra structural assumptions.

5. Proofs

Proof of Lemma 2.1. First we have by (2.6),

$$\|\mu_{\underline{\theta}}\|_{p_0} \le \|\mu_{\underline{\theta}}\|_\infty = O\left(\epsilon 2^{(j_1+\cdots+j_d)/2}\right) = O(1). \tag{5.1}$$

For $q < \infty$, we have

$$\begin{aligned}\|\mu_{\underline{\theta}}\|_{b^{r_i}_{i,p_i,q}} &\le 2^{1/q}\Big(\int_0^{2^{-j_i}} |h|^{(s_i-r_i)q-1}\Big\|\Delta^2_{i,h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu_{\underline{\theta}}\Big)\Big\|^q_{L_{p_i}(g_{i,h})}\,dh\Big)^{1/q}\\ &\quad + 2^{1/q}\Big(\int_{2^{-j_i}}^1 |h|^{(s_i-r_i)q-1}\Big\|\Delta^2_{i,h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu_{\underline{\theta}}\Big)\Big\|^q_{L_{p_i}(g_{i,h})}\,dh\Big)^{1/q}\\ &= I_1 + I_2, \end{aligned} \tag{5.2}$$

say. Since, for $h \le 2^{-j_i}$,

$$\operatorname{supp}\Big(\Delta^2_{i,h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu_{k_1,\ldots,k_d}\Big)\Big) \cap \operatorname{supp}\Big(\Delta^2_{i,h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu_{k'_1,\ldots,k'_d}\Big)\Big) = \emptyset$$

holds if $|k_i - k'_i| > C$ or $k_j \ne k'_j$ for any $j \ne i$, we obtain

$$\begin{aligned}\Big\|\Delta^2_{i,h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu_{\underline{\theta}}\Big)\Big\|^{p_i}_{p_i} &= O\Big(2^{j_1+\cdots+j_d}\epsilon^{p_i}\Big\|\Delta^2_{i,h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu_{k_1,\ldots,k_d}\Big)\Big\|^{p_i}_{p_i}\Big)\\ &= O\Big(\epsilon^{p_i}2^{j_i s_i p_i}2^{p_i(j_1+\cdots+j_d)/2}\Big\|\Delta^2_{i,2^{j_i}h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu\otimes\cdots\otimes\mu\Big)\Big\|^{p_i}_{p_i}\Big).\end{aligned}$$

This implies

$$\begin{aligned} I_1 &= O\Big(\epsilon 2^{j_i s_i}2^{(j_1+\cdots+j_d)/2}\\ &\quad \Big(\int_0^{2^{-j_i}} |h|^{(s_i-r_i)q-1}\Big\|\Delta^2_{i,2^{j_i}h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu\otimes\cdots\otimes\mu\Big)\Big\|^q_{L_{p_i}(g_{i,h})}\,dh\Big)^{1/q}\Big)\\ &= O\Big(\epsilon 2^{j_i s_i}2^{(j_1+\cdots+j_d)/2}2^{j_i(r_i-s_i)}\\ &\quad \Big(\int_0^1 |h|^{(s_i-r_i)q-1}\Big\|\Delta^2_{i,h}\Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu\otimes\cdots\otimes\mu\Big)\Big\|^q_{L_{p_i}(g_{i,h})}\,dh\Big)^{1/q}\Big)\\ &= O\left(\epsilon 2^{j_i s_i}2^{(j_1+\cdots+j_d)/2}2^{j_i(r_i-s_i)}\right) = O(1). \end{aligned} \tag{5.3}$$

Since the $\mu_{k_1,\ldots,k_d}$ have disjoint support, we get

$$\Big\|\frac{\partial^{s_i}}{\partial x_i^{s_i}}\mu_{\underline{\theta}}\Big\|_{p_i} = O\left(\epsilon 2^{(j_1+\cdots+j_d)/2}2^{j_i s_i}\right).$$

Hence, we obtain

$$\begin{aligned} I_2 &= O\Big(\epsilon 2^{(j_1+\cdots+j_d)/2}2^{j_i s_i}\Big(\int_{2^{-j_i}}^1 |h|^{(s_i-r_i)q-1}\,dh\Big)^{1/q}\Big)\\ &= O\left(\epsilon 2^{(j_1+\cdots+j_d)/2}2^{j_i r_i}\right) = O(1). \end{aligned} \tag{5.4}$$

From (5.1) to (5.4) we obtain, for $q < \infty$ and C_0 sufficiently small, that

$$\|\mu_{\underline{\theta}}\|_{B^{\underline{r}}_{\underline{p},q}} \le K.$$

The proof for $q = \infty$ is analogous.

Proof of Theorem 2.2. By (2.13), we only have to study the decay of the functional $\Omega_\epsilon((\lambda_I^{opt}), \Theta)$ given by (2.14) as $\epsilon \to 0$, where $\Theta = \{(\theta_I) \mid \sum_I \theta_I \psi_I \in B^r_{\underline{p},q}(K)\}$. We proceed from the decomposition

$$\begin{aligned}
&\Omega_\epsilon\left((\lambda_I^{opt}), \Theta\right) \\
&\le \sum_{(j_1,\ldots,j_d):\ j_i \le j_i^* \,\forall\, i} \ \sum_{k_1,\ldots,k_d} \epsilon^2 \\
&\quad + \sum_{i=1}^d \sum_{j_i=j_i^*+1}^{\infty} \sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d): j_k r_k \le j_i r_i \,\forall\, k} \ \sum_{k_1,\ldots,k_d} \epsilon^2\Big(\frac{\lambda_I^{opt}}{\epsilon} + 1\Big)\varphi\Big(\frac{\lambda_I^{opt}}{\epsilon}\Big) \\
&\quad + \sum_{i=1}^d \sum_{j_i=j_i^*+1}^{\infty} \sup_{(\theta_I)\in\Theta} \Big\{ \sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d): j_k r_k \le j_i r_i \forall k} \ \sum_{k_1,\ldots,k_d} \min\Big\{(\lambda_I^{opt})^2, \theta_I^2\Big\} \Big\} \\
&= S_1 + S_2 + S_3, \qquad (5.5)
\end{aligned}$$

say. By (2.16), we have

$$\begin{aligned}
S_1 &= O\left(\epsilon^2 2^{j_1^*+\cdots+j_d^*}\right) \\
&= O\left(\epsilon^{2[1-(1/r_1+\cdots+1/r_d)/(1/r_1+\cdots+1/r_d+2)]}\right) = O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right). \qquad (5.6)
\end{aligned}$$

Fix i and $j_i > j_i^*$. Note that

$$\begin{aligned}
\#\{I \mid \ j_k r_k \le j_i r_i \ \text{ for all } \ k\} &= O\left(2^{j_i r_i (1/r_1+\cdots+1/r_d)}\right) \\
&= O\left(2^{j_1^*+\cdots+j_d^*} \ 2^{(j_i-j_i^*) r_i (1/r_1+\cdots+1/r_d)}\right). \qquad (5.7)
\end{aligned}$$

This implies that

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

Since $\log(2) - \kappa^2/2 < 0$, we find

$$S_2 = O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right). \qquad (5.8)$$

Choose $\widehat{p}_i$ such that $1 \le \widehat{p}_i \le p_i$, $\widehat{p}_i < 2$ and $\widehat{p}_i > (1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)$. It follows immediately from the definition of anisotropic Besov spaces that $B^{\underline{r}}_{\underline{p},q} \subseteq B^{\underline{r}}_{(p_0,\widehat{p}_1,\ldots,\widehat{p}_d),q}$. Hence, we obtain by Lemma A.1 that

$$\begin{aligned}&\sup_{(\theta_I)\in\Theta}\Big\{\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\, j_kr_k\le j_ir_i\,\forall\,k}\ \sum_{k_1,\ldots,k_d}|\theta_I|^{\widehat{p}_i}\Big\}\\ &\le C2^{-j_ir_i\widehat{p}_i}2^{j_ir_i(1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)}.\end{aligned}\tag{5.9}$$

This implies, for $j_i > j_i^*$,

$$\begin{aligned}&\sup_{(\theta_I)\in\Theta}\Big\{\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\, j_kr_k\le j_ir_i\,\forall\,k}\ \sum_{k_1,\ldots,k_d}\min\{(\lambda_I^{opt})^2,\theta_I^2\}\Big\}\\ &\le (\lambda_I^{opt})^{2-\widehat{p}_i}\sup_{(\theta_I)\in\Theta}\left\{\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\, j_kr_k\le j_ir_i\,\forall\,k}\ \sum_{k_1,\ldots,k_d}|\theta_I|^{\widehat{p}_i}\right\}\\ &= O\left(\epsilon^{2-\widehat{p}_i}(j_i-j_i^*)^{1-\widehat{p}_i/2}\quad 2^{-j_ir_i\widehat{p}_i}2^{j_ir_i(1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)}\right)\\ &= O\left(\epsilon^2 2^{j_i^*r_i(1/r_1+\cdots+1/r_d)}\right)O\left(\epsilon^{-\widehat{p}_i}2^{-j_i^*r_i\widehat{p}_i}2^{-j_i^*r_i\widehat{p}_i/2(1/r_1+\cdots+1/r_d)}\right)\\ &\quad\times O\left(2^{(j_i-j_i^*)r_i[(1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)-\widehat{p}_i]}(j_i-j_i^*)^{1-\widehat{p}_i/2}\right).\end{aligned}\tag{5.10}$$

From (2.16) we get $\epsilon^2 2^{j_i^*r_i(1/r_1+\cdots+1/r_d)} = \epsilon^2 2^{j_1^*+\cdots+j_d^*} \asymp \epsilon^{2\vartheta(r_1,\ldots,r_d)}$ and $\epsilon^{-\widehat{p}_i}2^{-j_i^*r_i\widehat{p}_i}2^{-j_i^*r_i\widehat{p}_i/2(1/r_1+\cdots+1/r_d)} \asymp 1$. Since $[(1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)-\widehat{p}_i] < 0$ we obtain

$$\begin{aligned}S_3 &= O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right)\sum_{i=1}^d\sum_{j_i=j_i^*+1}^{\infty}O\left(2^{(j_i-j_i^*)r_i[(1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)-\widehat{p}_i]}(j_i-j_i^*)^{1-\widehat{p}_i/2}\right)\\ &= O\left(\epsilon^{2\vartheta(r_1,\ldots,r_d)}\right).\end{aligned}\tag{5.11}$$

Collecting the estimates in (5.5), (5.6), (5.8), and (5.11) the assertion follows.

Proof of Lemma 2.2. By (5.9) we obtain, for $i = 1,\ldots,d$,

$$\sum_{k_1,\ldots,k_d}|\theta_I|^2 \le \Big(\sum_{k_1,\ldots,k_d}|\theta_I|^{\widetilde{p}_i}\Big)^{2/\widetilde{p}_i} = O\left(2^{-2j_ir_i+(j_1+\cdots+j_d)(2/\widetilde{p}_i-1)}\right).\tag{5.12}$$

Suppose that i minimizes the term on the right-hand side of (5.12), that is,

$$2j_ir_i+(j_1+\cdots+j_d)(1-2/\widetilde{p}_i) \ge 2j_kr_k+(j_1+\cdots+j_d)(1-2/\widetilde{p}_k), \qquad \text{for all}\quad k.$$

Dividing both sides by r_k and summing up over $k = 1,\ldots,d$ we get

$$\begin{aligned}&\left[2j_ir_i+(j_1+\cdots+j_d)(1-2/\widetilde{p}_i)\right](1/r_1+\cdots+1/r_d)\\ &\ge (j_1+\cdots+j_d)\left\{2+\left[(1-2/\widetilde{p}_1)/r_1+\cdots+(1-2/\widetilde{p}_d)/r_d\right]\right\},\end{aligned}$$

which is equivalent to

$$2j_ir_i + (j_1 + \cdots + j_d)(1 - 2/\widetilde{p}_i) \geq (j_1 + \cdots + j_d)\gamma(r_1, \ldots, r_d, p_1, \ldots, p_d). \quad (5.13)$$

(5.12) and (5.13) imply

$$\begin{aligned}
&\|f - \mathrm{Proj}_{\widetilde{V}_{J^*}} f\|_{L_2}^2 \\
&= \sum_{J>J^*} \sum_{j_1+\cdots+j_d=J} \sum_{k_1,\ldots,k_d} \theta_I^2 \\
&\leq \sum_{J>J^*} \sum_{i=1}^{d} \sum_{j_i:\, 2j_ir_i \geq J(\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)+2/\widetilde{p}_i-1)} \\
&\qquad \#\{(j_1, \ldots, j_{i-1}, j_{i+1}, \ldots, j_d) :\ j_1 + \cdots + j_d = J \\
&\qquad\quad \text{and } 2j_ir_i + J(1-2/\widetilde{p}_i) \geq 2j_kr_k + J(1-2/\widetilde{p}_k) \forall\ k\}\ O\left(2^{-2j_ir_i + J(2/\widetilde{p}_i-1)}\right) \\
&= \sum_{J>J^*} O\left(2^{-J\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)}\right) = O\left(2^{-J^*\gamma(r_1,\ldots,r_d,p_1,\ldots,p_d)}\right).
\end{aligned}$$

Proof of Theorem 2.3. As before, we have that

$$\sup_{f \in B^{\underline{r}}_{\underline{p},q}(K)} \left\{ E\|\widehat{f}_\epsilon^{univ} - f\|_{L_2}^2 \right\} \leq C\Omega((\lambda_\epsilon^{univ}), \Theta), \quad (5.14)$$

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

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

$$\begin{aligned}
\Omega\left((\lambda_\epsilon^{univ}), \Theta\right) \leq &\sum_{I \in \mathcal{I}_\epsilon} \epsilon^2 \Big(\frac{\lambda_\epsilon^{univ}}{\epsilon} + 1 \Big) \varphi\Big(\frac{\lambda_\epsilon^{univ}}{\epsilon} \Big) \\
&+ \sup_{(\theta_I) \in \Theta} \Big\{ \sum_{\substack{j_1+\cdots+j_d \leq J_\epsilon^* \\ j_i \leq j_i' \text{ for all } i=1,\ldots,d}} \sum_{k_1,\ldots,k_d} \min\left\{ (\lambda_\epsilon^{univ})^2, \theta_I^2 \right\} \Big\} \\
&+ \sup_{(\theta_I) \in \Theta} \Big\{ \sum_{\substack{j_1+\cdots+j_d \leq J_\epsilon^* \\ j_i > j_i' \text{ for some } i}} \sum_{k_1,\ldots,k_d} \min\left\{ (\lambda_\epsilon^{univ})^2, \theta_I^2 \right\} \Big\} \\
&+ \sup_{(\theta_I) \in \Theta} \Big\{ \sum_{I \notin \mathcal{I}_\epsilon} \theta_I^2 \Big\}. \quad (5.15)
\end{aligned}$$

The first term on the right-hand side is obviously of order $\epsilon^2\sqrt{\log(\epsilon^{-1})}$. The second term can be majorized by $C(\lambda_\epsilon^{univ})^2 2^{j_1' + \cdots + j_d'}$, which is $O((\epsilon^2 \log(\epsilon^{-1}))^{\vartheta(r_1,\ldots,r_d)})$. As in the proof of Theorem 2.2, choose $\widehat{p}_i$ such that $1 \leq \widehat{p}_i \leq p_i$, $\widehat{p}_i < 2$ and

$\widehat{p}_i > (1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)$. By (5.9), the third term can be estimated by

$$\begin{aligned}
&\sup_{(\theta_I)\in\Theta}\Big\{\sum_{i=1}^{d}(\lambda_\epsilon^{univ})^{2-\widehat{p}_i}\sum_{j_i>j_i'}\sum_{(j_1,\ldots,j_{i-1},j_{i+1},\ldots,j_d):\ j_k r_k\le j_i r_i\,\forall\, k}\ \sum_{k_1,\ldots,k_d}|\theta_I|^{\widehat{p}_i}\Big\}\\
&=\sum_i O\left((\epsilon^2\log(\epsilon^{-1}))^{1-\widehat{p}_i/2}\right)\sum_{j_i>j_i'}O\left(2^{j_i r_i[(1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)-\widehat{p}_i]}\right)\\
&=\sum_i O\left((\epsilon^2\log(\epsilon^{-1}))^{1-\widehat{p}_i/2}2^{j_i' r_i[(1-\widehat{p}_i/2)(1/r_1+\cdots+1/r_d)-\widehat{p}_i]}\right)\\
&=O\left((\epsilon^2\log(\epsilon^{-1}))^{\vartheta(r_1,\ldots,r_d)}\right).
\end{aligned}$$

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 $0\le r_1,\ldots,r_d\le r$. Define $\mathcal{I}=\{I\mid j_1+\cdots+j_d=J_\epsilon\}$. We denote by $\langle\cdot,\cdot\rangle$ the inner product in L_2. Using Lemma A.2(i) we obtain

$$\begin{aligned}
&\Big\|\sum_{I\in\mathcal{I}}\theta_I\mu_I^{(r_1,\ldots,r_d)}\Big\|_{L_2}^2\\
&=\sum_{I,I'\in\mathcal{I}}\theta_I\theta_{I'}\langle\mu_I^{(r_1,\ldots,r_d)},\mu_{I'}^{(r_1,\ldots,r_d)}\rangle\\
&\le\sqrt{\sum_{I,I'\in\mathcal{I}}\theta_I^2\left|\langle\mu_I^{(r_1,\ldots,r_d)},\mu_{I'}^{(r_1,\ldots,r_d)}\rangle\right|}\sqrt{\sum_{I,I'\in\mathcal{I}}\theta_{I'}^2\left|\langle\mu_I^{(r_1,\ldots,r_d)},\mu_{I'}^{(r_1,\ldots,r_d)}\rangle\right|}\\
&\le\sup_{I'\in\mathcal{I}}\left\{\sum_{I\in\mathcal{I}}\left|\langle\mu_I^{(r_1,\ldots,r_d)},\mu_{I'}^{(r_1,\ldots,r_d)}\rangle\right|\right\}\times\sum_{I'\in\mathcal{I}}\theta_{I'}^2,\\
&=O\left(2^{2J_\epsilon r}\quad 2^{J_\epsilon}J_\epsilon^{d-1}\epsilon^2\right)=O(1).
\end{aligned}\tag{5.16}$$

Proof of Lemma 3.2. To prove this lemma, we use the idea of the proof of Theorem 4.2.3 in Young (1980, pp.156/157). We consider an arbitrary subset $A\subseteq\{1,\ldots,d\}$ and define

$$\mathcal{I}_A=\{I:\quad j_i\ge l\ \text{ for all }\ i\in A,\ \ j_i=l-1\ \text{ for all }\ i\notin A\}.$$

By (A2)(i) we have $\psi_{j_i,k_i}^{(-s)}(0)=\psi_{j_i,k_i}^{(-s)}(1)=0$ for $j_i\ge l$ and $s=0,\ldots,r$. Hence we obtain by integration by parts, for $I\in\mathcal{I}_A$, that

$$\theta_I\ =\ (-1)^{r\#A}\langle f^{(r,A)},\psi_I^{(-r,A)}\rangle,$$

where $g^{(s,A)}=g^{(s_1,\ldots,s_d)}$ with $s_i=s$ if $i\in A$ and $s_i=0$ if $i\notin A$.

Now we choose an arbitrary finite index set $\mathcal{I}_{A,fin} \subseteq \mathcal{I}_A$. Then

$$\begin{aligned}\Big| \sum_{I\in\mathcal{I}_{A,fin}} 2^{2(j_1+\cdots+j_d)r}\theta_I^2\Big|^2 &= \Big| \sum_{I\in\mathcal{I}_{A,fin}} 2^{2(j_1+\cdots+j_d)r}\theta_I \langle f^{(r,A)}, \psi_I^{(-r,A)}\rangle\Big|^2 \\ &= \Big|\langle f^{(r,A)}, \sum_{I\in\mathcal{I}_{A,fin}} 2^{2(j_1+\cdots+j_d)r}\theta_I \psi_I^{(-r,A)}\rangle\Big|^2 \\ &= \Big\|f^{(r,A)}\Big\|_{L_2}^2 \Big\| \sum_{I\in\mathcal{I}_{A,fin}} 2^{2(j_1+\cdots+j_d)r}\theta_I \psi_I^{(-r,A)}\Big\|_{L_2}^2.\end{aligned}$$

Furthermore, we obtain in complete analogy to (5.16) that

$$\begin{aligned}&\Big\| \sum_{I\in\mathcal{I}_{A,fin}} 2^{2(j_1+\cdots+j_d)r}\theta_I \psi_I^{(-r,A)}\Big\|_{L_2}^2 \\ &\le \sup_{I'\in\mathcal{I}_{A,fin}} \Big\{ \sum_{I\in\mathcal{I}_{A,fin}} 2^{[(j_1+\cdots+j_d)+(j_1'+\cdots+j_d')]r} \Big|\langle \psi_I^{(-r,I)}, \psi_{I'}^{(-r,I')}\rangle\Big|\Big\} \\ &\quad \times \sum_{I'\in\mathcal{I}_{A,fin}} 2^{2(j_1'+\cdots+j_d')r}\theta_{I'}^2.\end{aligned}$$

This implies by Lemma A.2(ii) that

$$\begin{aligned}&\sum_{I\in\mathcal{I}_{A,fin}} 2^{2(j_1+\cdots+j_d)r}\theta_I^2 \\ &\le \Big\|f^{(r,A)}\Big\|_{L_2}^2 \sup_{I'\in\mathcal{I}_{A,fin}} \Big\{ \sum_{I\in\mathcal{I}_{A,fin}} 2^{[(j_1+\cdots+j_d)+(j_1'+\cdots+j_d')]} \Big|\langle \psi_I^{(-r,I)}, \psi_{I'}^{(-r,I')}\rangle\Big| \Big\} \\ &\le C(A)\Big\|f^{(r,A)}\Big\|_{L_2}^2.\end{aligned}$$

Since this relation is true for any arbitrary finite subset $\mathcal{I}_{A,fin}$ we obtain

$$\sum_I 2^{2(j_1+\cdots+j_d)r}\theta_I^2 = \sum_{A:A\subseteq\{1,\ldots,d\}} \sum_{I\in\mathcal{I}_A} 2^{2(j_1+\cdots+j_d)r}\theta_I^2 \le \sum_{A:A\subseteq\{1,\ldots,d\}} C(A)\Big\|f^{(r,A)}\Big\|_{L_2}^2,$$

which completes the proof.

Proof of Theorem 3.3. We conclude from (2.13) and Lemma 3.2 that

$$\begin{aligned}&E\|\widehat{f}_\epsilon^{univ} - f\|_{L_2}^2 \\ &\le \epsilon^2 \#\mathcal{I}_\epsilon \Big(\frac{\lambda_\epsilon^{univ}}{\epsilon}+1\Big)\varphi\Big(\frac{\lambda_\epsilon^{univ}}{\epsilon}\Big) + \sum_{I\in\mathcal{I}_\epsilon} \min\Big\{(\lambda_\epsilon^{univ})^2, \theta_I^2\Big\} + \sum_{I\notin\mathcal{I}_\epsilon}\theta_I^2 \\ &= O\left(\epsilon^2\sqrt{\log(\epsilon^{-1})}\right) + O\left(\sum_J \min\Big\{(\lambda_\epsilon^{univ})^2 2^J J^{d-1}, 2^{-2Jr}\Big\}\right) + O\left(2^{-2J_\epsilon^* r}\right).\end{aligned}$$

To estimate the second term on the right-hand side, we choose J' such that the balance relation

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

is satisfied. This implies $J' \asymp \log(\epsilon^{-1})$ and, therefore, $\epsilon^2[\log(\epsilon^{-1})]^d \asymp 2^{-(2r+1)J'}$. Hence, we get

$$\sum_J \min\left\{(\lambda_\epsilon^{univ})^2 2^J J^{d-1}, 2^{-2Jr}\right\} = O\left(2^{-2J'r}\right) = O\left((\epsilon^2[\log(\epsilon^{-1}]^d)^{2r/(2r+1)}\right),$$

which completes the proof.

Proof of Theorem 3.4. By (2.13), the proof of the theorem is reduced to estimating $\Omega_\epsilon((\lambda_I^*), \Theta)$, where, according to (3.8), $\Theta = \{(\theta_I) | \sup_J \{2^{J(r-1/2)} J^{-(d-1)/2} \sum_{j_1+\cdots+j_d=J} \sum_{k_1,\ldots,k_d} |\theta_I|\} \le K\}$. We have

$$\begin{aligned}
\Omega_\epsilon\left((\lambda_I^*), \Theta\right) \le & \sum_{j_1+\cdots+j_d \le J_\epsilon} \sum_{k_1,\ldots,k_d} \epsilon^2 \\
& + \sum_{J=J_\epsilon+1}^{\infty} \sum_{j_1+\cdots+j_d=J} \sum_{k_1,\ldots,k_d} \epsilon^2\Big(\frac{\lambda_I^*}{\epsilon} + 1\Big)\varphi\Big(\frac{\lambda_I^*}{\epsilon}\Big) \\
& + \sup_{(\theta_I)\in\Theta} \Big\{ \sum_{J=J_\epsilon+1}^{\infty} \sum_{j_1+\cdots+j_d=J} \sum_{k_1,\ldots,k_d} \min\left\{(\lambda_I^*)^2, \theta_I^2\right\} \Big\} \\
= & \, T_1 + T_2 + T_3. \qquad (5.17)
\end{aligned}$$

From (3.5) and (3.6) we see that

$$T_1 = O\left(\epsilon^2 2^{J_\epsilon} \ J_\epsilon^{d-1}\right) = O\left((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}\right). \qquad (5.18)$$

Since

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

we get

$$\begin{aligned}
T_2 &= O\left(\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1}\right) \sum_{J>J_\epsilon} O\left(2^{J-J_\epsilon} (J/J_\epsilon)^{d-1} \left(\frac{\lambda_I^*}{\epsilon} + 1\right) \varphi\left(\frac{\lambda_I^*}{\epsilon}\right)\right) \\
&= O\left(\epsilon^2 2^{J_\epsilon} J_\epsilon^{d-1}\right) = O\left((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}\right). \qquad (5.19)
\end{aligned}$$

Finally, we have

$$\sum_{j_1+\cdots+j_d=J} \sum_{k_1,\ldots,k_d} \min\left\{(\lambda_I^*)^2, \theta_I^2\right\}$$

$$\begin{aligned}
&\le \lambda_I^* \sum_{j_1+\cdots+j_d=J} \sum_{k_1,\ldots,k_d} |\theta_I| \\
&= O\left(\epsilon\sqrt{J-J_\epsilon}\quad 2^{-J(r-1/2)} J^{(d-1)/2}\right) \\
&= O\left(\epsilon 2^{-J_\epsilon(r-1/2)} J_\epsilon^{(d-1)/2}\right) O\left(2^{-(J-J_\epsilon)(r-1/2)}\sqrt{J-J_\epsilon}(J/J_\epsilon)^{(d-1)/2}\right),
\end{aligned}$$

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

$$T_3 = O\left((\epsilon^2[\log(\epsilon^{-1})]^{d-1})^{2r/(2r+1)}\right). \tag{5.20}$$

Acknowledgement

I thank the referees and an Associate Editor for a very careful check of former versions of this paper, and for numerous helpful comments.

Appendix

Lemma A.1. *Let $\Theta = \{(\theta_I) | \sum \theta_I \psi_I \in B^r_{\underline{p},q}(K)\}$. Then there exists some K' such that, for all $j_1, \ldots, j_d$ and all i,*

$$\sup_{(\theta_I)\in\Theta} \Big\{ \sum_{k_1,\ldots,k_d} |\theta_I|^{p_i} \Big\} \le K' 2^{-j_i r_i p_i} 2^{(j_1+\cdots+j_d)(1-p_i/2)}.$$

Proof. Let $P_{(j_1,\ldots,j_d)}$ be the projector onto the subspace spanned by $\{\psi_{(j_1,\ldots,j_d;k_1,\ldots,k_d)}\}_{k_1,\ldots,k_d}$, that is,

$$\Big(P_{(j_1,\ldots,j_d)} g\Big)(\underline{x}) = \int K(\underline{x},\underline{y}) g(\underline{y}) d\underline{y},$$

where $K(\underline{x},\underline{y}) = \sum_{k_1,\ldots,k_d} \psi_{(j_1,\ldots,j_d;k_1,\ldots,k_d)}(\underline{x}) \psi_{(j_1,\ldots,j_d;k_1,\ldots,k_d)}(\underline{y})$.

Let $f = \sum_I \theta_I \psi_I$. Since $\theta_{(j_1,\ldots,j_d;k_1,\ldots,k_d)} = \int P_{(j_1,\ldots,j_d)} f(\underline{x}) \psi_{(j_1,\ldots,j_d;k_1,\ldots,k_d)}(\underline{x}) d(\underline{x})$ we obtain by Hölder's inequality

$$|\theta_I| \le \left(\int |(P_{(j_1,\ldots,j_d)} f)(\underline{x})|^{p_i} |\psi_I(\underline{x})|\, d\underline{x}\right)^{1/p_i} \left(\int |\psi_I(\underline{x})|\, d\underline{x}\right)^{1-1/p_i},$$

which implies

$$\begin{aligned}
&\sum_{k_1,\ldots,k_d} |\theta_I|^{p_i} \\
&\le \sup_{\underline{x}} \Big\{ \sum_{k_1,\ldots,k_d} |\psi_I(\underline{x})| \Big\} \int |(P_{(j_1,\ldots,j_d)} f)(\underline{x})|^{p_i} d\underline{x} \sup_{k_1,\ldots,k_d} \Big\{ \Big(\int |\psi_I(\underline{x})| d\underline{x}\Big)^{p_i-1} \Big\}. \qquad \text{(A.1)}
\end{aligned}$$

We readily obtain

$$\sup_{\underline{x}} \Big\{ \sum_{k_1,\ldots,k_d} |\psi_I(\underline{x})| \Big\} = O\left(2^{(j_1+\cdots+j_d)/2}\right) \tag{A.2}$$

and

$$\sup_{k_1,\ldots,k_d} \Big\{ \Big(\int |\psi_I(\underline{x})|\, d\underline{x} \Big)^{p_i - 1} \Big\} = O\left(2^{((j_1+\cdots+j_d)(1/2-p_i/2)} \right). \tag{A.3}$$

It remains to derive an upper estimate for the second term on the right-hand side of (A.1). To this end, we have to use a Littlewood-Paley decomposition; see, for example, Härdle, Kerkyacharian, Picard and Tsybakov (1998, Section 9.3) for a convenient description in the univariate case.

In our multivariate context, we introduce such a decomposition in the direction of x_i. Since the following definitions involve also values of f for $\underline{x}$ outside the unit cube, we have to extend f appropriately. As described in Besov, Il'in and Nikol'skii (1979b, Theorem 18.5), we can extend f on $(0,1)^{i-1} \times \mathbb{R} \times (0,1)^{d-i}$ in such a way that, for $q < \infty$,

$$\Big(\int |h|^{(s_i - r_i)q - 1} \Big\| \Delta_{i,h}^2 \Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f \Big) \Big\|^q_{L_{p_i}((0,1)^{i-1} \times \mathbb{R} \times (0,1)^{d-i})} \, dh \Big)^{1/q} \leq C, \tag{A.4}$$

and, for $q = \infty$,

$$\sup_h \Big\{ |h|^{s_i - r_i} \Big\| \Delta_{i,h}^2 \Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f \Big) \Big\|_{L_{p_i}((0,1)^{i-1} \times \mathbb{R} \times (0,1)^{d-i})} \Big\} \leq C. \tag{A.5}$$

Let G be a symmetric kernel function whose Fourier transform $\widehat{G}$ satisfies, for some $A > 0$,

$$\operatorname{supp}(\widehat{G}) \subseteq [-A, A],$$

$$\widehat{G}(\xi) = 1 \quad \text{for } \xi \in [-3A/4, 3A/4].$$

We decompose f as

$$f = G_{-1}^{[i]} f + \sum_{l=0}^{\infty} G_l^{[i]} f, \tag{A.6}$$

where

$$\left(G_l^{[i]} f \right)(\underline{x}) = \int G_l(y) f(x_1, \ldots, x_{i-1}, x_i - y, x_{i+1}, \ldots, x_d)\, dy,$$

$G_{-1}(y) = G(y)$, and, for $l \geq 0$, $G_l(y) = 2^{l+1} G(2^{l+1} y) - 2^l G(2^l y)$.

We have

$$\left\| P_{(j_1,\ldots,j_d)} f \right\|_{p_i} \leq \sum_{l=-1}^{\infty} \left\| P_{(j_1,\ldots,j_d)} (G_l^{[i]} f) \right\|_{p_i}. \tag{A.7}$$

Next we show that

$$\left\| G_l^{[i]} f \right\|_{p_i} = O(2^{-l r_i}). \tag{A.8}$$

This relation is obviously fulfilled for any fixed l, in particular for $l = -1$. Therefore it suffices to prove (A.8) for $l \geq 0$. Since $\int G_l(y)dy = 0$ for $l \geq 0$ and $G_l(y) = G_l(-y)$, we get

$$\left(G_l^{[i]} f\right)(\underline{x}) = \frac{1}{2}\int G_l(y)\Delta_{i,y}^2 f(x_1, \ldots, x_{i-1}, x_i - y, x_{i+1}, \ldots, x_d)\, dy.$$

This yields, by s_i-fold integration by parts, that

$$G_l^{[i]} f(\underline{x}) = (-1)^{s_i}/2 \int G_l^{(-s_i)}(y)\Delta_{i,y}^2 \left(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f\right)(x_1, \ldots, x_{i-1}, x_i - y, x_{i+1}, \ldots, x_d)\, dy.$$

Hence, we obtain by the Generalized Minkowski inequality (see, e.g., Besov, Il'in and Nikol'skii (1979a, p.24) that

$$\left\|G_l^{[i]} f\right\|_{p_i} \leq \frac{1}{2}\int |G_l^{(-s_i)}(y)| \left\|\Delta_{i,y}^2 \Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f\Big)\right\|_{p_i} dy.$$

In the case $q < \infty$ we get (here $\widetilde{q}$ is chosen such that $1/q + 1/\widetilde{q} = 1$)

$$\begin{aligned}
\|G_l^{[i]} f\|_{p_i} &\leq \frac{1}{2}\Big(\int |y|^{(s_i - r_i)q - 1} \left\|\Delta_{i,y}^2 \Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f\Big)\right\|_{p_i}^q dy\Big)^{1/q} \\
&\qquad \times \Big(\int |G_l^{(-s_i)}(y)|^{\widetilde{q}} |y|^{(r_i - s_i)\widetilde{q} + \widetilde{q}/q}\, dy\Big)^{1/\widetilde{q}} \\
&= O(1) \times \Big(\int |2^{-l s_i} 2^l G_0^{(-s_i)}(2^l y)|^{\widetilde{q}} |y|^{(r_i - s_i)\widetilde{q} + \widetilde{q}/q}\, dy\Big)^{1/\widetilde{q}} \\
&= O(2^{-l r_i}),
\end{aligned}$$

while in the case of $q = \infty$ we get

$$\|G_l^{[i]} f\|_{p_i} \leq \frac{1}{2}\int |G_l^{(-s_i)}(y)||y|^{r_i - s_i} dy \times \sup_y \left\{|y|^{s_i - r_i} \left\|\Delta_{i,y}^2 \Big(\frac{\partial^{s_i}}{\partial x_i^{s_i}} f\Big)\right\|_{p_i}\right\} = O(2^{-l r_i}).$$

Hence, (A.8) is proved.

With the definition $F(\underline{x}) = \sup_{\underline{y}}\{|K(\underline{y}, \underline{y} + \underline{x})|\}$, we get the inequality $|K(\underline{x}, \underline{y})| \leq F(\underline{x} - \underline{y})$, where $\|F\|_1 = O(1)$. Applying Young's inequality (see, e.g., Besov, Il'in and Nikol'skii (1979a, p.26)) we obtain

$$\left\|P_{(j_1, \ldots, j_d)}(G_l^{[i]} f)\right\|_{p_i} \leq \left\|F * |G_l^{[i]} f(\cdot)|\right\|_{p_i} \leq \|F\|_1 \|G_l^{[i]} f\|_{p_i} = O(2^{-l r_i}). \qquad \text{(A.9)}$$

Now

$$\begin{aligned}
\frac{\partial^{t_i}}{\partial x_i^{t_i}} G_l^{[i]} f(\underline{x}) &= \int \frac{\partial^{t_i}}{\partial x_i^{t_i}} G_l(x_i - y) f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d)\, dy \\
&= 2^{l t_i} \int 2^l G_0^{(t_i)}(2^l (x_i - y)) f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d)\, dy,
\end{aligned}$$

which yields, analogously to (A.8) (with $G_0^{(t_i)}$ instead of G_0),

$$\left\| \frac{\partial^{t_i}}{\partial x_i^{t_i}} G_l^{[i]} f \right\|_{p_i} = O\left(2^{lt_i} 2^{-lr_i}\right). \tag{A.10}$$

Using the Taylor series expansion

$$\begin{aligned}
&(G_l^{[i]} f)(\underline{y}) \\
&= \sum_{k=0}^{t_i-1} \frac{1}{k!} \frac{\partial^k}{\partial x_i^k} (G_l^{[i]} f)(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_d)(y_i - x_i)^k \\
&\quad + \int_0^1 (y_i - x_i)^{t_i} \frac{(1-u)^{t_i-1}}{(t_i-1)!} \frac{\partial^{t_i}}{\partial x_i^{t_i}} (G_l^{[i]} f)(y_1, \ldots, y_{i-1}, x_i + u(y_i - x_i), y_{i+1}, \ldots, y_d) du,
\end{aligned}$$

we obtain

$$\begin{aligned}
&P_{(j_1,\ldots,j_d)} (G_l^{[i]} f)(\underline{x}) \\
&= \sum_{k_1,\ldots,k_d} \psi_I(\underline{x}) \int_{\mathbb{R}^d} \psi_I(\underline{y}) \times \Big(\int_0^1 (y_i - x_i)^{t_i} \frac{(1-u)^{t_i-1}}{(t_i-1)!} \\
&\qquad\qquad \frac{\partial^{t_i}}{\partial x_i^{t_i}} (G_l^{[i]} f)(y_1, \ldots, y_{i-1}, x_i + u(y_i - x_i), y_{i+1}, \ldots, y_d)\, du \Big)\, d\underline{y} \\
&= \int_0^1 \int_{\mathbb{R}^d} K(\underline{x}, \underline{y}) (y_i - x_i)^{t_i} \frac{(1-u)^{t_i-1}}{(t_i-1)!} \\
&\qquad\qquad \times \frac{\partial^{t_i}}{\partial x_i^{t_i}} (G_l^{[i]} f)(y_1, \ldots, y_{i-1}, x_i + u(y_i - x_i), y_{i+1}, \ldots, y_d)\, d\underline{y}\, du.
\end{aligned}$$

Hence,

$$\begin{aligned}
\left| P_{(j_1,\ldots,j_d)} (G_l^{[i]} f)(\underline{x}) \right| &\le \int_0^1 du \int_{\mathbb{R}^d} F(\underline{x} - \underline{y}) |y_i - x_i|^{t_i} \\
&\qquad \left| \frac{\partial^{t_i}}{\partial x_i^{t_i}} (G_l^{[i]} f)(y_1, \ldots, y_{i-1}, x_i + u(y_i - x_i), y_{i+1}, \ldots, y_d) \right| d\underline{y},
\end{aligned}$$

which implies

$$\left\| P_{(j_1,\ldots,j_d)} (G_l^{[i]} f) \right\|_{p_i} = O\left(2^{-j_i t_i} 2^{l(t_i - r_i)}\right). \tag{A.11}$$

From (A.7), (A.8) and (A.11) we obtain

$$\left\| P_{(j_1,\ldots,j_d)} f \right\|_{p_i} = O\left(2^{-j_i r_i}\right). \tag{A.12}$$

This implies, in conjunction with (A.1) to (A.3), the assertion of the lemma.

Lemma A.2. (*Near-orthogonality of certain families of functions*)

(i) *Let ζ be an r-times boundedly differentiable function with finite support and $\int \zeta^{(s)}(x)\,dx = 0$ for all $0 \le s \le r$. Define $\zeta_I(\underline{x}) = 2^{(j_1+\cdots+j_d)/2}\zeta(2^{j_1}x_1 - k_1)\cdots\zeta(2^{j_d}x_d - k_d)$. Then, for $0 \le r_1, \ldots, r_d \le r$,*

$$\sup_{I'}\Big\{\sum_{I} 2^{-[(j_1+j_1')r_1+\cdots+(j_d+j_d')r_d]}\left|\langle\zeta_I^{(r_1,\ldots,r_d)}, \zeta_{I'}^{(r_1,\ldots,r_d)}\rangle\right|\Big\} \le C.$$

(ii) *Assume* (A2). *For $A \subseteq \{1,\ldots,d\}$, define $g^{(-r,A)} = g^{(-r_1,\ldots,-r_d)}$, where $r_i = r$ if $i \in A$ and $r_i = 0$ if $i \notin A$. Let $\mathcal{I}_A = \{I : \ j_i \ge l$ for all $i \in A$ and $j_i = l-1$ for all $i \notin A\}$. Then*

$$\sup_{I'\in\mathcal{I}_A}\Big\{\sum_{I\in\mathcal{I}_A} 2^{[(j_1+\cdots+j_d)+(j_1'+\cdots+j_d')]r}\left|\langle\psi_I^{(-r,A)}, \psi_{I'}^{(-r,A)}\rangle\right|\Big\} \le C.$$

Proof.

(i) For $j_i > j_i'$, we get from $\int \zeta_{j_i,k_i}^{(r_i)}(x)\,dx = 0$ that

$$\sum_{k_i}\left|\int \zeta_{j_i,k_i}^{(r_i)}(x)\zeta_{j_i',k_i'}^{(r_i)}(x)dx\right| = O\left(\left\|\zeta_{j_i,k_i}^{(r_i)}\right\|_1 TV\left(\zeta_{j_i',k_i'}^{(r_i)}\right)\right) = O\left(2^{(j_i+j_i')r_i}2^{(j_i'-j_i)/2}\right).$$

For $j_i \le j_i'$, we have

$$\sum_{k_i}\left|\int \zeta_{j_i,k_i}^{(r_i)}(x)\zeta_{j_i',k_i'}^{(r_i)}(x)\,dx\right| = O\left(\left\|\zeta_{j_i,k_i}^{(r_i)}\right\|_\infty \left\|\zeta_{j_i',k_i'}^{(r_i)}\right\|_1\right) = O\left(2^{(j_i+j_i')r_i}2^{(j_i-j_i')/2}\right).$$

Using the product structure of $\zeta_I^{(r_1,\ldots,r_d)}$ we obtain

$$\sum_{j_1,\ldots,j_d=-\infty}^{\infty} 2^{-[(j_1+j_1')r_1+\cdots+(j_d+j_d')r_d]}\sum_{k_1,\ldots,k_d}\left|\langle\zeta_I^{(r_1,\ldots,r_d)}, \zeta_{I'}^{(r_1,\ldots,r_d)}\rangle\right|$$
$$= \prod_{i=1}^{d} O\left(\sum_{j_i=-\infty}^{\infty} 2^{|j_i-j_i'|/2}\right) = O(1).$$

(ii) For $i \in A$ we proceed as above. For $j_i > j_i'$, we get from (A2) that $\int \psi_{j_i,k_i}^{(-r)}(x)\,dx = 0$, which implies

$$\sum_{k_i}\left|\int 2^{j_i r}\psi_{j_i,k_i}^{(-r)}(x)\; 2^{j_i' r}\psi_{j_i',k_i'}^{(-r)}(x)\,dx\right| = O\left(2^{(j_i'-j_i)/2}\right).$$

For $j_i \le j_i'$, we have

$$\sum_{k_i}\left|\int 2^{j_i r}\psi_{j_i,k_i}^{(-r)}(x)\,2^{j_i' r}\psi_{j_i',k_i'}^{(-r)}(x)\,dx\right| = O\left(2^{(j_i-j_i')/2}\right).$$

Finally, for $i \notin A$ and $j_i = j_i' = l-1$, we get immediately that

$$\sum_{k_i} \left| \int 2^{j_i r} \psi_{j_i,k_i}(x)\, 2^{j_i' r} \psi_{j_i',k_i'}(x)\, dx \right| = O(1).$$

Hence, we obtain

$$\sum_{I \in \mathcal{I}_A} 2^{[(j_1+\cdots+j_d)+(j_1'+\cdots+j_d')]r} \left| \langle \psi_I^{(-r,A)}, \psi_{I'}^{(-r,A)} \rangle \right|$$

$$= \prod_{i \in A} O\left(\sum_{j_i \geq l} 2^{|j_i - j_i'|/2} \right) \prod_{i \notin A} O(1) = O(1).$$

References

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39**, 930–945.

Besov, O. V., Il'in, V. P. and Nikol'skii, S. M. (1979a). *Integral Representations of Functions and Imbedding Theorems I.* Wiley, New York.

Besov, O. V., Il'in, V. P. and Nikol'skii, S. M. (1979b). *Integral Representations of Functions and Imbedding Theorems II.* Wiley, New York.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque mimimax. *Z. Wahr. verw. Gebiete* **47**, 119-137.

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transform. *Appl. Comput. Harmonic Anal.* **1**, 54-81.

Dahlhaus, R., Neumann, M. H. and von Sachs, R. (1999). Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli* **5**, 873-906.

Daubechies, I. (1988). Orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41**, 909-996.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

Delyon, B. and Juditsky, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmonic Anal.* **3**, 215-228.

Donoho, D. L. (1997). CART and Best-Ortho-Basis: a connection. *Ann. Statist.* **25**, 1870-1911.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion) *J. Roy. Statist. Soc. Ser. B* **57**, 301-369.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. B. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statist.* **129**. Springer, New York.

Kamont, A. (1994). Isomorphism of some anisotropic Besov and sequence spaces. *Studia Mathematica* **110**, 169-189.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93-101.

Meyer, Y. (1991). Ondelettes sur l'intervalle. *Revista Mathematica Ibero-Americana* **7**, 115-133.

Neumann, M. H. (1995). Discussion to the paper "Wavelet shrinkage: asymptopia?" by Donoho *et al. J. Roy. Statist. Soc. Ser. B* **57**, 346-347.

Neumann, M. H. (1996). Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *J. Time Ser. Anal.* **17**, 601-633.

Neumann, M. H. and von Sachs, R. (1995). Wavelet thresholding: beyond the Gaussian i.i.d. situation. In *Lecture Notes in Statistics: Wavelets and Statistics* (Edited by A. Antoniadis and G. Oppenheim), 301-329.

Neumann, M. H. and von Sachs, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.* **25**, 38-76.

Neumann, M. H. and Spokoiny, V. G. (1995). On the efficiency of wavelet estimators under arbitrary error distributions. *Math. Methods Statist.* **4**, 137-166.

Nussbaum, M. (1982). Optimal L_p-convergence rates for estimates of a multiple regression function. Preprint P-Math 07/82, Institut für Mathematik, Akademie der Wissenschaften der DDR.

von Sachs, R. and Schneider, K. (1996). Wavelet smoothing of evolutionary spectra by nonlinear thresholding. *Appl. Comput. Harmonic Anal.* **3**, 268-282.

Schmeißer, H.-J. and Triebel, H. (1987). *Topics in Fourier Analysis and Function Spaces.* Geest & Portig, Leipzig.

Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualisation.* Wiley, New York.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689-705.

Tjøstheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89**, 1398-1409.

Tribouley, K. (1995). Practical estimation of multivariate densities using wavelet methods. *Statistica Neerlandica* **49**, 41-62.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Walsh, J. B. (1986). Martingales with a multidimensional parameter and stochastic integrals in the plane. In *Lectures in Probability and Statistics* (Edited by A. Dold and B. Eckmann). *Lecture Notes in Math.* **1215**, 329-491. Springer, Berlin.

Young, R. M. (1980). *An Introduction to Nonharmonic Fourier Series.* Academic Press, New York.

Sonderforschungsbereich 373, Humboldt-Universität zu Berlin, Spandauer Straße 1, D - 10178 Berlin, Germany.

E-mail: neumann@wiwi.hu-berlin.de

(Received November 1997; accepted June 1999)