****

PENALIZED WAVELET ESTIMATION WITH BESOV REGULARITY CONSTRAINTS

Lubomir T. Dechevsky

Département de mathématiques et de statistique

Université de Montréal

dechevsk@dms.umontreal.ca

James O. Ramsay

Department of psychology

McGill University

ramsay@psych.mcgill.ca

Research Report/Rapport de recherche CRM-2602, April/avril 1999

Centre de recherches mathématiques, Université de Montréal

http://www.CRM.UMontreal.CA/

****

Abstract

A wavelet counterpart to Wahba's spline smoothing technique is developed for the case of fitting less regular curves. Our approach is a realization of the smoothness penalty method of Tikhonov regularization of ill-posed inverse stochastic problems within a wavelet setting. The penalized cost functional to be minimized is Peetre's K-functional between Besov spaces. The regularity of the curve is discussed in terms of the size of its seminorm in homogeneous Besov spaces $\dot{B}^s_{pq}$ with a relatively small value of the smoothness index $s > 0$. Penalized L_2-estimation with Besov-type constraints, considered in the literature, is included as a partial case. The optimal solution of the penalization problem is in the form of a wavelet expansion whose coefficients are obtained by appropriate level- and/or space-dependent shrinking of the empirical wavelet coefficients. Thanks to the use of wavelets, both density and regression estimation can be treated in a somehow unified way. In the case of regression-function estimation, an enhanced version of cross validation (generalized full cross validation) is implemented for the choice of the smoothing parameter. Numerical examples illustrate the advantage of our procedures in comparison to more standard wavelet methods when the regularity is small and sample sizes are moderate. The approach is very flexible and allows for a diversity of extensions. Some first extensions can be found in Section 7, among which are the iterative individual shrinking estimator and the self-similar fractal estimator. The other extensions considered are grouped in Appendix B, which is in a sense the most advanced part of the paper, and can be considered as an outline of a plan for further research. For conciseness of presentation, only the univariate case is considered in detail, but all statements and their proofs do admit multidimensional generalization.

Mathematics Subject Classification: Primary 42C15, 46E35, 62G05, 65D10; Secondary 26A30, 26D10, 35K10, 35K30, 39B12, 41A63, 42C20, 46N30, 47B10, 47D06, 49M45, 62J07, 65K10, 65R30, 65U05

Key Words and Phrases: penalized nonparametric regression and density estimation, consistency of estimation and cross validation, asymptotic-minimax rate of estimation, atomic decomposition of Besov spaces by orthonormal wavelets, K-functional Tikhonov regularization of ill-posed inverse stochastic problems, iterative individual shrinking of empirical wavelet coefficients, self-similar fractal estimator.

Résumé

Un analogue de la technique de Wahba de lissage avec des fonctions splines est développé ici dans le cas des ondelettes, pour traiter des courbes moins régulières. Notre méthode est une réalisation dans le cadre d'ondelettes de la régularisation de Tikhonov de problèmes stochastiques inverses mal posés. La fonction non-négative pénalisée à optimiser est la K-fonction de Peetre entre des espaces de Besov. La régularité de la courbe est mesurée dans les espaces Bésoviens homogènes $\dot{B}^s_{pq}$, pour une valeur de l'index de régularité $s > 0$ relativement petite. L'estimation pénalisée dans L_2 avec des contraintes de type de Besov, considérée dans la littérature, est un cas partiel. La solution optimale du problème avec pénalisation est une expansion en ondelettes orthogonales dont les coefficients sont obtenus par contraction des coefficients empiriques. Grâce à l'utilisation d'ondelettes, on peut développer une méthode quasi-unifiée pour traiter l'estimation de densité et le cas de régression. Dans le contexte de régression, une version perfectionnée de la validation croisée (la validation croisée complète généralisée) du paramètre de régularisation est élaborée. Des exemples numériques illustrent l'avantage de nos procédures en comparaison aux méthodes standardisées contractantes d'ondelettes puisque la régularité est minime et le volume de données est moyen. Notre méthode est très flexible et a des extensions et des généralisations diverses. Les premières d'entre elles sont décrites dans Section 7; parmi elles, l'estimateur de contraction individuelle itérative et l'estimateur fractal auto-similaire. Les autres extensions considérées sont groupées dans Supplément B. Ce supplément est, en un certain sens, la partie du travail la plus avancée et peut être vu comme l'esquisse d'un plan de recherche future. Afin d'abréger l'exposition, seulement le cas unidimensionnel est considéré en détail, mais toutes assertions et leurs démonstrations peuvent être généralisées dans le cas multidimensionnel.

1. INTRODUCTION

This paper consists of three parts: main body, Appendix A and Appendix B. References to parts of the appendices are denoted in the text by ^{A2} or ^{A7,B3,B8}, and so on.

Our purpose is to suggest a wavelet-based approach that parallels spline smoothing techniques, as described by Wahba (1990), and to apply it for estimating spatially inhomogeneous curves. Our method is thus an extension of the smoothness penalty approach to wavelets. The penalized cost functional to be minimized is Peetre's K-functional between Besov spaces. Its minimizer is in the form of a wavelet expansion with coefficients obtained by level- and/or space-dependent shrinking of the empirical wavelet coefficients. In the case of nonparametric regression, an enhanced version of cross validation (generalized full cross validation) is implemented. Measuring the estimation rates and the regularity of the curve in Besov spaces allows for optimizing the cross-validation criterion not only with respect to the penalizing smoothing parameter, but also with respect to the smoothness index of the Besov space. This brings about control over fractional-order derivatives^{A1} of the curve and makes the new estimator superior to Wahba's smoothing natural spline estimator when estimating less regular curves. To quote Hall and Patil (1996): "we do not suggest that wavelet methods should supplant existing techniques for estimating very smooth curves, but we do agree that they have ready application to problems where curves change relatively sharply. Thus, they complement existing methodology for non-parametric curve estimation, rather than replace it". In the specific context of the smoothness penalty approach there are two additional advantages of the utilization of wavelets. (a) While the optimization problem in Wahba's approach is intrinsically *non-parametric* (consisting of a large number of boundary-value problems for differential equations plus subsequent comparison of the boundary values), its analogue in the wavelet setting is *parametric*^{A2}. (b) While the penalized smoothing spline technique is designed specifically for regression-function estimation, its wavelet analogue allows treating *both* non-parametric regression and density estimation in a somehow *unified* way. More precisely, we shall be considering the following regression and density models.

The regression model is: $Y_i = f(x_i) + \epsilon_i,\ i = 1, 2, \dots, n$, where $x_i = a_1 + (a_2 - a_1)i/n$, $[a_1, a_2]$ being the definition interval. For the i.i.d. errors ϵ_i we assume $E\epsilon_1 = 0, E\epsilon_1^2 = \delta^2$.

In the case of density estimation, the sample consists of n i.i.d. observations $X_i,\ i = 1, 2, \dots, n$ with unknown density $f(x)$.

It is outside the scope of the paper to compare the merits of our solutions of these problems with the many alternative wavelet-based approaches (we can easily list at least ten alternatives) since this would require performing a very long simulation study and spacious discussion. This is why only one, well-known, method - the soft thresholding - was singled out for the sake of graphical comparison. The numerical simulations and graphical results presented in the sequel illustrate the advantages of the shrinking over the thresholding procedure when regularity is low and sample sizes are moderate. We should underscore though that our primary goal in the present paper is not demonstration of the superiority of the fits we get in comparison to soft thresholding and other standard wavelet methods. The main merits of our approach at this stage can rather be summarized as follows. (a) This approach offers an excellent trade-off between good estimation and *simplicity* of the estimator. (b) While satisfactory practical estimation with wavelet estimators (in particular, soft- and hard-thresholded) which are asymptotic-minimax within the "asymptopia" paradigm (Donoho, Johnstone, Kerkyacharian and Picard (1995)) can in general be only achieved for large sample sizes ($n \approx 10^4$), the penalization approach considered here yields comparable quality of the fit already for moderate samples ($n \approx 10^3$), under virtually no additional assumptions compared to soft thresholding provided that the Besov smoothness index is being *also optimized* in the cross-validation criterion. (c) The present approach is very flexible and can be extended simply and elegantly in diverse directions, thus allowing enriching and upgrading the features of the model, the expected ultimate reward being to get good fits for sample sizes as low as $n \approx 10^2$ and, for some specific estimation problems, even less. Some of these extensions are briefly addressed in appropriate places (notably, Section 7, [A3], Appendix B).

In the sequel only the univariate case is considered in detail for conciseness of presentation. However, all the main results can be generalized to the *multivariate* case. We shall make more detailed comments about this generalization in appropriate places later in the text.[B21]

Finally, a few more words on the organization of the paper. Sections 2 and 3 are introductory. Section 4 gives a brief exposition of our concept of *K-functional Tikhonov regularization* and explains why the K-functional approach is of fundamental importance for deterministic and nondeterministic smoothing and, in particular, for nonparametric penalized wavelet estimation. In Sections 5 and 6 a model case (the case of Hilbert spaces) is presented, and a model procedure for estimating the regularization parameter - cross validation - is studied in detail. In Section 7 numerical examples are discussed, together with some immediate extensions of the K-functional model. Appendix A contains the proofs, as well as "second reading" remarks to the text in the main body. Appendix B is essentially a list of further extensions and generalizations which can be considered as a program for further research on the topic of K-functional penalized estimation. In a sense, Appendix B is the most important and advanced part of this paper.

2. THRESHOLDING

The study of statistical applications of wavelets received its initial boost in 1992-1994 with the development of an elegant general asymptotic-minimax theory in a series of papers by Donoho, Johnstone, Kerkyacharian and Picard. In this theory thresholding of the empirical wavelet coefficients emerged as a key tool for smoothing and denoising. The loss functional with respect to which the estimation rates were measured was an L_p-norm. That essentially meant that only the function was being consistently estimated but not its derivatives. One reason for this was that in many cases the initially proposed "universal" threshold turned out to be locally too high or too low and did not allow for an optimal trade-off between bias and variance locally along the curve. Since then, Donoho and Johnstone have proposed several more refined threshold strategies for Gaussian white noise (notably the SURE shrink, WaveJS and oracle inequalities) that allow for better adaptivity (and have had important impact on the development of the currently most advanced adaptive thresholding strategies NeighBlock and NeighCoeff (see Cai and Silverman (1998b,c))). Further methodological advance in the asymptotic theory was achieved, in both density and regression estimation context, with the introduction of level-dependent thresholds by Donoho and Johnstone (1995) for the case of nonparametric regression and by Delyon and Juditsky (1996) for both the regression-function and density estimation problem. This refinement contributed to bringing down the additional log-factor appearing in the asymptotic rates when estimating sufficiently regular curves using wavelets. The loss functional in which these rates were estimated by Delyon and Juditsky was a Besov (quasi-)norm with possibly positive smoothness index. This meant that also derivatives of the function up to a certain order were being consistently estimated. Moreover, the estimation rates for the derivatives were also asymptotically optimal within the "asymptopia" paradigm of Donoho, Johnstone, Kerkyacharian and Picard (1995). Despite these improvements, for practical purposes wavelet-based estimation still remained an essentially *large-sample* technique. This was part of the price to pay for making these estimators almost minimax in a broad range of function spaces simultaneously, as suggested in the "asymptopia"[A4].

Recently, several thresholding approaches have been proposed to generate wavelet estimators with improved performance on moderate samples. Variants of Bayesian approach have been suggested by several authors (Vidakovic (1994), Abramovich, Sapatinas and Silverman (1997), Chipman, Kolaczyk and McCulloch (1997)). Another approach was Nason's (Nason (1996)). He considered the integrated square error of the thresholded estimator as a function of the threshold and tried to find its minimum. One problem there appeared to be that the cross-validation criterion depended non-smoothly on the threshold value and had numerous local extrema. This made it very difficult, if possible at all, to develop advanced rigorous theory for this estimator. A further refinement of the level-dependent approach of Delyon and Juditsky was suggested in the so-called block-thresholding procedure by Hall, Kerkyacharian and Picard (1995) and Hall, Penev, Kerkyacharian and Picard (1997) (followed by MacGibbon and von Sachs (1998), Cai (1998), Cai and Silverman (1998b,c)), which helped to bring down the log-factor in the rates and to further improve the estimation of derivatives by combining wavelet coefficients into groups and using information about each one to assist in assessing its neighbours.

Yet, even after all these efforts, thresholded wavelet estimators still have their problems, especially when the sample sizes are small to moderate, the noise-to-signal ratio is relatively large, the white noise is not Gaussian, and the estimated function is not very smooth. The only realistic opportunity to systematically get good fits for small samples seems to be *to introduce in the model all available additional information about essential features of the curve.* Together with Bayesian approach, the *variational* approach seems to be the natural choice in such upgrading. However, optimization with respect to threshold values generally leads to non-linear, non-smooth, non-convex, multi-extremal problems with high computational complexity. An additional problem when estimating functions with low regularity and fractals is that thresholding methods tend to oversmooth the curve because they are well adapted for functions which gather their value on relatively few, relatively large wavelet coefficients only. Continuous fractals and functions with jumps do not generally fall into this class: they gather their value relatively uniformly from many wavelet coefficients on infinitely many levels and the vector of their wavelet coefficients does not necessarily look sparse. Due to these facts, for such types of functions a non-threshold shrinking procedure is appropriate. An alternative approach to thresholding, leading to smooth optimization criteria for determining the regularization parameter, was proposed by Amato and Vuza (1994) and, independently, by Antoniadis (1996). Their methods were implicitly based on a technique for explicit computation of K-functionals between Hilbert spaces, developed earlier by Dechevski (1989) and Dechevsky (1993, 1993a). Although of non-threshold type, the parameter settings in the methods of Amato and Vuza, and Antoniadis, made these methods appropriate for estimating very smooth and relatively spatially homogeneous functions only. On the contrary, in the context of spatially inhomogeneous functions with low regularity our penalized non-threshold shrinking wavelet estimator and its various extensions are at their best, and we recommend them as a valuable complement to existing thresholding techniques.

3. SPLINES, WAVELETS, FUNCTION SPACES

By the end of the 1960s the penalized smoothing-spline (or spline-smoothing) technique had already evolved into a very general penalization approach for solving *deterministic* smoothing problems (Anselone and Laurent (1968)). Its *statistical* applications - to non-parametric regression - began approximately at that time. They were based on an intuitively appealing technique for selection of the penalizing (or smoothing) parameter: cross validation (notably *ordinary* and *generalized*) (see Wahba (1990)). For the smoothing version of Schoenberg's problem (Wahba (1990), formula (0.0.3)) there are examples where the performance of Wahba's natural-spline smoothing estimator is spectacular, even for sample sizes as small as 100 or so. However, getting such a good fit is only likely for sufficiently smooth functions, since the penalization term contains a homogeneous Sobolev semi-norm with integer smoothness index. Even if the curve is very smooth, it should not be too spatially inhomogeneous because the cross-validation criterion can only be optimized with respect to a single *global* smoothing parameter. It is for more spatially inhomogeneous and/or less regular curves that *wavelet* methods show their strength, in *both* regression and density estimation problems.

Let $f(x)$, $x \in \mathbf{R}^1$ be a locally integrable function. Its wavelet expansion is

$$f(x) = \sum_{k\in\mathbf{Z}} \alpha_{j_0k}\varphi_{j_0k}(x) + \sum_{j=j_0}^{\infty}\sum_{k\in\mathbf{Z}} \beta_{jk}\psi_{jk}(x) \tag{1}$$

where $j_0 \in \mathbf{Z}$, $\{\varphi_{j_0k}$ and $\psi_{jk}, j = j_0, j_0+1, j_0+2, \ldots;\ k \in \mathbf{Z}\}$ form an orthonormal basis of $L_2(\mathbf{R})$. Here $\varphi_{jk}(x) = 2^{j/2}\varphi(2^jx-k)$, $\psi_{jk}(x) = 2^{j/2}\psi(2^jx-k)$, j describes the resolution level. For better spatial and level localization, φ and ψ are assumed to be compactly supported. Convergence of the series in (1) is in the (quasi-)norm topology (Bergh and Löfström, Section 3.10) of *Besov* spaces. The Besov scale $B^s_{pq}(\mathbf{R}^d)$ (resp. $\dot{B}^s_{pq}(\mathbf{R}^d)$) of *inhomogeneous* (resp.

homogeneous) Besov spaces, $0 < p \leq \infty$, $0 < q \leq \infty$, $s \in \mathbf{R}$, is the typical class of translation-invariant function spaces in which spatial inhomogeneity can be described. (For equivalent definitions of Besov spaces and their properties we refer to Bergh and Löfström (1976), Triebel (1983, 1992, 1997).) In statistical context, Besov spaces have been introduced by Donoho for the purposes of asymptotic-minimax theory of wavelet estimators (see Donoho, Johnstone, Kerkyacharian and Picard (1995) for an overview). This approach is founded on a relatively recent discovery (Sickel (1990)) of a remarkable family of equivalent (quasi-)norms in Besov spaces in terms of *wavelet coefficients*. In brief, the idea is as follows. In addition to the already made assumptions about φ, ψ, suppose that for $r > 0$ the linear span of $\varphi(. - k)$, $k \in \mathbf{Z}$, contains all algebraic polynomials P with $deg\ P \leq [r]$ and $\varphi \in B^r_{\infty\infty}$. Then φ_{j_0k}, ψ_{jk}, $k \in \mathbf{Z}$, $j = j_0, j_0+1, \ldots$, form a Riesz basis simultaneously for all $B^s_{pq}(\mathbf{R})$, $0 < p \leq \infty$, $0 < q \leq \infty$, $\max(0, 1/p-1) < s < r$. Therefore, for the above range of parameters considered, if $f \in B^s_{pq}(\mathbf{R})$, then the series in (1) is always convergent in the (quasi-)norm topology of the space and

$$J_{spq}(\alpha,\beta) = \left(\| \alpha_{j_0.} \|_p^q + \sum_{j=j_0}^{\infty} (2^{j(s+(1/2)-(1/p))} \| \beta_{j.} \|_p)^q \right)^{1/q} \tag{2}$$

is an equivalent norm in $B^s_{pq}(\mathbf{R})$. Here the notation

$$\| \alpha_{j_0.} \|_p = \left(\sum_{k\in\mathbf{Z}} |\alpha_{j_0k}|^p\right)^{1/p}; \ \| \beta_{j.} \|_p = \left(\sum_{k\in\mathbf{Z}} |\beta_{jk}|^p\right)^{1/p} \tag{3}$$

has been used. Henceforward, we shall always make the aforementioned assumption that the analyzing wavelet has smoothness index $r > s$. For the purposes of Section 6 we also need to suppose that the wavelet basis is obtained via multiresolution analysis (see Daubechies (1992)) and in Subsection 6.2 an additional explicit assumption is required that the first $[r]$ moments of ψ vanish. In Section 6 it is also implicitly assumed that φ and ψ have bounded variation $\bigvee \varphi$, $\bigvee \psi$, but this is true even for the Haar wavelet.

All notations related to wavelets and Besov spaces are the same as in Delyon and Juditsky (1996).[A5]

For simplicity, we shall always be working with wavelets obtained via multiresolution analysis on the whole $\mathbf{R}$ (or $\mathbf{R}^d$), and will be assuming that the support of the function in consideration is compact on its domain Ω, because in this way we are working with spaces which are simultaneously closed subspaces of the function space over Ω, and of the function space over $\mathbf{R}^d$. Because of their closedness, these subspaces enjoy all the essential properties of the entire function spaces. This trick is done only for simplicity of presentation; it is quite straightforward to extend our constructions to periodic functions in one and several dimensions by considering periodic wavelets and (when the sample size is a power of two) periodic orthogonal discrete wavelet transform (ODWT) (see, e.g., Donoho and Johnstone (1995) and the references therein). It is also easy to further extend our considerations also for the finite interval (see Chui and Quak

(1992), Cohen, Daubechies and Vial (1994)), as well as for a finite hyperrectangle in d dimensions. Moreover, it is even possible to extend our constructions, as described in more detail in [B20,B21], to function classes on Lipschitz-graph domains Ω in d dimensions, by utilizing wavelets generated via multiresolution analysis on Ω, as proposed by Cohen, Dahmen and DeVore (1997) in their wavelet-based approach to deriving Whitney-type extension theorems.

4. THE K-FUNCTIONAL

The K-functional has been introduced by Peetre (1963) as a tool for measuring smoothness in abstract spaces. Since then, it has had a rich variety of important applications, for example: interpolation of operators and interpolation spaces (see Bergh and Löfström (1976)); applications to approximation theory via equivalence with moduli of smoothness (Butzer and Berens (1967), Johnen and Scherer (1977), Dechevski (1988,1989), Dechevsky (1993); general characterization of best approximation (Petrushev and Popov (1987)), and many others. For the purposes of these applications it has also been generalized in several aspects. Here we consider one of these generalizations - the K_ρ-functional between two (quasi-semi-)normed spaces A, B:

$$K_\rho(t,f;A,B) = \inf_{f=a+b}(||a||_A^\rho + t^\rho.||b||_B^\rho)^{1/\rho}, 0 < t < \infty, 0 < \rho \leq \infty \tag{4}$$

(with max-modification for $\rho = \infty$), $f \in A + B$ (the sum of A and B as linear vector spaces)[A6].

Taking in consideration the clear resemblance between Wahba's statistical spline-smoothing model and the deterministic model of the K-functional, it is amazing that, ever since the 1960s, two distinct and very *complementary* theories related to the two models have been developed, without almost any methodological interconnection. In the last several years a number of interesting papers on penalized L_2-estimation have appeared, some of them considering also Besov penalties (see DeVore and Lucier (1992), DeVore, Kyriazis, Leviatan and Tikhomirov (1993), Birgé and Massart (1994), Barron, Birgé and Massart (1995), Antoniadis (1996), Amato and Vuza (1997)), but, to the best of our knowledge, so far the only one who has essentially exploited the analogy between Wahba's model and the K-functional has been Cox (1988), Theorems 3.3,4. In a sequence of papers (see also Dechevsky and Penev (1998a,1999), we intend to pursue a study of the applications of the theory of the K-functional to statistical estimation.

To illustrate the conceptual power of the K-functional, and for further use, let us discuss the "smoothing paradigm" in a general K-functional setting. Our consideration will be an extension, generalization and further development of the approaches of Anselone and Laurent (1968) and Cox (1988). Let A, B be quasi-Banach spaces with $B \hookrightarrow A$ (i.e., $B \subset A$ and there exists $c < \infty$ such that $||b||_A \leq c.||b||_B$ for any $b \in B$). If for any $a \in A + B = A$ there exists $b^* = b^*(a)$ so that $K_\rho(t,a;A,B) = \min_{b \in B}(||a-b||_A^\rho + t^\rho||b||_B^\rho)^{1/\rho} = (||a-b^*||_A^\rho + t^\rho||b^*||_B^\rho)^{1/\rho}$, then the K-functional measures the quality of the approximation of a by $b \in B$ for a preassigned (through the t-value) trade-off between fidelity of the data and smoothness of the solution. The embedding $B \hookrightarrow A$ ensures that the presence of the penalization term has smoothing effect. This general smoothing model can be further enhanced by considering a *homogeneous version* of B. Assume that the quasi-norm in B is of the particular form $||b||_B = (||b||_A^\rho + ||Tb||_{B_1}^\rho)^{1/\rho}$, where T is a densely defined on A, linear, possibly unbounded, closed operator from A to the quasi-Banach space B_1. Then, by the closedness of T, B is quasi-Banach; clearly, $B \hookrightarrow A$ and $||.||_B$ is the *graph* quasi-norm of T. The homogeneous version of B is the quasi-(semi-)normed complete space $\dot{B}$ with $||.||_{\dot{B}} = ||T.||_{B_1}$. Smoothing with respect to $K_\rho(t, a; A, \dot{B})$ is generally not identical but always equivalent to smoothing via $K_\rho(t, a; A, B)$ in the sense that

$$K_\rho(t, .; A, B) \asymp K_\rho(t, .; A, \dot{B}) + \min(1, t).||.||_A, \ 0 < t < \infty, \tag{5}$$

with equivalence constants *independent* of the smoothing parameter t.[A7] In Wahba's model B and $\dot{B}$ are an inhomogeneous and a homogeneous Sobolev space; in our wavelet model $B = B^s_{pq}$, $\dot{B} = \dot{B}^s_{pq}$, $A = B^\sigma_{\pi u}$, where the relation between (π, u, σ) and (p, q, s) must be such that: (a) for the selected r (see Section 3) (2,3) hold for both triples; (b) $B \hookrightarrow A$. Hence, by *embedding* theorems for Besov spaces (see Bergh and Löfström (1976), Triebel (1983)), the range of the two triples can be shown to *exactly coincide* with their respective range in Delyon and Juditsky (1996), Theorem 1.[A8] For a fixed preassigned value of t, the role of the parameters π, p, u, q, σ, s in the K-functional model is essentially the same as in Delyon and Juditsky (1996). In particular, the estimated function is assumed to be in B^s_{pq}. The loss functional is the quasi-norm in $B^\sigma_{\pi u}$. If, however, the problem is inverse, i.e., t has to be *estimated* in order to recover f, then the meaning of π, p, u, q, σ, s is different and these parameters may eventually also be subject to estimation. To explain their meaning in this case, let us recall the general inhomogeneous K-functional model. According to it, the estimated f belongs to X - an *intermediate* quasi-Banach space between A and B (see Bergh and Löfström (1976)). In our case $B \hookrightarrow A$, therefore, $B \hookrightarrow X \hookrightarrow A$. For our purposes it suffices to assume that $X = (A, B)_{\theta,\tau}$ or $X = (A, B)_{[\theta]}$, $0 < \theta < 1$, $0 < \tau \leq \infty$, i.e., X is an interpolation space obtained by the real method of Lions and Peetre, or by the complex method of Calderon, Krein and Lions, respectively. We shall concentrate on the real method which is the one directly related to the K-functional. Assume, additionally, that B is *dense* on A. (All the above assumptions, including the latter one, are fulfilled for Sobolev and Besov spaces.) Because of $B \hookrightarrow A$, by saturation (cf. [A7]) argument, $||f||_X = ||f||_{(A,B)_{\theta,\tau}} \asymp ||f||_A + [\int_0^C (t^{-\theta} K_\rho(t, f; A, B))^\tau dt/t]^{1/\tau}$, for any, henceforward fixed, $C : 0 < C \leq \infty$. Because B is dense in A, for any $\epsilon > 0$ one can choose $f_\epsilon \in B$ so that $||f - f_\epsilon||_A < \epsilon$. It can be seen that $K(t, f_\epsilon; A, B) \sim \min(1, t).||f_\epsilon||_B$. Because of the latter, the *Generalized First Mean-value Theorem*, valid for improper integrals (see Fikhtengol'ts (1969), Section 487) can be applied, which yields

$$||f_\epsilon||_X \asymp ||f_\epsilon||_A + \frac{K_\rho(\tilde{t}, f_\epsilon; A, B)}{\min(1, \tilde{t})} [\int_0^C (t^{-\theta}. \min(1, t))^\tau \frac{dt}{t}]^{1/\tau}. \tag{6}$$

The integral is finite and its value does not exceed $[(1-\theta)\theta\tau]^{-1/\tau}$ for any $C : 0 < C \leq \infty$.

(The case $\tau = \infty$ is analogous, but simpler.) Here $\tilde{t} = \tilde{t}_\epsilon : 0 \le \tilde{t} \le C \le \infty$ is the *target value* of the smoothing parameter subject to estimation in the statistical smoothness-penalty models. In these models $\epsilon = \epsilon(n) = o(1)$ as $n \to \infty$, where n is the sample size. The density of B on A ensures that the problem for penalized estimation of f is *correctly posed in the sense of Tikhonov.* For our particular model, however, we need further enhancement. Assuming X to be fixed, let A', B' be quasi-Banach spaces, such that $B \hookrightarrow B' \hookrightarrow X \hookrightarrow A' \hookrightarrow A$. For $X = (A, B)_{\theta,\tau}$ one can choose $A' = (A, B)_{\theta_0,\tau}$, $B' = (A, B)_{\theta_1,\tau}$, where $0 < \theta_0 < \theta < \theta_1 < 1$. By the Reiteration theorem (see Bergh and Löfström (1976)), $X = (A', B')_{\theta',\tau}$, where $\theta' = \frac{\theta-\theta_0}{\theta_1-\theta_0} \in (0, 1)$. This means that (6) can be applied again, this time with A, B, θ and $\tilde{t}$ replaced by A', B', θ' and $\tilde{t}'$, respectively. Therefore, estimation of f_ϵ, hence, of f, can be *improved* by estimating t not only for A and B ($\theta_0 = 0, \theta_1 = 1$), but also for A', B' corresponding to any $\theta_0, \theta_1 : 0 < \theta_0 < \theta < \theta_1 < 1$. By (5), the above considerations are also valid, mutatis mutandis, for the homogeneous version, too.

Returning to the interpretation of the Besov indices (π, u, σ) and (p, q, s), we now see that penalized estimation via $K_\rho(\tilde{t}, f; B^\sigma_{\pi u}, B^s_{pq})$ is specialized for $f \in B^{s'}_{p'q'}$, where $s' = (1-\theta)\sigma + \theta s$, $\frac{1}{p'} = \frac{1-\theta}{\pi} + \frac{\theta}{p}$, $\frac{1}{q'} = \frac{1-\theta}{u} + \frac{\theta}{q}$, $0 < \theta < 1$, $q' = \tau \in (0, \infty]$. In particular, if $\pi = p$, then σ and s are always a *lower* and an *upper bound* for the exact s', respectively.[A9,A10,A11] The methods for penalized L_2 estimation with Besov penalty (see, e.g., Antoniadis (1996), Amato and Vuza (1997)) deal with the case $\pi = u = 2$.

5. EXACT SOLUTION FOR THE CASE $\pi = u = p = q = \rho = 2,\ \sigma = 0$

In this and next section we shall be considering the Hilbert case $\pi = u = p = q = 2$ where comparison can be made with the spline-smoothing model. σ is fixed to 0, that is, we shall be estimating the function itself, but not its derivatives. Besides t, only one additional continuous parameter - s - will be optimized most of the time. For the general case about π, p, u, q, σ, s, see [B9]. The estimator has the form

$$\tilde{f}(x) = \sum_{k\in\mathbf{Z}} \tilde{\alpha}_{j_0 k} \cdot \varphi_{j_0 k}(x) + \sum_{j=j_0}^{j_1} \sum_{k\in\mathbf{Z}} \tilde{\beta}_{jk}\psi_{jk}(x), \tag{1'}$$

with a reasonable choice of the resolution levels j_0 and j_1, variants of which will be discussed below.

The Hilbert case is of special importance, because then the estimator (1') is *linear* in the data for any fixed t in the K-functional. For the Hilbert case, a very general formula for the explicit computation of the K_2-functional between any semi-Hilbert spaces was obtained in Dechevski (1989), Dechevsky (1993, 1993a), where some deterministic applications to the theory of real and complex interpolation spaces and to approximation theory were considered. For our purposes here it suffices to invoke only a partial case of these general results[A13], as follows. The spaces A and B in the definition of the K_ρ-functional are any two Hilbert spaces such that $A \cap B$ is dense in A and B, and the cardinality of B does not exceed that of A. Let L be any, henceforward fixed, isometric operator acting from $A \cap B$ to A, with $\langle La, La\rangle_A = \langle a, a\rangle_B$, $a \in A \cap B$, where $\langle .,. \rangle_A$ and $\langle .,. \rangle_B$ are the scalar products in A, B, respectively. Then (Dechevsky (1993a), Theorem 2.1, (i,ii)), L is a densely defined, possibly unbounded, closed linear (DDPUCL) operator on A; L^* (the Hilbert adjoint of L) is also a DDPUCL operator with $(L^*)^* = L$; L^*L is a densely defined, possibly unbounded, positive definite self-adjoint linear operator on A. Moreover (Dechevsky (1993a), Corollary 2.1 and Remark 2.6), $K_2(t, a; A, B) = Sa$, where S is the extension onto $A + B$ of the densely defined continuous sublinear functional $\overline{S} : A \cap B \to \mathbf{R}_+$, defined by

$$\overline{S}a := \Big(\int_{Sp(L^*L)} \frac{t^2\lambda}{1+t^2\lambda}\, d\langle \mathcal{E}_\lambda a, a\rangle_A\Big)^{1/2}, \quad a \in A \cap B,$$

where $\mathcal{E}_\lambda$ is a spectral function of L^*L (that is, $\mathcal{E}_\lambda$ is a spectral resolution of the identity operator I on $A \cap B$, generated by L^*L), $Sp(L^*L) \subset \mathbf{R}_+$ is the spectrum of L^*L.

An intermediate step to obtaining the above integral representation for $K_2(t, a; A, B)$ is the computation of the optimal $b^* \in B$ on which $\min_{b\in B}$ in the definition of the K_2-functional is attained (Dechevsky (1993a), formula (5.2.13)):

$$b^* = (I + t^2 L^*L)^{-1} a = \int_{Sp(L^*L)} \frac{1}{1+t^2\lambda}\, d\mathcal{E}_\lambda,$$

where the integration is in the sense of the abstract Stieltjes integral with respect to the spectral measure $d\mathcal{E}_\lambda$.

In the case when $||b||_B = (||b||_A^2 + ||Tb||_A^2)^{1/2}$, where T is defined in Section 4, a homogeneous semi-Hilbert version $\dot{B}$ of the Hilbert space B can be considered, as discussed in Section 4. Then, the formulae about $\overline{S}a$ and b^* continue to hold (Dechevsky (1993a), Theorems 3.1.1 and 3.1.1a) also for $K_2(t, a; A, \dot{B})$, with $L = T$. For the homogeneous case, L^*L is positive *semi-definite* ($0 \in Sp(L^*L)$). The case when $A = B_{22}^{\sigma}$, $B = B_{22}^{s}$, $\dot{B} = \dot{B}_{22}^{s}$, $0 \le \sigma < s < r$, with the equivalent wavelet-coefficient norms (2,3), starting from level j_0, is a very simple partial case of the above general consideration. For example, let us consider the homogeneous model. In the present context, the operator $L = T$ is self-adjoint itself: $L = L^*$, therefore, $L^*L = L^2$ and it suffices to study the spectral properties of L. The spectrum $Sp(L)$ of L is discrete; every $\lambda \in Sp(L)$ is an eigenvalue; the eigenvalues are: $\lambda_0 = 0$, $\lambda_j = 2^{js}$, $j = j_0, j_0+1, \dots$. The respective eigenspaces are the subspaces in the multiresolution analysis generated by φ and ψ (with the usual notations $V_{j+1} = V_j \oplus W_j$): the eigenspace of λ_0 is V_{j_0}, that of λ_j is W_j, $j = j_0, j_0+1, \dots$. The formula about b^* now implies the following

Lemma 1.[A13,A14] *Let $s > 0$ be fixed, $t > 0$ be fixed and the quality of fit be measured by the K-functional $K_2(t, \hat{f}; B_{2,2}^0, \dot{B}_{2,2}^s)$. Then, the K-functional is attained for the minimizing estimator (1') with coefficients*

$$\tilde{\alpha}_{j_0k} = \hat{\alpha}_{j_0k},\ k \in \mathbf{Z},\ \tilde{\beta}_{jk} = \frac{\hat{\beta}_{jk}}{1+t^2.2^{2js}},\ j = j_0, \dots, j_1. \tag{7}$$

The empirical wavelet coefficients are: $\hat{\alpha}_{j_0k} = \frac{1}{n}\sum_{i=1}^{n}\varphi_{j_0k}(X_i)$, $\hat{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(X_i)$ *in the density estimation case and* $\hat{\alpha}_{j_0k} = \frac{a_2-a_1}{n}\sum_{i=1}^{n}\varphi_{j_0k}(x_i)y_i$, $\hat{\beta}_{jk} = \frac{a_2-a_1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)y_i$ *in the case of regression.*

Formula (7) tells us that the resulting estimator (1') is of *level-dependent non-threshold shrinkage* type.

The parametrization of the nonparametric problem achieved by the atomic decomposition via orthogonal wavelets makes it very easy to give a direct proof of Lemma 1 (see Appendix A). This simple proof of (7) has been obtained, independently from Dechevski (1989) and Dechevsky (1993, 1993a), by Amato and Vuza (1994) and Antoniadis (1996). To the best of our knowledge, the non-threshold shrinking model (7) has not been explored in the case of density estimation. In the case of regression-function estimation, Amato and Vuza (1994, 1997, 1997a, 1998), and Antoniadis (1996) seem to be the first to propose concrete statistical procedures for the estimation of the regularization parameter t^2 for (7) (see also Amato, Occorsio and Vuza (1998)). The methods of Antoniadis and of Amato and Vuza are both specialized for work with sample sizes which are a power of two, when ODWT can be applied. In both methods the value of j_1 is fixed on the maximal possible value in the ODWT model: $j_1 = \log_2 n - 1$. As we shall see, this turns out to be one major weakness of both of these methods when estimating continuous functions with fractal graphs and spatially inhomogeneous functions with relatively low regularity, in particular, piecewise smooth functions with discontinuities. We suggest that in all cases (i.e., when estimating fractal functions, functions with jumps, and even smooth functions, for both the regression and density estimation models) j_1 should be a controlled parameter and that $2^{j_1} = o(n)$ should hold. The choice $2^{j_1} \asymp n$ can be acceptable for very smooth functions only. For such very smooth functions, however, thresholding techniques seem to offer quite a competitive alternative to (7) for consistent estimation of the function together with its derivatives. Thus, the methods of Antoniadis and of Amato and Vuza, which are not well adapted for work with small values of the smoothness index s, essentially miss the most important range of s for which non-threshold shrinking methods outperform threshold ones.

Considerations related to asymptotic minimax optimality theory (see Delyon and Juditsky (1996)) suggest $(\tilde{C}n)^{1/(2s+1)} \le 2^{j_0} \le 2(\tilde{C}n)^{1/(2s+1)}$, where $\tilde{C} = \tilde{C}(L)$ is defined via (16) and (21) in Delyon and Juditsky (1996) for density and for regression-function estimation, respectively. (The reader is cautioned that there is a misprint in formula (16) in Delyon and Juditsky (1996).) Here L is an *upper bound* for the B^s_{22}-norm of the estimated function f. Again by (16) and (21) in the same paper, $n/\ln n \le 2^{j_1} \le 2n/\ln n$.

Remark 1.^{A15} The proofs of the theorems in Section 6 (see Appendix A) show that the cross-validation method for estimating t is consistent, for a broad range of s, including small values, under relaxed assumptions: $2^{j_1} = o(n)$ (for both the regression and density case); $2^{j_0} = O(1)$ or $j_0 \to \infty$, $j_0 \le j_1$ for the regression case, $2^{-j_0} = O(n^{-\frac{1}{2(1+s)}})$ in the density case. If no additional information about the function is available, we recommend the choice $2^{j_1} \asymp \frac{n}{\ln n}$ because the estimator performs well for both smooth and piecewise smooth functions with isolated points of singularity. The type of 1-d singularity may be, e.g., a "horn" or a "cusp" (discontinuity of the first derivative, which in the case of a cusp point is unbounded in a neighbourhood of this point), a "jump" (the left-hand and right-hand limits of the function are different at this point), or a "chirp" (the function does not have a left-hand and/or a right hand limit and its variation is unbounded in a neighbourhood of this point). If additional information is available that f is smooth, then choices of the type, say, $2^{j_1} = \frac{n}{\ln \ln n}$ are acceptable. If further additional information is available that f is not very spatially inhomogeneous and s is 'large' (at least $s > 1$ but typically $s \gg 1$), then $2^{j_1} \asymp n$ is also an admissible choice. If, on the contrary, it is known additionally that $s \leq 1/2$ and that the function is a 'monstrous' one (e.g., has a countable set of discontinuities with at least one density point), then choices $2^{j_1} = o(\frac{n}{\ln n})$ should be considered. The option to control the performance of the estimator by varying j_1 (that is, by applying a non-threshold shrinking approach via (7) in combination with a levelwise 'kill-or-keep' thresholding is one realization of a simple partial case of the general idea about composite shrinking/thresholding estimators, considered in ^{B8}.

6. CROSS VALIDATION IN L_2

Besides asymptotic-minimax methods, there is also a variety of other methods for estimation of the smoothing parameter(s) in penalized estimation (e.g., the method of Antoniadis (1996), SURE and James-Stein estimation for Gaussian white noise, etc.) but one of the oldest and most general methods, working also for non-Gaussian white noise, available also in the case of density estimation, is cross validation. Its theory poses serious challenges and asymptotic results hold sometimes in a weaker sense than in asymptotic minimax theory, but its performance on moderate samples is good most of the time. This method copes well even with correlated noise, at least when the covariance matrix is band-limited or its entries tend to vanish relatively fast away from the main diagonal. That is why, of all methods for determining the smoothness parameter in the general setting of (6), here we would like to consider in detail cross validation as a model example. In principle, cross-validation procedures can be formulated for non-Hilbertian loss functions, too, but the resulting optimization problems are easier to analyze in the Hilbert setting. For this reason, it is desirable to study the Hilbert case in quite a detail, before looking at more general situations. This is going to be the direction of our research in this section, where we concentrate on work with the non-threshold shrinking estimator (7), adapted for the Hilbert case. For the generalization of (7) to the general quasi-Banach case, see Appendix B, formulae (B4-B6). Section 6 contains a more detailed and updated discussion of some earlier results (see Dechevsky and Penev (1998a)), as well as new results on consistency of the estimator and existence and asymptotic behaviour of the smoothing parameter. Subsection 6.1 includes also a comparison of our method for the regression case with the approach of Amato and Vuza (1994, 1997a, 1998) and the approach of Antoniadis (1996).

Cross validation (CV) is performed in a different way for regression estimation and for density estimation. The reason for the difference is that in the latter case there are no output observations $Y_i,\ i = 1, 2, \ldots, n$ that could be utilized to evaluate the predictive power of the model. Another important difference is that in the latter case MISE(v)$= E\int(\tilde{f}_v(x) - f(x))^2 dx$ is being directly estimated, while in the former one the estimation is of a quadrature formula approximating MISE(v). However, there is an important common feature between the two models: the dependence of the CV-criteria for both models is analytic in $v = t^2$ on the compact $[0, C]$ for any $C : 0 < C < \infty$. This is an essential difference between the shrinking and the thresholded wavelet estimator (cf. Jansen, Malfait and Bultheel (1997)). Because of it, existence and asymptotic behaviour of the v-minimizer of the CV-functional for the shrinking estimator can be studied in quite a detail. In general, consistency of the cross-validation procedure in L_2 does not automatically imply consistency of the resulting estimator $\tilde{f}_v$ in L_2; for this to happen, the v-minimizer must have a specific asymptotic behaviour: $v \to 0$ as $n \to \infty$, at least when $j_0 = O(1)$. If this is not fulfilled, the resulting estimator $\tilde{f}_v$ cannot be consistent in L_2, because β_{jk} would not be consistently estimated. For an arbitrary wavelet-based estimator, consistent estimation of the coefficients in the wavelet expansion of f is only a very weak necessary condition for consistency of the estimator itself: it is not sufficient for the estimator's consistency even in the weak topology on L_2. We show that the shrinking wavelet estimator $\tilde{f}_v$ enjoys a remarkable property: when $j_0 = O(1)$, consistent estimation of the wavelet coefficients is for it not only necessary, but also *sufficient* for consistency of $\tilde{f}_v$ in L_2 and even in B^{σ}_{22}, $0 < \sigma < s$. The range of σ depends on s and the choice of j_0 and j_1.

The wavelet setting permits also explicit characterization of all functions for which there is no v-minimizer of the CV-criterion for arbitrarily large n. It turns out that any such function must be in V_j - the closure of the linear span of φ_{jk}, $k \in \mathbf{Z}$, for some $j \in \mathbf{Z} : j < j_0$.

We shall be utilizing the homogeneous K-functional $K_2(\sqrt{v}, f; L_2, \dot{B}^s_{22})$ throughout, with $f \in B^{s'}_{22}$, $0 < s' < s$. Under these conditions, $B^{s'}_{22} \hookrightarrow \dot{B}^{s'}_{22} \hookrightarrow L_2 + \dot{B}^s_{22}$ (see also Section 4). For j_0 we shall be considering two cases: $j_0 = O(1)$ and $j_0 \to \infty$. In the proofs we shall be giving the details for the second case. In the first case the proof is analogous, but simpler. To obtain quantitative results about the second case we shall be considering only those $f \in B^{s'}_{22}$ for which the index s' is sharp, that is, $f \notin B^s_{2\infty}$ for any $s > s'$. Note that, for any compactly supported $f \in B^{s_0}_{22} \setminus C^\infty$ for some s_0, there exists $s' \geq s_0$ such that $f \in B^{s'}_{22}$ and s' is sharp.[A16] The case $f \in C^\infty$ is possible to incorporate, of course, but in K-functional context it is not very interesting.

For the purposes of estimation in regression problems, when f has low regularity ($0 < s' \leq 1/2$), we discuss in subsection 6.1 the importance of the concept of *A-spaces* and *average moduli of smoothness* [B12b)].

Throughout the section we implicitly assume that f is compactly supported, as discussed in the end of Section 3.

6.1. Regression estimation

Cross validation is usually numerically intensive unless there are some updating formulae that allow to calculate the "leaving-out-one" predictions on the basis of the "full" predictions only. In this respect, very helpful is the "leaving-out-one" lemma (Lemma 4.2.1 of Wahba (1990)) which allows deriving such updating formulae. This lemma's conclusion can only be achieved under the so-called "compatibility condition" (see also Droge (1996), equation (3.2)). Although the compatibility condition is easy to derive for projection-type estimators, it might be difficult to prove in other situations. In fact, it very often fails to hold. This is also the case with our estimator (1'). Remedy from this situation is to pretend that the compatibility condition "almost holds" and work further by bearing into consequence the imprecision that occurs due to the approximation error. This approach to cross-validation in wavelet regression has been adopted, in threshold setting, by Jansen, Malfait and Bultheel (1997). Similar is the situation with the cross-validation method of Nason (1996) which is based on an ingenious exploitation of the symmetries existing only when n is a power of two. This method essentially relies on the brute force of the adopted "leave-out-half" strategy. At present, it is hard to say what price, in terms of, say, asymptotic rates, is being paid for leaving out half of the sample in this context, because this estimator has hardly any theory developed for it, partly due to its non-smooth optimization criterion. (We discuss Nason's estimator in more detail in ^{B20}.) A similar approach to overcoming the failure of the compatibilty condition is that of Amato and Vuza (1997a, 1998). Since for the estimator (1') with (7) the compatibility condition is violated, they have chosen to work directly with a wavelet analogue of formula (4.3.1) in Wahba (1990) for $n = 2^N$ and, by doing so, have essentially adopted the same approach as Jansen, Malfait and Bultheel (1997), but in the setting of (7). On its part, the method of Antoniadis (1996), which is not of cross-validation type, completely avoids the compatibility problem. One big price to be paid for this is that the estimator depends explicitly on the noise variance. To deal with this problem, Antoniadis incorporates an estimator of the noise variance within the definition of his estimator, but this comes at the price of spoiling the performance of the estimator on less regular functions. This is one of several reasons why this estimator is only appropriate for estimating smooth functions.

Here we suggest to follow another possibility which leads to a stringent solution of the compatibility problem within the cross-validation setting, namely, to modify the cross-validation criterion itself by replacing it with the so-called *full cross-validation criterion (FCV)* (see Droge (1996)).

Denote

$$\tilde{f}_v(x) = \sum_k \hat{\alpha}_{j_0 k}\varphi_{j_0 k}(x) + \sum_{j=j_0}^{j_1}\sum_k \hat{\beta}_{jk}\frac{1}{1+v2^{2js}}\psi_{jk}(x),$$

$$\tilde{f}_v^{(-i)}(x) = \sum_k \hat{\alpha}_{j_0 k(-i)}\varphi_{j_0 k}(x) + \sum_{j=j_0}^{j_1}\sum_k \hat{\beta}_{jk(-i)}\frac{1}{1+v2^{2js}}\psi_{jk}(x),$$

with $\hat{\alpha}_{jk(-i)} = \sum_{l=1,l\neq i}^{n} \frac{a_2-a_1}{n-1}\varphi_{jk}(x_l)Y_l$ and $\hat{\beta}_{jk(-i)} = \sum_{l=1,l\neq i}^{n} \frac{a_2-a_1}{(n-1)(1+t^2 2^{2js})}\psi_{jk}(x_l)Y_l$. (For the sake of convenient comparison with our main reference source here (Wahba (1990)), the notation $y_i := Y_i$ will be used henceforward.) Then, the typical *ordinary cross-validation* (OCV) functional to be minimized with respect to $t^2 = v$ is given by (Wahba (1990), p. 47):

$$V_0(v) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \tilde{f}_v^{(-i)}(x_i)]^2 \tag{8}$$

Computational reduction in minimizing (8) is achieved if the "compatibility condition"

$$\hat{y}_{(-i)}(v) = \tilde{f}_{n,v}(x_i; y_1, \ldots, y_{i-1}, \hat{y}_{(-i)}(v), y_{i+1}, \ldots, y_n)) \tag{9}$$

holds, where $\hat{y}_i = \tilde{f}_v(x_i)$, $\hat{y}_{(-i)}(v) = \tilde{f}_v^{(-i)}(x_i) := \tilde{f}_{n-1,v}(x_i; y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$.

It is easy to show that, under the above condition, the OCV-functional can be expressed in terms of the ordinary residuals:

$$V_0(v) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \tilde{f}_v(x_i)]^2/(1-h_{ii}(v))^2, h_{ii}(v) = \frac{\partial \tilde{f}_v}{\partial y_i}(x_i; y_1, \ldots, y_n) \tag{10}$$

To explain our notation, let us note that because of the linearity of our estimator, we can write $\hat{\mathbf{y}}(v) = \mathbf{H}(v)\mathbf{y}$, $\hat{\mathbf{y}}(v) = [\hat{y}_1(v), \hat{y}_2(v), \ldots, \hat{y}_n(v)]'$, $\mathbf{y} = [y_1, y_2, \ldots, y_n]'$ with certain $n \times n$ matrix $\mathbf{H}(v)$ (called *influence matrix* (Wahba (1990)). In that case, $\frac{\partial \tilde{f}_v}{\partial y_l}(x_i; y_1, \ldots, y_n)$ would indeed be the (i,l)-th element of the matrix $\mathbf{H}(v)$.

Unfortunately, for a shrinking-type estimator like (1'), the compatibility condition is violated and a formula like (10) does not hold precisely. Simple observation shows that (10) still can be considered as approximately true, i.e., $h_{il}(v)$ can still be calculated as $h_{il}(v) \approx \frac{\partial \tilde{f}_v}{\partial y_l}(x_i; y_1, \ldots, y_n)$, but the error of this approximation has yet to be evaluated. We opt for another alternative here by changing the cross-validation criterion itself. Note that standard cross-validation is defined entirely with respect to samples of size $n-1$. One can adjust it for samples of size n as suggested by Bunke, Droge and Polzehl (1993). In their approach, the value of y_i is replaced by $\hat{y}_i(v)$ instead of leaving it out in defining the prediction of the i-th design point. The resulting functional is called the FCV (*full* cross-validation) functional:

$$FCV(v) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \tilde{y}_i(v)]^2 \tag{11}$$

with $\tilde{y}_i(v) = \tilde{f}_{n,v}(x_i; y_1, \ldots, y_{i-1}, \hat{y}_i(v), y_{i+1}, \ldots, y_n)$. Then it turns out that under the condition of *linearity only*, one gets in terms of the ordinary residuals:

$$FCV(v) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \tilde{f}_v(x_i)]^2.(1+h_{ii}(v))^2, \quad h_{il}(v) = \frac{\partial \tilde{f}_v}{\partial y_l}(x_i; y_1, \ldots, y_n) \tag{12}$$

An easy calculation gives the following value for $h_{il}(v)$:

$$h_{il}(v) = \frac{a_2 - a_1}{n}[\sum_k \varphi_{j_0k}(x_i)\varphi_{j_0k}(x_l) + \sum_{j=j_0}^{j_1}\sum_k \frac{1}{1+v2^{2js}}\psi_{jk}(x_i)\psi_{jk}(x_l)], \qquad (13)$$

and the matrix $\mathbf{H}(v)$ is symmetric positive semi-definite.

As it is often done in the cross-validation approach, OCV is replaced by *generalized cross validation* (GCV) (see Wahba (1990), Section 4.3). The reason for this is that the generalized version has desirable invariance properties that do not generally hold for OCV. Numerically, going over from the ordinary to the generalized version means that one replaces the values of $h_{ii}(v),\ i = 1, 2, \ldots, n$ in (10) by one value only: $\overline{h}(v) = \frac{1}{n}\sum_{i=1}^{n} h_{ii}(v) = \frac{1}{n}$ tr $\mathbf{H}(v)$. Also, proofs related to consistency of the cross-validation procedure are easier to derive for generalized, rather than ordinary cross-validation. Following the same idea, we introduce *generalized FCV* (*GFCV*) by replacing $h_{ii}(v),\ i = 1, 2, \ldots, n$ in (12) with $\overline{h}(v)$.

It is informative here to compare our approach to that of Amato and Vuza (1997a), who have tried to develop a straightforward analogue of the cross-validation procedure, as suggested by Wahba (1990). Because the compatibility condition fails for OCV when the shrinking is via (7), an analogue of our formula (12) and of formula (4.2.13) in Wahba (1990) does not hold precisely. However, one can hope, in the spirit of Jansen, Malfait and Bultheel (1997), that for OCV such an analogous formula holds at least approximately. Under this assumption, following Wahba, it is possible to consider a generalized version of the approximate formula for OCV by replacing h_{ii} with $\overline{h}$. Following this line of reasoning for the case $n = 2^N$, and making use of the discrete wavelet transform, Amato and Vuza have proposed a method which they call generalized cross validation (GCV). In the sequel we shall refer to their method as quasi-GCV (or QGCV, for short) because it is not an exact generalization of OCV.

In the rest of this subsection we shall study the properties of GFCV and shall compare them to the respective properties of the existing non-threshold shrinking techniques QGCV of Amato and Vuza and WAVREG of Antoniadis (1996).

We are in a position to show that GFCV is a consistent method in the traditional sense. To formulate precisely our results, let us introduce the mean square error $ER(v) = E\frac{1}{n}\sum_{i=1}^{n}(\tilde{f}_v(x_i) - f(x_i))^2$ (the unknown quantity we are interested in controlling) and compare it to $E(GFCV(v)) = E\frac{1}{n}\sum_{i=1}^{n}\{y_i - \tilde{f}_v(x_i)\}^2(1+\bar{h}(v))^2$. Further, let us define $\mu(v) = \frac{1}{n}$ tr $\mathbf{H}'(v)\mathbf{H}(v) = \frac{1}{n}$ tr $\mathbf{H}(v)^2$. It can be seen that $\mu(v) \geq \overline{h}(v)^2$ holds. ($\bar{h}(v)$ and $\mu(v)$ are also functions of n but this dependence has been suppressed in the notation.) We denote asymptotic equivalence by '$\sim$'.

Theorem 1. *(Sufficient condition for existence of the v-minimizer.) Let* $0 < s_1 \leq s' < s < r,\ s_1 < 1$. *Assume that the penalized model is via* $K_2(\sqrt{v}, f; L_2, \dot{B}^s_{22})$ *and that* $f \in B^{s'}_{22} \cap B^{s_1}_{\infty\infty}$. *Assume also that* $j_0 \leq j_1,\ j_1 \to \infty$ *with* $2^{j_1} = O(\frac{n}{\ln n})$. *If* $j_0 = O(1)$ *and* $f \notin V_j$ *for any* $j < j_0$, *or if* $j_0 \to \infty,\ f \notin V_j$ *for any* $j \in \mathbf{Z}$, *the index* s' *is sharp and* $2^{j_0} = o(n^{\min(\frac{s_1}{(s-1/2)_+}, \frac{2}{1+2s})})$, *then, for sufficiently large n there exist $\tilde{v}$: $E(GFCV(\tilde{v})) = \min_v E(GFCV(v))$, v^* : $ER(v^*) = \min_v ER(v)$, and $0 \le \tilde{v} < C$, $0 \le v^* < C$, where $C < \infty$ depends on f and ψ only.*

The problem about existence of the minimizer has not been explored for QGCV and WAVREG and Theorem 1 has no analogue in the theory developed for these estimators.

Asymptotic optimality of the minimizer of the smoothing parameter is established by the following

Theorem 2. *(Consistency of GFCV - cf. also Dechevsky and Penev (1998a).) Assume that $0 < s' \le s < r$, the penalized model is via $K_2(\sqrt{v}, f; L_2, \dot{B}^s_{22})$ and $f \in B^{s'}_{22}$. Assume also that $j_0 \le j_1$, $j_1 \to \infty$ and $2^{j_1} = o(n)$ holds. Then, for any $v \ge 0$, $E(GFCV(v)) \sim ER(v) + \delta^2$ holds, δ^2 being the error variance. If $\tilde{v}$ and v^* exist for sufficiently large n, then $\frac{ER(\tilde{v})}{ER(v^*)} \searrow 1$, as $n \to \infty$.*

Note that the asymptotic optimality of the minimizer $\tilde{v}$ of GFCV is achieved without knowing the value of δ or the type of the white noise. The respective result in Amato and Vuza (1997a), p.486, is also for unknown δ, and is valid for $s > 1/2$. The respective results in Antoniadis (1996) (Corollary 3.2 for the case of known noise variance δ^2 and Theorem 3.3 for the general case (δ unknown)) are weaker: they describe the behaviour of WAVREG's v-minimizer only by measuring the relative error with respect to the asymptotic minimax rate $n^{-\frac{2s}{1+2s}}$ instead of the relative error with respect to the true value of the estimated quantity.

Theorem 3. *(Consistency of $\tilde{f}_{v^*}$ in B^{σ}_{22}, $0 \le \sigma < s' \le s < r$.) Assume that the conditions of Theorem 1 hold, without the sharpness conditions on s'. Then, if $j_0 = O(1)$ and if $v = v_n$ is bounded, the condition $v \to 0$ as $n \to \infty$ is necessary for $E||\tilde{f}_v - f||^2_{B^{\sigma}_{22}} \to 0$ to hold; moreover, in both cases $j_0 = O(1)$ and $j_0 \to \infty$, if $2^{j_1} = o(n^{\min(\frac{s_1}{\sigma}, \frac{1}{1+2\sigma})})$ holds, then $v \to 0$ is also a sufficient condition for $E||\tilde{f}_v - f||^2_{B^{\sigma}_{22}} \to 0$. In particular, if s' is sharp, then: the above necessity claim is true for $v = v^*_n$; if $2^{j_1} = o(n^{\min(\frac{s_1}{\sigma}, \frac{1}{1+2\sigma})})$ holds, too, the above sufficiency claim is also true for $v = v^*_n$.*

Theorem 3 has no analogue for QGCV and WAVREG. One important reason for this is that in those two techniques j_1 is fixed at $2^{j_1} \asymp n$. It should be noted, though, that if $v \to 0$ and if the rate of v tending to zero is known, the restriction on j_1 in Theorem 3 can be relaxed and, if the rate is high enough, $2^{j_1} \asymp n$ becomes an admissible choice, too.

Theorem 4. *(Asymptotic behaviour of v^* and $\tilde{v}$.) Under the premises of Theorem 1, $v^* \to 0$ and $\tilde{v} \to 0$ as $n \to \infty$.*

Theorem 4 has no analogue for WAVREG. For QGCV, Amato and Vuza study only the case about $\tilde{v}$, for j_0 and j_1 fixed at 0 and $\log_2 n - 1$, respectively, where n is a power of two. The restriction on s is $s > 1/2$ (the range where QGCV has been shown to be consistent), and s' is set equal to s. Under these assumptions Amato and Vuza (1998) find also the sharp quantitative rates for $\tilde{v}$. In our case, for GFCV, the sharp quantitative rates for $v^* \to 0$ and $\tilde{v} \to 0$ can be found in terms of s', s, j_0 and j_1 for any $j_0 \leq j_1$, $j_1 \to \infty$, and any s' and s, such that $0 < s' \leq s < r$. One of the interesting extensions of this topic is to analyze for which pairs (j_0, j_1) the expected rate for v^* and $\tilde{v}$ is the asymptotic minimax one, i.e., $n^{-2s'/(1+2s')}$ (see also [B14]).

Corollary 1. *(Consistency of $\tilde{f}_{\tilde{v}}$ in B^{σ}_{22}, $0 \leq \sigma < s' \leq s < r$.)). Under the premises of Theorem 1, $MISE(\tilde{v}) = E||\tilde{f}_{\tilde{v}} - f||^2_{L_2} \to 0$ holds as $n \to \infty$. If, additionally, $2^{j_1} = o(n^{\min(\frac{s_1}{\sigma}, \frac{1}{1+2\sigma})})$ is fulfilled, then also $E||\tilde{f}_{\tilde{v}} - f||^2_{B^{\sigma}_{22}} \to 0$ holds.*

The respective result for WAVREG (Antoniadis (1996), Theorem 3.2) is valid in a weaker sense: convergence is in probability, for $\sigma = 0$ only, under stronger restrictions on the white noise. Amato and Vuza (1998) prove an analogue of Corollary 1 for QGCV when $j_0 = 0$ and $2^{j_1} = n-1$, by a detailed study of the convergence rate for $\tilde{v}$ in their model. An analogous study for GFCV [B14] additionally allows to determine for which pairs (j_0, j_1) the rate of $E||\tilde{f}_v - f||^2_{B^{\sigma}_{22}} \to 0$ (with $v = E\tilde{v}$ or $v = Ev^*$) is asymptotically minimax-optimal.

Remark 2. If $f \in V_j$ for some $j < j_0$, then $\{v^*_n\}$ and $\{\tilde{v}_n\}$ can have more than one density point in $[0, \infty)$. It is also possible for the infima of GFCV(v) and ER(v) to get attained for $v \to \infty$ for infinitely many n. In all these cases the estimator $\tilde{f}_v$ with the cross-validated choice of v can still be consistent in B^{σ}_{22}. It is important that for the *inhomogeneous* model (see Sections 4, 5) asymptotic behaviour $v \to 0$ is guaranteed (for sharp s') *in all cases* except $f \equiv 0$ a.e.

Remark 3. Let us summarize the results of the comparison of GFCV, QGCV and WAVREG, all of them based on non-threshold shrinking via (7). (a) In the theory of QGCV and WAVREG it is assumed that $s' = s$; the considerations in our Section 4 show that $s' = s$ is in fact only a limiting case, while the true relation between s' and s is $0 \leq \sigma < s' \leq s$. Therefore, in order to apply penalized shrinking, the true index s' of the estimated function does not have to be known exactly, but only approximately, which is, of course, a very important improvement of the penalization model. This also suggests that the formulations and proofs of all theorems in the theory of QGCV and WAVREG need to be modified so as to hold in the general case $s' \leq s$. Moreover, we believe that some of the key results in Antoniadis (1996) can be improved considerably even in the case $s' = s$. (b) The three methods yield comparable results for smooth functions ($s > 1$). The high order of vanishing moments of the coiflets used in WAVREG reduces the bias considerably when $s \gg 1$, but the variance term is generally larger than that of GFCV (with $2^{j_1} = o(n)$) and comparable to that of QGCV, which makes the use of coiflets efficient only for very small noise variance and very spatially homogeneous smooth curves. The performance of WAVREG on very smooth but spatially inhomogeneous curves can be visibly improved if j_1 is reduced from its present value $\log_2 n - 1$ to $o(\log_2 n)$ because this would reduce the contribution of the variance term to the overall error of the method. In its present version proposed in Antoniadis (1996), the distance between j_0 and j_1 is simply too large for a single smoothing parameter to handle for moderate sample size and a spatially inhomogeneous curve, even if the curve is very smooth. For the case of non-smooth functions ($s \leq 1$, especially $s \leq 1/2$) GFCV is an easy winner over QGCV and both GFCV and QGCV are easy winners over REGWAV. The most important (though not the only) reason for the domination of GFCV in this range of the smoothness index is quite simple: for QGCV and WAVREG $2^{j_1} \asymp n$ while for GFCV $2^{j_1} = o(n)$ holds (see also Remark 4 below). If QGCV and WAVREG be modified so that $2^{j_1} = o(n)$ then these estimators, especially QGCV, would become more competitive for small s. Let us support this claim by a more detailed comparison between GFCV and each of QGCV and WAVREG. In order to compare GFCV and QGCV, assume for a moment the (very unfavourable for GFCV) setting $j_1 = \log_2 n - 1$ and $j_0 = 0$. Assume also that $n = 2^N$ and ODWT is applied for both GFCV and QGCV. Under these very favourable for QGCV conditions, it is easy to compute

$$\overline{h}(v) = \frac{1}{n}(1 + \sum_{j=0}^{j_1} \frac{2^j}{1 + v2^{2js}}), \quad j_1 = \log_2 n - 1,$$

$$QGCV_s(v) = \frac{1}{(1 - \overline{h}(v)^2)^2} GFCV_s(v).$$

Taking in consideration that the asymptotic behaviour of the v-minimizer for both criteria is $v \to 0$ as $n \to 0$, and that, clearly, $\overline{h}(v) \to 1$ as $v \to 0$, it is seen that, with the increase of n, $QGCV_s$ becomes increasingly numerically unstable and the problem for optimization of v becomes increasingly ill-conditioned. It is also seen that when s decreases the instability problems with QGCV become ever more severe, which means that the smaller s, the worse the performance of $QGCV_s$ is, compared to that of $GFCV_s$. By replacing the sequence $\{2^j\}$ in QGCV by another increasing sequence $\{a_j\}$ (which is an additional justification for calling this method *quasi*-GCV), Amato and Vuza have reduced the instability problems, because for an appropriate choice of a_j QGCV remains bounded as $n \to \infty$. However, even in this form, $\overline{h}(v) > 0$ remains bounded away from zero as $n \to \infty$ which means that the variability of QGCV is asymptotically bigger than that of GFCV. Returning to the present setting for j_1 ($2^{j_1} \asymp n$ for QGCV, $2^{j_1} = o(n)$ for GFCV) leads to total domination of GFCV. If we now improve QGCV by resetting j_1 in its definition to be such that $2^{j_1} = o(n)$ holds, then $\overline{h}(v) \to 0$ as $n \to \infty$, and the two methods become asymptotically equivalent. However, the variability of QGCV is still larger than that of GFCV (cf. also Droge (1996)), and, when s is small, GFCV performs better than the improved QGCV for moderate samples. If s is large, the performance of GFCV and the improved QGCV is very similar already for moderate sample sizes. These theoretical observations were confirmed by our numerical studies. With the increase of the noise variance the superiority of GFCV becomes more visible. GFCV is also more robust against outliers in the data. Summing up, GFCV can be considered as a stabilized version of the improved QGCV. Compared to the original version of QGCV, GFCV is visibly superior not only for moderate, but also for relatively large sample sizes. In brief, we feel safer, both theoretically and computationally, when applying the FCV (GFCV) approach rather than the OCV (GCV) type approach. In fact, due to the above-mentioned reasons, we recommend it for Wahba's spline-smoothing model, too.

Comparing GFCV to WAVREG, we note that, in contrast to most wavelet estimation methods, *no estimator of the noise variance* is involved in the definition of our estimator. This is a very essential advantage: the definition of the estimator does not depend on the structure and parameters of the noise. On the contrary, WAVREG is a typical estimator involving noise variance. The explicit dependence of WAVREG (via the estimator $\hat{\Lambda}_n$ (Antoniadis (1996), p.323)) on the noise variance is eliminated by passing on to $\tilde{\Lambda}_n$ (Antoniadis (1996), p.323), but only at the price of additional worsening of the performance of WAVREG for non-smooth functions. There is also an important error in Lemma 3.1 of Antoniadis (1994, 1996). We shall be discussing the consequences of this error in more detail in the end of this subsection.

In the proof of Theorem 2 the following lemma is of crucial importance.[A17]

Lemma 2. *(cf. Dechevsky and Penev (1998a).) Assume that for the estimator (1') $j_0 \leq j_1$, $j_0 = O(1)$ and $2^{j_1} = o(n)$ hold as $n \to \infty$. Then, for any $v \geq 0$,*

$$\lim_{n \to 0} \frac{\overline{h}(v)^2}{\mu(v)} = 0. \tag{14}$$

Remark 4. An analogue of (14) is crucial for the consistency of L_2-cross validation for any estimator - spline- or wavelet-based, thresholded or shrinking. Essentially, in (14) two invariants of $\mathbf{H}(v)$ are compared. They are two distinct *Schatten - von Neumann S_γ-norms* of the linear compact finite-rank operators with matrices $\mathbf{H}(v) = \mathbf{H}_n(v)$, $n \in \mathbf{N}$ (see Dechevsky and Persson (1996)). It is the choice of the L_2-metric that determines the two S_γ-norms to be the *Hilbert-Schmidt* norm ($\gamma = 2$) and the *Fredholm* (or *nuclear*) norm ($\gamma = 1$) (for other metrics, see [B1]). Our proof is essentially different from the "classical" approach to proving (14) for the spline-smoothing model (see Wahba (1990), pp. 57-58, for a short outline). The latter can only be applied under additional restrictive assumptions about the s-values of $\mathbf{H}(v)$ (Utreras (1979), p. 204, conditions (i) and (ii); Cox (1988), p. 701, Remark after Theorem 2.4; Wahba (1990), pp. 57-59). The same is true for the "classical" approach to proving (14) adopted in wavelet context by Jansen, Malfait and Bultheel (1997) for thresholding via cross validation. The nature of the thresholding procedure requires that in this case - unlike ours - $\mathbf{H}(v)$ also *depends on the estimated function f via the sample* $\mathbf{y}$. In this case the additional restrictions on the s-numbers of $\mathbf{H}(v)$ result in an essential restriction on the regularity of the estimated f, formulated in Assumption 1 of Jansen, Malfait and Bultheel (1997). This assumption virtually excludes from consideration continuous f with fractal graphs and also very spatially inhomogeneous $f \in C^\infty$. The wavelet version of the Riemann-Lebesgue Lemma also implies that Assumption 1 fails for piecewise smooth f with discontinuities. The "classical" approach *ignores the key fact on which our proof relies:* $2^{j_1} = o(n)$. In the "classical" setting there is no parameter j_1 to control, or it is ignored, which in our case is equivalent to the implicit assumption $2^{j_1} = O(n)$. Under the latter condition on j_1, we would also need restrictive assumptions on the s-numbers in order to prove (14). The above consideration shows that it is the *frequency localization* which is crucial here: it is the one that provides the controllable parameter j_1. As a consequence, the new approach to proving (14) can in principle be adapted to relax the restrictions on the s-numbers of $\mathbf{H}(v)$ by a controlled truncation of the highest frequencies also for a rather wide subclass of spline-smoothing estimators, thereby improving their performance for less regular curves. Summing up: the non-threshold shrinking wavelet estimator with $2^{j_1} = o(n)$ is expected to perform better than smoothing-spline and threshold-wavelet estimators, and also better than non-threshold shrinking estimators based on (7) for which $2^{j_1} \asymp n$, when the estimated f has low regularity. In this aspect, its main advantage to standard smoothing splines is the simultaneous spatial and frequency localization; its main advantage to thresholded-wavelet estimators is that its influence matrix $\mathbf{H}$ does not depend on f or the sample $\mathbf{y}$. One of its main advantages to non-threshold estimators based on (7) with fixed $j_1 : 2^{j_1} \asymp n$ is essentially the same as with the classical smoothing-spline technique, namely, with GFCV j_1 is also a controlled parameter and is set at a lower value, with $2^{j_1} = o(n)$ (see also Remark 3 (b)).

Remark 5. In Theorem 2 we considered the case of fixed s. This was done mainly for simplicity of the exposition. In practice, one would consider minimizing the FCV (or GFCV) functional as a function of both parameters s and v, together with the discrete parameter j_0. For both the homogeneous and inhomogeneous model, if $f \in V_j$ for some $j \in \mathbf{Z}$, then for n large enough the expected cross-validated value for j_0 is $j_0 = j$ (see also Remark 7).

Finally, let us note that the restriction $s > 1/2$ in the analogues of our Theorem 2 for QGCV and WAVREG essentially eliminates discontinuous functions f from consideration, because curves f which have at least one jump point do not belong to B^s_{22} for any $s > 1/2$. In this aspect, the performance of WAVREG is additionally worsened by the necessity to estimate the noise variance. It should be noted that there is a common error in the following results: Antoniadis (1994), Lemma 3.1; Antoniadis (1996), Lemma 3.1, Theorem 3.1, Theorem 3.2 and Corollary 3.2; Amato and Vuza (1997a), Theorem 3; Amato and Vuza (1998), Theorem 1; in all these results it is assumed that $s > 1/4$, while the correct range for s in all of them is $s > 1/2$. The error consists in the assumption that the error of the quadrature formula admits the bound $|\beta_{jk} - E\hat{\beta}_{jk}| = O(2^{-sj})$ for any $f \in B^s_{22}$, for any $s > 0$, thus including also the range $0 < s \leq 1/2$. In fact, the correct range of s for which the above bound holds is $s > 1/2$ (see Donoho (1992), section 2.3, pp. 11-12, section 2.4, p.13, section 6.1, pp. 29-30). A counterexample is given in Ivanov (1985). For further discussion, see $^{B12(b)}$.

What about FCV and GFCV? Theorem 2 is clearly valid also for discontinuous functions. However, the other results in this subsection are valid under the constraint $f \in B^{s_1}_{\infty\infty}$, $s_1 > 0$, which also eliminates discontinuous functions. That is why it is important to note here that in all these theorems the assumption $f \in B^{s'}_{22} \cap B^{s_1}_{\infty\infty}$ can be relaxed by replacing it with $f \in A^{s'}_{22}$, where $A^{s'}_{pq}$ is the A-space of V. A. Popov (see $^{B12(b)}$). In this improvement of the results in this section, the condition

$$2^{j_1} = o(n^{\min(\frac{s_1}{\sigma}, \frac{1}{1+2\sigma})})$$

is replaced by

$$2^{j_1} = o(n^{\frac{2\min(s',1/2)}{1+2\sigma}}), \ \ 0 \le \sigma < s' \le s < r.$$

When $s > 1/p$, or $s = 1/p$ and $q \le \min(1,p)$, A^s_{pq} and B^s_{pq} are isomorphic, with equivalence of the (quasi-)norms (Ivanov (1985), Dechevski (1988)), while if $s < 1/p$ or $s = 1/p$ and $q > \min(1,p)$, the essential embedding $A^s_{pq} \hookrightarrow B^s_{pq}$ holds. For $0 < p \le \infty,\ 0 < q \le \infty,\ s > 0,\ s_1 > 0$, $B^s_{pq} \cap B^{s_1}_{\infty\infty} \subset A^s_{pq} \subset B^s_{pq}$ holds. All the elements of A^s_{pq} are bounded measurable functions defined pointwise everywhere on their domain, and for $s < 1/p$ or $s = 1/p$ and $q = \infty$, A^s_{pq} contains also discontinuous functions. In fact, all interesting fractal curves are contained in some A^s_{pq}-space with $s > 0$; for example, the fractal dimension of the function's graph is equal to $2 - s$, where s is the sharp smoothness index of the space $A^s_{1,\infty}$ to which the function belongs. As for the set $B^s_{pq} \setminus A^s_{pq}$, $s < 1/p$, $p < \infty$, it contains unbounded and 'monstrous' bounded functions (see Ivanov (1985) for details).

Summarizing, our method GFCV is very well adapted for estimation of continuous functions with fractal graphs, as well as discontinuous functions.

6.2. Density estimation.

The usual cross validation approach in density estimation is to replace $V_0(v)$ in (8) by the $ISE(v) = \int(\tilde{f}_v(x) - f(x))^2 dx = \int \tilde{f}^2_v(x)dx - 2\int \tilde{f}_v(x)f(x)dx + \int f^2(x)dx$. (For the sake of simpler and less spacious theoretical derivations, in this section the parameter s is considered to be fixed). In practice, one minimizes the empirical version of the ISE (up to a summand not depending on v), i.e.,

$$M(\tilde{f}_v) = \int \tilde{f}^2_v(x)dx - \frac{2}{n}\sum_{i=1}^{n} \tilde{f}^{(-i)}_v(X_i) =$$

$$\sum_k \hat{\alpha}^2_{j_0k} + \sum_{j=j_0}^{j_1}\sum_k \frac{\hat{\beta}^2_{jk}}{(1+t^2 2^{2js})^2} - \frac{2}{n}\sum_{i=1}^{n}[\sum_k \hat{\alpha}_{j_0k(-i)}\varphi_{j_0k}(X_i) + \sum_{j=j_0}^{j_1}\sum_k \hat{\beta}_{jk(-i)}\frac{\psi_{jk}(X_i)}{1+t^2 2^{2js}}].$$

After a simple manipulation, we get (up to summands not depending on v) the following expression to be minimized:

$$\tilde{M}(\tilde{f}_v) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{j=j_0}^{j_1}\sum_k \frac{\psi_{jk}(X_i)\psi_{jk}(X_{i'})}{(1+t^2 2^{2js})^2} - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i'=1, i'\ne i}^{n}\sum_{j=j_0}^{j_1}\sum_k \frac{\psi_{jk}(X_i)\psi_{jk}(X_{i'})}{1+t^2 2^{2js}}.$$

Theorem 5. *(Sufficient conditions for existence of the v-minimizer.) Let $0 < s' < s < r$. Assume that the penalized model is via $K_2(\sqrt{v}, f; L_2, \dot{B}^s_{22})$, $f \in B^{s'}_{22} \cap L_\infty$ and that $j_0 \le j_1$, $j_1 \to \infty$ and $2^{j_1} = o(n)$. If $j_0 = O(1)$ and $f \notin V_j$ for any $j < j_0$, or if $j_0 \to \infty$, $f \notin V_j$ for* *any $j \in \mathbf{Z}$ and the index s' is sharp, then, for sufficiently large n there exists $\tilde{v} : MISE(\tilde{v}) = \min_v MISE(v)$, with $0 < \tilde{v} < C$. Under the same conditions, if in the case $j_0 \to \infty$ additionally $2^{j_0} = O(n^{\frac{1}{1+2s}})$, $0 < s < r$, holds, and, in particular, for $0 < s \leq 1/2$ $2^{j_0} \leq 2^{j_1} \leq \mathrm{const}.2^{j_0}$ is additionally fulfilled, then, with probability tending to 1 as $n \to \infty$, there exists $v^* : M(\tilde{f}_{v^*}) = \min_v M(\tilde{f}_v)$, and $0 < v^* < C$ holds. Here $C < \infty$ depends on f and ψ only.*[A18]

Remark 2 is valid for $\tilde{f}_{\tilde{v}}$ and $\tilde{f}_{v^*}$ in the density case, too. However, it is not clear whether this remark is relevant when estimating compactly supported densities, because, although for any $j \in \mathbf{Z}$ there exist functions in V_j which are non-trivially non-negative on $\mathbf{R}$ (Walter and Shen (1997)), the functions in the examples available are not *compactly supported*.

Analogously to Theorem 2, we now show consistency of cross validation for the case of density, i.e., asymptotic equivalence of minimizing $M(\tilde{f}_v)$ and $MISE(v)$: $\frac{MISE(v^*)}{MISE(\tilde{v})} \overset{P}{\searrow} 1$, $n \to \infty$. Corollary 1 in Tribouley (1995) implies that this follows from the next

Theorem 6. *(Consistency of CV - cf. also Dechevsky and Penev (1988a).) Assume that $0 < s' \leq s < \infty$, the penalized model is via $K_2(\sqrt{v}, f; L_2, \dot{B}^s_{22})$, $f \in L_\infty \cap B^{s'}_{22}$ and that the compactly supported $\psi \in B^r_{\infty\infty}$, $r > s$ is orthogonal to algebraic polynomials P with deg $P \leq [r]$. Assume also that $j_0 \leq j_1$, $j_1 \to \infty$, $2^{j_1} = o(n)$, and $2^{-j_0} = O(n^{-\frac{1}{2(1+s')}})$. Then, for any $v \geq 0$,*

$$\frac{M(\tilde{f}_v) - MISE(v) + T_n}{MISE(v)} \underset{n\to\infty}{\overset{P}{\to}} 0,$$

where $T_n = \int f(x)^2 dx + \frac{2(n+1)}{n} \sum_k \hat{\alpha}_{j_0k}\alpha_{j_0k} - \frac{2(n+1)}{n} \sum_k \alpha^2_{j_0k}$ is a quantity that does not depend on v and $T_n \underset{n\to\infty}{\overset{P}{\to}} \int f(x)^2 dx$ holds.

Remark 6. The condition $f \in L_\infty$ in Theorem 6 is essential for $s \leq 1/2$, that is, for less regular and possibly discontinuous curves. For $s > 1/2$, the *Sobolev embedding* (see, e.g., Bergh and Löfström (1976)) ensures that $||f||_{L_\infty} \leq c_{s'}||f||_{B^{s'}_{22}}$, where $c_{s'}$ depends on s' only. Readers who utilize the reference source Tribouley (1995) should be cautioned that, for the same reasons, the additional constraint $f \in L_\infty$ must be imposed in subsection 2.3 there to validate the proof of the bound for $Var(T_{12})$ on p.50 when $0 < s < d/2$, d being the dimension (in our case, $d = 1$).

Theorem 7. *(Consistency of $\tilde{f}_{\tilde{v}}$ in B^σ_{22}, $0 \leq \sigma < s' \leq s < r$.) Assume that the conditions of Theorem 5 hold, without the sharpness condition on s'. Then, if $j_0 = O(1)$, and if $v = v_n$ is bounded, the condition $v \to 0$ as $n \to \infty$ is necessary for $E||\tilde{f}_v - f||^2_{B^\sigma_{22}} \to 0$ to hold; moreover, in both cases $j_0 = O(1)$ and $j_0 \to \infty$, if $2^{j_1} = o(n^{\frac{1}{1+2\sigma}})$ holds, then $v \to 0$ is also a sufficient condition for $E||\tilde{f}_v - f||^2_{B^\sigma_{22}} \to 0$. In particular, if s' is sharp, then: the above necessity claim is true for $v = \tilde{v}_n$; if $2^{j_1} = o(n^{\frac{1}{1+2\sigma}})$ is fulfilled, too, the above sufficiency claim is also true for $v = \tilde{v}_n$.*

Theorem 8. *(Asymptotic behaviour of $\tilde{v}$.) Under the premises of Theorem 5, $\tilde{v} \to 0$.*

Corollary 2. *(Consistency of $\tilde{f}_{v^*}$ in $B^0_{22} = L_2$ in probability). Under the premises of Theorem 5, $MISE(v^*) \xrightarrow{P} 0$ as $n \to \infty$ holds.*

The method of proof of Theorem 8 allows obtaining precise quantitative results, too: for any admissible choice of j_0 and j_1 the sharp asymptotic rates can be found for $\tilde{v} \to 0$. A complete analysis of the rates for $\tilde{v}$ for any admissible j_0 and j_1, the respective rates for $E||\tilde{f}_{\tilde{v}} - f||^2_{B^\sigma_{22}} \to 0$, and comparison of these rates to the asymptotically minimax-optimal ones, will be carried out elsewhere.[B14] Here we only include the following simple model corollary.

Corollary 3. *(Asymptotic-minimax rates in B^σ_{22}-norm, $0 \le \sigma < s' = s$.) Under the premises of Theorem 5, assume that $s = s'$ and $c.n^{\frac{1}{1+2s}} \le 2^{j_0} \le 2^{j_1} \le C.n^{\frac{1}{1+2s}}$. Then, $\tilde{f}_{\tilde{v}} = O(n^{-\frac{2s}{1+2s}})$ and $E \parallel \tilde{f}_{\tilde{v}} - f \parallel^2_{B^\sigma_{22}} = O(n^{-\frac{2(s-\sigma)}{1+2s}})$, i.e., for this choice of s, j_0 and j_1 $\tilde{f}_{\tilde{v}}$ achieves the asymptotic-minimax rate.*

Of course, Theorem 6 is not strong enough to imply, together with Corollary 3, that $\tilde{f}_{v^*}$ also achieves the asymptotic-minimax rate, but the choice of s, j_0 and j_1 in Corollary 3 is clearly of interest also for $\tilde{f}_{v^*}$.

Remark 7. The constraint on j_0 in Theorem 6 is a lower bound depending on s'. It means that, if j_0 is so selected that $(\tilde{C}n)^{\frac{1}{2(s_0+1)}} \le 2^{j_0} < 2(\tilde{C}n)^{\frac{1}{2(s_0+1)}}$, for some $s_0 : 0 < s_0 < r$, then the theorem's statement holds uniformly in $s' : s_0 \le s' < r$. Consideration of a bound s_0 can be avoided by simultaneously optimizing the CV-functional in v and s, hence, also in $j_0 = j_0(n, s)$. In practice, for moderate samples, the relation $j_0 = j_0(n, s)$ can be ignored and optimization can be in v, s, and j_0 as an independent discrete parameter. This is an analogue of our practical approach to the regression problem (see Remark 5), as well as of the practical cross-validation approach for estimating j_0 by Tribouley (1995).

Finally, comparing our results for the density and the regression case, we note that the condition $f \in B^{s'}_{22} \cap L_\infty$ in the theory of density estimation is less restrictive than its counterpart $f \in B^{s'}_{22} \cap B^{s_1}_{\infty\infty}$ (or the more general condition $f \in A^{s'}_{22}$) in the case of regression. The reason is that in the density case $\hat{\alpha}_{jk}$, $\hat{\beta}_{jk}$ are *unbiased* estimators of α_{jk}, β_{jk}, respectively. For the same reasons, the same relaxation of the assumptions on f occurs for the Bowman-Rudemo type approach to cross validation for regression problems with random design (see [B1], Dechevsky and Penev (1999), and also Delyon and Juditsky (1996)).

7. EXAMPLES AND EXTENSIONS

In this section we consider some model examples of estimation of a regression function or a density f, together with some first extensions of the penalized shrinking wavelet model (see also Appendix B). The comparison is between penalized shrinking and soft thresholding; numerical

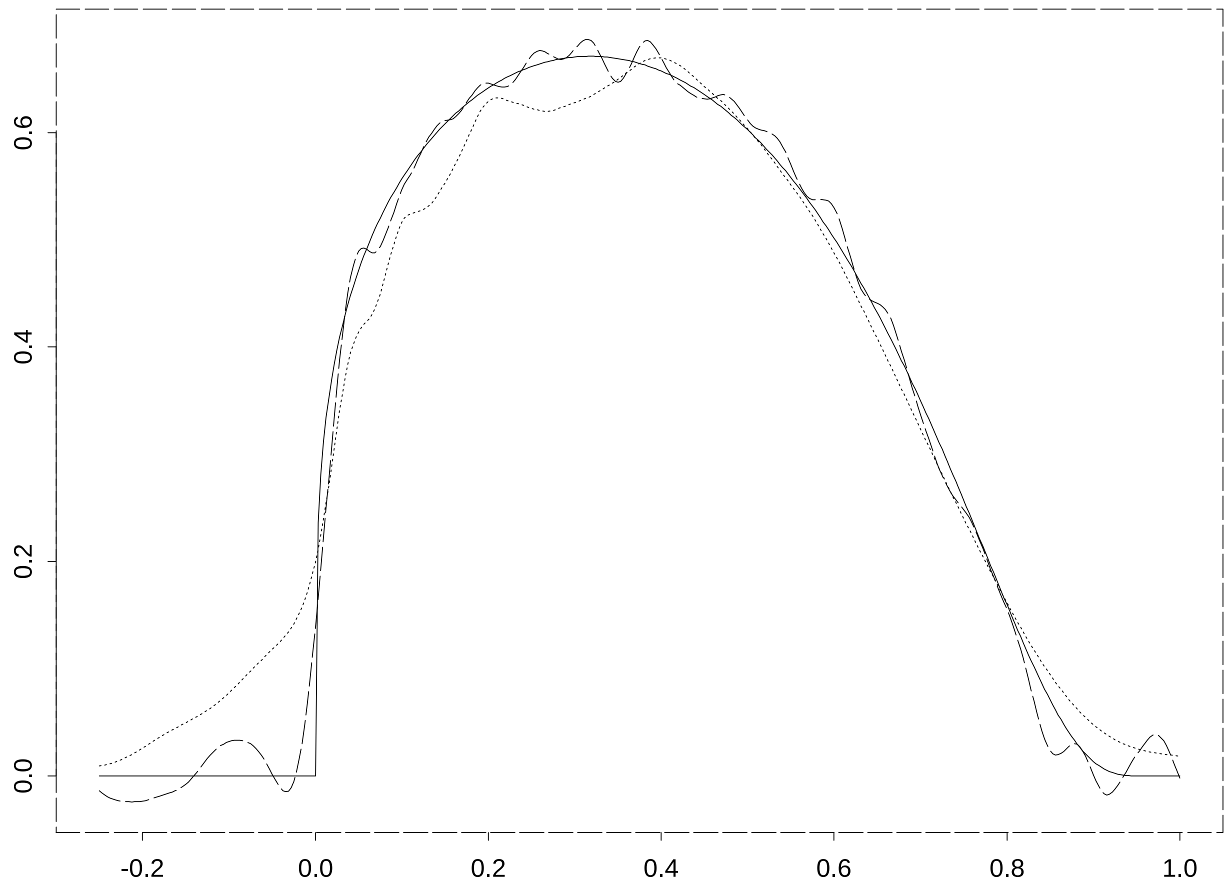

Figure 1: "λ-tear": $\lambda = 0.25$, $n = 1024$, $y_i = f[-\lambda + (1+\lambda)i/n] + \epsilon_i$, $\epsilon_i \sim N(0, 0.1^2)$. True curve: solid line. Estimators: global penalized shrinking: dashed line; soft thresholding: dotted line.

data will be added for hard thresholding in case that its performance turns out to be better than that of soft thresholding. To avoid introducing wavelets for a domain with boundary, f is always assumed defined on $\mathbf{R}$, with compact support. The combined requirement for orthogonality and compact support of the wavelet excludes the use of spline-wavelets; throughout Daubechies' extremal wavelets (Daubechies (1992), p. 195) of order M=6 or 8 are utilized, calculated to considerable accuracy. The loss is quadratic and the values of π, p, u, q, σ, s are as in Section 5. The GFCV-functional is being minimized simultaneously in v, s and j_0 (see Remarks 5, 7)^{A19}; $2^{j_1} \asymp \frac{n}{\ln n}$ (which, by way of Corollary 1, means that $\sigma = 0$, i.e., only the function is being consistently estimated, but not its fractional derivatives of any positive order).

Example 1. *The "λ-tear"* (Figure 1). $f(x) = x_+^\lambda \exp(-\frac{x^2}{1-x^2})$, $\lambda > 0$, where $x_+ = \max(x, 0)$, $\lambda = 0.25$. $f \in C^0(\mathbf{R}) \bigcap C^\infty(\mathbf{R} \setminus \{0\})$, supp $f = [0,1]$. For $1 \le p \le \infty$, $0 < q \le \infty$ and $\lambda + 1/p > 0$, $f \in B^s_{pq}$ if and only if $s < \lambda + 1/p$ or $s = \lambda + 1/p$, $q = \infty$ (Brenner, Thomée and Wahlbin (1975), Proposition 2.4.2).

Figure 1 displays clearly that around the critical point $x = 0$ the shrinking estimator shows better adaptivity. Its graph can be made less "wiggly" by selecting a smaller value for j_1, as indicated in Corollary 1. (In this case, $s_1 = \lambda = 1/4$, $s' = \lambda + 1/2 = 3/4$, and, by Corollary 1, a

Figure 2: Weierstrass curve: $\tau = 0.5$, $n = 1024$, $y_i = f(i/n) + \epsilon_i$, $\epsilon_i \sim N(0, 0.2^2)$. True curve: solid line. Estimators: global penalized shrinking: dashed line, estimated ISE $\approx 3.442 \times 10^{-2}$; soft thresholding: dotted line, estimated ISE $\approx 6.691 \times 10^{-2}$; hard thresholding: estimated ISE $\approx 4.013 \times 10^{-2}$. Data for f_1 - discontinuous Weierstrass-type curve with vertical displacement at 0.5: estimated ISE: 3.450×10^{-2} (global penalized shrinking); 6.683×10^{-2} (soft thresholding).

good selection for j_1 is, e.g., $j_1 : 2^{j_1} \asymp n^{\min(\frac{1}{4\sigma}, \frac{1}{1+2\sigma})} / \ln \ln n$, where $\sigma : 0 < \sigma < s' = 3/4$ is small enough for $2^{j_0} \leq 2^{j_1}$ to hold.) Fot further improvements, see [B1,B6,B8].

Example 2. *The Weierstrass curve* (Figure 2). $f(x) = \sum_{k=0}^{\infty} 1.5^{-\tau k} \sin(1.5^k \times 5x)$, $\tau > 0$, $x \in [0,1]$, where $\tau = 0.5$. As usual, we shall consider its restriction $\bar{f}$ on $[0,1]$ (see also Wang (1997)). The graph of f is a typical *self-similar monofractal*: it has constant local Hölder index τ and constant local fractal dimension $2 - \tau$ (which is also its global fractal dimension on $[0,1]$). For such functions the universal thresholding and global penalized shrinking are expected to be at their best. For any compactly supported $\chi \in C^\infty(\mathbf{R})$ such that $[0,1] \subset$ supp f and $\chi \equiv 1$ on $[0,1]$, $\chi.f \in B^s_{pq}(\mathbf{R})$, $1 \leq p \leq \infty$, $< q \leq \infty$, if and only if $s < \tau$ or $s = \tau$, $q = \infty$ (Brenner, Thomée and Wahlbin (1975), Proposition 2.4.1, the imaginary part of G_τ with additional rescaling). By the Whitney-type trace theorem for Besov spaces (see, e.g., Penev and Dechevsky (1997)), the same is true for $\bar{f}$ with respect to $B^s_{pq}([0,1])$.

The shrinking estimator on Figure 2 shows better adaptivity to abrupt changes in the curve's graph than the soft- (and hard-) thresholded. This is true for discontinuities, too (numerical data are included for $f_1(x) := f(x) + 0.4 \times (x - 0.5)_+^0$ in the caption to Figure 2).

Extension: kernel-regularization of the shrinking wavelet estimator. So far, we have been paying most attention to applications of the shrinking wavelet estimator in the case of non-smooth functions. What about very smooth functions? Recall that, for the choice $j_1 : 2^{j_1} \asymp n/\ln n$, $\tilde{f}_{\tilde{v}}$ in Section 6.1 and $\tilde{f}_{v^*}$ in section 6.2 have been shown to be consistent in B^σ_{22} for $\sigma = 0$ only, i.e., they estimate consistently the function, but not its derivatives of any positive order. Example 1 clearly illustrates this. Several ways can be suggested to upgrade the shrinking wavelet estimator to make it estimate derivatives consistently. The simplest solution suggested by Theorem 3, Corollary 1 and Theorem 7 is to decrease j_1: $2^{j_1} = o(n^{\frac{1}{1+2\sigma}})$ should hold, rather than the standard $2^{j_1} = O(\frac{n}{\ln n})$. Other possibilities will be discussed in [B1,B6,B8]. Here we consider an alternative based on kernel-regularization. Let $\Phi \in B^r_{\infty\infty}(\mathbf{R})$, supp$\Phi = \{x : |x| \le 1\}$, $\int \Phi(x)dx = 1$ and, optionally, $\Phi \ge 0$ on $\mathbf{R}$. Denote $\Phi_\varepsilon := \frac{1}{\varepsilon}\Phi(\frac{\cdot}{\varepsilon})$, $\varepsilon > 0$. Consider $\tilde{f}_{v,\varepsilon}(x) := (\Phi_\varepsilon * \tilde{f}_v)(x)$. Let v_{opt} be the v-minimizer with respect to any of the CV-functionals in Section 6. For the bandwidth ε we suggest the choice $\varepsilon = O(v_{opt}^{1/2})$. This topic is currently under investigation and results obtained so far indicate that $\tilde{f}_{v,\varepsilon}$ with $v = v_{opt}$ and $\varepsilon = O(v_{opt}^{1/2})$ is a very competitive alternative to thresholded wavelet estimators when f is smooth and sample sizes are moderate. For a typical graphical result we refer to Example 4 below. For numerical integration of the convolution integral we recommend Romberg integration. This quadrature method is excellently suited to dyadic wavelets and can in fact be considered as an extension of Mallat's recursive algorithm "on sub-coefficient level".

The study of consistency of $\tilde{f}_{v,\varepsilon}$ in Besov spaces is much facilitated by the equivalent definition of these spaces, considered, e.g., in Sickel (1990), Definition 1 (ii), which is valid for a broader range of p, q and s than the orthonormal-wavelet atomic decomposition of these spaces (see section 3).

Extension: vector parameters of optimization. There is an ongoing discussion in the literature about passing over from universal to *level-dependent*, *block* and even *individual* thresholding of the empirical wavelet coefficients. Level-dependent thresholding has been proposed by Donoho and Johnstone (1995) and Delyon and Juditsky (1996). The block-thresholding approach was proposed and developed by Hall, Kerkyacharian and Picard (1995) and its numerical performance was studied in Hall, Penev, Kerkyacharian and Picard (1997). Individual shrinking was discussed, e.g., in Härdle, Kerkyacharian, Picard and Tsybakov (1998). Other relevant development and discussion of these topics can be found in MacGibbon and von Sachs (1997), Johnstone and Silverman (1997), Cai (1998), Cai and Silverman (1998 b, c)). The aims of these enhancements of the thresholding approach are twofold: (a) achieving better spatial and frequency adaptivity while preserving the smoothing and denoising properties of the wavelet estimator; (b) studying regression problems with correlated-noise error structure. This discussion can be readily extended to penalized shrinking, too.[A20] In the context of penalized shrinking, the new enhanced strategies are equivalent to replacing t (and, eventually, s) in (7) by t_{jk} (s_{jk}), and then relaxing the global constraint $t_{jk} = const = t$ ($s_{jk} = const = s$), $j = j_0, \ldots, j_1$, $k \in \mathbf{Z}$, by imposing a less stringent condition on t_{jk} (s_{jk}). The resulting extended version of the K-functional is an equivalent quasi-(semi-) norm in the quasi-Banach sum of *closed subspaces* of two (or more) Besov spaces.[B12(a),B17] The resulting cross-validation functionals for both the regression and the density case are now to be optimized in a vector parameter $\mathbf{v}$ (in the optimization for the level-dependent case the parameter $\mathbf{s}$ is eliminated by a change of variable, rescaling $\mathbf{v}$).

Level-dependent smoothing parameters. Optimization of the cross-validation functional is in $\mathbf{v} = (v_{j_0}, \ldots, v_{j_1})'$, $v_j = t_j^2$. The expected improvement is better adaptivity in the *frequency* domain. Our first test was on the Weierstrass curve with white-noise error structure (Example 2), where the expected outcome for the optimal $\mathbf{v}^*$ and v^* from Example 2 was $v_j^* \approx const = v^*$, because of the global self-similarity of the curve. Indeed, the new fit virtually coincided with the old fit from Example 2, ISE($\mathbf{v}^*$) coinciding with ISE(v^*) up to the fourth decimal place. There was a slight change in favour of the level-dependent estimator when the white-noise model was replaced by fractional Brownian noise with correlation structure $\rho(k) = 0.5[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}]$, $k \geq 0$ and $\rho(k) = \rho(-k)$ for $k < 0$, and with standard deviation of 0.2. This type of noise was simulated with $H = 0.95$ by using the S-plus routine `simFGN0` in Chapter 12 of Beran (1994). In one of the typical numerical results obtained, ISE(v) was 4.633×10^{-2} and ISE($\mathbf{v}$) was 4.621×10^{-2}.

Block-penalized shrinking. The constraints on t_{jk} are as follows: if supp ψ_{jk} and supp $\psi_{j'k'}$ are contained in supp $\psi_{j_0k_0}$ for the *same* value of k_0, then $t_{jk} = t_{j'k'}$. The expected improvement is better *spatial* adaptivity - see the following

Example 3. *Double chirp* (Figure 3). $f(x) = \sqrt{x}\exp(-\frac{x^2}{1-x^2})\sin[64\pi x(1-x)]$. $f \in C^0(\mathbf{R}) \bigcap C^\infty(\mathbf{R} \setminus \{0\})$, $f^{(\nu)}(1) = 0$, $\nu \in \mathbf{N} \bigcup \{0\}$, $f(0) = 0$. The graph of f contains two single chirps (at 0 and 1) of very different nature. In a neighbourhood of 0 f is with unbounded variation; in a neighbourhood of 1 (not containing 0) f is absolutely continuous. Its Besov regularity is bounded from above by that of its "profile": the "λ-tear" for $\lambda = 0.50$ (cf. Example 1).[A21]

Iterative individual penalized shrinking. In this extreme case all t_{jk}, $j = j_0, \ldots, j_1$, $k \in \mathbf{Z}$, are independent parameters in which the cross-validation functional is being optimized. This extension is of interest for both the regression and density model. Here we shall consider density estimation. Then, optimization of $\tilde{M}(\tilde{f}_{\mathbf{v}})$, $\mathbf{v} = \{v_{jk}, j = j_0, \ldots, j_1, k \in \mathbf{Z}\}$, in $v_{jk} = t_{jk}^2$ yields

$$\frac{[\sum_{i=1}^n \psi_{jk}(X_i)]^2}{n^2(1 + v_{jk}2^{2js})^3} = \frac{[(\sum_{i=1}^n \psi_{jk}(X_i))^2 - \sum_{i=1}^n \psi_{jk}(X_i)^2]}{n(n-1)(1 + v_{jk}2^{2js})^2}.$$

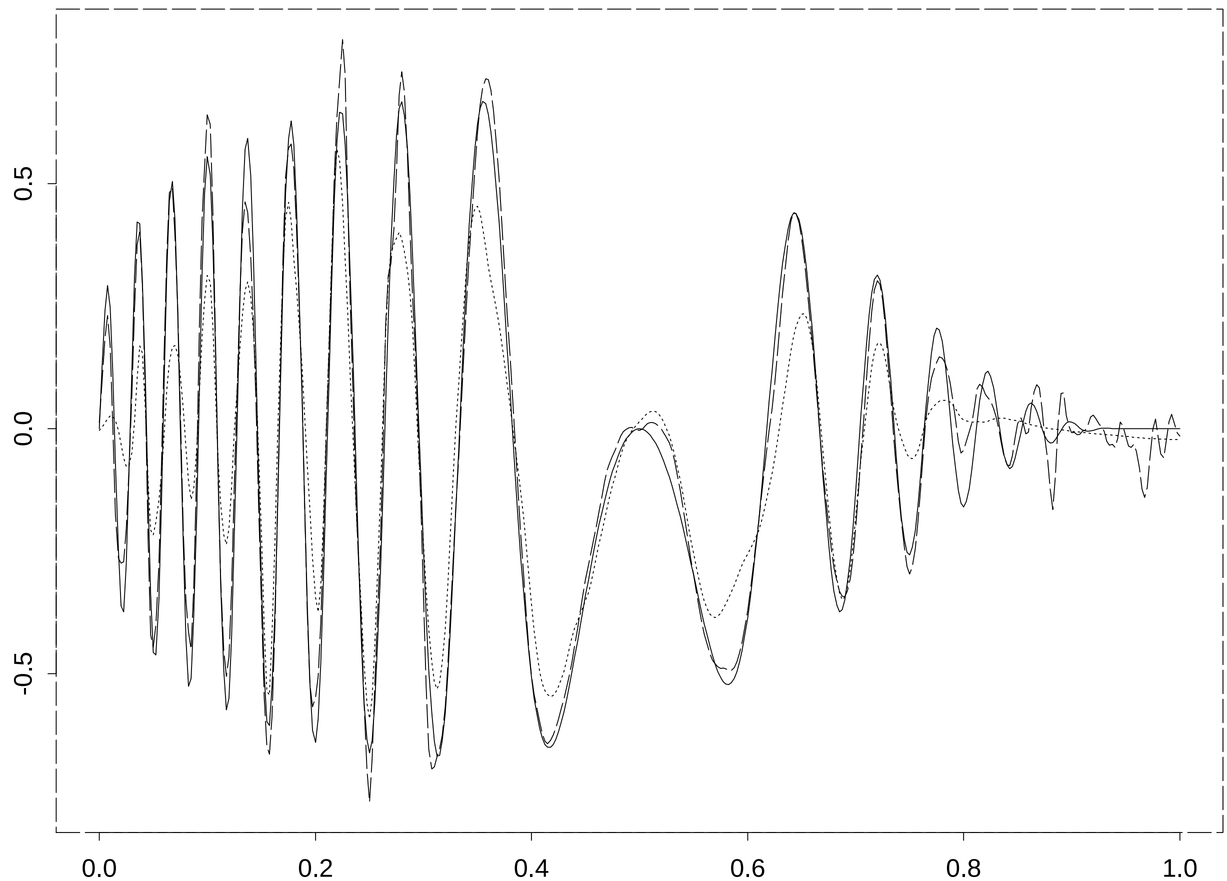

Figure 3: Double chirp: $n = 1024$, $y_i = f(i/n) + \epsilon_i$, $\epsilon_i \sim N(0, 0.2^2)$. True curve: solid line. Estimators: block penalized shrinking: dashed line; soft thresholding: dotted line.

Hence, the following analogue of (7) is obtained:

$$\tilde{\beta}_{jk} = \frac{\hat{\beta}_{jk}}{1 + v_{jk}.2^{2js}} = \hat{\beta}_{jk}.[1 - \frac{1}{n-1}\frac{\overline{\psi^2_{jk}} - (\overline{\psi_{jk}})^2}{(\overline{\psi_{jk}})^2}],$$

where $\overline{\psi^2_{jk}} = \frac{1}{n}\sum_{i=1}^n \psi^2_{jk}(X_i)$, $\overline{\psi_{jk}} = \frac{1}{n}\sum_{i=1}^n \psi_{jk}(X_i)$. Now the estimator (1') is very undersmoothed. To smooth it more, we reapply the individual penalization with $\hat{\beta}^{(1)}_{jk} = \tilde{\beta}_{jk}$. Repeating this procedure leads to ever smoother, highly adaptive estimators. Performing $O(n)$ iterations has approximately the same effect for large n as the following *individually shrinking* rule:

$$\tilde{\beta}_{jk} = \hat{\beta}_{jk} C_1(s, \mathbf{v}) \exp\left(-C_2(s, \mathbf{v})\frac{\overline{\psi^2_{jk}} - (\overline{\psi_{jk}})^2}{(\overline{\psi_{jk}})^2}\right) \tag{15}$$

where the parameters $C_1 : 0 < C_1 < \infty$ and $C_2 : 0 < C_2 < \infty$ are explicit functions of s and v_{jk}, $j = j_0, \ldots, j_1$, $k \in \mathbf{Z}$. Substituting the quantities $\frac{1}{1+v_{jk}2^{2js}} = \frac{\tilde{\beta}_{jk}}{\hat{\beta}_{jk}} = C_1 \exp\left(-C_2\frac{\overline{\psi^2_{jk}} - (\overline{\psi_{jk}})^2}{(\overline{\psi_{jk}})^2}\right)$

back in the expression for $\tilde{M}(\tilde{f}_{\mathbf{v}})$ yields a cross-validation functional of the form:

$$\sum_{j=j_0}^{j_1} \sum_k \{(\overline{\psi}_{jk})^2 C_1^2 \exp(-2C_2 \frac{\overline{\psi_{jk}^2} - (\overline{\psi_{jk}})^2}{(\overline{\psi_{jk}})^2}) - \frac{2}{n-1}[n(\overline{\psi_{jk}})^2 - \overline{\psi_{jk}^2}] C_1 \exp(-C_2 \frac{\overline{\psi_{jk}^2} - (\overline{\psi_{jk}})^2}{(\overline{\psi_{jk}})^2})\}$$

which is to be minimized with respect to *two unknowns only*: C_1 and C_2, with expected asymptotic behaviour $C_1 \to 1$, $C_2 \to 0$. Hence, the strategy of choosing individual shrinkers is reduced to a minimization problem with respect to two variables only and can be tackled easily.

Formula (15) leads to an extremely adaptive estimator. Since it is generally still undersmoothed, additional smoothing is recommended, e.g., by convolving it with a smooth density kernel. The estimator is so adaptive that even the use of a simple constant-bandwidth kernel suffices to produce fits which are superior to the fits by standard thresholding techniques (for a typical example, see Figure 4). The iterative approach is currently under investigation, both with respect to kernel smoothing[B3,B4] and with respect to composite shrinking/thresholding estimation[B6,B8], and we hope to report some more detailed results soon. So far, we only note that the (constant) kernel bandwidth must be $O(\sqrt{v^*})$, $n \to \infty$, where $v^* = v_n^*$ is the minimizer from Theorem 5.

Another interesting opportunity here, in particular, in the case of Gaussian white noise, is to consider, levelwise and blockwise within each level, a composite shrinking/thresholding strategy by SURE (cf. [B8] (the second approach)).

Figure 4 shows that the new estimator adapts better to the spatial irregularities of f along the curve, including the discontinuity points $x = -2\pi/3$ and $x = \pi/3$ where the Gibbs effect is much smaller than with soft thresholding.

Example 4. Sinusoidal density (Figure 4). $f(x) = \begin{cases} \frac{1}{2}|\sin x| & \text{for } x \in [-2\pi/3, \pi/3], \\ 0 & \text{elsewhere.} \end{cases}$

Extension: self-similar fractal estimator. If no additional information about self-similarity of f is available, between the α- and β-coefficients only the relations from Mallat's recursive algorithm hold. For the wavelet estimators of f, $j_1 < \infty$ holds, which essentially means that f is being only perceived as a linear combination of wavelets and all information about the fractal structure of the graph of f comes from the fractal structure of the wavelet itself. Under these general assumptions, there are no forced relationships between the β-coefficients on neighbouring levels. Assume now that f is dyadically self-similar, i.e., it satisfies a dyadic functional equation

$$f(x) = \sum_{k=-L}^{L} c_k f(2x-k), \ x \in \mathbf{R}, \tag{16}$$

where $L \in \mathbf{N}$. Then f is compactly supported at $[-L, L]$ (see, e.g., Daubechies (1992), Dechevsky and Dubuc (1998), for details about the implied properties of f) and there exists additional relationship between the betas on every two neighbouring levels. This relationship is given by one and the same linear combination of the betas on the finer level, having for coefficients the c_k's

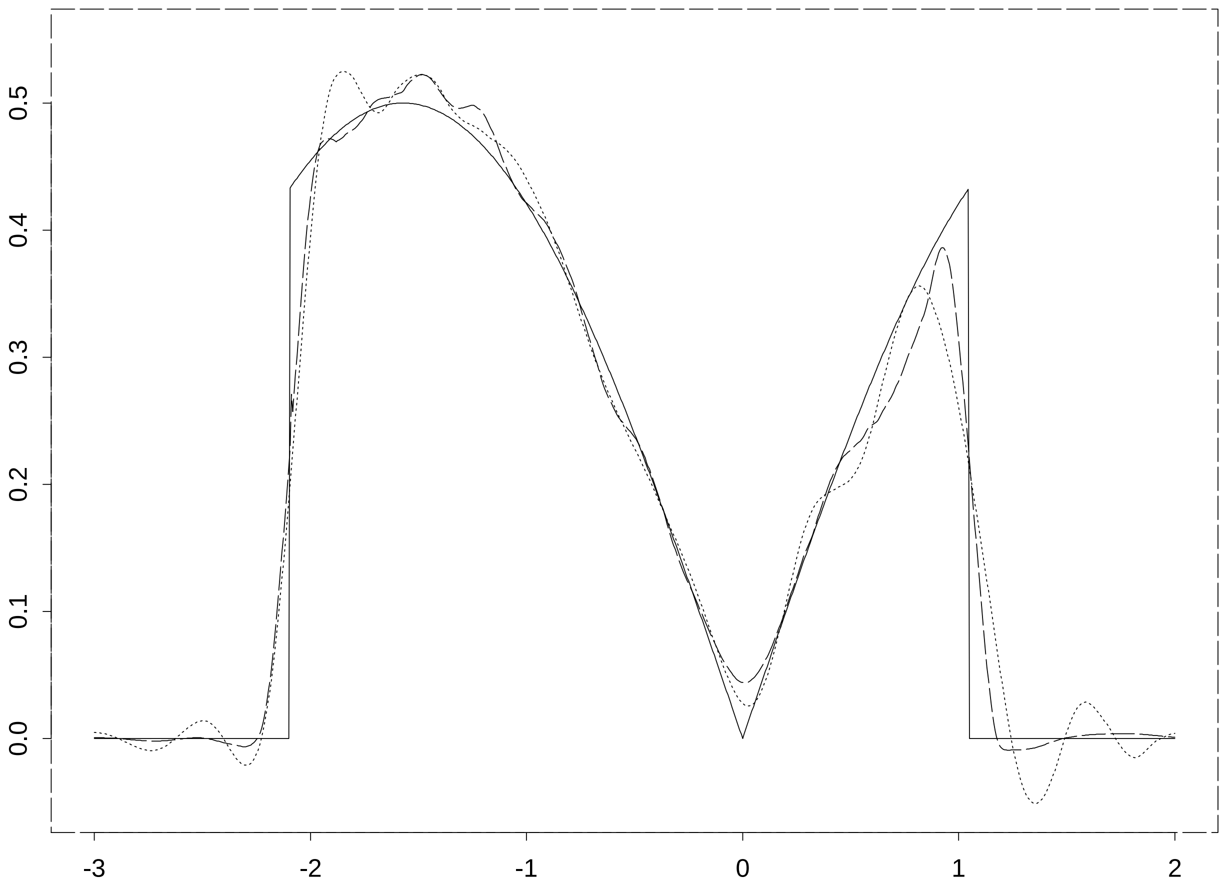

Figure 4: Sinusoidal density: $n = 2000$. True curve: solid line. Estimators: iterative individual penalized shrinking: dashed line; soft thresholding: dotted line. Kernel smoothing was by convolving with $h_\epsilon(x) = \frac{1}{\epsilon}h(\frac{x}{\epsilon})$, $h(x) = K.\exp[-(\frac{x^2}{1-x^2})_+]$, where $K = 1/1.2069$ is a norming constant. The kernel bandwidth ϵ was 0.2. The approximate computation of the convolution integral was via the trapezoidal rule.

from (16). If f is being estimated in the non-parametric regression or density model, we see that for fixed v the problem about finding the optimal $\tilde{\alpha}$- and $\tilde{\beta}$-coefficients of $\tilde{f}_v$ is now upgraded to an optimization problem with restrictions of type equality to which Lagrange-multiplier technique can be applied. The unknowns now are the $\tilde{\alpha}$'s and $\tilde{\beta}$'s and, additionally, the c_k's from (16). The equality constraints are *bilinear* - they are linear with respect to two groups of unknowns in separate: the coefficients of $\tilde{f}_v$ and the coefficients of the functional equation. This means that the optimization problem can be reduced to *convex*. It can be solved, e.g., via the *iterative quadratic penalty method* (Bertsekas (1982), Section 2.1). This is a simple method, but the resulting problem about unconstrained multivariate optimization, solved in each iteration, may become very ill-conditioned. To avoid this, the *iterative Lagrange-multiplier method* (Bertsekas (1982), Section 2.2) can be utilized. At every iteration it leads to an unconstrained optimization problem which is well-conditioned and, besides, its iterations converge superlinearly, much faster than with the quadratic-penalty method (Bertsekas (1982), Subsection 2.2.5). More recently developed advanced superlinear iterative methods (of quadratic, of Newton type, etc.) can be found in literature from the last few years (see, e.g., Boggs and Tolle (1995)).

Since the optimization problem is (can be reduced to) convex, any vector of c_k's is an admissible initial solution for starting the iterative process. To reduce the number of iterations, we suggest to obtain the initial solution in the following way: the initial $\tilde{\alpha}$- and $\tilde{\beta}$-coefficients are computed by (7), ignoring the information about self-similarity. The smoothing parameter is obtained as earlier, by cross validation. With the so-obtained values of the smoothing parameter and the respective $\tilde{\alpha}$'s and $\tilde{\beta}$'s, generating the initial values of the c_k's is by *linear regression* (i.e., least-squares method) based on the equality constraints.

Although the class of integrable functions f that satisfy (16) is relatively small compared to, say, the whole of L_1, it contains functions which are interesting with respect to both applications and theory. For instance, such are all Daubechies wavelets as well as orthogonal spline wavelets (for the latter, $L = \infty$). Besides, our approach can be extended to all f which are solutions of *non-stationary dyadic refinement schemes* and form a much larger functional class. Here the c_k's depend on the level j involved in the computation. A larger system of equations will result in this way, but the reward is that we can thus treat fractals with much more general structure. Having estimated the c_k's means that we can construct a *fractal estimator* of f via the respective refinement scheme. In the stationary case (c_k independent of j), this estimator is *globally self-similar*, like, say, the Weierstrass function. In the non-stationary case it is *self-similar locally in the frequency domain.* When the observations are noisy, the additional information about certain - global or local - self-similarity of f results in an estimator whose graph is a fractal as irregular as the noise itself, and yet it can be claimed that it is denoised to a considerable extent. Moreover, since self-similarity is a very spatially informative feature, it can be hoped that the fractal estimator will have good performance already for moderate samples, in particular, in the symmetric case, i.e., when it is known additionally that $c_k = c_{-k}$ for all k.

For Gaussian and Poissonian white noise there are already publications discussing estimation of the local Hölder index and more detailed multifractal characteristics (see Arneodo, Bacry and Muzy (1998), Istas and Lang (1997), Benassi, Cohen and Istas (1998)). Under assumption of dyadic self-similarity and Gaussian white noise, the above suggested fractal estimator can now be upgraded to produce good fits already for small to moderate samples. The idea is to consider a penalized version of the NeighBlock and NeighCoeff estimators of Cai and Silverman (1998b) which takes into account dyadic self-similarity (with $L_0 = 1$ in the notations of Cai and Silverman). Assume first that the support of the dyadically self-similar function, that is, the value of L in (16) is known. (This is usually the case in related applications.) Then, the oracle inequality of NeighBlock is no longer needed; the length of each block, on every level, is $d = 2L + 1$; the shrinking estimate is by the classical James-Stein estimator (JS)

$$\tilde{\beta}_{jk} = (1 - \frac{d-2}{\sum_{\nu=-L}^{L} \hat{\beta}^2_{j,k-\nu}})_+ \hat{\beta}_{jk},$$

for any $j = j_0, ..., j_1$, and any $k : \hat{\beta}_{jk} \neq 0$, assuming that the white noise is $N(0,1)$.

We observe that JS can be obtained by a penalized model in the following way.

1) For any $\theta \geq 0$, minimize

$$\Phi(\tau_{-L}, \ldots, \tau_0, \ldots, \tau_L) = \sum_{l=-L}^{L} (\hat{\beta}_{j,k-l} - \tau_l)^2 + 2\theta \frac{\sum_{l=-L}^{L} |\hat{\beta}_{j,k-l}||\tau_l|}{\sum_{\nu=-L}^{L} \hat{\beta}_{j,k-\nu}^2}$$

with respect to τ_l, $l = -L, \ldots, 0, \ldots, L$. The solution is

$$\tau_l = (1 - \frac{\theta}{\sum_{\nu=-L}^{L} \hat{\beta}_{j,k-\nu}^2})_+ \hat{\beta}_{j,k-l}.$$

2) Select optimal θ by SURE, applied to the τ_l's as functions of the $\hat{\beta}_{j,k-\nu}$'s. The result is $\theta = d - 2 = 2L - 1$.

3) Set $\tilde{\beta}_{jk} = \tau_0$; ignore the other τ_l: continue in the same way for the next block, corresponding to $\tilde{\beta}_{j,k+1}$, and so on.

Now we shall demonstrate how the penalized model interpretation of JS allows to upgrade the estimator, so that the new version takes dyadic self-similarity into account.

Consider the problem of minimizing the function

$$\sum_{j=j_0}^{j_1-1} \sum_{k} \{(\hat{\beta}_{jk} - \tilde{\beta}_{jk})^2 + 2(d-2) \frac{|\hat{\beta}_{jk}|.|\tau_{jk,0}|}{\sum_{\nu=-L}^{L} \hat{\beta}_{j,k-\nu}^2} +$$

$$+ \sum_{l=-L, l\neq 0}^{L} [(\hat{\beta}_{j,k-l} - \tau_{jk,l})^2 + 2(d-2) \frac{|\hat{\beta}_{j,k-l}|.|\tau_{jk,l}|}{\sum_{\nu=-L}^{L} \hat{\beta}_{j,k-\nu}^2}]\},$$

with respect to the variables

$$\tau_{jk,l},\ l = -L, \ldots, -1, 1, \ldots, L,\ \tilde{\beta}_{jk},\ j = j_0, \ldots, j_1 - 1,\ k : \hat{\beta}_{jk} \neq 0,\ c_\nu,\ \nu = -L, \ldots, 0, \ldots, L,$$

with the following constraints of type equality

$$\tilde{\beta}_{jk} = \frac{1}{\sqrt{2}} \sum_{\nu=-L}^{L} c_\nu \tilde{\beta}_{j+1,k+\nu},\ k : \hat{\beta}_{jk} \neq 0,\ j = j_0, \ldots, j_1 - 1,\ \tilde{\beta}_{j_1 k} = (1 - \frac{d-2}{\sum_{l=-L}^{L} \hat{\beta}_{j_1,k-l}^2})_+ \hat{\beta}_{j_1 k},$$

i.e., the finest level is still estimated by the (unconstrained) JS.

If additional information about the relative importance of the different coefficients is available, a w_{jk}-weighted version of the cost functional can be considered.

The equality constraints come from (16), by argument similar to the one used in Mallat's reconstruction recursive algorithm. If these constraints are disabled, that is, if (16) is not fulfilled and, hence, there is no global dyadic self-similarity, it can be seen that the groups of variables $\{\tilde{\beta}_{jk}, \tau_{jk,l}\}_{l=-L}^{L}$ (each one corresponding to the respective sliding block at position (j,k)) become disconnected from each other, which leads to the classical James-Stein estimator for every (j,k).

The problem can be solved numerically by a superlinear iterative method for nonsmooth convex optimization based on Clarke's subdifferential. The same is true if, for example, additional information about nonnegativity of f in (16) is available. In this case, additional constraints $c_l \geq 0$, of type inequality, are added to the optimization problem, which remains convex.

Further refinements of this model can be obtained by replacing the "hard" value $d-2$ in the functional by a θ_{jk} which varies with n in a small neighbourhood of $d-2$, and is reset by SURE-type optimization on every iteration of the optimization algorithm. However, developing a rigorous theory requires a much more detailed and elaborate consideration than is possible here.

If d is not known a priori, one good compromise between performance and simplicity of estimation of d is to apply a *global* oracle inequality, obtained by summing up in j the LHS and the RHS, respectively, of the levelwise oracle inequality (11) in Cai (1998), and thus select the same block size simultaneously for all levels $j = j_0, \ldots, j_1$.

If the white noise is not Gaussian, this model can still be helpful for large samples, when the central limit theorem takes over. For moderate samples, however, the initially proposed variant via cross validation is expected to yield better results, especially if the original noise distribution is far from Gaussian.

The requirement that the functional equation be dyadic is not essential. For example, one can consider triadically self-similar fractals, but then the wavelets φ and ψ must also be triadic.

ACKNOWLEDGEMENT

This research has been supported by the Natural Sciences and Engineering Research Council of Canada.

The first author would like to thank Spiridon Penev (UNSW, Sydney, Australia) for many valuable discussions and for informing us about some important relevant references which have been later included in the text.

The numerical and graphical results in this paper have been taken from the data files of Dechevsky and Penev (1998a).

We also wish to thank Brenda MacGibbon (UQAM and GERAD) for her support and encouragement.

APPENDIX A: PROOFS^{A0} AND DETAILS

A0. In all proofs, wherever this is relevant, we assume, with no loss of generality, that $a_1 = 0,\ a_2 = 1$, in order to shorten the mathematical expressions.

A1. Fractional-order derivatives of f are (a) for the inhomogeneous model (see Sections 3, 4): in the sense of the *Bessel potential* of f, i.e., convolution of f with the Bessel-MacDonald kernel (see Samko, Kilbas and Marichev (1987), Section 27); (b) for the homogeneous model (see Sections 3, 4): in the sense of the *Riesz potential* of f, i.e., convolution of f with the kernel in Samko, Kilbas and Marichev (1987), Section 25, formula (25.25). Since we shall be estimating only functions f which are within the range of equivalence of the two models [A14], we can consider the Bessel-MacDonald kernel only.

A2. The reason for this dramatic simplification is that atomic decomposition in Besov spaces (see Feichtinger and Gröchenig (1989), Frazier and Jawerth (1990)) leads to equivalent *sequence* norms (or semi-norms); in particular, when the atoms are orthonormal compactly supported wavelets (Sickel (1990)), the resulting sequences are comprised of the respective wavelet coefficients. In contrast to this, atomic decomposition in Triebel-Lizorkin spaces (including the Sobolev spaces associated with the smoothing spline technique) gives rise to norms that are not sequence norms except for those values of the indices for which the Besov and Triebel-Lizorkin space scales coincide.

A3. Here we single out only one of them: estimating functions whose graphs are self-similar fractals. Based on our approach, a new *self-similar fractal estimator* of noisy fractal curves can be constructed. Since it takes into account self-similarity, satisfactory estimation of simple fractal characteristics like, say, fractal dimension (Wang (1997)) is expected to be achieved already for *moderate* samples. Moreover, considering a non-stationary, locally self-similar fractal estimator allows subsequent estimation of much more informative fractal characteristics (e.g., Hölder spectrum, multifractal characteristics (Arneodo, Bacry, Jaffard and Muzy (1997), Meyer (1998)), discontinuity points, etc.).

A4. Indeed, *for small and moderate samples*, the minimax estimators corresponding to two different metrics can be so far away from each other that any admissible estimator which is "near both of them" would necessarily have unsatisfactory performance with respect to *both* metrics!

A5. (1-3) can be generalized for $B^s_{pq}(\mathbf{R}^d)$, with preservation of the resolution-level expansion via a *tensor-product* d-variate wavelet basis (see Delyon and Juditsky (1996) for a short overview; see also Daubechies (1992)). The range of the scale parameters is $0 < p \leq \infty$, $0 < q \leq \infty$, $\max(0, d(1/p-1)) < s < r$, which is exactly the range for which Besov spaces contain *only regular distributions*, i.e., *only locally integrable functions*.

A6. (For two different ρ-values the respective K_ρ's are equivalent but the extra parameter ρ provides useful additional flexibility.) For all t and all ρ $K_\rho(t, .; A, B)$ is an equivalent (quasi-semi-)norm on $A + B$, the standard norm in this space being $K_1(1, .; A, B)$ (cf. Bergh and Löfström (1976)). Under very general assumptions the infimum in (4) is attained for a unique pair $a^* \in A$, $b^* \in B$, $a^* = \alpha - b^*$, for any $\alpha \in A + B$. (For example, this is true if A and B are *uniformly convex* Banach spaces.)

A7. Note that the second term on the right (sometimes known in approximation theory as a *saturation* term) is always subordinate to $K_\rho(t, .; A, \dot{B})$ which is typically $c.t^\theta$ for some $\theta \in [0, 1]$.

A8. In fact, Delyon and Juditsky additionally require compact support of the estimated density or regression function. For such functions we can show that the range of admissible (π, u, σ) and (p, q, s) in the K-functional model is considerably *larger*.

A9. Then, the constraint on τ via u and q can be removed. For $\pi \neq p$ this is possible only if the model be refined by considering *four-indexed* Besov spaces.

A10. A comparison with the spline-smoothing model shows that the latter is essentially limited to the Hilbert range only ($\pi = p = u = q = 2$). Outside this range the resulting Euler-Lagrange equations are *nonlinear differential* equations. The parametrization provided by the wavelet model makes it possible to compute the coefficients of the minimizing wavelet in an essentially *explicit* way for the *whole quasi-Banach range*.[B9] But even for $\pi = p = u = q = 2$ the wavelet estimator is more adaptive: not only t, but also σ, s and j_0 (see (1)) can be optimized. (In principle, optimization in σ and s can be carried out for Wahba's spline model, too, but the problem is nonparametric - the *spectral theorem* has to be invoked. Then σ and s have the meaning of *fractional powers of a self-adjoint differential operator*. In practice, this means that a large number of eigenvalues and eigenfunctions have to be computed to a considerable precision, which can be a very ill-conditioned and numerically intensive problem.)

A11. The equivalence constants between the sequence norm (2) and any equivalent norm in the Besov space depend on j_0. This is important, because in some of our applications $j_0 \to \infty$ as $n \to \infty$. This requires that resolution levels between, say, $j = 0$, and $j = j_0$ be paid attention, too. However, the coefficients in these levels do not depend on the smoothing parameter, and thus these additional considerations only lead to additive rescaling of the relevant quantities and do not have essential impact on the statistical applications considered.

A12. These heuristic considerations are *pointwise*; this means that, in the context of Delyon and Juditsky's results, we must take $p = 2$, $\pi = \infty$. This brings about the constraint $s > 1/2$ (or $s > d/2$ in the d-dimensional case).

Proof of Lemma 1. (See also Dechevski (1989), Dechevsky (1993, 1993a).) Follows directly from the fact that in order to obtain the value of the K_2-functional, one needs to minimize the following expression:

$$\sum_{k\in Z}(\hat{\alpha}_{j_0k} - \alpha_{j_0k})^2 + \sum_{j=j_0}^{j_1}\sum_{k\in Z}(\hat{\beta}_{jk} - \beta_{jk})^2 + t^2\sum_{j=j_0}^{j_1}\sum_{k\in Z}2^{2js}\beta_{jk}^2 + (1+t^2)\sum_{j>j_1}\sum_{k\in Z}2^{2js}\beta_{jk}^2$$

with respect to $\alpha_{j_0k}, k \in Z, \beta_{jk}, j \geq j_0, k \in Z$. $\square$

A13. For the case when $A \cap B$ may not be dense in A and/or B, see Dechevsky (1993a), Lemma 2.1. For the case when A and/or B are semi-Hilbert spaces, see Dechevsky (1993a), Theorem 3.1.1. For the case when the cardinality of A is less than that of B, see Dechevsky (1993a), Theorem 2.1a (for Hilbert spaces), Theorem 3.1.1a (for semi-Hilbert spaces), and Remark 2.3. However, the generalizations with non-dense $A \cap B$ are not needed for the purposes of Tikhonov regularization (see Section 4).

A14. By (5), the homogeneous and inhomogeneous models are equivalent on $A \bigcap B = A \bigcap \dot{B} = B^s_{22}$. For functions f with *heavy tails*, such that $f \in \dot{B}^s_{22} \setminus B^s_{22}$, the inhomogeneous model is more precise (see also Dechevsky and Penev (1998), Remark 2.2.4.). In the sequel of this paper, compactly supported f will be considered only, so we shall be within the range of equivalence of the two models. We have chosen the homogeneous model because results about it are easier to directly compare to corresponding results for thresholding techniques, since in the homogeneous model, like with standard thresholding techniques, the empirical α's are left unchanged (see (7)).

A15. Asymptotic theory is beyond the scope of this paper,[B2] but we note that for $\pi = p$ linear estimators do achieve the minimax rates (see Delyon and Juditsky (1996), Theorems 3 and 4, case $\epsilon > 0$, and the sequence of two papers Dechevsky and Penev (1997,1998)). Moreover, it is possible that our estimators may achieve the asymptotic-minimax (within the "asymptopia") rates for functions in Besov spaces with $p = q = 2$, $2 \leq \pi \leq \infty$, $0 < u \leq \infty$, and also when $\pi \neq p$ and π, p vary in a *neighbourhood of* 2. (A (very weak) necessary condition for this is that the estimator be *nonlinear*. In our case this condition is fulfilled: the minimizer of the smoothing parameter t with respect to cross validation (see Section 6) is nonlinear, hence the estimator (1'), with t estimated via cross validation, is also nonlinear.)

A16. If $f \notin V_j$ for any $j \in \mathbf{Z}$ and $f \in B^{s'}_{22}$, but $f \notin B^s_{2\infty}$ for any $s > s'$, then for any fixed $s > s'$ there exists $c_0 = c_0(f, \psi, s - s') > 0$ with $\lim_{s \to s'+} c_0 = 0+$, such that for any N there exists $(j,k) \in \mathbf{Z}^2$ with $j \geq N$ so that $\beta_{jk} \neq 0$ and $|\beta_{jk}| \geq c_0 2^{js}$ hold. Here is an outline of the proof of this claim. Assume that its statement is not true. If for some N there is no (j,k) with $j \geq N$ and $\beta_{jk} \neq 0$, this would contradict to $f \notin V_j$ for any $j \in \mathbf{Z}$. If for any $c_0 > 0$ there exists N such that for any (j,k) with $j \geq N$ $|\beta_{jk}| < c_0 2^{js}$ holds, then it can be shown that $\| f \|_{B^s_{2\infty}} < \infty$, which is in contradiction with $f \notin B^s_{2\infty}$. Finally, if there exists a decreasing sequence $\{s_n\}$ such that $c_0(f, \psi, s_n - s')$ remains bounded away from zero as $n \to \infty$, then it can be proved that there exists $c_0(f, \psi, 0) > 0$ so that for any N there exists (j,k) with $j \geq N$ and with $|\beta_{jk}| \geq c_0(f, \psi, 0) 2^{-js'}$. But then $f \notin B^{s'}_{22}$, which again is a contradiction with the claim's assumptions on f. This completes the proof of the claim.

Proof of Theorem 1. We give the proof for v^*. The statement about $\tilde{v}$ can be obtained analogously, or can be derived from the asymptotic identity $E(GFCV(v)) \sim ER(v) + \delta^2$ (see Theorem 2). Under the theorem's assumptions, it can be shown that for any $v \in [0, \infty)$ $(\frac{d}{dv} ER)(v) \sim w'(v)$, where

$$w'(v) = 2 \sum_{j=j_0}^{j_1} \frac{2^{2js}}{(1 + v 2^{2js})^3} \sum_k [(v 2^{2js} + 1 - \frac{1}{n} \sum_{i=1}^n \psi_{jk}(x_i)^2) \overline{\beta}^2_{jk} - \frac{\delta^2}{n} (\frac{1}{n} \sum_{i=1}^n \psi_{jk}(x_i)^2)^2] \quad (A1),$$

as $n \to \infty$. Here, $E\hat{\alpha}_{j_0 k} = \overline{\alpha}_{j_0 k} = \frac{1}{n} \sum_{i=1}^n f(x_i) \varphi_{j_0 k}(x_i)$, $E\hat{\beta}_{jk} = \overline{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^n f(x_i) \psi_{jk}(x_i)$, $E\hat{\alpha}^2_{j_0 k} = \overline{\alpha}^2_{j_0 k} + \frac{\delta^2}{n} . \frac{1}{n} \sum_{i=1}^n \varphi_{jk}(x_i)^2$, $E\hat{\beta}^2_{jk} = \overline{\beta}^2_{jk} + \frac{\delta^2}{n} . \frac{1}{n} \sum_{i=1}^n \psi_{jk}(x_i)^2$. We outline the basic facts in this proof, as follows. (i) it holds

$$|\overline{\beta}_{jk} - \beta_{jk}| \leq \| f \|_{B^{s_1}_{\infty\infty}} 2^{-j/2} n^{-s_1} (\| \psi \|_{L_1} + \frac{2^{j/2}}{n} \bigvee \psi) + \frac{2^{j/2}}{n} \| f \|_{L_\infty} \qquad (A2)$$

(for $\hat{\alpha}_{j_0 k}$ and $\alpha_{j_0 k}$ ψ is replaced by φ). An improvement of (A2) in terms of the average moduli of smoothness and A-spaces is considered in [B12(b)]. (ii) Orthogonality of the wavelet basis implies

$$|\langle\psi_{jk},\psi_{j'k'}\rangle - \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)\psi_{j'k'}(x_i)| \leq \frac{2^{(j+j')/2}}{n}(\bigvee\psi)^2 \tag{A3}$$

(for $\varphi_{j_0k}.\varphi_{j_0k'}$ and $\varphi_{j_0k}.\psi_{j'k'}$ $(\bigvee\psi)^2$ is replaced by $(\bigvee\varphi)^2$ and $\bigvee\varphi.\bigvee\psi$, respectively). (iii) Almost-diagonality (decorrelation) property of the wavelet transform with respect to the inner product $\langle g,h\rangle_n := \frac{1}{n}\sum_{i=1}^{n}g(x_i)h(x_i)$. Its essence is that the Gramian matrix of $\{\varphi_{j_0k},\psi_{jk}\}$ with respect to $\langle .,.\rangle_n$ is sparse because of the compactness of the wavelet support.

Let us compare the results obtained without and with the decorrelation property. The number N_j of non-zero β_{jk} on a fixed level does not exceed $\sum_{-k\in\mathrm{supp}\psi-2^j.\mathrm{supp}f}1 \leq c.2^j$, where c depends on the lengths of the supports of f and ψ only. For fixed j and j' this estimate implies that the number $N_{jj'}$ of non-zero products $\beta_{jk}.\beta_{j'k'}.\langle\psi_{jk},\psi_{j'k'}\rangle_n$ would be $O(2^{j+j'})$. On the other hand, by the decorrelation property, $\langle\psi_{jk},\psi_{j'k'}\rangle_n$ can be non-zero only if $-k\in\mathrm{supp}\psi-2^{j-j'}.\mathrm{supp}\psi-2^{j-j'}.k'$, i.e., only if $|k-2^{j-j'}.k'|\leq c.\max(2^{j-j'},1)$, c depending on the length of suppψ only. This yields the more refined bound $N_{jj'}=O(2^{\max(j,j')})$. Estimating the quantity $\sum_{j=j_0}^{j_1}\sum_{j'=j_0}^{j_1}N_{jj'}$ needed in the proof of (A1) via the former bound on $N_{jj'}$ yields $O(2^{2j_1})$, while the latter bound implies $\sum_{j=j_0}^{j_1}\sum_{j'=j_0}^{j_1}N_{jj'}\leq c[4(j_1-j_0-1/2)2^{j_1}+2^{j_0+1}+1]=O(j_1.2^{j_1})$.

The proof of (A1) follows from (i-iii), $2^{j_1}=O(\frac{n}{\ln n})$ and $s_1>0$.

Recalling that $1=\int\psi_{jk}(x)^2dx$ and estimating the error of the quadrature formula for the latter integral yields (see (A3)) $|1-\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)^2|\leq\frac{2^j}{n}\bigvee(\psi^2)$. Fix $v>0$. For n large enough, $v2^{2js}+1-\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)^2\geq\frac{1}{2}v2^{2js}$, uniformly in $j=j_0,\ldots,j_1$, where, with no loss of generality, $j_0\geq 0$. By the lemma's assumptions about f, there exists (j',k') with $j'\geq j_0$, such that $\beta_{j'k'}\neq 0$. We shall be considering here the more difficult case $j_0\to\infty$. It can be seen that, without loss of generality, one may assume that $j'=j_0$. By the sharpness of s' and $s'<s$, there exists[A16] $c_0: 0<c_0<\infty$, depending on f and ψ only, such that $|\beta_{j_0,k'}|\geq c_0.2^{-j_0s}$. From here and (A2), $|\overline{\beta}_{j_0k'}|\geq|\beta_{j_0k'}|-|\overline{\beta}_{j_0k'}-\beta_{j_0k'}|\geq c_0.2^{-j_0s}-O(2^{-j_0/2}.n^{-s_1}+2^{j_0/2}.n^{-1})=c_0.(1-o(1)).2^{j_0s}>\frac{1}{2}.c_0.2^{-j_0s}$ for sufficiently big n, by the theorem's assumptions on j_0. For n large enough $w'(v)$ can now be bounded from below by

$$w'(v)/2 > c_1.\frac{2^{4j_0s}v}{(1+v2^{2j_0s})^3}.2^{-2j_0s}-\frac{\delta^2}{nv^3}.\sum_{j=j_0}^{j_1}2^{-4js}.N_j.(1+O(\frac{2^j}{n}))^2=$$

$$=\frac{1}{v^2 2^{4j_0s}}(\frac{c_1}{(1+v^{-1}2^{-2j_0s})^3}-\frac{c_2}{nv}\sum_{j=j_0}^{j_1}2^{j-4(j-j_0)s})\geq$$

$$\geq\frac{1}{v^2 2^{4j_0s}}(\frac{c_1}{(1+v^{-1}2^{-2j_0s})^3}-\frac{c_3}{v}.\frac{2^{j_1}}{n})=\frac{1}{v^2 2^{4j_0s}}(\frac{c_1}{(1-o(1))^3}-\frac{1}{v}.o(1))>0$$

for sufficiently large n. Therefore, by continuity and asymptotic equivalence to $(\frac{d}{dv}ER)(v)$ of $w'(v)$, there exists $C = C(f,\psi) \in (0,\infty)$, such that for n large enough $\inf_{v\in\mathbf{R}} w(v)$ and $\inf_{v\in\mathbf{R}} ER(v)$ are both achieved on $[0,C]$. By continuity of $ER(v)$ and compactness of $[0,C]$, v^* exists and, moreover, $v^* \in [0,C)$ and either $v^*=0$ or $(\frac{d}{dv}ER)(v^*)=0$. $\square$

Proof of Theorem 2. The proof is similar in spirit to the derivations in Section 4.4 of Wahba (1990) (cf. Dechevsky and Penev (1998a)). First of all,

$$E(GFCV(v)) = (1+\overline{h}(v))^2[\delta^2 + \frac{1}{n}E\sum_{i=1}^{n}(f(x_i)-\tilde{f}_v(x_i))^2 - \frac{2}{n}E\sum_{i=1}^{n}(y_i - f(x_i)(\tilde{f}_v(x_i)-f(x_i))] =$$

$$= (1+\overline{h}(v))^2[\delta^2 + ER(v) - \frac{2}{n}E\sum_{i=1}^{n}\{\epsilon_i(\tilde{f}_v(x_i)-f(x_i))\}]. \qquad (A4)$$

Now we show that $\overline{h}(v) \to 0$ as $n\to\infty$, uniformly in $v\geq 0$. Invoking again a familiar bound for the error of the quadrature formulae involved (cf. the proof of Theorem 1), by the orthonormality of the wavelet basis and the compactness of the support of f,

$$0 \leq \overline{h}(v) \leq c[\frac{2^{j_0}}{n}(1+\frac{2^{j_0}}{n}\bigvee(\varphi^2)) + \frac{1}{n}\sum_{j=j_0}^{j_1}2^j(1+\frac{2^j}{n}\bigvee(\psi^2))] \leq c_1.\frac{2^{j_1}}{n} = o(1),$$

as $n\to\infty$, uniformly in $v\geq 0$. Here $c_1 = c_1(\varphi,\psi,f)$.

Fix $v\geq 0$ and denote by $\tilde{\epsilon}_i(v) = \tilde{f}_v(x_i) - f(x_i)$. Since $E\epsilon_i = 0$ and ϵ_i is independent of any $y_l, l\neq i$ we can evaluate:

$$E[\epsilon_i\tilde{\epsilon}_i(v)] = E\{\epsilon_i[\sum_k\sum_{l=1}^{n}\frac{\varphi_{j_0k}(x_l)y_l}{n}\varphi_{j_0k}(x_i) + \sum_k\sum_{j=j_0}^{j_1}\sum_{l=1}^{n}\frac{\psi_{jk}(x_l)\psi_{jk}(x_i)y_l}{n(1+v2^{2js})}]\} =$$

$$= \sum_k E\frac{\varphi_{j_0k}(x_i)^2 y_i\epsilon_i}{n} + \sum_k\sum_{j=j_0}^{j_1}E\frac{\psi_{jk}(x_l)^2 y_i\epsilon_i}{n(1+v2^{2js})} =$$

$$= \frac{\delta^2}{n}[\sum_k\varphi_{j_0}(x_i)^2 + \sum_{j=j_0}^{j_1}\sum_k\frac{\psi_{jk}(x_i)^2}{1+v2^{2js}}] = \delta^2 h_{ii}(v) \qquad (A5)$$

Substitution from (A5) into (A4) yields $E(GFCV(v)) = (1+\overline{h}(v))^2(\delta^2 + ER(v) - 2\delta^2\overline{h}(v))$ from which also $E(GFCV(v)) \sim ER(v) + \delta^2$ follows.

To show the required consistency, we first note that

$$\frac{ER(v) - E(GFCV(v)) + \delta^2}{ER(v)} = -2\overline{h}(v) - \overline{h}(v)^2 + \frac{\overline{h}(v)^2\delta^2}{ER(v)}(3+2\overline{h}(v)) \qquad (A6)$$

holds. Denoting $\mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_n)]'$, $\mathbf{I}$ - the identity on $\mathbf{R}^n$, $\| \, . \, \|$ - the usual Hilbert norm on $\mathbf{R}^n$, $\mathbf{f}_v = [\tilde{f}_v(x_1), \tilde{f}_v(x_2), \ldots, \tilde{f}_v(x_n)]'$, $\epsilon = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]'$ we have

$$ER(v) = E\frac{1}{n} \| \mathbf{f}_v - \mathbf{f} \|^2 = E\frac{1}{n} \| (\mathbf{I} - \mathbf{H}(v))\mathbf{f} - \mathbf{H}(v)\epsilon \|^2 =$$

$$= \frac{1}{n} \| (\mathbf{I} - \mathbf{H}(v))\mathbf{f} \|^2 + \delta^2 \mu(v) \geq \delta^2 \mu(v).$$

Therefore, (A6) yields:

$$\frac{|ER(v) - E(GFCV(v)) + \delta^2|}{ER(v)} \leq \frac{\overline{h}(v)^2}{\mu(v)}(3 + 2\overline{h}(v)) = \xi(v).$$

Now, by Lemma 2, $\lim_{n\to\infty} \frac{h(v)^2}{\mu(v)} = 0$, hence $\xi(v) \to 0$. Then it is easily seen that, by definition of $\tilde{v}$ and v^*,

$$(1 - \xi(\tilde{v}))ER(\tilde{v}) \leq E(GFCV(\tilde{v})) - \delta^2 \leq E(GFCV(v^*)) - \delta^2 \leq (1 + \xi(v^*))ER(v^*);$$

again by definition of $\tilde{v}$ and v^*, $\frac{ER(\tilde{v})}{ER(v^*)} \geq 1$, for any $n \in \mathbf{N}$. Hence, $1 \leq \frac{ER(\tilde{v})}{ER(v^*)} \leq \frac{1+\xi(v^*)}{1-\xi(\tilde{v})}$ holds. $\square$

Proof of Lemma 2. Fix $v \geq 0$. Denote $m := \text{rank}\mathbf{H}(v)$; $m \leq n$. On the one hand, $m = O(2^{j_1}) = o(n)$ holds. Indeed, while the domain of the linear operator $\mathcal{L}_v(\mathbf{y}) := \mathbf{H}(v)\mathbf{y}$ is $\mathbf{R}^n$, the range of $\mathcal{L}_v$ is the linear span of the set of n-dimensional vectors

$$V_{j_1}(\text{supp}f) := \{(\varphi_{j_0 k}(x_1), \ldots, (\varphi_{j_0 k}(x_n))', \ k : \text{supp}\varphi_{j_0 k} \cap \text{supp}f \neq \emptyset\} \cup$$

$$\cup \{(\psi_{jk}(x_1), \ldots, (\psi_{jk}(x_n))', \ j = j_0, \ldots, j_1, \ k : \text{supp}\psi_{jk} \cap \text{supp}f \neq \emptyset\},$$

which has dimension $O(2^{j_1})$.

On the other hand,

$$\frac{\overline{h}(v)^2}{\mu(v)} = \frac{(\text{tr}\mathbf{H}(v))^2}{n.\text{tr}(\mathbf{H}'(v)\mathbf{H}(v))} = \frac{m}{n} . \frac{(\frac{1}{m}\sum_{i=1}^m s_i)^2}{\frac{1}{m}\sum_{i=1}^m s_i^2},$$

where s_i, $i = 1, \ldots, n$, are the s-numbers of $\mathbf{H}(v)$ (see, e.g., Marshall and Olkin (1979)), and $s_{m+1} = \cdots = s_n = 0$. Therefore, by the inequality between the arithmetic and the quadratic mean, $\frac{\overline{h}(v)^2}{\mu(v)} \leq \frac{m}{n} = o(1)$ as $n \to \infty$, uniformly in $v \geq 0$. $\square$

Proof of Theorem 3. We prove necessity when $j_0 = O(1)$ first. For any fixed (j,k) $j \geq j_0$, consistency of $\tilde{f}_v$ in B^{σ}_{22}, $\sigma \geq 0$, implies $E(\tilde{\beta}_{jk} - \beta_{jk})^2 \to 0$, $n \to 0$. Choose $(j', k') : \beta_{j'k'} \neq 0$. Simple computations show that

$$E(\tilde{\beta}_{jk} - \beta_{jk})^2 =$$

$$= \frac{1}{(1+v2^{2js})^2}[(\bar{\beta}_{jk}-\beta_{jk})^2 + \frac{\delta^2}{n}.\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)^2 - 2v2^{2js}\beta_{jk}(\bar{\beta}_{jk}-\beta_{jk}) + v^2 2^{4js}\beta_{jk}^2].$$

By Riemann integrability of $f\psi_{jk}$, $\bar{\beta}_{jk}-\beta_{jk}=o(1)$; by Riemann integrability of ψ_{jk}^2 (or by the stronger (A3)), $\frac{\delta^2}{n}.\frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)^2 = O(\frac{1}{n}) = o(1)$; by the boundedness of v, $2v2^{2js}\beta_{jk}(\bar{\beta}_{jk}-\beta_{jk}) = v.o(1) = o(1)$ and $\frac{1}{(1+v2^{2js})^2}$ is bounded away from zero. Therefore,

$$E(\tilde{\beta}_{jk}-\beta_{jk})^2 = \frac{1}{(1+v2^{2js})^2}(o(1) + v^2 2^{4js}\beta_{jk}^2)$$

is only possible if $v \to 0$, which proves the necessity claim. Admissibility of the choice $v = v_n^*$ here follows from the boundedness of v_n^*, proved in Theorem 1.

Now we prove sufficiency. After some computations, utilizing, as usual, the wavelet-coefficient equivalent norm of B_{22}^{σ}, and after a simple bound from above, by the inequality between the arithmetic and quadratic mean,

$$E \parallel \tilde{f}_v - f \parallel_{B_{22}^{\sigma}}^2 = E \parallel \tilde{f}_v - E\tilde{f}_v \parallel_{B_{22}^{\sigma}}^2 + \parallel E\tilde{f}_v - f \parallel_{B_{22}^{\sigma}}^2 \le$$

$$\le \frac{\delta^2}{n}\sum_k \frac{1}{n}\sum_{i=1}^{n}\varphi_{j_0k}(x_i)^2 + \frac{\delta^2}{n}\sum_{j=j_0}^{j_1}\frac{2^{2j\sigma}}{(1+v2^{2js})^2}\sum_k \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i)^2 +$$

$$+\sum_k(\overline{\alpha}_{j_0k}-\alpha_{j_0k})^2 + 2\sum_{j=j_0}^{j_1}\frac{2^{2j\sigma}}{(1+v2^{2js})^2}\sum_k(\overline{\beta}_{jk}-\beta_{jk})^2 +$$

$$+2\sum_{j=j_0}^{j_1}2^{2j\sigma}\frac{v^2 2^{4js}}{(1+v2^{2js})^2}\sum_k\beta_{jk}^2 + \sum_{j=j_1+1}^{\infty}2^{2j\sigma}\sum_k\beta_{jk}^2.$$

Applying (A1,A2), the inequality between the arithmetic and quadratic mean, $v \ge 0$, utilizing $\frac{s'-\sigma}{2s} \in (0,1/2) \subset (0,1)$, which yields the inequality

$$\frac{v^2 2^{4js}}{(1+v2^{2js})^2} \le \Big(\frac{v^2 2^{4js}}{1}\Big)^{\frac{s'-\sigma}{2s}}.\Big(\frac{v^2 2^{4js}}{v^2 2^{4js}}\Big)^{1-\frac{s'-\sigma}{2s}} = v^{\frac{s'-\sigma}{s}}.2^{2j(s'-\sigma)},$$

invoking also the bound

$$\sum_{j=j_1+1}^{\infty}2^{2j\sigma}\sum_k\beta_{jk}^2 \le 2^{2(\sigma-s')}2^{2j_1(\sigma-s')}\sum_{j=j_1+1}^{\infty}2^{2js'}\sum_k\beta_{jk}^2 \le c2^{2j_1(\sigma-s')}||f||_{B_{22}^{s'}}^2,$$

as well as the familiar upper bound on N_j, after obtaining

$$\sum_{j=j_0}^{j_1}2^{2j\sigma}\frac{v^2 2^{4js}}{(1+v2^{2js})^2}\sum_k\beta_{jk}^2 \le v^{\frac{s'-\sigma}{s}}\sum_{j=j_0}^{j_1}2^{2js'}\sum_k\beta_{jk}^2 \le v^{\frac{s'-\sigma}{s}}||f||_{B_{22}^{s'}}^2,$$

we arrive at

$$||\tilde{f}_v - f||_{B^\sigma_{22}} \le c(f,\varphi)\frac{\delta^2}{n}2^{j_0}(1+\frac{2^{j_0}}{n}) + c(f,\psi)\frac{\delta^2}{n}\sum_{j=j_0}^{j_1} 2^{j(1+2\sigma)}(1+\frac{2^j}{n})+$$

$$+c_1(f,\varphi)2^{j_0}[\frac{2^{-j_0}}{n^{2s_1}}(1+\frac{2^{j_0}}{n^2})+\frac{2^{j_0}}{n^2}] + c_1(f,\psi)\sum_{j=j_0}^{j_1} 2^{j(1+2\sigma)}[\frac{2^{-j}}{n^{2s_1}}(1+\frac{2^j}{n^2})+\frac{2^j}{n^2}]+$$

$$+c_{s'}(f)v^{\frac{s'-\sigma}{s}} + c_{s',\sigma}(f)2^{-2j_1(s'-\sigma)} \le$$

$$\le c_{s',\sigma,\delta}(f,\varphi,\psi)(\frac{2^{j_0}}{n} + \frac{2^{j_1(1+2\sigma)}}{n} + \frac{2^{2j_1\sigma}}{n^{2s_1}} + v^{\frac{s'-\sigma}{s}} + 2^{-2j_1(s'-\sigma)}) = o(1),$$

when $2^{j_1} = o(n^{\min(\frac{s_1}{\sigma}, \frac{1}{1+2\sigma})})$, $j_0 \le j_1$, $j_1 \to \infty$ and $v \to 0$.

Admissibility of the choice $v = v_n^*$ follows from Theorem 4. □

Proof of Theorem 4. We give the proof for v^*. The statement about $\tilde{v}$ follows from the asymptotic equivalence $E(GFCV(v)) - \delta^2 \sim ER(v)$, $v \ge 0$ (see the proof of Theorem 2). Indeed, by the minimizing properties of $\tilde{v}_n$ and v_n^* and the continuous dependence of $E(GFCV(v))$ and $ER(v)$ on v, this asymptotic equality implies (e.g., by proving contradiction with the opposite) that $\tilde{v}_n$ and v_n^* have the same density points.

From the proof of Theorem 1 it follows that, for n large enough, $v^* = v_n^*$ exists and $v_n^* \in S_{n,1} := \{0\} \cup \{v = v_n \in [0,C) : (\frac{d}{dv}ER)(v) = 0\}$, where C is defined in the same proof. Consider also $S_{n,2} := \{0\} \cup \{v = v_n \in [0,C) : w'(v) = 0\}$. By Bolzano-Weierstrass theorem, every sequence $v_n \in S_{n,\nu}$ has at least one density point and every such point is in $[0,C]$, $\nu = 1,2$. Consider $S_\nu = \bigcup_{n\in\mathbf{N}} S_{n,\nu}$, $\nu = 1,2$. By definition of $S_{n,\nu}$, $v = 0$ is a density point of $S_{n,\nu}$, $\nu = 1,2$. By (A1), S_1 and S_2 have the same density points. To prove $v^* \to 0$, it now suffices to show that the only density point of S_2 is $v = 0$.

As in the proof of Theorem 1, we can assume that there exists (j_0,k_0) such that $\beta_{j_0k_0} \ne 0$. Also, without loss of generality, $C \ge 1$ and $j_0 \ge 0$. Assume that $v^* = v_n^*$ remains bounded away from 0 for infinitely many values of n. Then, for these particular values of n, the same argument as in the proof of Theorem 1 yields

$$\frac{w'(v)}{2} > c_1\frac{v2^{2j_0s}}{(1+v2^{2j_0s})^3} - \frac{c_2}{nv^3}\sum_{j=j_0}^{j_1} 2^{j(1-4s)} \ge \frac{c_1}{8C^3}2^{-4j_0s}v - \frac{c_2}{nv^3}\sum_{j=j_0}^{j_1} 2^{j(1-4s)} = w_0(v).$$

The equation $w_0(v) = 0$ has a unique real positive root

$$v_0 : v_0^4 = \frac{c_3}{n}\sum_{j=j_0}^{j_1} 2^{j-4(j-j_0)s} \le c_4.\frac{2^{j_1}}{n} = o(1)$$

as $n \to \infty$. Since $w'(v) > 0$ for $v \geq C$ and all sufficiently large n, for the same n the largest real zero of any *lower* bound of $w'(v)$ is necessarily an upper bound for all real zeros of $w'(v)$. In particular, v_0 is an upper bound for the zeros of $w'(v)$ for any n considered. Therefore, contradiction has been achieved with the assumption that v^* remains bounded away from $v = 0$ for infinitely many n. $\square$

Proof of Corollary 1. Follows from Theorem 4 and the sufficiency part of Theorem 3, applied to $\tilde{v}$. $\square$

A17. The proof of Lemma 2 can be simplified by relying on *selfadjointness* (cf. Dechevsky and Penev (1998)). Indeed, in our case the influence matrix $\mathbf{H}(v)$ is symmetric positive semi-definite, therefore, its s-numbers are just its eigenvalues (all nonnegative), arranged in order of decreasing magnitude. We have chosen a more general consideration because of envisaging an extension of our model for *biorthogonal wavelets* and *wavelet packets*.[B18] In these generalized cases the problem becomes non-selfadjoint. Our general proof shows that Lemma 2 is valid in this more general context, too.

Proof of Theorem 5. We outline briefly the proof for $v = \tilde{v}$ and give a detailed proof of the more difficult case $v = v^*$. As earlier, we consider the more difficult case for j_0: $j_0 \to \infty$.

For any $n = 1, 2, \ldots$, $(\frac{d}{dv}MISE)(0) = -\frac{2}{n}\sum_{j=j_0}^{j_1} 2^{2js}\sum_k \int \psi_{jk}^2 f < 0$. Since $s > s'$ and the index s' is sharp, one may assume that there exists (cf. the proof of Theorem 5 and Theorem 4) (j_0, k_0): $\beta_{j_0k_0} \neq 0$ and $|\beta_{j_0k_0}| \geq c_0 2^{-j_0 s}$ for some $c_0 > 0$. Hence, utilizing also $\int \psi_{jk}^2 f \leq \| f \|_{L_\infty}$, and invoking the bound $N_j \leq c2^j$ again,

$$\frac{1}{2}(\frac{d}{dv}MISE)(v) \geq \frac{c_0^2 2^{2j_0 s} v}{(1+v2^{2j_0 s})^3} - \frac{\| f \|_{L_\infty}}{n}\sum_{j=j_0}^{j_1}\frac{2^{j(1+2s)}}{(1+v2^{2js})^3} \geq$$

$$\geq \frac{1}{v^2 2^{4j_0 s}}(\frac{c_1}{(1+v^{-1}2^{-2j_0 s})^3} - \frac{c_2}{nv}\sum_{j=j_0}^{j_1} 2^{j-4(j-j_0)s}) \geq$$

$$\geq \frac{1}{v^2 2^{4j_0 s}}(\frac{c_1}{(1+v^{-1}2^{-2j_0 s})^3} - \frac{c_3}{nv}2^{j_1}) = \frac{1}{v^2 2^{4j_0 s}}(\frac{c_1}{(1+v^{-1}2^{-2j_0 s})^3} - \frac{o(1)}{v}) > 0$$

for v and n large enough. Now existence of $C < \infty$ and of $\tilde{v} : 0 < \tilde{v} < C$ follows in the same way as in Theorem 1.

The basic idea of the proof about v^* is the same. Denote $w(v) := M(\tilde{f}_v)$. For any $n = 2, 3, \ldots$, by the inequality between the arithmetic and quadratic mean,

$$w'(0) = -\frac{2}{n-1}\sum_{j=j_0}^{j_1} 2^{2js}\sum_k[\frac{1}{n}\sum_{i=1}^n \psi_{jk}(X_i)^2 - (\frac{1}{n}\sum_{i=1}^n \psi_{jk}(X_i))^2] \leq 0$$

and this inequality is strict almost surely. We proceed to study the behaviour of $w'(v)$ for large v. By $s' < s$, there exists $s_0 : 0 < s' < s_0 < \frac{s'+s}{2} < s$. By sharpness of s', one can find (j_0, k_0)

with $|\beta_{j_0k_0}| \geq c_0 2^{-j_0 s_0}$, for some $c_0 = c_0(f, \psi, s_0 - s') > 0$. Then, after computations, omitting positive summands, bounding positive terms from below and negative ones from above, one obtains

$$\frac{1}{2}Ew'(v) \geq \frac{v2^{4j_0 s}}{(1+v2^{2j_0 s})^3}\beta_{j_0k_0}^2 - \frac{c \parallel f \parallel_{L_\infty}}{n}\sum_{j=j_0}^{j_1}\frac{2^{j(1+2s)}}{(1+v2^{2js})^2} \geq$$

$$\geq \frac{1}{v^2 2^{2j_0(s+s_0)}}\left(\frac{c_1}{(1+v^{-1}2^{-2j_0 s})^3} - \frac{c_2}{n}2^{2j_0(s+s_0)}\sum_{j=j_0}^{j_1}2^{j(1-2s)}\right).$$

Case $0 < s < 1/2$. By conditions on j_0 and j_1, $n^{-1}2^{2j_0(s+s_0)}\sum_{j=j_0}^{j_1}2^{j(1-2s)} = O(n^{-\frac{2(s-s_0)}{1+2s}}) = o(1)$.

Case $s = 1/2$. Again by the conditions on j_0 and j_1, $j_1 - j_0 + 1 = o(\ln n)$ and $n^{-1}2^{2j_0(s+s_0)}(j_1 - j_0 + 1) = O(n^{-\frac{1-2s_0}{1+2s}}).o(\ln n) = o(1)$, since $s_0 < s = 1/2$.

Case $s > 1/2$. Now $n^{-1}2^{2j_0(s+s_0)}\sum_{j=j_0}^{j_1}2^{j(1-2s)} = O(n^{-\frac{2(s-s_0)}{1+2s}}) = o(1)$.

Therefore, in all three cases,

$$Ew'(v) \geq \frac{c_1}{v^2 2^{2j_0(s+s_0)}} \tag{A7}$$

with certain positive constant c_1 holds for sufficiently large v and n big enough.

On the other hand, by the triangle inequality, and after some computations,

$$\frac{1}{2}E|w'(v) - Ew'(v)| \leq \sum_{j=j_0}^{j_1}\frac{2^{2js}}{(1+v2^{2js})^2}\times$$

$$\times\sum_k[(E|\hat{\beta}_{jk}^2 - \beta_{jk}^2|)(\frac{v2^{2js}}{1+v2^{2js}} + \frac{1}{n-1}) + (\int \psi_{jk}^2 f).(\frac{v2^{2js}}{n(1+v2^{2js})} + \frac{2+1/n}{n-1})]. \tag{A8}$$

By Cauchy-Schwartz inequality, and by $(|a|+|b|)^{1/2} \leq |a|^{1/2} + |b|^{1/2}$,

$$E|\hat{\beta}_{jk}^2 - \beta_{jk}^2| \leq (E(\hat{\beta}_{jk} - \beta_{jk})^2)^{1/2}(E(\hat{\beta}_{jk} + \beta_{jk})^2)^{1/2} \leq \frac{1}{\sqrt{n}}(\int \psi_{jk}^2 f)^{1/2}(2|\beta_{jk}| + \frac{1}{\sqrt{n}}(\int \psi_{jk}^2 f)^{1/2}).$$

Substituting this into (A8), after computations utilizing $\sum \psi_{jk}^2 f \leq \parallel f \parallel_{L_\infty}$, yields

$$E|w'(v) - Ew'(v)| \leq c_4\sum_{j=j_0}^{j_1}\frac{2^{2js}}{(1+v2^{2js})^2}\sum_k(\frac{v2^{2js}}{1+v2^{2js}}\frac{|\beta_{jk}|}{\sqrt{n}} + \frac{1}{n}) \leq$$

$$\leq \frac{c_4}{v^2}(\sum_{j=j_0}^{j_1}2^{-2js}\sum_k\frac{|\beta_{jk}|}{\sqrt{n}} + \frac{c_5}{n}\sum_{j=j_0}^{j_1}2^{j(1-2s)}) = \frac{c_4}{v^2}(I_1 + I_2),$$

where $c_4 = c_4(f, \psi)$. Therefore, by (A7) and Markov inequality,

$$P\{w'(v) < \frac{c_1}{2v^2 2^{2j_0(s+s_0)}}\} \le P\{|w'(v) - Ew'(v)| > \frac{c_1}{2v^2 2^{2j_0(s+s_0)}}\} \le \frac{2c_4}{c_1} 2^{2j_0(s+s_0)}.O(I_1 + I_2).$$

By Cauchy-Schwartz inequality, and by the bound on N_j (cf. the proof of Theorem 1), $I_1 \le \frac{c_6}{\sqrt{n}}(\sum_{j=j_0}^{j_1} 2^{j(1-2(2s+s'))})^{1/2} \parallel f \parallel_{B^{s'}_{22}}$. Considering the three cases $2s + s' < 1/2$, $2s + s' = 1/2$, $2s + s' > 1/2$, and noting that in the first two of these cases necessarily $s < 1/2$ holds, it turns out that in all three cases $2^{2j_0(s+s_0)} I_1 = O(n^{2\frac{s_0 - (s+s')/2}{1+2s}})$ holds, and this tends to zero as $n \to \infty$ by the choice of s_0. As for the term $2^{2j_0(s+s_0)} I_2$, it was already shown while deriving the lower bound for $Ew'(v)$ that this term tends to zero, too. Let $C : 0 < C < \infty$, be the constant obtained in the proof of the case about $\tilde{v}$. Now we see that for $v \ge C$

$$P\{w'(v) \le 0\} \le P\{w'(v) < \frac{c_1}{2v^2 2^{2j_0(s+s_0)}}\} \to 0$$

as $n \to \infty$. This implies the existence of v^* by already familiar continuity/compactness argument. □

A18. In the case $0 < s \le 1/2$ (the range of s for which B^s_{22} contains discontinuous functions), the condition on j_1 in Theorem 5 seems to be restrictive: $2^{j_0} \le 2^{j_1} < C_1 2^{j_0}$ indicates a rather narrow range for j_1. But first, Theorem 5 only gives sufficient conditions for the existence of the minimizer v^*. Second, a close inspection of the proof of this theorem shows that its statement in the case $0 < s \le 1/2$ remains true also for a broader range of j_1: it suffices to have $2^{j_1} = o(n^{\frac{1-2s_0}{1+2s}}.2^{j_0})$ for any s_0 such that $s' < s_0 < \frac{s'+s}{2}$. Third, the theorem may be sharpened by admitting a broader range for j_1, if one considers a more stringent notion of sharpness of s', namely, f is such that $f \in B^{s'}_{22}$ and $f \notin B^{s'}_{2q}$ for any $q < 2$. The proof of this sharpened version is similar.

Proof of Theorem 6. In this proof, C denotes a positive constant which may vary along the lines. Fix $v \ge 0$. First, let us evaluate $MISE(v)$ from below. Using Parseval's identity and neglecting some nonnegative summands, we get:

$$MISE(v) \ge \frac{1}{n^2} \sum_k E\{\sum_{i=1}^n \varphi_{j_0 k}(X_i)\}^2 + \sum_{j=j_0}^{j_1} \sum_k \frac{1}{n^2} E\{\sum_{i=1}^n \psi_{jk}(X_i)\}^2 \frac{1}{(1 + v2^{2js})^2}$$

$$- \sum_k \alpha^2_{j_0 k} - \sum_{j=j_0}^{j_1} \sum_k \frac{\beta^2_{jk}(1 - v2^{2js})}{1 + v2^{2js}}$$

Since $\frac{1-v2^{2js}}{1+v2^{2js}} < \frac{1}{(1+v2^{2js})^2}$, we can continue the above inequality to obtain:

$$MISE(v) \ge \sum_k \frac{1}{n} \{\int \varphi^2_{j_0 k} f - (\int \varphi_{j_0 k} f)^2\} + \sum_{j=j_0}^{j=j_1} \sum_k \frac{1}{n(1 + v2^{2js})^2} \{\int \psi^2_{jk} f - (\int \psi_{jk} f)^2\}$$

$$\geq \frac{1}{n}\{\sum_k \int \varphi_{j_0k}^2 f + \sum_{j=j_0}^{j_1} \sum_k \frac{1}{(1+v2^{2js})^2} \int \psi_{jk}^2 f - \| f \|_2^2\}$$

Now we invoke Meyer's lemma (Lemma 1 in Tribouley (1995) - for the application of this lemma the periodic extension onto $\mathbf{R}$ of all relevant compactly supported functions is considered, with period equal to or greater than the support's diameter). By this lemma, and in view of $\int f = 1$,

$$2^{-j_0} \int f(x) \sum_k \varphi_{j_0k}^2(x)dx = \int f(x) \sum \varphi^2(2^{j_0}x - k)dx \rightarrow_{j_0 \rightarrow \infty} \sum_k \int_0^1 \varphi^2(x-k)dx = C,\ C = C(\varphi).$$

Utilizing the compactness of supp$f$ and $j_0 \rightarrow \infty$ in a similar consideration of the expressions of type $2^{-j} \int f(x) \sum_k \psi_{jk}^2(x)dx$, we arrive at

$$MISE(v) \geq \frac{2^{j_0}C}{n}\{1 + \sum_{j=j_0}^{j_1} \frac{2^{j-j_0}}{(1+v2^{2js})^2}\},\ C = C(\varphi, \psi), \tag{A9}$$

for sufficiently large n, so that $2^{-j_0} \parallel f \parallel_{L_2}^2$ is smaller than, say, 1/2. Next, we consider $\xi_v = M(\tilde{f}_v) - MISE(v) + T_n$ with T_n as defined in the premises of the theorem. Note that $E\xi_v = 0$ and that T_n does not depend on v. After further computations, we obtain

$$\xi_v = -\frac{2}{n(n-1)} \sum_k [\{\sum_{i=1}^n \varphi_{j_0k}(X_i)\}^2 - \sum_{i=1}^n \varphi_{j_0k}^2(X_i)]$$

$$-\frac{2}{n(n-1)} \sum_{j=j_0}^{j_1} \sum_k \frac{\{\sum_{i=1}^n \psi_{jk}(X_i)\}^2 - \sum_{i=1}^n \psi_{jk}^2(X_i)}{1+v2^{2js}} - \frac{n+2}{n} \sum_k \alpha_{j_0k}^2 + \sum_{j=j_0}^{j_1} \sum_k \{\frac{2\beta_{jk}^2}{1+v2^{2js}} - \frac{\beta_{jk}^2}{(1+v2^{2js})^2}\}$$

$$+\sum_k \hat{\alpha}_{j_0k}^2 + \sum_{j=j_0}^{j_1} \sum_k \frac{\hat{\beta}_{jk}^2}{(1+v2^{2js})^2} - \sum_k Var(\hat{\alpha}_{j_0k}) - \sum_{j=j_0}^{j_1} \sum_k \frac{Var(\hat{\beta}_{jk})}{(1+v2^{2js})^2} + \frac{2(n+1)}{n} \sum_k \hat{\alpha}_{j_0k}\alpha_{j_0k}$$

Substituting $Var(\hat{\alpha}_{j_0k}) = \frac{1}{n}\{\int \varphi_{j_0k}^2 f - (\int \varphi_{j_0k} f)^2\}, Var(\hat{\beta}_{jk}) = \frac{1}{n}\{\int \psi_{jk}^2 f - (\int \psi_{jk} f)^2\}$ leads to

$$\xi_v = (-\frac{2}{n(n-1)} + \frac{1}{n^2}) \sum_k \sum_{i=1}^n \sum_{l=1, l\neq i}^n \varphi_{j_0k}(X_i)\varphi_{j_0k}(X_l) +$$

$$\sum_{j=j_0}^{j_1} \{-\frac{2}{n(n-1)} + \frac{1}{n^2(1+v2^{2js})}\} \sum_k \sum_{i=1}^n \sum_{l=1, l\neq i}^n \frac{\psi_{jk}(X_i)\psi_{jk}(X_l)}{1+v2^{2js}} - \frac{n+1}{n} \sum_k (\int \varphi_{j_0k} f)^2 +$$

$$\sum_{j=j_0}^{j_1} \sum_k (\int \psi_{jk} f)^2 \{\frac{2}{1+v2^{2js}} - \frac{n-1}{n(1+v2^{2js})^2}\} + \frac{2(n+1)}{n} \sum_k \hat{\alpha}_{j_0k}\alpha_{j_0k} +$$

$$+\frac{1}{n^2} \sum_{i=1}^n \sum_k \{\varphi_{j_0k}^2(X_i) - \int \varphi_{j_0k}^2 f\} + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=j_0}^{j_1} \sum_k \frac{\psi_{jk}^2(X_i) - \int \psi_{jk}^2 f}{(1+v2^{2js})^2} \tag{A10}$$

Denoting

$$A_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_k \{\varphi_{j_0k}^2(X_i) - \int \varphi_{j_0k}^2 f\},\ A_2 = \sum_{j=j_0}^{j_1} A_{2j}, A_{2j} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_k \frac{\psi_{jk}^2(X_i) - \int \psi_{jk}^2 f}{(1+v2^{2js})^2},$$

we see that $E(A_i) = 0,\ i = 1,2$ holds. With the above notation introduced, we can write for the RHS of (A10):

$$A_1 + A_2 - \frac{n+1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{l=1,l\neq i}^{n}\sum_k\{\varphi_{j_0k}(X_i) - \int \varphi_{j_0k}f\}\{\varphi_{j_0k}(X_l) - \int \varphi_{j_0k}f\}+$$

$$\sum_{j=j_0}^{j_1}\{-\frac{2}{n(n-1)(1+v2^{2js})}+\frac{1}{n^2(1+v2^{2js})^2}\}\sum_k\sum_{i=1}^{n}\sum_{l=1,l\neq i}^{n}(\psi_{jk}(X_i)-\int \psi_{jk}f)(\psi_{jk}(X_l)-\int \psi_{jk}f)$$

$$+\sum_{j=j_0}^{j_1}\{-\frac{2}{n(n-1)(1+v2^{2js})}+\frac{1}{n^2(1+v2^{2js})^2}\}2(n-1)\sum_k\int \psi_{jk}f.\sum_{i=1}^{n}\psi_{jk}(X_i)$$

Now we invoke the decorrelation property of the wavelet transform with respect to the inner product $\langle g,h\rangle_f := \int g(x)h(x)f(x)dx$ (cf. the proof of Theorem 1). This property, together with the bound on N_j, implies that, in particular, $\sum_k\sum_{k'}(\int \varphi_{j_0k}\varphi_{j_0k'}f)^2 \le C2^{j_0}$ holds, rather than $\le C2^{2j_0}$. Here, again, $C = C(\varphi)$. This observation is implicitly used in Tribouley (1995), p.49. Applying the above observation, let us evaluate the order of magnitude of

$$B_1 = -\frac{n+1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{l=1,l\neq i}^{n}h(X_i,X_l)$$

with

$$h(x,y) = \sum_k\{\varphi_{j_0k}(x) - \int \varphi_{j_0k}f\}\{\varphi_{j_0k}(y) - \int \varphi_{j_0k}f\}.$$

Note that

$$Eh(X_i,X_l) = E\big(h(X_i,X_l)|X_i\big) = E\big(h(X_i,X_l)|X_l\big) = 0 \qquad (A11)$$

holds. Invoking (A11), the decorrelation property and the Cauchy-Schwartz inequality, we obtain like in Tribouley (1995), p.49 that $Var(B_1) \le C\frac{2^{j_0}}{n^2}$ and, hence, $P\{B_1 MISE(v)^{-1} > \epsilon\} \le \epsilon^{-2}C2^{-j_0}$. Let now $B_2 = \sum_{j=j_0}^{j_1} B_{2j}$,

$$B_{2j} = \{-\frac{2}{n(n-1)(1+v2^{2js})}+\frac{1}{n^2(1+v2^{2js})^2}\}\sum_k\sum_{i=1}^{n}\sum_{l=1,l\neq i}^{n}(\psi_{jk}(X_i)-\int \psi_{jk}f)(\psi_{jk}(X_l)-\int \psi_{jk}f).$$

Applying the quasi-triangle inequality to the sum in j and continuing by the same type of argument as for B_1, one obtains

$$E(B_2^2) = VarB_2 \le 2(j_1 - j_0 + 1)\sum_{j=j_0}^{j_1}E(B_{2j}^2) \le C(f,\psi)j_1\frac{2^{j_0}}{n^2}[1+\sum_{j=j_0}^{j_1}\frac{2^{j-j_0}}{(1+v2^{2js})^2}],$$

and, on applying Chebyshev inequality,

$$P\{B_2 MISE(v)^{-1} > \epsilon\} \leq \epsilon^{-2} C j_1 2^{-j_0} = C(f, \psi, \epsilon) 2^{-j_0} . o(\log_2 n) \to 0,$$

by the choice of j_0.

Next, we show that under the theorem's assumptions

$$Var[\sum_{j=j_0}^{j_1} \{-\frac{2}{n(n-1)(1+v2^{2js})} + \frac{1}{n^2(1+v2^{2js})^2}\} 2(n-1) \sum_k \int \psi_{jk} f . \sum_{i=1}^{n} \psi_{jk}(X_i)] =$$

$$= Var(D_1) \leq \frac{C2^{-2j_0 s'}}{n} . \gamma_{j_0}$$

where $\sum_{j=j_0}^{\infty} \gamma_j^2 < \infty$. To prove this, denote $c_{j,n}(v) = [-\frac{2}{n(n-1)(1+v2^{2js})} + \frac{1}{n^2(1+v2^{2js})^2}] 2(n-1)$. It is easy to see that $|c_{j,n}(v)| < 4/n$, uniformly in v, s and j.

After computations,

$$Var(D_1) = n \sum_{j=j_0}^{j_1} \sum_{j'=j_0}^{j_1} c_{j,n}(v) c_{j',n}(v) \sum_k \sum_{k'} \beta_{jk} \beta_{j'k'} \int \psi_{jk} \psi_{j'k'} f - n(\sum_{j=j_0}^{j_1} c_{j,n}(v) \sum_k \beta_{jk}^2)^2 \leq$$

$$\leq n \sum_{j=j_0}^{j_1} \sum_{j'=j_0}^{j_1} c_{j,n}(v) c_{j',n}(v) \sum_k \sum_{k'} \beta_{jk} \beta_{j'k'} \int \psi_{jk} \psi_{j'k'} f.$$

Now we apply consecutively a change of the order of summation and integration, Hölder's inequality taking into consideration that $f \in L_\infty$, Parseval's identity and the upper bound for $|c_{j,n}(v)|$ to obtain

$$Var(D_1) \leq n \int (\sum_{j=j_0}^{j_1} c_{j,n}(v) \sum_k \beta_{jk} \psi_{jk})^2 f \leq n \parallel f \parallel_{L_\infty} \sum_{j=j_0}^{j_1} \sum_k c_{j,n}(v)^2 \beta_{jk}^2 \leq$$

$$\leq \frac{16}{n} \parallel f \parallel_{L_\infty} \sum_{j=j_0}^{j_1} \sum_k \beta_{jk}^2 = \frac{16}{n} \parallel f \parallel_{L_\infty} \parallel \sum_{j=j_0}^{j_1} \sum_k \beta_{jk} \psi_{jk}(.) \parallel_{L_2}^2 .$$

By using a theorem due to Kerkyacharian and Picard (compare the citation in Theorem 1 in Tribouley (1995)) we can now claim that

$$Var(D_1) \leq \frac{C2^{-2j_0 s}}{n} . \gamma_{j_0},$$

where $C = C_s \parallel f \parallel_{L_\infty}$, C_s depending on s only, and where $\{\gamma_{j_0}\}_{j_0=0}^{\infty} \in l_2$ and, therefore, $\gamma_{j_0} \to 0$ as $j_0 \to \infty$, hence $P(D_1 MISE(v)^{-1} > \epsilon) \leq \epsilon^{-2} C 2^{-2j_0(1+s')} n \gamma_{j_0}$. Having in mind that $2^{-j_0} = O(n^{-\frac{1}{2+2s'}})$, we see that $\frac{D_1}{MISE(v)}$ also tends to zero in probability.

It remains to evaluate A_1 and A_2. A_1 is a sum of i.i.d. random variables and their variances can be bounded from above. Using again Meyer's lemma and applying the decorrelation argument (see also Tribouley (1995), p.50) yields

$$Var(A_1) = E(A_1^2) \leq \frac{C2^{2j_0}}{n^3}.$$

The argument about A_2 is analogous to the one with B_2, with additional invoking of the inequality between the l_2 and l_1-norm:

$$Var(A_2) = E(A_2^2) \leq 2(j_1 - j_0 + 1)[\sum_{j=j_0}^{j_1} E(A_{2j}^2)] \leq \frac{C}{n^3}.j_1.2^{2j_0}.\sum_{j=j_0}^{j_1} \frac{2^{2(j-j_0)}}{(1+v2^{2js})^4} \leq$$

$$\leq \frac{C}{n^3}.j_1.2^{2j_0}.[\sum_{j=j_0}^{j_1} \frac{2^{(j-j_0)}}{(1+v2^{2js})^2}]^2$$

for a suitably chosen $C = C(f, \psi) > 0$. Hence, by virtue of (A9), using the Chebyshev inequality, we can claim that $P\{A_1 MISE(v)^{-1} > \epsilon\} \leq \epsilon^{-2}\frac{C}{n}$, $P\{A_2 MISE(v)^{-1} > \epsilon\} \leq \epsilon^{-2}\frac{C}{n}o(\log_2 n)$.

Finally, let us note that $E(T_n) = \int f(x)^2 dx$ holds. Neglecting negative summands, interchanging the order of integration and summation and applying Hölder's inequality and Parseval's identity, we bound the variance of T_n from above:

$$Var(T_n) = \frac{4(n+1)^2}{n^3}[\sum_k \sum_{k'} (\int \varphi_{j_0k}\varphi_{j_0k'} f).\alpha_{j_0k}\alpha_{j_0k'} - (\sum_k \alpha_{j_0k}^2)^2] \leq$$

$$\leq \frac{4(n+1)^2}{n^3} \int (\sum_k \alpha_{j_0k}\varphi_{j_0k})^2 f \leq \frac{4(n+1)^2}{n^3}(\sum_k \alpha_{j_0k}^2)||f||_{L_\infty} \leq \frac{4(n+1)^2}{n^3}.||f||_{L_2}^2 ||f||_{L_\infty}.$$

Thus, we arrive at $Var(T_n) \leq \frac{C}{n}$ and $T_n \xrightarrow{P} \int f(x)^2 dx$ as $n \to \infty$, by Chebyshev inequality. □

Proof of Theorem 7. The idea of the proof of the necessity part is essentially the same as the necessity part of Theorem 3. The admissibility of the choice $v = \tilde{v}_n$ in the necessity part of the theorem follows from the boundedness of $\tilde{v}_n$ proved in Theorem 5. We outline the proof of the sufficiency part. Proceeding first analogously to the proof of the sufficiency part of Theorem 5, and then applying the Cauchy-Schwartz inequality and interchanging the order of summation and integration, we obtain

$$E \parallel \tilde{f}_v - f \parallel_{B_{22}^\sigma}^2 = E \parallel \tilde{f}_v - E\tilde{f}_v \parallel_{B_{22}^\sigma}^2 + \parallel E\tilde{f}_v - f \parallel_{B_{22}^\sigma}^2 =$$

$$= \frac{1}{n}\sum_k [(\int \varphi_{j_0k}^2 f) - \alpha_{j_0k}^2] + \frac{1}{n}\sum_{j=j_0}^{j_1} \frac{2^{2j\sigma}}{(1+v2^{2js})^2} \sum_k [(\int \psi_{jk}^2 f) - \beta_{jk}^2] +$$

$$+ \sum_{j=j_0}^{j_1} 2^{2j\sigma} \frac{v^2 2^{4js}}{(1+v2^{2js})^2} \sum_k \beta_{jk}^2 + \sum_{j=j_1+1}^{\infty} 2^{2j\sigma} \sum_k \beta_{jk}^2 \leq$$

$$\leq \frac{1}{n}\int[(\sum_k \varphi_{j_0k}^2)f] + \frac{1}{n}\sum_{j=j_0}^{j_1} 2^{2j\sigma}\frac{v^2 2^{4js}}{(1+v2^{2js})^2}\int[(\sum_k \psi_{jk}^2)f]+$$

$$+\sum_{j=j_0}^{j_1} 2^{2j\sigma}\frac{v^2 2^{4js}}{(1+v2^{2js})^2}\sum_k \beta_{jk}^2 + \sum_{j=j_1+1}^{\infty} 2^{2j\sigma}\sum_k \beta_{jk}^2.$$

Bounding each of the last two terms from above in the same way as in the proof of Theorem 3, and applying Meyer's lemma to each of the first two terms, we obtain

$$E \parallel \tilde{f}_v - f \parallel_{B_{22}^\sigma}^2 =\leq c(\varphi)\frac{2^{j_0}}{n} + \frac{c(\psi)}{n}\sum_{j=j_0}^{j_1}\frac{2^{j(1+2\sigma)}}{(1+v2^{2js})^2} + (v^{\frac{s'-\sigma}{s}} + c_{s',\sigma}2^{-2j_1(s'-\sigma)})||f||_{B_{22}^{s'}}^2 \leq$$

$$\leq c_{s',\sigma}(f,\varphi,\psi)(\frac{2^{j_0}}{n} + \frac{2^{j_1(1+2\sigma)}}{n} + v^{\frac{s'-\sigma}{s}} + 2^{-2j_1(s'-\sigma)}) = o(1),$$

when $2^{j_1} = o(n^{\frac{1}{1+2\sigma}})$, $j_0 \leq j_1$, $j_1 \to \infty$ and $v \to 0$.

The admissibility of $\tilde{v}_n$ in the sufficiency part follows from Theorem 8. □

Proof of Theorem 8. (*Outline.*) We follow closely the idea of proof of Theorem 4. In this case, the resulting equation for the upper bound v_0 is

$$v = \frac{\parallel f \parallel_{L_\infty}}{c_0^2}.\frac{1}{n}.\sum_{j=j_0}^{j_1} 2^{2(j-j_0)s}\Big(\frac{1+v2^{j_0 s}}{1+v2^{2js}}\Big)^3.2^j, \tag{A12}$$

where c_0 is the sharpness constant. The proof can be completed by invoking the bound $v < C$ again and proceeding as in the proof of Theorem 4. However, in this case, due to the unbiased estimation of the wavelet coefficients when the design is random, we can directly obtain the quantitative rates of $\tilde{v} \to 0$, for any $s' > 0$. In fact, *for any* $j_0, j_1 : j_0 \leq j_1$ we can obtain the *sharp rate* in $\tilde{v} = o(1)$ corresponding to these j_0 and j_1 in the following way: instead of utilizing the rough bound $v \leq C$, we replace the RHS of (A12) by

$$\frac{\parallel f \parallel_{L_\infty}}{c_0^2}.\frac{1}{n}.\sum_{j=j_0}^{j_1} 2^{2(j-j_0)s}\Big(\frac{\max(1, v2^{j_0 s})}{\max(1, v2^{2js})}\Big)^3.2^j.$$

The new RHS of (A12) is equivalent to the old one, with absolute equivalence constants. From the new expression, the sharp rate can be computed by considering the three cases: (i) $v > 2^{-2j_0 s}$; (ii) $v \leq 2^{-2j_1 s}$; (iii) $2^{-2(j'+1)s} \leq v < 2^{-2j's}$, $j' \in \mathbf{Z} : j_0 \leq j' < j_1$. By solving (A12) for v for each of the cases (i-iii) and checking whether the solution v_0 falls into the range of the respective case, we are able to eliminate two of the cases, and accept the remaining third case. If case (iii) is the accepted one, the substitution $v = 2^{-j's}$, where $j' = (1-\theta)j_0 + \theta j_1$, allows locating j' within $[j_0, j_1)$ by solving the resulting equation for $\theta \in [0,1)$. □ ^{B14}

Proof of Corollary 2. Theorem 8 and the sufficiency part of Theorem 7 are applied for $v = \tilde{v}$ and $\sigma = 0$, and then the proof is completed via Theorem 6. $\square$

Proof of Corollary 3. The rate $v_0 = O(\frac{2^{j_1}}{n})$ follows immediately from (A12) with $j_0 = j_1$ and, therefore, $cn^{\frac{1}{1+2s}} \le 2^{j_1} \le Cn^{\frac{1}{1+2s}}$ and $\tilde{v} \le v_0$ together imply $\tilde{v} = O(n^{-\frac{2s}{1+2s}})$. The proof of Lemma 4 shows that, for this choice of j_1, $E \parallel \hat{f} - f \parallel^2_{B^\sigma_{22}} = O(n^{-\frac{2(s-\sigma)}{1+2s}})$. Now it follows from the proof of Theorem 7 that $E \parallel \tilde{f}_{\tilde{v}} - f \parallel^2_{B^\sigma_{22}} = O(n^{-\frac{2(s-\sigma)}{1+2s}})$. $\square$

A19. (See also Dechevsky and Penev (1998a).) For the purpose of optimizing the GFCV- (FCV-) functional, the excellent subroutine `amoeba` from Numerical Recipes has been utilized. It is based on the downhill simplex method of Nelder and Mead. Since the minimization is performed only once, without iterations, so the computational burden is small, the above procedure has been selected because of its simplicity and robustness. It is also derivative-free and we intend to use this for some of the extensions in Appendix B.

A20. There follow some model examples where the universal-threshold and global-penalization shrinking strategies are *not* expected to yield equally satisfactory estimation of every part of the curve: Marron and Wand's "smooth comb" (see Penev and Dechevsky (1997)), Examples 3 and 4, and, on fractal level, any *multifractal* function (Arneodo, Bacry, Jaffard and Muzy (1997)).

A21. Notice the bad performance of our non-threshold estimator near $x = 1$, where the true signal is negligible compared to the noise level and the thresholded estimator is at its best. One optional modification of the shrinking estimator in the case of regression is to threshold it (after shrinking) with hard threshold equal to the noise variance of a purely noisy coefficient. This threshold level (typically, $\frac{2\delta}{\sqrt{n}}$), is very low compared to levels used in threshold methods, and is aimed at removing pure noise for moderate samples. This is a most primitive example of a composite shrinking/thresholding estimator which will be discussed in more detail in ^{B8}.

APPENDIX B: FURTHER EXTENSIONS AND GENERALIZATIONS

In Section 7 we discussed some extensions of the K-functional technique of penalized wavelet estimation. Here we give a (non-exhaustive) list of further developments and applications, together with short descriptions of some underlying ideas. Again, we assume that $a_1 = 0,\ a_2 = 1$.

B1. *Cross-validation in Besov and potential spaces.* There is a common opinion expressed in literature that cross-validation in L_2 may in some cases lead to a slightly underestimated value of the smoothing parameter leading to a slight tendency to overfit the curve. Analytically, this can be explained with the fact that in general L_2-cross validation leads to consistent estimation of f itself, but not of its fractional derivatives of any positive order. This can be overcome by considering cross-validation with respect to the B^σ_{22}-norm such that available results for L_2-cross validation correspond to the partial case $\sigma = 0$. The key fact for this extension is to notice that L_2-cross validation, as performed in the density case (see subsection 6.2) is of Bowman–Rudemo type and is equivalent to applying the cross-validation rule $\tilde{\alpha}_{j_0k}\alpha_{j_0k} \mapsto \frac{1}{n}\sum_{i=1}^{n} \tilde{\alpha}_{j_0k(-i)}\varphi_{j_0k}(X_i)$,

$\tilde{\beta}_{jk}\beta_{jk} \mapsto \frac{1}{n}\sum_{i=1}^n \tilde{\beta}_{jk(-i)}\psi_{jk}(X_i)$, for every $(j,k):\ j_0 \le j \le j_1,\ \mathrm{supp}\varphi_{j_0k}\cap\mathrm{supp}f \ne \emptyset,\ \mathrm{supp}\psi_{jk}\cap \mathrm{supp}f \ne \emptyset$, where the notations are the same as in Subsection 6.2 (recall that the $\tilde{\beta}$'s depend on v). This leads to optimization in v of the CV-criterion

$$CV_\sigma(v) = CV_{\sigma,s,j_0}(v) =$$

$$= \sum_k \tilde{\alpha}_{j_0k}(v)^2 + \sum_{j=j_0}^{j_1} 2^{2j\sigma}\sum_k \tilde{\beta}_{jk}(v)^2 - \frac{2}{n}\sum_{i=1}^n (\sum_k \tilde{\alpha}_{j_0k(-i)}\varphi_{j_0k}(X_i) + \sum_{j=j_0}^{j_1} 2^{2j\sigma}\sum_k \tilde{\beta}_{jk(-i)}\psi_{jk}(X_i)),$$

with $CV_0(v) = M(\tilde{f}_v)$. For $\sigma > 0$ $CV_\sigma(v)$ is the cross-validated version of $\| \tilde{f}_v - f \|^2_{B^\sigma_{22}}$ up to summands independent of v. For the regression case we suggest *the same Bowman–Rudemo approach*, based on the formulae $\tilde{\alpha}_{j_0k}\alpha_{j_0k} \mapsto \frac{1}{n}\sum_{i=1}^n \tilde{\alpha}_{j_0k(-i)}y_i\varphi_{j_0k}(x_i)$, $\tilde{\beta}_{j_0k}\beta_{j_0k} \mapsto \frac{1}{n}\sum_{i=1}^n \tilde{\beta}_{j_0k(-i)}y_i\psi_{j_0k}(x_i)$. This yields

$$\tilde{CV}_\sigma(v) = CV_{\sigma,s,j_0}(v) =$$

$$= \sum_k \tilde{\alpha}_{j_0k}(v)^2 + \sum_{j=j_0}^{j_1} 2^{2j\sigma}\sum_k \tilde{\beta}_{jk}(v)^2 - \frac{2}{n}\sum_{i=1}^n (\sum_k \tilde{\alpha}_{j_0k(-i)}y_i\varphi_{j_0k}(x_i) + \sum_{j=j_0}^{j_1} 2^{2j\sigma}\sum_k \tilde{\beta}_{jk(-i)}y_i\psi_{jk}(x_i)),$$

where the $\tilde{\alpha}$'s and $\tilde{\beta}$'s have meaning as of Subsection 6.1. For $\sigma = 0$ (see Dechevsky and Penev (1998a)) this is an important and, as it seems, quite new alternative to Wahba's model, with "pointwise" estimation (i.e., estimation of the linear interpolation functionals $L_i(f) = f(x_i)$) being replaced by "weak" estimation (i.e., estimation of the linear interpolation functionals $f \mapsto \alpha_{j_0k}$ and $f \mapsto \beta_{jk}$). Besides admitting an extension for $\sigma > 0$ in a natural way, the Bowman–Rudemo type approach to the regression case has some additional advantages compared to the model in Subsection 6.1: (a) it is available also for the model where $x_i = X_i,\ i = 1,\dots,n$, are i.i.d. random variables, uniformly distributed in a neighbourhood of $\mathrm{supp}f$; (b) this model admits *analogous theory of iterative individual penalized shrinking*, as considered in Section 7 for the density case.

It should be noted that in the case of deterministic design $\{x_i\}_{i=1}^n$ theoretical derivations are easier to carry through if the *periodized* (Meyer (1991), Dahmen (1997)) wavelets $\varphi^{per}_{j_0k}$ and ψ^{per}_{jk} are considered. (Computing the α_{j_0k}'s and the β_{jk}'s in (2,3) with respect to $\varphi^{per}_{j_0k}$ and ψ^{per}_{jk} can be shown to yield an equivalent quasi-norm of $J_{spq}(\alpha,\beta)$.)

As for the model in Section 6.1, it can be extended for $\sigma > 0$ by utilization of the fact that B^σ_{22} is isomorphic to the potential (Sobolev) space H^σ_2 (see Bergh and Löfström (1976), 6.4.4). The CV-functional is now the expectation of a quadrature formula for the integral in $\| \tilde{f}_v - f \|^2_{H^\sigma_2} = \int [G_\sigma * (\tilde{f}_v - f)(x)]^2 dx$, where G_σ is the Bessel-McDonald kernel[A1]. The order of the quadrature formula can be higher than one. (In comparison, the optimization criteria of GCV type like QGCV and GFCV are essentially quadrature formulae of first order.) The higher order of the quadrature formula is helpful in reducing the bias for more smooth functions, and plays essentially the same role as the order of vanishing moments of the coiflet used in the method WAVREG of Antoniadis (1996). This approach is more technically involved because it includes regularization of the singular kernel G_σ depending on the sample size n. We conjecture that the analogues of $\bar{h}(v)^2$ and $\mu(v)$ (see (14)) are in the new case the norms in *weighted* (both with the same weight) Schatten–von Neumann S_γ-norms, $\gamma = 2, 1$, respectively, the weight depending on σ and tending to 1 as $\sigma \to 0+$.

The results in Section 6 show that consistency of cross validation for $\sigma > 0$ is sufficient for consistency of the resulting estimator $\tilde{f}_v$ in B^σ_{22}, i.e., with better estimation of the derivatives and less overfitting of the curve. One way to determine a best selection of σ for the concrete f, if simultaneous optimization in v and s is performed, is to increase σ (e.g., by using the dyadic-bisection trial-and-error method) until for the optimal value $s^* = s^*(\sigma)$ the equality $s^*(\sigma) \approx \sigma$ holds with sufficient precision. For $f \notin C^\infty$ this is possible to achieve because $s^*(\sigma)$ remains in a neighbourhood of the true smoothness index s' of the curve for any $\sigma > 0$. Another approach here, for both the regression and density case, is to consider a *relative cross-validation criterion*:

$$CW_\sigma(v) = CW_{\sigma,s,j_0}(v) = (b^\sigma_{22}(\hat{f}))^{-1} CV_\sigma(v),$$

where $b^\sigma_{22}(\hat{f}) = b^\sigma_{22}(\hat{f})_{j_0,j_1} = \| \hat{\alpha}_{j_0 \cdot} \|_2^2 + \sum_{j=j_0}^{j_1} (2^{j\sigma} \| \hat{\beta}_{j \cdot} \|_2)^2$ (cf. (2)). The optimal value of σ is obtained by $CW_{\sigma^*} = \min_{0 \le \sigma \le s^*} CW_{\sigma,s^*,j_0^*}(v^{opt})$, where s^*, j_0^* and v^{opt} are obtained as discussed in Remarks 5 and 7, for the regression and density case, respectively. An interesting modification is to minimize $CW_{\sigma,s,j_0}(v)$ simultaneously in σ, s, j_0 and v.

In the case of regression with deterministic design, the method of Antoniadis (1996) can also be extended, similarly to the Bowman-Rudemo approach, so as determine the optimal value of the smoothing parameter on the basis of an estimate of the upper bound for $||\tilde{f}_v - f||_{B_{22\sigma}}$ for $0 \le \sigma \le s$, $j_1 : 2^{j_1} = o(n)$, which bound is the RHS of

$$||\tilde{f}_v - f||_{B_{22\sigma}} \le ||(I - P_{j_1})f||^2_{B^\sigma_{22}} + 2||(P_{j_1} f - \Pi_{j_1})f||^2_{B^\sigma_{22}} + 2||(\Pi_{j_1} f - \tilde{f}_v||^2_{B^\sigma_{22}} = RHS,$$

where P_{j_1} is the orthogonal projection of f onto V_{j_1}, and Π_{j_1} is defined in Antoniadis (1996), p.318. An unbiased estimate $CV\sigma, s, j_0(v)$ of $||\Pi_{j_1} f - \tilde{f}_v||_{B^\sigma_{22}}$ can be obtained for $\sigma > 0$ in the way it has been proposed by Antoniadis (1996) for the case $\sigma = 0$. The use of coiflets makes this approach quite promising, especially when the estimated function $f \in B^{s'}_{22}$, $1/2 < \sigma + 1/2 < s' \le s < r$, is very smooth (i.e., $s' \gg 1$) and $\sigma \approx s' - 1/2$. However, the support of the coiflet is larger than that of the respective Daubechies' minimal-support wavelet (see Daubechies (1992)), which reduces the adaptivity of the method for spatially inhomogeneous curves, especially, when the true smoothness index s' does not exceed 1. In order to improve the spatial adaptivity of the method, we suggest to replace the coiflets with 'multi-coiflets', that is, multi-scaling functions $^{B18(c)}$ which are orthogonal to the same monomials as the coiflets and have approximately the same regularity, but are better spatially localized. The so proposed generalization of WAVREG of Antoniadis (1996) (with the additional improvement that j_1 is not fixed, but controlled, together with j_0 and with a coiflet or 'multi-coiflet') will be referred to as $WAVREG_\sigma$ later in this text. As far as determining an optimal value of σ^* (based on $WAVREG_\sigma$) is concerned, it does not seem possible to develop a relative optimization criterion of the type of $CW_\sigma(v)$ defined for the Bowman-Rudemo approach. However, the first method considered above (i.e., finding a fixed point of the mapping $\sigma \mapsto s^*(\sigma)$) is still available.

Finally, let us note that in the models of cross validation in B^σ_{22} with subsequent optimal choice of σ it is a good idea to optimize the criterion not only in j_0, but also in j_1. (For the usual L_2-cross validation optimization in j_1 is not advisable, because the optimal j_1 would tend to be too large most of the time, thus leading to overfitting the curve.)

B2. *Asymptotic-minimax theory in Besov spaces.*[B9] As with thresholded estimators, it can be expected that in the theoretical study of the asymptotic-minimax optimal choice of the smoothing parameter exhaustive quantitative results can be obtained . We expect that such exhaustive results can be obtained for the level-dependent shrinking estimator $\tilde{f}_{\mathbf{v}}$, $\mathbf{v} = (v_{j_0}, \dots, v_{j_1})'$, for any (π, u, σ) and (p, q, s) such that $B^\sigma_{\pi u} \hookleftarrow B^s_{pq}$. Although not identical, the proof in the shrinking case can be based on the same basic ideas as the one for the threshold case. We do not expect major technical challenges in obtaining asymptotic-minimax results analogous to the results of Delyon and Juditsky (1996), i.e., including optimal rates of estimation of the derivatives of f, too. However, our aim will be to remove some of the logarithmic factors in the rates and to obtain *more precise bounds for the constant factors of the rates*, depending on the concrete choice of (π, u, σ) and (p, q, s).

B3. *Kernel regularization via quasi-linearization of K-functionals.* Following Peetre (1963), we call the pair of quasi-normed spaces A, B quasi-linearizable if there exists a family of (sub)linear (see Bergh and Löfström (1976), 3.10, 3.11, and Dechevski (1988), Section 5) operators $V(t) : A + B \to A + B$ such that $V(t)(A + B) \subset B$, $(I_{A+B} - V(t))(A + B) \subset A$ and

$$c[\| (I_{A+B} - V(t))f \|^\rho_A + t^\rho \| V(t)f \|^\rho_B]^{1/\rho} \leq K_\rho(t, f; A, B) \leq$$

$$\leq C[\| (I_{A+B} - V(t))f \|^\rho_A + t^\rho \| V(t)f \|^\rho_B]^{1/\rho}$$

for any $f \in A + B$ and any $t \in (0, \infty)$. Here $0 < c \leq C < \infty$ and c, C may depend on A, B and $\rho : 0 < \rho \leq \infty$, but are independent of f and t.

Let $K(x)$ be the kernel associated with any r-th order kernel estimator, that is, $\int K(x)dx = 1$ and $\int K(x)x^l dx = 0$, $l = 1, \dots, [r]$ (with obvious modifications in the multivariate case, when l becomes a multiindex). Denote $K_t = K(\frac{\cdot}{t})$. Then, utilizing the real and complex interpolation methods (see Section 4), it can be shown that $(B^\sigma_{pu}, B^s_{pq})$ and $(B^\sigma_{pu}, \dot{B}^s_{pq})$ are quasi-linearizable pairs with $V(t) : f \mapsto K_{t^{\frac{1}{s-\sigma}}} * f$, uniformly in $0 < p \leq \infty$, $0 < u \leq \infty$, $0 < q \leq \infty$, $\max\{0, d(\frac{1}{p} - 1)\} < \sigma < s < r$, where d is the dimension. Here c and C depend on the concrete choice of K, p, q, u, σ, s and ρ. The same is true if Besov spaces are replaced by potential (Sobolev) spaces for the same values of the parameters σ, s and $p : 1 \leq p \leq \infty$. A similar result holds for wavelet estimators with ψ having vanishing first $[r]$ moments. In fact, for the case $p = u = q = 2$ and $\rho = 2$ the β-coefficients of $V(t)f$ are computed from the coefficients of f via (7) (for other values, see ^{B9}). These general quasi-linearization results show that the role of the parameter $v = t^{\rho}$ is that of a smoothing parameter in a penalization model induced by the K-functional and, simultaneously, $v^{\frac{1}{\rho(s-\sigma)}}$ has the role of a kernel estimator's bandwidth. From this point of view, the difference between kernel and wavelet estimators is mainly in the choice of the type of equivalent quasi-norms in the Besov spaces involved in the K-functional. In fact, rigorous results can be obtained, linking the optimal smoothing parameter $\tilde{v}$ for the wavelet model with the bandwidth v^* of a kernel estimator. The relation is very simple: $c_1\tilde{v} \leq v^* \leq C_1\tilde{v}$, where c_1 and C_1 depend only on the constants of equivalence between the K-functional, involving the coefficient Besov quasi-norms, and its quasi-linearized version involving the kernel K. Taking into account that under very general assumptions $\tilde{v} = \tilde{v}_n \to 0$ as the sample size $n \to \infty$, we see that a "good bandwidth" v^* can be found by only searching within $[c_1\tilde{v}_n, C_1\tilde{v}_n]$, which interval becomes narrower with the increase of n. Certainly this technique can be applied also for variable-kernel estimators, provided that one starts from the block-shrinking wavelet estimator (see also Examples 3 and 4). Of course, rigorous formulations depend on the specific statistical model (density estimation, regression-function estimation (Wahba's model or the Bowman–Rudemo approach), etc.) but in all cases the underlying general idea is as outlined above. In particular, one or both spaces in the K-functional may depend on its step t: $K(t, f) = K(t, f; A_t, B_t)$ (see, e.g. ^{B10} and Dechevski (1988)).In the most practically important statistical applications one faces the situation $K(t, f) = K(t_n, \hat{f}_n; A_n, B_n)$, where n and $\hat{f}_n$ are, as usual, the sample size and the empirical estimator, respectively. In these more general situations t may still be equivalent to the bandwidth of a kernel estimator, but the simple rescaling $t \mapsto t^{\frac{1}{s-\sigma}}$ considered above may be replaced now by the more general $t \mapsto w(t)$, where $w(.)$ is a warping function which depends on the type of dependence of A and/or B on the parameter. For some cases of dependence of A and B on the parameter, it may happen that $w(.)$ is not right- or left-invertible, which may be an indication that the dependence of the spaces A and/or B on the parameter takes control over the smoothing process, the penalization parameter t playing then the less significant role. The typical case when A and B depend on n is $j_0 \to \infty$, $j_1 = o(\log_2 n)$.

B4. *Kernel estimation via solutions of evolutionary differential equations.* Here we outline a new, very general approach to kernel estimation is statistics. This approach falls into the general quasi-linearization scheme considered in ^{B3}, but the method of generating the kernel estimator is very different and seems to be complementary to traditional kernel-estimation techniques.

We return to the considerations in Section 4 leading to formula (5). In the notation there, assume that $B_1 = A$ where A is a Banach space. To conform with commonly accepted notation, we denote by $B = D(T)$ the domain of T (see Bergh and Löfström (1976), Section 6.7), where

T has the properties assumed in Section 4. Consider the following Cauchy problem for the evolutionary differential equation in the Banach space A

$$\frac{d}{dt}g(t) = Tg(t),\ g(0) = a \in A,\ t > 0. \qquad (B1)$$

To ensure smoothing effect, $\| g(t_2) \|_A \leq \| g(t_1) \|_A$, $t_2 > t_1$, should be fulfilled. A necessary assumption to ensure consistency is $\lim_{t\to 0+} \| g(t) - a \|_A = 0$. Both these assumptions are fulfilled if T is such that $g(t) = G(t)a$, where $G(t)$ is an *equibounded, strongly continuous operator semigroup* (ESCOSG) on A (see Bergh and Löfström (1976), Section 6.7, (i-iii)), which, additionally, is *contractive*, that is, in the notations of Bergh and Löfström, $M \leq 1$. In this case T is the infinitesimal generator (IG) of the semi-group $\{G(t),\ t > 0\}$. For necessary and sufficient conditions on T under which T is the IG of a contractive ESCOSG on A, we refer to Dunford and Schwartz (1967), Yosida(1974), Krein (1967), Henry (1981), Hoppenstadt (1969). A simple necessary (but not sufficient) condition is that the spectrum of T be contained in $\{z \in \mathbf{C} : \mathrm{Re}\ z \leq 0\}$. The key result is (see Peetre (1963), Butzer and Berens (1967)): if $\{G(t),\ t > 0\}$ is an ESCOSG on A with IG T and $\nu \in \mathbf{N}$ is fixed, then for any $\rho \in (0, \infty]$ there exist $c_\nu = c_\nu(\rho),\ C_\nu = C_\nu(\rho) :\ 0 < c_\nu \leq C_\nu < \infty$, such that

$$c_\nu K_\rho(t, a; A, D(T^\nu)) \leq \omega_\nu(a; t)_A + \min(1, t) \| a \|_A \leq C_\nu K_\rho(t, a; A, D(T^\nu)), \qquad (B2)$$

for any $t > 0$ and any $a \in A$, where $\omega_\nu(a; t)_A = \sup_{0<s\leq t} \| (G(s) - I_A)^\nu a \|_A$. We see from (B2) that the parameter t plays a twofold role again: on the one side, it is the smoothing parameter in a penalization model; on the other, it denotes the length of the "time interval" in which a has been subject to smoothing via the process described by (B1).

Let us consider one simple example. Assume that $A \in L_2$, $T_1 = k^2 \frac{d^2}{dx^2}$, $k \in \mathbf{R}$, $x \in \mathbf{R}$, $\nu = 1$, $\sigma = 0$, $s = 2 = \rho$, $a = f \in L_2(\mathbf{R})$, T is the closed self-adjoint densely defined extension of T_1 on $L_2(\mathbf{R})$. In other words, (B1) is the Cauchy problem for the unidimensional heat and diffusion equation in homogeneous media with conductivity coefficient k^2. T is the IG of a contractive ESCOSG on $L_2(\mathbf{R})$ Then T has only continuous spectrum, and it is contained in $\{z : \mathrm{Re} z \leq 0\}$ (if $\mathrm{supp} f \subset [0, l]$, $l < \infty$, then we can consider $L_2([0, l])$ instead of $L_2(\mathbf{R})$ and the spectrum of T is discrete; it consists of the simple eigenvalues $\lambda_\mu = -k^2\pi^2\mu^2/l^2$). Then, by (5), $K_2(t, a; A, D(T^\nu))$ is equivalent to $K_2(\sqrt{v}, f; B^0_{22}, \dot{B}^s_{22}) + \min(1, \sqrt{v}) \| f \|_{B^0_{22}}$, with equivalence constants independent of $t = \sqrt{v}$ and $a = f$. Besides, $[G(t)f](x) = K_{t^{1/s}} * f(x)$, $x \in \mathbf{R}$, where $K(\xi) = \frac{|k|}{2\sqrt{\pi t}} e^{-\frac{\xi^2}{4k^2}}$, and we observe that this is indeed a partial case of the kernel-estimation model considered in B3. This last observation is in fact correct under quite general assumptions on T in (B1) (see, e.g., the example considered in Bergh and Löfström (1976), Section 6.7, pp. 157,158, and Theorem 6.7.4., as well as Thomée (1969,1971)).[B21]

This new approach to kernel estimation is of considerable practical importance. Indeed, in the last several decades an extensive theory and rich variety of software products has been developed for the solution of evolutionary differential equations via *finite-element* and *boundary-element* methods (FEM and BEM). The new approach described here allows applying the full computational power of FEM and BEM to statistical kernel-estimation, which is of particular benefit in the multivariate case.[B15,B21]

B5. *Estimation with linear constraints.* The procedure of shrinking the wavelet coefficients described by (7) can be generalized in a natural way if the K-functional penalization model considered in Section 5 be upgraded to take into account additional information about constraints of: (a) *interpolatory* type: $f^{(\nu)}(\xi_\mu) = c_{\mu\nu}$, $\nu = 0, \ldots, \nu_\mu$, $\mu = 0, \ldots, m$, where $\xi_{\mu_1} \neq \xi_{\mu_2}$ for $\mu_1 \neq \mu_2$; (b) *isoperimetric* type: e.g., $\int f^{(\nu)}(x) g_\nu(x) dx = c_\nu$, $\nu = 0, \ldots, \nu_0$, where $g_\nu(x)$ are known functions. If the constraints are to be obeyed *strictly*, then *Lagrange-multiplier* technique can be applied. If only an *approximate* tendency to obey the constraints is being assumed (which is often the case in statistical applications), then *quadratic penalization* is recommended. In the "mixed" case there will be a Lagrange multiplier for each strict constraint and a quadratic penalization term (multiplied by a respective large parameter) for every non-strict constraint. Since all constraints, in both the interpolatory and isoperimetric case, are linear, so will be the system of equations for $\tilde{\alpha}_{j_0 k}$, $\tilde{\beta}_{jk}$ and the eventual Lagrange multipliers. The unique solution of this system provides the analogue of (7) for the so-upgraded model. The value of v utilized in the computations can be chosen as the optimal value obtained by cross validation or by asymptotic-minimax considerations for the unconstrained model (7).

B6. *Iterative application of the K-functional criterion.* One instance of such iterative smoothing was considered in Section 7, related to the individual shrinking model in density estimation. In fact, this approach is of interest for all penalization models discussed in Section 6 and 7. For example, consider the GFCV model in Subsection 6.1. Since only linearity of the estimator is needed for GFCV, after determining $\tilde{f}_{\tilde{v}}$ from $\hat{f}$, on a second iteration, one may replace $\hat{f}$ by $\tilde{f}_{\tilde{v}}$ and, denoting $\tilde{v}_1 = \tilde{v}$, apply GFCV again to obtain $(\tilde{\tilde{f}}_{\tilde{v}_1})_{\tilde{v}_2}$, and so on. In general, the latter estimator will be smoother than $\tilde{f}_{\tilde{v}_1}$. For the case of a single global smoothing parameter v, it can be expected that the iterative process will be quickly convergent (i.e., the iteration number l increasing, $\tilde{v}_l$ will tend quickly to zero.) For the level-dependent and block-shrinking models the iterative process is expected to converge more gradually, leading to a more visible improvement in the adaptivity of the final iterated estimator. For individual shrinking the iterative process may not converge, but an optimal number of iterations can be determined by cross validation (cf. Section 7).

B7. *Estimation with smoothness constraints.* Consider $\tilde{f}_v$, with wavelet coefficients defined via (7), where s is assumed to be fixed. Suppose that for a certain p, q, $s' \leq s$ it is additionally known for the true function f that $\| f \|_{B^{s'}_{pq}} = c_f$ holds, where c_f is a *known* positive constant (there is a modification for the homogeneous model). Then, it is possible to determine a *unique* non-negative value v_n of the smoothing parameter, such that $\| \tilde{f}_{v_n} \|_{B^{s'}_{pq}} = c_f$ for any $n \in \mathbf{N}$.

(If $\| \hat{f} \|_{B^{s'}_{pq}} \leq c_f$, then $v_n = 0$.) Thus, additional information about such a type of smoothness constraint on f proves to be decisive for the choice of the smoothing parameter, and we no longer need additional (cross-validational, asymptotic-minimax optimal, etc.) statistical procedure for determining v. (For the level-dependent model in Section 7, under the additional assumption $q = \infty$, one may determine in a unique way the smallest v_j on every level (i.e., the least regular $\tilde{f}_{\mathbf{v}}$), so that $\| \tilde{f}_{\mathbf{v}} \|_{B^{s'}_{pq}} = c_f$.) A lower (upper) bound on c_f leads to an upper (lower) bound on the smoothing parameter, respectively. In [B9(c)] utilizing a smoothness constraint will be important. In all these cases, the statistical aspect of the model can be transferred to estimating c_f itself, and obtaining v_n from the estimate $\hat{c}_f = \hat{c}_{f,n}$ is then achieved by the above deterministic procedure. In principle, this procedure of smoothing is available for very general type of noise (even correlated).

B8. *Composition and comparison with other consistent estimators.* [B6] provided an example of a composite strategy of smoothing and denoising involving several estimators. Here we discuss two other composite strategies.

(a) *Shrinking by a combined cross-validation/asymptotic-minimax approach.* For $\tilde{f}_v$ defined by (1',7), let $v^*_{1,n}$ be the optimal value obtained for sample size n via cross validation; let $v^*_{2,n}$ be the respective value recommended by asymptotic-minimax theory. Consider $v^*_n = (1-\tau_n)v^*_{1,n} + \tau_n v^*_{2,n}$, where $\tau_n \in [0,1]$. For small to moderate samples, take $\tau_n = 0$; for very large n, choose $\tau_n = 1$; for intermediate n let τ_n have intermediate values. The reason for this type of choice is that for very large n the asymptotic-minimax choice is both computationally cheaper and more reliable than the cross-validated one, most of the time. If, additionally, there is a biased perception of f being a smooth function, then the choice $v^* = \max(v^*_1, v^*_2)$ is recommended; if, on the contrary, f is perceived to have an irregular graph (e.g., fractal), then $v^* = \min(v^*_1, v^*_2)$ can be suggested (cf. also [B11]).

(b) *Combined shrinking/thresholding approach.* This is a modification of the iterative approach[B6] where on some of the iterations thresholding is performed rather than shrinking. The most important case is that of two iterations - one shrinking, one thresholding, or vice versa (in fact, these iterations commute modulo rescaling the shrinking and/or thresholding parameter). Because the strategy is composite, estimation of the smoothness parameter for general types of noise is an open problem. We propose two approaches to solve it. The first one (for arbitrary white or weakly correlated noise, and for density estimation) is by estimation with smoothness constraints[B7]: for $\tilde{f}_v$, as defined via (7), we apply the CV-techniques of Section 6 to obtain an optimal value v^*_0 and optimal s^*. Thus, it is assumed that $f \in B^{s^*}_{22}$, and $\| f \|_{B^{s^*}_{22}}$ is estimated by the coefficient norm $\| \tilde{f}_{v^*_0} \|_{B^{s^*}_{22}}$. For a fixed $v \in [0, v^*_0)$ the β-coefficients of $\tilde{f}_v$ are being thresholded with a threshold value such that the $B^{s^*}_{22}$-norm of the resulting estimator is approximately equal to $\| \tilde{f}_{v^*_0} \|_{B^{s^*}_{22}}$. This is repeated for a mesh of values of v in $[0, v^*_0]$. An optimal value $v^* \in [0, v_0^*]$ is finally selected by a customized criterion.

It is possible to start from the other side, too. Begin by determining an optimal threshold value for the "pure" thresholded estimator, compute from there an estimate for $\| f \|_{B^s_{22}}$ for some s, then decrease the threshold and start trials with shrinking via (7) where $\hat{\alpha}_{j_0 k}$ and $\hat{\beta}_{jk}$ are the coefficients of $\hat{f}$ after thresholding at the current threshold level. Select the optimal proportion between shrinking and thresholding by a customized criterion.

The second approach, in the regression case, for Gaussian white noise, is to consider, in the level-dependent case or blockwise, the composite penalized "shrink/threshold" estimator with a SURE strategy of determining both the smoothing parameter v_j and the threshold level λ_j. Denote by $SURE_{v,\lambda}(v, \lambda, \mathbf{x})$ and $SURE_{\lambda,v}(v, \lambda, \mathbf{x})$ the SURE functionals corresponding to the "first shrink, then threshold" and the "first threshold, then shrink" estimators, respectively. In the notations and under the assumptions of Donoho and Johnstone (1995), subsection 2.3, with $\tau = v_j 2^{2js}$, j - fixed, $t = \lambda_j$, $d = c2^j$ (the number of non-zero empirical coefficients x_i, $i = 1, \ldots, d$ at level j),

$$SURE_{v,\lambda}(\tau, t, \mathbf{x}) = \frac{1-\tau}{1+\tau} d - \frac{2}{1+\tau} . \#\{i : |x_i| \leq t(1+\tau)\} + \sum_{i=1}^{d} [\min(t + \frac{\tau}{1+\tau}|x_i|, |x_i|)]^2,$$

$$SURE_{\lambda,v}(\tau, t, \mathbf{x}) = SURE_{v,\lambda}(\tau, \frac{t}{1+\tau}, \mathbf{x}).$$

The limiting case $\tau = 0$ corresponds to Donoho and Johnstone's SURE(t,$\mathbf{x}$); the other limiting case $t = 0$ - to linear estimators (cf. Li (1985)). Minimizing with respect to both τ and t clearly gives more flexibility than in the limiting cases.

B9. *Shrinking in the general quasi-Banach case.* Formula (7) refers to the case $p = q = \pi = u = 2$, $\sigma = 0$ only. Now we consider the general case, as discussed in the end of Section 4.

(a) The penalization model is via $L(v, \hat{f}; B^\sigma_{\pi\pi}, \dot{B}^s_{pp}) := K_1(v, \hat{f}; (B^\sigma_{\pi\pi})^\pi, (\dot{B}^s_{pp})^p)$, where p, π, s, σ are such that $p < \infty$, $\pi < \infty$ and $B^\sigma_{\pi\pi} \hookleftarrow B^s_{pp}$. (Here A^τ is the space A endowed with the quasi-norm $\| \cdot \|_{A^\tau} := \| \cdot \|^\tau_A$.) From the proof of Theorem 5.5.1 in Bergh and Löfström (1996) it follows that the formula about $\tilde{\beta}_{jk}$ in (7) is now generalized to $\tilde{\beta}_{jk} = \mu_{jk} . \hat{\beta}_{jk}$, where $\mu_{jk}(v)$:

$$N(|1 - \mu_{jk}|^\pi, v2^{j\varepsilon/2} |\hat{\beta}_{jk}|^{p-\pi} |\mu_{jk}|^p) = \min_{\mu \in \mathbf{R}} N(|1-\mu|^\pi, v2^{j\varepsilon/2} |\hat{\beta}_{jk}|^{p-\pi} |\mu|^p), \qquad (B3)$$

$N(x_1, x_2)$ being any fixed norm (or, more generally, any fixed quasi-norm equivalent to any norm) in $\mathbf{R}^2$, and $\varepsilon = 2ps - 2\pi\sigma + p - \pi$ being the *critical-regularity* index, well-known in asymptotic-minimax theory (see Delyon and Juditsky (1996), Theorem 1, for $d = 1$). Then, there exist c, $C : 0 < c \leq 1 \leq C < \infty$, depending on $N(\cdot, \cdot)$ only, such that

$$c.L(v, \hat{f}; B^\sigma_{\pi\pi}, \dot{B}^s_{pp}) \leq \| \hat{f} - \tilde{f}_v \|^\pi_{B^\sigma_{\pi\pi}} + v \| \tilde{f}_v \|^p_{\dot{B}^s_{pp}} \leq C.L(v, \hat{f}; B^\sigma_{\pi\pi}, \dot{B}^s_{pp}),$$

for any $v > 0$ and any $\hat{f} \in B^{\sigma}_{\pi\pi} + \dot{B}^{s}_{pp}$. Moreover, $c = C = 1$ if and only if $N(x_1, x_2) = |x_1| + |x_2|$.

If, in particular, $N(x_1, x_2) = N_\eta(x_1, x_2) = (|x_1|^\eta + |x_2|^\eta)^{1/\eta}$, $\max\{1/p, 1/\pi\} < \eta \le \infty$, then μ_{jk} exists, is unique, and is the only root in $[0, 1]$ of the equation for μ

$$\pi^{\frac{1}{\eta}}(1-\mu)^{\pi-\frac{1}{\eta}}|\hat{\beta}_{jk}|^{\pi-p} = v2^{\frac{j\varepsilon}{2}}p^{\frac{1}{\eta}}\mu^{p-\frac{1}{\eta}}. \tag{B4}$$

(Note that $B^{\sigma}_{\pi\pi} \hookleftarrow B^{s}_{pp}$ is only possible if $\pi \ge p$.) In general, (B4) cannot be solved explicitly. However, for any (j, k) it can be solved by very quickly convergent iterations of the dyadic or Fibonacci bisection method. Moreover, in many important partial cases (B4) can be solved explicitly. The most important such case is $\pi = p < \infty$, $s > \sigma$. Then,

$$\mu_{jk} = \mu_j = \frac{1}{1 + v^{\frac{1}{p-\frac{1}{\eta}}}.2^{j(s-\sigma)\frac{p}{p-\frac{1}{\eta}}}}. \tag{B5}$$

It can be seen that (7) corresponds to the partial case $p = 2$, $\eta = 1$. The case when $\pi \le 1/\eta$ and/or $p \le 1/\eta$ can be handled via the same argument, but this case requires special considerations and will not be discussed here.

(b) $\pi = p \le \infty$. In this case the penalization model is via $K_p(t, f; B^{\sigma}_{pp}, \dot{B}^{s}_{pp})$, and the embedding $B^{\sigma}_{pu} \hookleftarrow B^{s}_{pq}$ holds, $\sigma < s$, $0 < u \le \infty$, $0 < q \le \infty$. Assuming that in the computation of the shrinking factor μ_{jk} the quasi-norm $N_\eta(x_1, x_2)$ is being utilized (see case (a)), $\eta : 1/p < \eta \le \infty$, we see that μ_{jk} is given by (B5) with $v = t^p$, and *the case $p = \infty$ is also included*:

$$\mu_{jk} = \mu_j = \frac{1}{1 + (t2^{j(s-\sigma)})^{\frac{p}{p-\frac{1}{\eta}}}}, \tag{B6}$$

which yields $\mu_{jk} = \mu_j = \frac{1}{1+t2^{j(s-\sigma)}}$ for $p = \infty$. Denote by $\tilde{f}_{1,t}$ the estimator $\tilde{f}$ defined by (1'), with (7) replaced by (B6). Then,

$$c.K_p(t, f; B^{\sigma}_{pp}, \dot{B}^{s}_{pp}) \le (\| \hat{f} - \tilde{f}_{1,t} \|^p_{B^{\sigma}_{pp}} + t^p \| \tilde{f}_{1,t} \|^p_{\dot{B}^{s}_{pp}})^{1/p} \le C.K_p(t, f; B^{\sigma}_{pp}, \dot{B}^{s}_{pp}) \tag{B7}$$

(with max-modification for $p = \infty$), where c and C depend on η and $c = C = 1$ if and only if $\eta = 1$ or $p = \infty$.

Remark. It is seen from (B4-B6) that, the closer p is to $1/\eta$, the more sensitive the model is to variations of s, σ, t or v. The case $p = 1/\eta$ will be dealt with in [B10].

Remark. The inequalities

$$\min(1, 2^{\frac{1}{\eta_2}-\frac{1}{\eta_1}})(|a|^{\eta_1} + |b|^{\eta_1})^{1/\eta_1} \le (|a|^{\eta_2} + |b|^{\eta_2})^{1/\eta_2} \le \max(1, 2^{\frac{1}{\eta_2}-\frac{1}{\eta_1}})(|a|^{\eta_1} + |b|^{\eta_1})^{1/\eta_1}$$

show that in (B7) $c = \min(1, 2^{\frac{1}{p}(\frac{1}{\eta}-1)})$, $C = \max(1, 2^{\frac{1}{p}(\frac{1}{\eta}-1)})$ holds, and, unless $p = \infty$, C becomes much larger than c as η approaches 0. In statistical context this means that the smaller $\eta : 0 < \eta < 1$, the larger the sample sizes n for which $\tilde{f}_{1,t}$ is expected to perform well. On the other hand, for $\eta : 1 \leq \eta \leq \infty$ the upper constant C is 1 and the smallest possible value for c is $1/2$, so $\tilde{f}_{1,t}$ is expected to perform well already for moderate sample sizes. The properties of the estimator for $p = \infty$ deserve special attention.

(c) *Determining the smoothness parameter.* In order to estimate v or t when utilizing the general shrinking formulae in (a) and (b), one may apply asymptotic-minimax considerations. In particular, all comments in (B2) are valid here, for $1/\eta < p = \pi \leq \infty$ (see case (b)). This approach is expected to be effective for large samples. To improve the performance of $\tilde{f}$, defined by (1') and (B4), (B5) or (B6), on moderate samples, we suggest estimation with smoothness constraints^{B7}. To this end, cross validation is performed as proposed in $^{B8(b)}$. This yields the estimate $\| \tilde{f}_{v_0^*} \|_{B_{22}^{s^*}}$ for $\| f \|_{B_{22}^{s^*}}$, as discussed there. The smoothing parameter t of $\tilde{f}_{1,t}$ (see case (b)) is determined by requiring that $\| \tilde{f}_{1,t} \|_{B_{22}^{s^*}} \approx \| \tilde{f}_{v_0^*} \|_{B_{22}^{s^*}}$. For Gaussian white noise, the method of $^{B8(b)}$ can be applied. An extension of ^{B6} for this case is also of interest.

B10. *Thresholding - partial case of the penalized regularization approach. Relationship between the smoothing parameter and the threshold value.* Amato and Vuza (1997) show that soft and hard thresholding, as well as other familiar thresholding methods can be obtained as partial cases of penalized L_2 estimation. In our more general parameter setting in ^{B9}, their considerations about soft thresholding correspond to studying the behaviour of the solution of (B5) when $\pi = 2$, $\eta = 1$ and $p \searrow 1+$. For hard thresholding they consider a penalized L_2 model which is not of K-functional type, hence, the underlying functional space cannot be precisely identified. Here we indicate how hard thresholding can be obtained within the K-functional setting. Looking at (4), we see that the abundance of admissible couples (a, b): $a \in A$, $b \in B$, $a + b = \alpha$, among which the minimum (infimum) is being sought, depends on how spacious $A \cap B$ is. The more spacious this intersection is, the larger the class of admissible (a, b). On the other extreme, if $A \cap B = \{0\}$, then the K-functional is an equivalent quasi-norm in the *direct sum* of A and B, and any $\alpha \in A + B = A \oplus B$ is a unique sum of the unique projections of α onto A and B, respectively. Thus, there is only one admissible couple (a, b) in this case and, of course, it is the optimal one, or, in other words, the variational nature of the model is *eliminated*. It turns out that classical hard thresholding strategies correspond to exactly such K-functionals with $A \cap B = \{0\}$, where the choice of B depends on the concrete thresholding strategy. Typically, $A \oplus B = C$ is a fixed finite-dimensional space with dimension $O(2^{j_1})$ in which $\{\varphi_{j_0 k} : \mathrm{supp}\varphi_{j_0 k} \cap \mathrm{supp}\hat{f} \neq \emptyset\} \cap \{\psi_{jk} : j = j_0, \ldots, j_1,\ \mathrm{supp}\psi_{jk} \cap \mathrm{supp}\hat{f} \neq \emptyset\}$ is a basis; $||.||_A$ and $||.||_B$ are any two fixed (quasi-)norms in finite dimensional space with dimension equal to that of C; B is the (closed) linear span of those ψ_{jk} in C for which $|\langle \hat{f}, \psi_{jk} \rangle|$ is above the threshold value. The threshold value is a monotone function of the regularization parameter t of the K-functional, and thus the space B and its complement A to C depend on t, too. This dependence is easiest to express for the K_∞-functional. Denoting by $\lambda = \lambda(t)$ the threshold level, and by $\tilde{f}_\lambda$ the thresholded estimator, it can be seen that $K_\infty(t, \hat{f}; A_t, B_t) = \inf_{\hat{f}=f_1+f_2} \max(||f_1||_{A_t}, t||f_2||_{B_t}) = \max(||\hat{f} - \tilde{f}_{\lambda(t)}||_{A_t}, t||\tilde{f}_{\lambda(t)}||_{B_t})$, where the threshold $\lambda(t)$ is selected so that $|\ ||\hat{f} - \tilde{f}_{\lambda(t)}||_{A_t} - t||\tilde{f}_{\lambda(t)}||_{B_t}\ |$ is minimal. Here: $\lambda(0) = 0$; $A_0 = \{0\}$; $B_0 = C$ (endowed with $||.||_B$); the first term $||\hat{f} - \tilde{f}_{\lambda(t)}||_{A_t}$ is zero at $t = 0$, increases with the increase of $t \geq 0$ and is $||\hat{f}||_A$ for sufficiently large t, while the second term $||\tilde{f}_{\lambda(t)}||_{B_t}$ is $||\hat{f}||_B$ at $t = 0$, decreases with the increase of t and is zero for t large enough.

In fact, the K-functional setting proposes also a new, unexplored thresholding method, different from the classical ones, with remarkable properties deserving much attention. We shall briefly describe it in item (b) below. Before this, in item (a) we shall complete our analysis of the case $p = \pi$, begun in ^{B9}.

(a) *The case $p = \pi = 1/\eta < \infty$ in (B4).* In this case formulae (B5) and (B6) degenerate, but μ_{jk} can still be computed explicitly from (B3) and it turns out that

$$\mu_{jk} = \mu_j = (2^{-pj(s-\sigma)} - v)^0_+. \tag{$B5'$}$$

We see that this is a classical *thresholding method localizing the spectral window and having band-limiting effect.* This shows that classical filtering methods can also be described in terms of appropriate choice of equivalent quasi-norms in the spaces involved in the K-functional. This clearly demonstrates once again that the K-functional setting provides a very general and *unified* methodological basis for the study, description and classification of smoothing, denoising and filtering techniques from variational point of view.

Completing our discussion of the case $p = \pi$ begun in $^{B9(b)}$, we note that, together with other applications of theirs, the shrinking formulae obtained have potential applications in Bayesian statistics. For example, the values μ_j in (B6) can be used to generate the parameters π_j as defined in Abramovich, Sapatinas and Silverman (1997), formula (10) (note that for μ_j in (B6) $\mu_j \leq \min(1, (2^{j\sigma}/t)^{p/(p-1/\eta)}.2^{-jsp/(p-1/\eta)}) \leq 2\mu_j$ holds). For instance, it would be interesting to test the performance of the Bayesian model proposed by Abramovich, Sapatinas and Silverman with $\pi_j = \min(1, \tilde{v}^{-1}.2^{-2j\tilde{s}})$, corresponding to $p = 2,\ \eta = 1,\ \sigma = 0$, with $\tilde{v}$ and $\tilde{s}$ obtained by optimization of the GFCV criterion (see Subsection 6.1).

(b) In $^{B9(b)}$ we saw that in the case of the embedding $B^\sigma_{pp} \hookleftarrow B^s_{pp},\ s > \sigma$, there is an explicit shrinking formula for the shrinking factor μ_{jk} (formula (B6)). Another major embedding within the Besov-space scale is *the Sobolev-type* embedding: $B^\sigma_{\pi\pi} \hookleftarrow B^s_{pp}$ if $\sigma - \frac{1}{\pi} = s - \frac{1}{p}$ and $0 < p \leq \pi \leq \infty$. From (B4) it can be seen that there is no explicit formula of the type of (B6) for this setting of $\sigma,\ s,\ \pi$ and p. This makes even more interesting the fact that the K-functional approach indicates that there is a remarkable new thresholding method narrowly specialized for: $\sigma,\ s,\ \pi,\ p:\ \sigma - \frac{1}{\pi} = s - \frac{1}{p} =: \tau \in \mathbf{R},\ 0 < p \leq \pi \leq \infty$. The essential additional fact here is that *τ is the same for both spaces in the K-functional.* Because of this invariance of τ we can invoke the following formula

$$K_1(v,f;L_\pi(U,d\mu),L_p(U,d\mu)) \sim (\int_{v^{-\lambda}}^{\infty} f^*(\xi)^\pi d\xi)^{1/\pi} + v(\int_0^{v^{-\lambda}} f^*(\xi)^p d\xi)^{1/p}, \qquad (B8)$$

where $(U,d\mu)$ is an arbitrary measure space ($d\mu$ can be also an atomic measure), $f^*(\xi)$ is the *decreasing rearrangement* of f with respect to the measure $d\mu$ (see Bergh and Löfström (1976), Section 1.3), and $\lambda : \frac{1}{\lambda} = \frac{1}{p} - \frac{1}{\pi}$. Formula (B8) cannot be found explicitly in our reference sources but it can be obtained from the Peetre-Krée formula (Bergh and Löfström (1976), Theorem 3.6.1, together with 3.14.5,6 and Theorem 5.2.1 (2) for $q=p$ in their notation).

Suppose, as usual, that f and ψ are compactly supported. For simplicity of presentation, assume that $j_0 = 0$ for any sample size n. Formula (B8) implies the following strategy for thresholding the β-wavelet coefficients of $\hat{f} = \tilde{f}_0$ (see (1', 7) for $v=0$):

(b1) Consider all $(j,k) :$ supp$\psi_{jk} \cap$ supp$f \neq \emptyset$. Denote the set of all such (j,k) by $I(f,\psi)$. We already know (see the proof of Theorem 1 in Appendix A) that the number M of elements of $I(f,\psi)$ does not exceed $c.2^{j_1}$, where $c = c(f,\psi)$.

(b2) Consider the decreasing rearrangement $\{b_\nu,\ \nu = 1,\dots,M\}$ of the finite set $\{2^{j(\tau+1/2)}|\hat{\beta}_{jk}| : (j,k) \in I(f,\psi)\}$. By definition, $b_\nu \geq b_{\nu+1}$, $\nu = 1,\dots,M-1$; for any $\nu = 1,\dots,M-1$, there exists a unique $(j_\nu,k_\nu) \in I(f,\psi)$, such that $2^{j(\tau+1/2)}|\hat{\beta}_{jk}| = b_\nu$; $(j_{\nu_1},k_{\nu_1}) \neq (j_{\nu_2},k_{\nu_2})$ if and only if $\nu_1 \neq \nu_2$. The new, hard-threshold wavelet estimator f_v^* is defined by

$$f_v^*(x) = \sum_k \hat{\alpha}_{0k}\varphi_{0k}(x) + \sum_{\nu=1}^{[v^{-\frac{\pi p}{\pi - p}}]} \hat{\beta}_{j_\nu k_\nu}\psi_{j_\nu k_\nu}(x), \quad x \in \mathbf{R}. \qquad (B9)$$

(b3) If $b_{[v^{-\frac{\pi p}{\pi-p}}]-1} = b_{[v^{-\frac{\pi p}{\pi-p}}]}$ up to some resolution threshold (tending to zero as $n \to \infty$ and much smaller than the threshold levels in standard thresholding methods), then the terms corresponding to all the last ν's included in the RHS of (B9) which fall into this resolution threshold (that is, are formally equal to $b_{[v^{-\frac{\pi p}{\pi-p}}]}$) are being *removed*. Step (b3) ensures the uniqueness of f_v^*. If the sequence $\{b_\nu\}$ is strictly decreasing (relative to the resolution threshold chosen), step (b3) is not needed.

Let us discuss (B9) in brief. The significance of every $|\hat{\beta}_{jk}|$ is being assessed with respect to its level j. As $v \to 0$ with $n \to \infty$, the most significant features appear first, and less significant details emerge only for sufficiently small v. The criterion for significance of the coefficients is the regularity assumption expressed in the value of τ. Notice the qualitative differences between the strategies for "more regular functions" ($\tau > -1/2$) and for "less regular functions" ($\tau < -1/2$). The boundary value $\tau = -1/2$ exactly corresponds to the critical regularity $\varepsilon = -2(\pi - p)(\tau + 1/2) = 0$ (cf. Delyon and Juditsky (1996), Theorem 1, $d=1$). We believe that the new thresholding approach suggested by (B9) can prove to be quite useful in applications involving composite estimators.[B8] It is also very appropriate for the case when n is a power of two and the orthogonal discrete wavelet transform (ODWT) is being applied.

(c) *Determining the smoothing parameter.* We suggest the same approaches as in [B9(c)], although the algorithmic realization of the same ideas would obviously be quite different, notably involving sorting procedures.

B11. *Estimation in intersections of Besov spaces.* There is a theorem about explicit computation of K-functionals between intersections of quasi-normed lattices, due to Karadzhov (1973), Theorem 1.2.51. Later results of Feichtinger and Gröchenig (1989), Frazier and Jawerth (1990) and Sickel (1990) on atomic decompositions of Besov and Triebel-Lizorkin spaces, in combination with Karadzhov's result, made it possible to compute the K-functional between intersections of such spaces (with norm $||.||_{B_0 \cap B_1} = \max(||.||_{B_0}, ||.||_{B_1})$). As a consequence one obtains the following statement. Fix $t > 0$. Consider $K(t, \hat{f}; B^{\sigma}_{\pi u}, B^{s_0}_{p_0 q_0})$, $K(t, \hat{f}; B^{\sigma}_{\pi u}, B^{s_1}_{p_1 q_1})$, where the Besov quasi-norms are given by (2), $\hat{f}$ is the noisy wavelet expansion (from j_0 to j_1) and let $\mu^{(\nu)}_{jk}(t) \in [0,1]$ be such that $\tilde{\beta}^{(\nu)}_{jk}(t) = \mu^{(\nu)}_{jk}(t).\hat{\beta}_{jk}$, where $\tilde{\beta}^{(\nu)}_{jk}(t)$ is the respective coefficient of the minimizer of the ν-th K-functional, $\nu = 0, 1$. If $\tilde{\beta}_{jk}(t)$ is the respective coefficient of the minimizer of $K(t, \hat{f}; B^{\sigma}_{\pi u}, B^{s_0}_{p_0 q_0} \cap B^{s_1}_{p_1 q_1})$, and if $\mu_{jk}(t)$ is such that $\tilde{\beta}_{jk}(t) = \mu_{jk}(t).\hat{\beta}_{jk}$, then $\mu_{jk}(t) = \min_{\nu} \mu^{(\nu)}_{jk}(t)$.

B12. *More general function spaces with coefficient norms.* Recent applications of wavelet theory to the study of Hölder, cusp, singularity and oscillation spectra and to the development of multifractal formalism for functions require the consideration of more general types of function spaces. Here we consider two model types of such spaces.

(a) *The Nikol'skii-Besov spaces* $B^{\omega(.)}_{pq}(\mathbf{R}^d)$, with norm

$$||f|| = ||f||_{L_p(\mathbf{R}^d)} + \{\int_0^1 [\frac{\omega_r(f;t)_{L_p}}{\omega(t)}]^q \frac{d\omega(t)}{\omega(t)}\}^{1/q} < \infty,$$

$r \in \mathbf{N}$, $1 \le p \le \infty$, $1 \le q \le \infty$, where $\omega : [0,\infty) \to [0,\infty)$ is a nondecreasing continuous function with $\omega(0) = 0$ and $\{\int_0^1 [t^r/\omega(t)]^q \frac{dt}{t}\}^{1/q} < \infty$, $\omega_r(f;t)_{L_p}$ being the L_p-modulus of smoothness of f of order r, with step t. The classical theory (embeddings, traces, etc.) of these spaces has been developed in the 1970s and 1980s, notably by M. L. Gol'dman. To our best knowledge, after the series of papers on atomic decompositions and Riesz bases of orthogonal wavelets in 1989-1990, it seems to have remained unnoticed that the Nikol'skii-Besov spaces (which are translation invariant) do have the same Riesz bases of sufficiently regular orthogonal wavelets as usual Besov spaces and, hence, also admit equivalent coefficient norms. The only difference with usual Besov spaces (which may be very useful in the context of penalized statistical estimation, in particular, level and block shrinking strategies) is that now the weights on each level j depend on the more general function ω.

(b) *The A-spaces of Vasil A. Popov* have initially been defined by V. A. Popov as analogues of Besov spaces, where the integral modulus of smoothness (or ω-modulus) $\omega_r(f;t)_{L_p}$ has been replaced by the so-called average modulus of smoothness (or τ-modulus) $\tau_r(f;t)_{L_p}$ introduced by Bl. Sendov (see, e.g., Sendov and Popov (1988), Petrushev and Popov (1987), Popov and Dechevski (1985), Ivanov (1985), Dechevski (1986, 1988, 1998a). More precisely, the inhomogeneous A-space A^s_{pq} has the quasi-norm

$$||f|| = ||f||_{L_p} + \{\int_0^1 [t^{-s}\tau_r(f;t)_{L_p}]^q \frac{dt}{t}\}^{1/q} < \infty,$$

$r \in \mathbf{N},\ s > 0,\ 0 < p \leq \infty,\ 0 < q \leq \infty$, while the quasi-seminorm in the homogeneous A-space $\dot{A}^s_{pq}$ is

$$||f|| = \{\int_0^\infty [t^{-s}\tau_r(f;t)_{L_p}]^q \frac{dt}{t}\}^{1/q} < \infty.$$

Although not directly discussed in the recent survey of Feichtinger (1997) (which addresses only issues related to harmonic analysis) A-spaces and the $A_{p,h}$ spaces mentioned below are perhaps the most practically important examples of Wiener amalgam spaces, due to their intrinsic relevance to pointwise approximation processes. In the 1980s these spaces proved to be very useful in the study of rates of convergence in approximation theory and numerical analysis, as well as for the characterization of best *one-sided* approximation. It is to be noted that (see Dechevski (1988)) $A^s_{pq} = B^s_{pq}$, $\dot{A}^s_{pq} = \dot{B}^s_{pq}$ (with equivalence of norms) when $0 < p \leq \infty,\ 0 < q \leq \infty,\ s > d/p$. Recently, Jaffard (1998) has rediscovered these spaces (the spaces $V^{s,p}$ (in his notations) are isomorphic to $\dot{A}^{s+d/p}_{p\infty}$) from quite a different viewpoint - the study of multifractal properties of functions. He has been able to find a new insight into the essence of A-spaces by obtaining a wavelet characterization for $\tau_r(f;t)_{L_p}$. These results of Jaffard show that clearly the A-spaces will be an important space scale in future study of deterministic fractals and trajectories of random processes. In fact, τ-moduli and A-spaces are of immediate use in nonparametric regression estimation with deterministic design, because the bias in estimating the wavelet coefficients, which is nothing else but the error of a quadrature formula, cannot be expressed in terms of the usual ω-moduli, but in terms of the τ-moduli (see, e.g, Sendov and Popov (1988), Dechevski (1986,1988a)). Hence, the bias cannot be measured in Besov spaces but only in terms of A-spaces. Therefore, estimation rates in the case of regression with deterministic design can be studied in Besov spaces only because these spaces are isomorphic to A-spaces when $s' > 1/p$ (for dimension 1). The importance of the bound $s' > 1/p$ has been discussed in a relevant context by Donoho (1992), sections 2.3 and 2.4. In section 6.1 Donoho (1992) discusses also the critical case $s' = 1/p$. For this case Donoho proposes that interpolation and sampling (hence, also quadrature formulae) be studied in the class V_p of Bergh and Peetre (1974), which includes the usual class with bounded Jordan variation ($p = 1$), and Wiener's quadratic variation ($p = 2$) (Wiener (1924)). We agree with Donoho's analysis, and make the following additional comments. (i) The definition of V_p is a partial case of the more general concept of the Φ-variation of Young (1937), and other, more refined, choices of $\Phi(t)$ (notably, $\Phi(t) = t^p \Phi_1(t)$, where Φ_1 has logarithmic size) are also of interest in the relevant context of error bounds for quadrature formulae. (ii) Considerations involving V_p are a partial case of the more general approach of τ-moduli, since $\tau_r(f;h)_{L_p} \leq c_r h^{1/p}(V^p f)^{1/p}$ holds, $1 \leq p < \infty$, where $V^p f$ is the p-variation of $f \in V_p$ (see Hristov (1981), Dechevski (1988a)). (iii) The isomorphism $\dot{A}^{1/p}_{p,\min(1,p)} = \dot{B}^{1/p}_{p,\min(1,p)}$ has been proved by Ivanov (1985) for the case $p \in [1,\infty)$, and can be proved also for $p \in (0,1)$ by a refinement of the technique in Dechevski (1988). It indicates that $\dot{B}^{1/p}_{p,\min(1,p)}$ is for fixed p the largest space in the Besov scale for which the error of a quadrature formula can be bounded uniformly in all elements of the Besov space. (iv) For the critical index $s' = 1/p$, in the case $1 \leq p < \infty$, the theory of A-spaces indicates that the natural space in which pointwise procedures (interpolation, sampling, etc.) should be studied is $\dot{A}^{1/p}_{p,\infty}$ (or $\dot{A}^{d/p}_{p,\infty}$ in the d-dimensional case). Note that $V_p \hookrightarrow \dot{A}^{1/p}_{p,\infty}$ holds. (v) The multivariate theory of τ-moduli and A-spaces has not been studied in such a detail as in the univariate case, since the great variety of interpolation problems in d dimensions, $d > 1$ (see Bojanov, Hakopian and Sahakian (1993)) can be studied in terms of several different extensions of the concept of τ-modulus to the multivariate case. Our opinion is that the the most important of these extensions is based on the K-functional involving $A_{p,h}$-space (see Dechevski (1988) for $d = 1$), because the space $A_{p,h}$ is easy to define on general domains in $\mathbf{R}^d$, and the resulting theory is analogous, though not identical, to the results of Johnen and Scherer (1977) for the ω-modulus. The above-mentioned isomorphism $A^s_{pq} = B^s_{pq}$, $\dot{A}^s_{pq} = \dot{B}^s_{pq}$, $s > d/p$, refers to this type of τ-moduli in the definition of A-spaces in d-dimensions. (vi) Summing up (i-v), the natural three-index space scale, in which approximation processes involving pointwise functionals can be considered *for any* $s > 0$, is $\dot{A}^s_{pq}$. This, in particular, refers to the evaluation of the expected rates of nonparametric regression estimation with deterministic (uniform or non-uniform) design.

To illustrate the use of τ-moduli and A-spaces, let us analyze the derivation of formula (A2) in detail.

Step 1. The error of the first-order quadrature formula in the LHS of (A2) is bounded in terms of the τ-modulus (see Sendov and Popov (1988)):

$$|E\hat{\beta}_{jk} - \beta_{jk}| \leq \tau_r(f\psi_{jk}; \frac{1}{n})_{L_p}, \;\; r = 1, \; p = 1.$$

Step 2. The Leibniz formula for τ-moduli (Dechevski (1988a), Lemma 3.8.1) is applied:

$$\tau_r(g_1 g_2; h)_{L_p} \leq \sum_{\nu=0}^{r} \binom{r}{\nu} \tau_\nu(g_1; \frac{2rh}{r_\nu})_{L_{p_\nu}} . \tau_\nu(g_2; 2h)_{L_{p'_\nu}},$$

where $h > 0$; $r_0 = r_1$, $r_\nu = \nu$, $\nu = 1, \ldots, r$; $0 < p \leq \infty$, $\frac{1}{p_\nu} + \frac{1}{p'_\nu} = \frac{1}{p}$, $\nu = 0, \ldots, r$; $\tau_0(g;h)_{L_p} := ||g||_{A_{p,h}}$; the definition of the $A_{p,h}$-spaces is given in Dechevski (1986, 1988,1988a).

In the case of the first-order quadrature formula considered here,

$$\tau_1(f\psi_{jk};\frac{1}{n})_{L_1} \le ||f||_{A_{p_0,1/n}}\tau_1(\psi_{jk};\frac{1}{n})_{L_{p_0'}} + \tau_1(f;\frac{1}{n})_{L_{p_1}}||f||_{A_{p_1',1/n}}.$$

Step 3. The bound

$$||g||_{A_{p,h}} \le ||g||_{L_p} + \tau_1(g;h)_{L_p}$$

(see Dechevski (1986), Lemma 1, and Dechevski (1988a)), is applied to the result in Step 2, for $g = f,\ p = p_0$, and for $g = \psi_{jk},\ p = p_1'$; $h = 1/n$.

Step 4. The properties of the τ-modulus (see the general references given above) are applied to obtain a bound of the error in terms of A-spaces, bounded p-variation and other relevant quantities.

Formula (A2) has been obtained via Steps 1-4, for $p_0 = p_1 = \infty$, in order to remain entirely within the Besov scale, since $A^s_{\infty q} = B^s_{\infty q}$ for any $s > 0$ and $0 < q \le \infty$. The more advantageous selection (which, however, brings us out of the Besov scale) is $p_0 = p_1 = 2$. For this choice, as discussed in the end of subsection 6.1, one obtains a relaxation of the assumption $f \in B^{s'}_{22} \cap B^{s_1}_{\infty\infty}$ by replacing it with $f \in A^{s'}_{22}$. Formula (A3) has been obtained following Steps 1-4, too.

It is also informative to apply the above Steps 1-4 to outline the proof of a correct version of Lemma 3.1 in Antoniadis (1994, 1996). In our notations, the original statement of the lemma claims that if the scaling function φ is a coiflet of accuracy order r (that is, if φ is orthogonal to the monomials x^ν of degree $\nu = 1, ..., r-1$) then $|E\hat{\alpha}_{j_0k} - \alpha_{jk}| \le c_s(f)2^{-2j_0 s}$, where $c_s(f) < \infty$ if $||f||_{B^s_{22}} < \infty$. Our observation is that for $s \le 1/2$ a counterexample can be constructed (Ivanov (1985)) which shows that this claim is not true. The correct version follows via Steps 1-4, with r in Step 1 being equal to the accuracy order of the coiflet, and with $p_\nu = 2,\ \nu = 0, \ldots, r$, in Step 2. The resulting sharp condition for the above bound on $|E\hat{\alpha}_{j_0k} - \alpha_{jk}|$ to hold for any $0 < s \le r$ is $f \in A^s_{2\infty},\ 0 < s < r$, and $f \in A^s_{22},\ s = r$. Hence, since $A^s_{22} \hookrightarrow A^s_{2\infty}$ and $B^s_{22} = A^s_{22}$ for $s > 1/2$, Lemma 3.1 in Antoniadis (1994,1996) is true if $1/2 < s \le r$.

B13. *K-functional estimation with constraints.* In some applications it is desirable to preserve certain informative spatial features (e.g., translation-invariance (Coifman and Donoho (1995), positivity and integral equal to one in density estimation (Pinheiro and Vidakovic (1997), Penev and Dechevsky (1997)) and others). In $B5$ and in Section 7 we already encountered examples of optimization with constraints in the K-functional setting. Further examples, with convex constraints, can be found in Delecroix, Simioni and Thomas-Agnan (1996) and the references therein. Utilizing the general results of Löfström (1995) about K-functionals with convex restrictions, it is now possible to solve nonparametric estimation problems with quite general convex constraints.

On the other hand, recall that wavelet atomic decomposition reduces nonparametric problems to essentially parametric ones. It is, therefore, interesting to find out to what extent the theory of parametric models (see Robertson, Wright and Dykstra (1988)) can be helpful for solving constrained optimization problems in a wavelet setting.

The constrained approximation approach can help improve the performance of nonparametric density and regression-function penalized estimators in the cases when additional information is a priori available about the shape of the curve (e.g., positivity, monotonicity, convexity, k-monotonicity, etc.). In the case of density estimation, imposing a non-negatitivity constraint on the solution of the K-functional penalized optimization problem, together with an isoperimetric type of constraint of type equality to 1 of its integral, would yield a solution which *is* a density. It is remarkable that this is achieved entirely within the framework of convex optimization, which means that finding the global extremum is guaranteed (compare with the approaches of Pinheiro and Vidakovic (1997), Penev and Dechevsky (1997), Dechevsky and MacGibbon (1998), Dechevsky, MacGibbon and Penev (1998)).

B14. *Rates for the smoothing parameter in cross validation.* Continuing the discussion begun in the proof of Theorem 8 (see (A12)), we can obtain, for the density case, by bounding $\frac{d}{dv}MISE(v)$ *from below* (under a sharpness assumption on s' if $j_0 \to \infty$), *the sharp upper rates* for the smoothing parameter $\tilde{v} = \text{argmin} MISE(v)$. On the other hand, by bounding $\frac{d}{dv}MISE(v)$ *from above*, we can obtain (by using $f \in B^{s'}_{22}$, without an assumption about sharpness of s') *the sharp lower rates* for $\tilde{v}$. (For each fixed choice of j_0 and j_1 there would be one upper and one lower rate for $\tilde{v}$ which are expected to coincide (up to a constant factor)). These rates can be then utilized to improve the performance of cross-validation (also with respect to choice of j_0 and j_1, in the spirit of the ideas of Johnstone and Hall (1992)). This approach can be transferred to Bowman–Rudemo type of cross validation in the case of regression (see Dechevsky and Penev (1999)). It can also be extended, for both the density and regression estimation, to cross validation in Besov spaces with $\sigma > 0$, as considered in ^{B1}. In principle, this method works also in the case of GFCV, but the difficulty in this case is that the rate of the optimal v gets "contaminated" by the error of the quadrature formula involved. In other words, varying s' within the interval $[0, 1/2]$ is no longer sufficient to influence the rate for v. If, however, it is assumed that $f \in A^{s'}_{22}$ rather than $f \in B^{s'}_{22}$, then s' controls the rate for v also when $s' \in [0, 1/2]$ (see $^{B12(b)}$). The same restriction appears, for the same reasons, in the theory of QGCV. In this case, the condition $s > 1/2$ is needed also to ensure consistency of QGCV (see Amato and Vuza (1997a)). For $s > 1/2$ and $s' = s$, Amato and Vuza (1998) found the sharp rates of $v^* \to 0$ for the QGCV criterion by a technique which refers to the case $j_0 = 0,\ j_1 = \log_2 n - 1$. We can generalize these results to include also the case $j_0 \to \infty$ (under assumption about sharpness of s'), as well as the case $0 < s' \leq 1/2$ (under the assumption that $f \in A^{s'}_{22}$). From the rates for $v \to 0$ it is, of course, easy to obtain the rates for $E||\tilde{f}_v - f||^2_{L_2} \to 0$. This approach can be extended to the generalizations considered in ^{B1}, too. Amato and Vuza (1998) showed that minimizing the QGCV criterion for $\sigma = 0$ leads to a consistent estimate of $E||\Pi_{j_1} f - \tilde{f}_{\tilde{v}}||^2_{B^{\sigma}_{22}} \to 0$, for $0 < \sigma < s' = s$, where Π_{j_1} is as in ^{B1}, $j_0 = 0,\ j_1 = \log_2 n - 1$. On the basis of this observation Amato and Vuza claim that QGCV is good enough to estimate the derivatives of the function (despite the fact that they proved that the minimizer $v = v_\sigma$ of $E||\Pi_{j_1} f - \tilde{f}_v||^2_{B^\sigma_{22}}$ (see [B1]) has rates which asymptotically split apart from the rates of the minimizer v^* of QGCV). We disagree with this opinion of Amato and Vuza, because it is based only on an asymptotic result about consistency of the estimator (cf. our Theorem 3 and Corollary 1). The more important result which has to be established here is, in our opinion, to derive an analogue of our Theorem 2 about consistency of the choice of the smoothing parameter when $0 < \sigma < s' \leq s$ (see also [B1]). It is for this type of result that the advantage of v_σ over v^* will show up. It is also clear that, if $s \geq s' > \sigma + 1/2 \gg 1$, the difference between the estimator $WAVREG_\sigma$ (see [B1]) and WAVREG of Antoniadis (1996), corresponding to $\sigma = 0$ will be quite visible for moderate samples. That is why it is important to study in detail the estimator $WAVREG_\sigma$ proposed in [B1], with controlled selection of j_0 and j_1.

The key technical result in the evaluation of the rates is

$$\sum_{j=j_0} \frac{2^{j\lambda}}{(1+v2^{j\mu})^\nu} \asymp v^{-\lambda/\mu} \int_{v^{1/\mu}2^{j_0}}^{v^{1/\mu}2^{j_1+1}} \min(\xi^\lambda, \xi^{\lambda-\mu\nu}) \frac{d\xi}{\xi}, \quad \lambda > 0,\ \mu > 0,\ \nu > 0,\ v \geq 0.$$

It follows by change of variable in the integral from the bounds

$$2^{-\lambda} \frac{2^{\eta\lambda}}{(1+v2^{\eta\mu})^\nu} \leq \frac{2^{j\lambda}}{(1+v2^{j\mu})^\nu} \leq \max(1, 2^{\mu\nu-\lambda}) \frac{2^{\eta\lambda}}{(1+v2^{\eta\mu})^\nu}$$

and $1 + v2^{\eta\mu} \asymp \max(1, v2^{\eta\mu})$, where $j \leq \eta \leq j+1$. The first of these two bounds follows from the strict monotonicity of $g_1(\zeta) = 2^\zeta$ and $g_2(\zeta) = \frac{\zeta}{1+\zeta}$ for $\zeta \geq 0$ (cf. Amato and Vuza (1998), Lemma 3.1, for another bound used in their computations).

The above technical result is sufficient for finding the sharp rates in the density case and the regression case with random design under the assumptions that $f \in B^{s'}_{22}$, $s' > 0$. If $j_0 \to \infty$, then also sharpness of s' is required. When the design is deterministic, the additional condition $s' > 1/2$ has to be imposed. The case $0 < s' \leq 1/2$ can be included by assuming that $f \in A^{s'}_{22}$, $s' > 0$, but an additional step in the proof is needed, involving the equivalence of the τ-modulus to a K-functional (see Dechevski (1988, 1986, 1988a)).

As for WAVREG of Antoniadis (1996), the sharp rates for $Ev^* \to 0$ (v^* being here the optimizer of $\tilde{\Lambda}_n$) and the respective estimation rates have not been studied at all. They, together with the respective rates for $WAVREG_\sigma$, can be obtained in a similar way.

B15. *Redesigning the experiment by K-functional techniques.* Recently, the development of discrete wavelet transform techniques requiring dyadic uniform design aroused renewed interest in methods for redesigning nonparametric regression models with random or irregular deterministic design (see Hall and Turlach (1997), Cai and Brown (1998,1998a)). Another important application of redesigning is in medical imaging, for the purposes of uniformization of data sets and registration of images (e.g., positron emission tomography (PET) images). In this case images are usually very smooth and not too spatially inhomogeneous, and the convolution method of Gasser and Müller (1979) seems to be appropriate. For this particular type of applications we suggest a modification of the convolution approach based on the K-functional. In [B4] we suggested a kernel estimation technique via solutions of evolutionary differential equations. The numerical solution proposed was by FEM. Now we propose to use a difference scheme over a uniform mesh. The resulting numerical solution provides the redesigned data set corresponding to the selected uniform mesh. The bias term in the estimation of the values on the mesh has been studied in detail (see Peetre and Thomée (1967), Widlund (1968, 1970), Thomée (1969,1971), Brenner, Thomée and Wahlbin (1975), Popov and Dechevski (1985), Dechevski (1986,1988a)).

B16. *Penalized regularization of non-K-functional type.* The K-functional is a very general functional characteristics. Most generally, the spaces A and B in (4) can be any quasi-normed abelian groups. Such groups are metrizable, and so the K-functional can be extended to any metric spaces A and B, such that the sum $a+b$, $a \in A$, $b \in B$, is meaningful. It is possible to go beyond that. Amato and Vuza (1997) give an example of penalized regularization where the penalty is of entropy type. This is one of their examples illustrating their general setting of solving $\min_{x\in K}(f(x-y)+g(x))$ (Formula (1) in Amato and Vuza (1997)). Here $K \subset X$, where X is a Banach lattice with order continuous norm; K is weakly closed; $y \in D \subset X$; $f : X \to \mathbf{R}$ is continuous for the norm topology and lower semi-continuous for the weak topology; $g : K \to \mathbf{R} \cup \{\infty\}$ is lower semi-continuous for the weak topology. Under these general (and some additional) assumptions they study in Theorems 2.1 and 2.2 the existence of a weak solution of the optimization problem when $y_n \to y$ in the norm topology, for some $y \in D$, as $n \to \infty$. Our experience with the diverse aspects of statistical estimation considered in the present paper tells us that this general optimization model needs to be extended in several directions: (a) one should consider *a sequence* of penalized optimization problems depending on n: $\min_{x\in K}(f_n(x-y_n)+g_n(x))$ as $y_n \to y$ (this corresponds to the case $j_0 \to \infty$ (see our Section 6)); (b) one should be able to characterize simultaneously the family $\{(f,g)\}$ such that $\min_{x\in K}(f_1(x-y)+g_1(x)) \asymp \min_{x\in K}(f_2(x-y)+g_2(x))$ for every admissible y and for every (f_1,g_1) and every (f_2,g_2) from this family (this is important, because some of the (f,g) in the family may have better properties (e.g., smoothness, strict convexity) than others, and for the worse ones there may not exist a solution to the respective optimization problem); (c) X should be allowed to be quasi-Banach (note that it was possible to derive formulae (B3-B6) only because quasi-Banach spaces were considered in the K-functional; the method would not work in the narrower Banach setting); (d) it should be possible to characterize the order of approximation to zero of $\min_{x\in K}(f_n(x-y)+g_n(x))$, $n\to\infty$, when the *intermediate* (cf. our Section 4) set $D : K\cap D \subset D \subset X$ varies from X to $K \cap D$ (this is necessary for studying the bias term when deriving asymptotic rates of statistical estimation).

In the K-functional setting, objectives (a-d) are achieved in a natural way.

Finally, it should be noted that if some of the general assumptions of Amato and Vuza on f and g be made more stringent, then others may be relaxed, or stronger results about existence and/or uniqueness of the solution of the optimization problem can be achieved (cf., e.g., the classical result in Theorem 28.1 of Vainberg (1979)).

B17. *K-functional with several penalization parameters.* The models of level-dependent and block-penalized shrinking are examples of generalization of the K-functional for the case of vector penalizing parameter. The respective properties of the K-functional and the associated real-interpolation theory have been studied (see, e.g., Sparr (1974)) and can offer further insight into level-dependent and block-shrinking strategies.

B18. *Biorthogonal wavelets, wavelet packets and multiwavelets.*

(a) It is possible to consider *biorthogonal* atomic decompositions of Besov and more general spaces.$^{B12(a)}$ There are some advantages of biorthogonal compared to orthogonal wavelet approximation. For example, in signal and image processing it is often observed that biorthogonally compressed images have somehow better quality than orthogonally compressed ones. Another instance is that in biorthogonal wavelet analysis one can utilize polynomial spline-wavelets with compact support, which is impossible in the orthogonal case.

(b) The utilization of *wavelet packets* becomes essential when studying local self-similarity and local periodicity in processes which are globally non-selfsimilar and non-periodic. (One typical example is *turbulence.*) The penalized model can be useful here, too, in particular, if the penalty is of *entropy* type (cf. Amato and Vuza (1997)). In the context of statistical nonparametric estimation, this allows combining adaptive smoothing and denoising with the Coifman-Wickerhauser recursive best-base selection algorithm. The advantages for fractal estimation can be considerable when *locally* self-similar fractals are studied because, as already noted in Section 7, the choice of the underlying wavelet is of crucial importance.

(c) *Multiwavelets* have more degrees of freedom than usual (mono)wavelets, thanks to which they enjoy a number of desirable properties that monowavelets do not have. This refers to both wavelets and scaling functions (mother and father wavelets). In particular, multiwavelets can combine sufficient regularity with short support, symmetry and orthogonality to polynomials up to a high degree, which is theoretically impossible for monowavelets. Besides their general computational advantages, also for statistical estimation (see, e.g., Downie and Silverman (1998) and the references therein), they have the potential to become a very effective tool in fractal-function estimation. Recalling the algorithm for generating the penalized self-similar fractal estimator (see Section 7), one can study in an analogous way fractals which are not self-similar but are comprised of a finite number of interlaced or superposed self-similar fractals. This may prove to be useful, e.g., in the study of growth of crystalline structures (cf. Arneodo, Argoul, Bacry, Elezgaray and Muzy (1995)).

B19. *Pattern recognition.* In our opinion, the most adaptive block-shrinking estimators can be obtained by introducing elements of pattern recognition into the block selection strategy. The pioneering work of Hall, Kerkyacharian and Picard (1995) and subsequent work of Cai (1998) and Cai and Silverman (1998b) (combining the block strategy of Hall, Kerkyacharian and Picard in the partial case of Gaussian white noise with oracle inequalities and the James-Stein estimator proposed by Donoho and Johnstone for level-dependent shrinking when the white noise is Gaussian) can be considered as prototypes of results in this direction.

B20. *The orthogonal discrete wavelet transform (ODWT), fractal estimation, and parallel computing architectures (PCA).* In this paper we have given preference to empirical wavelet coefficient estimation for sample sizes n which are not necessarily powers of 2 and, therefore, ODWT (called also *fast wavelet transform*) cannot be applied. The cross validation considered is also not adapted for ODWT inversion. One reason for our choice is that, when utilizing ODWT, one works with approximations of $\varphi_{j_0 k}$ and ψ_{jk} which depend on n. When n takes small or moderate values, and when the estimated function is not smooth, but a fractal, it is important to work with the "exact" $\varphi_{j_0 k}$ and ψ_{jk} (that is, with their values computed to sufficient precision independent of n, by the respective functional equations of type (16) by which φ and ψ are defined - see Daubechies (1992)). For example, this is the case when working with the self-similar fractal estimator considered in Section 7.

We also note that: (a) most of our considerations can be modified for sample sizes $n = 2^N$ and adapted for the use of ODWT, the resulting algorithms for $n = 2^N$ being fast, i.e., $O(n)$-methods (see the sequel of this paragraph); (b) while on sequential computers ODWT has essential advantages, being an $O(n)$-method, when utilizing PCA with $O(n)$ processors (this is realistic for moderate and small samples), both the ODWT and the method without use of ODWT are $O(\log_2 n)$-methods.

In order to justify our claim in (a), for sequential computing architecture, when n is a power of two, let us compare GFCV, as proposed in subsection 6.1, with the 'leave-half-out' or 'twofold' cross validation of Nason (1996), QGCV of Amato and Vuza (1997a,1998) and WAVREG of Antoniadis (1996). Nason's algorithm exploits in a nice way the symmetry on the real line to produce an $O(n)$ algorithm when $n = 2^n$, or, roughly, an $O(n^2)$-algorithm for general $n \in \mathbf{N}$ when the 'leave-one-out' CV algorithm has to be applied (see Nason (1996), p.570). Like Nason's twofold method, QGCV and WAVREG are specially designed for $n = 2^N$ and the use of ODWT, and it can be shown that they are also $O(n)$-methods. What about FCV and GFCV? For arbitrary $n \in \mathbf{N}$, without making use of the fulfillment of the compatibility condition, FCV, as defined in (11), is an $O(n^3)$-method. Because of the validity of the compatibility condition, however, FCV can be computed by (12,13), instead of (11), by $O(n^2)$ operations. For general n GFCV is also an $O(n^2)$-method. Now assume that $n = 2^N$ and that FCV and GFCV are applied in combination with ODWT. In this setting, x ranges over the discrete mesh $\{x_i\}_{i=1}^n$, the vectors $\{\varphi_{j_0 k}(x_i)\}_{i=1}^n$, $\{\psi_{jk_j}(x_i)\}_{i=1}^n$, $k_j = 1, \ldots, 2^j$, $j = j_0, \ldots, j_1$, are replaced by the respective first 2^{j_1+1} orthonormal basis vector-lines of the matrix $\mathbf{W}$ of the corresponding direct ODWT; the vector of noisy wavelet coefficients $\mathbf{w}$ is computed from the data vector $\mathbf{y}$ by $\mathbf{w} = \mathbf{W}\mathbf{y}$. (For simplicity of presentation, here we assume the periodic setting, the case of boundary-adjusted wavelets and corresponding wavelet transform being similar.) Then, it can be shown that FCV becomes an $O(n)$-method. GFCV becomes a 'faster' $O(n)$-method than FCV, the constant factor in the $O(n)$-rate being smaller because of the simple formula for $\overline{h}(v)$ in this case (see Remark

3). Thus, the numerical performance of FCV and GFCV is comparable with that of Nason's method in both cases of n being, or not being, a power of two. The advantages of full cross validation clearly show up in the multidimensional case, when the design is not rectangular and not uniform. The FCV criterion can be defined also when the design set $\{x_i\}_{i=1}^n$ is a (uniformly or non-uniformly) scattered set in $\mathbf{R}^d$, for any $n \in \mathbf{N}$. The construction is based on Voronoi simplectification (VS, for short) and is discussed in more detail in [B21]. Being able to handle the case of scattered design in the multivariate case is *a very important advantage* of FCV which makes the method well adapted to deal with the challenging computational geometric aspects of the rapidly developing wavelet theory for d-dimensional domains with general types of boundary. This theory has already a considerable achievement for Lipschitz-graph domains, due to Cohen, Dahmen and DeVore (1997), and it is also imminent that in the near future there will be also a complete wavelet-based analogue of the convolution/local Taylor expansion technique of Jonsson and Wallin (1984) of deriving Whitney-type extension theorems for Besov spaces on generalized Cantor sets with non-integer Hausdorff dimension in $\mathbf{R}^d$. VS is usually a computationally intensive procedure, but for fixed design VS has to be applied only once. Then, for any random vector $\{Y_i\}$ corresponding to the same fixed design set $\{x_i\}$, FCV (in its simplest version (11)) is an $O(n^3)$-method, for any dimension d. Moreover, the fulfillment of the compatibility condition for FCV makes it possible to obtain a computational reduction from $O(n^3)$ to $O(n^2)$ for any dimension $d \in \mathbf{N}$. Proving this reduction is based on the multivariate Hermite expansion for scattered data and the density of C^∞ in Besov spaces. A sufficiently detailed consideration of this topic (as well as of the possibility to define an analogue of GFCV in this case) would be too spacious and too technical to present here.

Finally, comparing the optimization algorithm's complexity, GFCV, FCV, QGCV and WAVREG are easy winners over Nason's twofold and 'leave-out-one' algorithms, because in the former group of algorithms the dependence on the regularization parameter is analytic. It is possible, however, to modify the definition of Nason's CV estimator by replacing thresholding with non-threshold shrinking via (7) or, more generally, (B5) or (B6). For this modification, the dependence on the smoothness parameter in Nason's twofold and 'leave-out-one' algorithms becomes analytic, and full-scale theoretical analysis of these algorithms can be carried out, in analogy to the results discussed in subsection 6.1 and [B14], with controlled j_0 and j_1: $j_0 = O(1)$ or $j_1 \to \infty$, $j_1 = o(\log_2 n)$ or $j_1 \asymp \log_2 n$, $j_0 \leq j_1$. After this, a theoretical comparison with FCV and GFCV can be made. In the multidimensional case clearly FCV is the more flexible method, available for much more general data sets.

B21. *The multivariate case.* All the main results for the basic penalized regression-function and density estimation models in Sections 2-6, as well as all extensions of these models considered in Section 7 and Appendix B, are also available in the d-dimensional case, $d > 1$. Here we give some additional details.

Besov spaces of multivariate functions and multivariate wavelets. There are several ways to define multivariate orthonormal wavelet bases. We consider here the tensor-product bases.[A5] With respect to these bases Besov spaces admit a coefficient quasi-norm[A5] which is a straightforward generalization of (2,3). In d dimensions the index of the dyadic level weights is $s+d(1/2-1/p)$. The Sobolev embedding is now true when the following inequality between the indices of two Besov spaces is fulfilled: $B^s_{pq} \hookrightarrow B^{s_1}_{p_1q_1}$, if $s-d/p \geq s_1 - d/p_1$, $0<p\leq p_1 \leq \infty$, $0<q\leq q_1\leq\infty$. (In this paper we have often been dealing with the case $d=1,\ p_1=q_1=\infty,\ s_1=0,\ B^{s_1}_{p_1q_1}=L_\infty,\ p=q=2$.) The restriction on s when utilizing $B^r_{\infty\infty}$-regular wavelets with compact support is[A5] $\max[0,d(1/p-1)]<s<r$.

Cross validation. The method for cross validation in the density estimation setting is extended to the multidimensional case without any essential changes[A5]. The same refers to the Bowman–Rudemo approach[B1] in the case of regression–function estimation with random design. This is also true for Bowman–Rudemo's and Wahba's approach (in the latter case we have in mind FCV and GFCV) when the design of the nonparametric regression estimation problem is deterministic, but in this case an explanation is needed how to generate such design in d dimensions. If the sample size is of the form $n=\Pi^d_{\nu=1}n_\nu$, where $n_\nu\to\infty,\ \nu=1,\ldots,d$, as $n\to\infty$, a rectangular mesh $\{\{x_{i\nu}\}^{n_\nu}_{i=1}\}^d_{\nu=1}$ can be considered (see [B20]). However, this can only be of interest for very specific problems. In general, for any $n\in\mathbf{N}$ consider the mesh $\{x_i\}^n_{i=1},\ x_{i_1}\neq x_{i_2},\ i_1\neq i_2,\ x_i\in\mathbf{R}^d$, and construct *the Voronoi simplectification of the convex hull* of $\{x_i\}^n_{i=1}$. Then, the empirical wavelet coefficients in the d-dimensional noisy wavelet expansion are of the form

$$\hat{\alpha}_{j_0k}=\sum_{i=1}^n w_iy_i\varphi_{j_0k}(x_i),\quad \hat{\beta}^{[l]}_{jk}=\sum_{i=1}^n w_iy_i\psi^{[l]}_{jk}(x_i),\quad l=1,\ldots 2^d-1,$$

where the weight w_i is defined by $w_i=\frac{1}{(d+1)V}\sum^{m_i}_{\nu=1}V_{i\nu}$, $V_{i\nu}$ being the d-dimensional volumes of those simplices which have x_i as a vertex, m_i being the number of all such simplices, V being the d-dimensional volume of the convex hull of $\{x_i\}^n_{i=1}$. The wavelets $\varphi_{j_0k},\ \psi^{[l]}_{jk}$ are now the elements of the tensor-product orthonormal wavelet basis on $L_2(\mathbf{R}^d)$, discussed in [A5]. The same weights appear in the quadrature formula of the GFCV (FCV) criterion. It can be seen that $\sum^n_{i=1}w_i=1$. If the mesh is uniform, all simplices are uniform and congruent to each other, hence, $w_i=const$ for all i such that x_i is in the interior of the convex hull. In the uniform case one can also construct a modification where *all* w_i are equal to $1/n$.

Applications in Section 7 can be extended to the d-variate case. In particular, this is true for the fractal estimator and its version for Gaussian white noise.

Kernel regularization. The kernel is usually taken radial (e.g., $\Phi(\mathbf{x})=K_d\exp[-(\frac{||\mathbf{x}||^2}{1-||\mathbf{x}||^2})_+]$, where $||\mathbf{x}||$ is the usual Hilbert norm of $\mathbf{x}\in\mathbf{R}^d$), or as a normalized tensor product of univariate kernels. In combination with the tensor-product wavelet basis[A5] the second alternative is computationally more advantageous.

The extensions in Appendix B can all be generalized to the multivariate case. Some remarks are due, as follows.

Multivariate periodized wavelets are defined as tensor products of univariate periodized wavelets, and yield equivalent quasi-norms in Besov spaces, with equivalence constants depending on d.

In the d-dimensional case, the most important class of operators T considered in (B1) are partial differential operators which are *parabolic in the sense of Petrovskii* (Thomée (1969,1971)). The respective partial finite difference schemes[B15] are *parabolic in the sense of John* (Thomée (1969,1971)).

Formulae (B3,B4) remain the same, but *the critical regularity index* ϵ involved in (B3,B4) now generalizes to $\epsilon = 2ps - 2\pi\sigma + d(p - \pi)$ (cf. Delyon and Juditsky (1996)). Formulae (B5,B5',B6) remain unchanged.

Formula (B9) also remains unchanged, as well as the definition of *decreasing rearrangement* in [B10b,(b2)], but the relevant value τ now generalizes to $\tau := \sigma - d/\pi = s - d/p$ (thus, the formula about ϵ generalizing to $\epsilon = -2(\pi - p)(\tau + d/2) = 0$ for $\tau = -d/2$).

The remarks about the critical value $s' = 1/p$ in [B12,B14] are now relevant when $s' = d/p$.

REFERENCES

Abramovich, F., Sapatinas, T. and Silverman, B. (1998). Wavelet Thresholding via a Bayesian Approach. *J. Royal Statist. Soc., Ser. B*, **60(4)**, 725-749.

Amato, U., Occorsio, M. R. and Vuza, D. T. (1998) Fully Adaptive Wavelet Regularization for Smoothing Data. Proc. 3-rd Intern. Conf. Functional Analysis and Approximation Theory, Acquafredda di Maratea (Potenza, Italy), 1996, vol. 1. *Suppl. Rendic. Circ. Mat. Palermo*, ser. 2, **52**, 207-222.

Amato, U. and Vuza, D. T. (1994) Wavelet Regularization for Smoothing Data. Techn. Report CNR 108/1994, Istituto per Applicazioni della Matematica.

Amato, U. and Vuza, D. T. (1997). Besov Regularization, Thresholding and Wavelets for Smoothing Data. *Numer. Funct. Anal. and Optimiz.*, **18(5&6)**, 461-493.

Amato, U. and Vuza, D. T. (1997a). Wavelet Approximation of a Function from Samples Affected by Noise. *Rev. Roumaine Math. Pures Appl.*, **42(7-8)**, 481-493.

Amato, U. and Vuza, D. T. (1998). Wavelet Simultaneous Approximation from Samples Affected by Noise. *Comput. Math. Appl.*, **36(5)**, 101-111.

Anselone, P. M. and Laurent, P. J. (1968). A General Method for the Construction of Interpolating or Smoothing Spline-functions. *Numerische Math.*, **12**, 66-82.

Antoniadis, A. (1994). Smoothing Noisy Data with Coiflets. *Statistica Sinica*, **4**, 651-678.

Antoniadis, A. (1996). Smoothing Noisy Data with Tapered Coiflet Series. *Scand. J. Statist.*, **23**, 313-330.

Arneodo, A., Argoul, F., Bacry, E., Elezgaray, J. and Muzy, J. F. (1995). Ondelettes, multifractales et turbulences: de l'ADN aux croissances cristallines. Diderot: Paris.

Arneodo, A., Bacry, E., Jaffard, S. and Muzy J.F. (1997). Oscillating Singularities on Cantor Sets: a Grand-canonical Multifractal Formalism. *J. Statist. Phys.*, **87(1/2)**, 179-209.

Arneodo, A., Bacry, E. and Muzy, J. F. (1998) Random Cascades on Wavelet Dyadic Trees. *J. Math. Phys.*, **39(8)**, 4142-4164.

Barron, A., Birgé, L. and Massart, P. (1995) Risk Bounds for Model Selection via Penalization. Preprint.

Benassi, A., Cohen, S. and Istas, J. (1998). Identifying the Multifractal Function of a Gaussian Process. *Statist. Probab. Lett.*. **39(4)**, 337-345.

Beran, J. (1994) *Statistics for Long Memory Processes.* Monogr. on Statist. and Appl. Probab., 61. Chapman and Hall: New York.

Bergh, J. and Löfström, J. (1976). *Interpolation Spaces. An Introduction.* Grundl. der math. Wiss., 223. Springer: Berlin-Heidelberg-New York.

Bergh, J. and Peetre, J. (1974). On the Spaces V_p, $(0 < p \leq \infty)$. *Bol. Unione Mat. Ital. (4)*, **10**, 632-648.

Bertsekas, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press: New York-London-Paris.

Birgé, L. and Massart, P. (1994) From Model Selection to Adaptive Estimation. Festschr. for Lucien Le Cam. Springer: New York, 1997, 55-87.

Boggs, P. T. and Tolle, J. W. (1995). Sequential Quadratic Programming. *Acta Numerica*, 1-51.

Bojanov, B. D., Hakopian, H. A., Sahakian, A. A. (1993). *Spline Functions and Multivariate Interpolations.* Math. and Its Appl., 248. Kluwer, Dordrecht.

Brenner, P., Thomée, V. and Wahlbin, L. (1975). *Besov spaces and Applications to Difference Methods for Initial Value Problems.* Lecture Notes in Mathematics, 434. Springer: Berlin-Heidelberg-New York.

Bunke, O., Droge, B. and Polzehl, J. (1993). Model Selection and Variable Transformations in Nonlinear Regression. *CORE Discussion paper.*

Butzer, P. L. and Berens, H. (1967). *Semi-groups of Operators and Approximation.* Grundl. der Math. Wiss., 145. Springer: Berlin-Heidelberg-New York.

Cai, T. T. (1998) Adaptive Wavelet Estimation: a Block Thresholding and Oracle Inequality Approach. Technical Report.

Cai, T. T. and Brown, L. D. (1998). Wavelet Shrinkage for Nonequispaced Samples. *The Annals of Statist.*, **26**, 1783-1799.

Cai, T. T. and Brown, L. D. (1998a). Wavelet Estimation for Samples with Random Uniform Design. To appear in *Statist. Probab. Lett.*

Cai, T. T. and Silverman, B. W. (1998b). Incorporating Information on Neighboring Coefficients into Wavelet Estimation. Technical Report.

Cai, T. T. and Silverman, B. W. (1998c) Finite-sample Performance of NeighBlock and Neigh-Coeff Estimators. Preprint.

Chipman, H., Kolaczyk, E. and McCulloch, R. (1997). Adaptive Bayesian Wavelet Shrinkage. *J. Amer. Statist. Assoc.*, **92(440)**, 1413-1421.

Chui, C. K. and Quak, E. (1992). Wavelets on a Bounded Interval. In: Braess, D. and Schumaker, L. L. (Eds.) *Numerical Methods of Approximation Theory*, vol. 9. Intern. Ser. Numer. Math., 105. Birkhäuser: Basel, 53-75.

Cohen, A., Dahmen, W. and DeVore, R. (1997). Multiscale Decompositions on Bounded Domains. *Trans. Amer. Math. Soc.* (To appear).

Cohen, A., Daubechies, I. and Vial, P. (1994). Wavelets and Fast Wavelet Transforms on the Interval. *Appl. Comput. Harmon. Anal.*, **1**, 54-81.

Coifman, R. R. and Donoho, D. L. (1995) Translation Invariant Denoising. In: Antoniadis, A. and Oppenheim, G. (Eds.) *Wavelets and Statistics.* Lecture Notes in Statist., 103. Springer: New York, 125-150.

Cox, D. D. (1988). Approximation of Method of Regularization Estimators. *The Annals of Statistics*, **16**, 694-713.

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

Dahmen, W. (1997). Wavelet and Multiscale Methods for Operator Equations. *Acta Numerica*, 55-228.

Dechevski, L. T. (1986) Network-norm Error Estimates of the Numerical Solution of Evolutionary Equations. *Serdica*, **12**, 53-64.

Dechevski, L. T. (1988). τ-moduli and Interpolation. In: Cwikel, M., Peetre, J., Sagher, Y. and Wallin, H. (Eds.). *Theory of Function Spaces and Applications.* Lecture Notes in Math., 1302. Springer: Berlin-Heidelberg-New York, 177-190.

Dechevski, L. T. (1988a). *Some Applications of the Theory of Function Spaces to Numerical Analysis.* Ph. D. Dissertation. Sofia University. (In Bulgarian.)

Dechevski, L. T. (1989). On the Constants of Equivalence between Some Functional Moduli and K-functionals. *Compt. Rend. Acad. Bulg. Sci.*, **42(2)**, 21-24.

Dechevsky, L. T. (1993). On the Sharp Constants of Equivalence between Integral Moduli of Smoothness and K-functionals. Research Report 288, Center for Approximation Theory, Dept. of Math., Texas A&M Univ.

Dechevsky, L. T. (1993a). Explicit Computation of the K-functional between Hilbert Spaces. Preprint 1579, Fachbereich Mathematik, TH Darmstadt.

Dechevsky, L. T. and Dubuc, S. (1998). Multidimensional Dyadic Iterative Interpolation and Fourier Multipliers on Lebesgue Spaces. Research Report CRM-2549. Centre de Recherches Mathématiques, Université de Montréal.

Dechevsky, L. T. and MacGibbon, B. (1998). Asymptotically Minimax Non-parametric Function Estimation with Positivity Constraints I. Preprint.

Dechevsky, L. T., MacGibbon, B. and Penev, S. I. (1998) Numerical Methods for Asymptotically Minimax Non-parametric Function Estimation with Positivity Constraints I. Preprint.

Dechevsky, L. T. and Penev, S. I. (1997). On Shape-preserving Probabilistic Wavelet Approximators. *Stochast. Anal. and Appl.*, **15(2)**, 187-215.

Dechevsky, L. T. and Penev, S. I. (1998). On Shape-preserving Wavelet Estimators of Cumulative Distribution Functions and Densities. *Stochast. Anal. and Appl.*, **16(3)**, 428-469.

Dechevsky, L. T. and Penev, S. I. (1998a). On Penalized Wavelet Estimation. Report S98-15, Dept. of Statist., School of Math., Univ. of New South Wales.

Dechevsky, L. T. and Penev, S. I. (1999). Weak Penalized Least Squares Wavelet Regression Estimation. Report S99-1, Dept. of Statist., School of Math., Univ. of New South Wales.

Dechevsky, L. T. and Persson, L.-E. (1996). On Sharpness, Applications and Generalizations of Some Carleman-type Inequalities. *Tôhoku Math. J.*, **48**, 1-22.

Delecroix, M., Simioni, M. and Thomas-Agnan, C. (1996). Functional Estimation under Shape Constraints. *Nonparam. Statist.*, **6**, 69-89.

Delyon, B. and Juditsky, A. (1996). On Minimax Wavelet Estimators. *Appl. Comput. Harmon. Anal.*, **3**, 215-228.

DeVore, R. A., Kyriazis, G., Leviatan, D. and Tikhomirov, V. M. (1993). Compression and Nonlinear n-widths. *J. Adv. Comp. Math.*, **1**, 197-214.

DeVore, R. A. and Lucier, B. J. (1992). Fast Wavelet Techniques for Near-optimal Image Processing. In: Proc. IEEE Military Communications Conf. IEEE Communications Society: New York.

Donoho, D. L. (1992). Interpolating Wavelet Transforms. Techn. Report, Dept. of Statistics, Stanford Univ.

Donoho, D. and Johnstone, I. (1995) Adapting to Unknown Smoothness via Wavelet Shrinkage. *J. American Statist. Assoc.*, **90(432)**, 1200-1224.

Donoho, D. L., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage: Asymptopia? (with discussion). *J. Royal Statist. Soc., Ser. B*, **57(2)**, 301-369.

Downie, T. R. and Silverman, B. W. (1998) The Discrete Multiple Wavelet Transform and Thresholding Methods. *IEEE Transact. on Signal Processing*, **46**, 2558-2561.

Droge, B. (1993) On Finite-sample Properties of Adaptive Least-squares Regression Estimates. *Statistics*, **24**, 181-203.

Droge, B. (1996) Some Comments on Cross Validation. Discussion paper 7. Sonderforshungsbereich 373, Humboldt University, Berlin.

Dunford, N. and Schwartz, J. T. (1967) *Linear Operators*, vol. 1-3. Wiley: New York.

Feichtinger, H. G. (1997). Amalgam Spaces and Generalized Harmonic Analysis. Norbert Wiener Centenary Congress, East Lansing, MI, 1994. *Proc. Sympos. Appl. Math.*, **52**. Providence, R. I.: AMS, 141-150.

Feichtinger, H. G. and Gröchenig, K.-H. (1989). Banach Spaces Related to Integrable Group Representations and their Atomic Decompositions II. *Monatshefte Math.*, **108**, 129-148.

Fikhtengol'ts, G. M. (1969). *A Course in Differential and Integral Calculus*, v. 2 (7th ed.). Nauka: Moscow. (In Russian.)

Frazier, M. and Jawerth, B. (1990). A Discrete Transform and Decomposition of Distribution Spaces. *J. Functional Analysis*, **93**, 34-170.

Gasser, T. and Müller, H.-J. Kernel Estimation of Regression Functions. In: Gasser, T. and Rosenblatt, M. (Eds.). *Smoothing Techniques for Curve Estimation.* Lecture Notes in Math.,

757. Springer: Heidelberg, 23-68.

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

Hall, P., Kerkyacharian, G. and Picard, D. (1995). Block Threshold Rules for Curve Estimation using Kernel and Wavelet Methods. *The Annals of Statist.*, **26**, 1998, 922-942.

Hall, P. and Patil, P. (1996). On the Choice of Smoothing Parameter, Threshold and Truncation in Nonparametric Regression by Non-linear Wavelet Methods . *J. Royal Statist. Soc., Ser. B*, **58(2)**, 361-377.

Hall, P., Penev, S., Kerkyacharian, G. and Picard, D. (1997). Numerical Performance of Block Thresholded Wavelet Estimators. *Statistics and Computing*, **7(2)**, 115-124.

Hall, P. and Turlach, B. A. (1997). Interpolation Methods for Nonlinear Wavelet regression with Irregularly Spaced Design. *The Annals of Statist.*, **25**, 1912-1925.

Heckman, N. and Ramsay, J. O. (1996). Some General Theory for Spline Smoothing. Preprint.

Henry, D. (1981). *Geometric Theory of Semilinear Parabolic Equations.* Springer: Berlin–Heidelberg–New York.

Hristov, V. H. (1981). On the Coefficients of Fourier–Lagrange. Preprint 5-81-318, JINR, Dubna. (In Russian.)

Hoppenstadt, E. (1969). Asymptotic Series Solutions of Some Nonlinear Parabolic Equations with a Small Parameter. *Arch. Rat. Mech. Anal.*, **35**, 284-298.

Istas, J. and Lang, G. (1997). Quadratic Variations and Estimation of the Local Hölder Index of a Gaussian Process. *Ann. Inst. Henry Poincaré Probab. Statist.*, **33(4)**, 407-436.

Ivanov, K. G. (1985). On the Behaviour of Two Moduli of Functions. *Compt. Rend. Acad. Bulg. Sci.*, **38(5)**, 539-542.

Jaffard, S. (1998) Oscillation Spaces: Properties and Applications to Fractal and Multifractal Functions. *J. Math. Phys.*, **39(8)**, 4129-4141.

Jansen, M., Malfait, M. and Bultheel, A. (1997). Generalized Cross Validation for Wavelet Thresholding. *Signal Processing*, **56**, 33-44.

Johnen, H. and Scherer, K. (1977). On the Equivalence of the K-functional and Moduli of Continuity and Some Applications. In: Schempp, W. and Zeller, K. (Eds.). *Constructive Theory of Functions of Several Variables.* Lecture Notes in Math., 571. Springer: Berlin-Heidelberg-New York, 119-140.

Johnstone, I. M. and Hall, P. G. (1992). Empirical Functionals and Efficient Smoothing Parameter Selection. (With discussion.) *J. Royal Statist. Soc., Ser. B*, **54(2)**, 475-530.

Johnstone, I. and Silverman, B. (1997) Wavelet Threshold Estimators for Data with Correlated Noise. *J. Royal Statist. Soc., Ser. B*, **59(2)**, 319-351.

Jonsson, A. and Wallin, H. (1984). *Function Spaces on Subsets of* $\mathbf{R}^N$. Math. Reports, vol. 2, part 1. Harwood: London.

Karadzhov, G. E. (1973). *The "Means" Interpolation Method for Quasi-normed Spaces and Its Applications.* Ph. D. Dissertation. St. Petersburg Univ. (In Russian.)

Kerkyacharian, G. and Picard, D. (1993) Density Estimation in Besov spaces, *Statistics and Probability Letters*, **18**, 327-336.

Krein, S. G. (1967). *Linear Differential Equations in Banach Space.* Nauka: Moscow. (In Russian.)

Li, K. C. (1985). From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-validation. *The Annals of Statist.*, **13**, 1352-1377.

Löfström, J. (1995). Real Interpolation with Constraints. *J. Approx. Theory*, **82(1)**, 30-53.

MacGibbon, B. and von Sachs, R. (1997). Nonparametric Curve Estimation by Wavelet Thresholding with Locally Stationary Errors. GÉRAD Preprint 97-58.

Meyer, Y. (1991). Wavelets and Operators. Cambridge Univ. Pr.: Cambridge.

Meyer, Y. (1998). *Wavelets, Vibrations and Scalings.* CRM Monogr. Ser., 9, Centre de Recherches Mathématiques, Université de Montréal. AMS: Providence, R. I.

Nason, G. (1996). Wavelet Shrinkage using Cross Validation. *J. Royal Statist. Soc., Ser. B*, **58(2)**, 463-479.

Peetre, J. (1963). *A Theory of Interpolation of Normed spaces.* Lecture Notes, Brasilia (1968). *Notas de matematica*, **39**, 1-86.

Peetre, J. and Thomée, V. (1967) On Rate of Convergence for Discrete Initial-value Problems. *Math. Scand.*, *21*, 159-176.

Penev, S. and Dechevsky, L. (1997). On Non-negative Wavelet-based Density Estimators. *Nonparam. Statist.* **7**, 365-394.

Petrushev, P. P. and Popov, V. A. (1987). *Rational Approximation of Real Functions.* Encycl. of Math. and its Appl., 28. Cambridge Univ. Press: Cambridge.

Pinheiro, A. and Vidakovic, B. (1997). Estimating the Square Root of a Density via Compactly Supported Wavelets. *Comput. Statist. Data Anal.*, **25(4)**, 399-415.

Popov, V. A. and Dechevski, L. T. (1985). On the Error of Numerical Solution of the Parabolic Equation in Network Norms. *Compt. Rend. Acad. Bulg. Sci.*, **36(4)**, 429-432.

Press, W., Teukolsky, S., Vettering, W. and Flannery, B. (1992). *Numerical Recipes in Fortran: the Art of Scientific Computing* (2nd ed.). Cambridge Univ. Press: Cambridge.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis.* Springer Ser. in Statistics. Springer: Berlin–Heidelberg–New York.

Robertson, T., Wright, F. T and Dykstra, R. L. (1988). *Order Restricted Statistical Inference.* Wiley: Chichester.

Samko, S. G., Kilbas, A. A. and Marichev, O. I. (1987). *Integrals and Derivatives of Fractional Order and some of their Applications.* Nauka i Tehnika: Minsk. (In Russian.)

Sendov, Bl. and Popov, V. A. (1988). *The Averaged Moduli of Smoothness. Applications in Numerical Methods and Approximation.* Wiley: Chichester.

Sickel, W. (1990). Spline Representations of Functions in Besov-Triebel-Lizorkin Spaces on $\mathbf{R}^n$. *Forum Math.*, **2**, 451-475.

Sparr, G. (1974). Interpolation of Several Banach Spaces. *Ann. Mat. Pura Appl.*, **99**, 247-316.

Thomée, V. (1969). Stability Theory for Partial Difference Operators. *SIAM Review*, **11(2)**, 152-195.

Thomée, V. (1971). Convergence Estimates in Discrete Initial Value Problems. *Actes Congrès Intern. Math.*, Nice'70, 321-329.

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

Triebel, H. (1983). Theory of Function Spaces. Monogr. in Math., 78. Birkhäuser: Basel-Boston-Stuttgart.

Triebel, H. (1992). Theory of Function Spaces II. Monogr. in Math., 84. Birkhäuser: Basel-Boston-Stuttgart.

Triebel, H. (1997). *Fractals and Spectra Related to Fourier Analysis and Function Spaces.* Monogr. in Math., 91. Birkhäuser: Basel-Boston-Berlin.

Utreras, F. D. (1979). Cross-validation Techniques for Smoothing Spline-functions in One or Two Dimensions. In: Gasser, T. and Rosenblatt, M. (Eds.). *Smoothing Techniques for Curve Estimation.* Lecture Notes in Math., 757. Springer: Heidelberg, 196-231.

Vainberg, M. M. (1979). *Functional Analysis.* Prosveshtenie: Moscow. (In Russian.)

Vidakovic, B. (1994) Nonlinear Wavelet Shrinkage with Bayes Rules and Bayes Factors. *Discussion Paper* 94-24. ISDS, Duke University, USA.

Wahba, G. (1990) *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics, 59. SIAM: Philadelphia.

Walter, G. G. and Shen, X. Continuous Non-negative Wavelets and their Use in Density Estimation. Preprint, 1997.

Wang, Y. (1997). Fractal Function Estimation via Wavelet Shrinkage. *J. Royal Statist. Soc., Ser. B*, **59(3)**, 603-613.

Widlund, O. B. (1968). On the Rate of Convergence for Parabolic Difference Schemes I. *Proc. AMS Sympos. Appl. Math.*, Durham, N. C., 60-73.

Widlund, O. B. (1968). On the Rate of Convergence for Parabolic Difference Schemes II. *Commun. Pure Appl. Math.*, **23**, 79-96.

Wiener, N. (1924). The Quadratic Variation of a Function and Its Fourier Coefficients. *Mass. J. Math. Phys.*, **3**, 72-94.

Yosida, K. (1974). *Functional Analysis.* Grundl. der math. Wiss., 123. Springer: New York-Heidelberg.

Young, L. C. (1937). Sur une généralisation de la notion de puissance p-ième bornée au sens de Wiener, et sur la convergence de series de Fourier. *Compt. Rend. Acad. Sci.*, **204(7)**, 470-472.