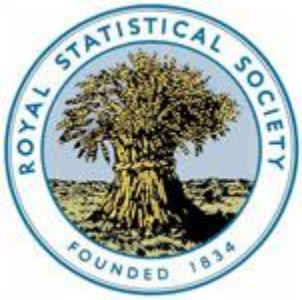

J. R. Statist. Soc. B (1995)
57, *No.* 2, *pp.* 301–369

Wavelet Shrinkage: Asymptopia?

By DAVID L. DONOHO and IAIN M. JOHNSTONE†,

Stanford University, USA

GÉRARD KERKYACHARIAN and DOMINIQUE PICARD

Université de Picardie, Amiens, France *Université de Paris VII, France*

[*Read at the Joint Congress of the Institute of Mathematical Statistics and the Bernoulli Society at an Ordinary Meeting of* The Royal Statistical Society *in Chapel Hill on Monday, June 20th, 1994, the President*, Professor D. J. Bartholomew, *in the Chair*]

SUMMARY

Much recent effort has sought asymptotically minimax methods for recovering infinite dimensional objects—curves, densities, spectral densities, images—from noisy data. A now rich and complex body of work develops nearly or exactly minimax estimators for an array of interesting problems. Unfortunately, the results have rarely moved into practice, for a variety of reasons—among them being similarity to known methods, computational intractability and lack of spatial adaptivity. We discuss a method for curve estimation based on n noisy data: translate the empirical wavelet coefficients towards the origin by an amount $\sqrt{(2 \log n)}\sigma/\sqrt{n}$. The proposal differs from those in current use, is computationally practical and is spatially adaptive; it thus avoids several of the previous objections. Further, the method is nearly minimax both for a wide variety of loss functions—pointwise error, global error measured in L^p-norms, pointwise and global error in estimation of derivatives—and for a wide range of smoothness classes, including standard Hölder and Sobolev classes, and bounded variation. This is a much broader near optimality than anything previously proposed: we draw loose parallels with near optimality in robustness and also with the broad near eigenfunction properties of wavelets themselves. Finally, the theory underlying the method is interesting, as it exploits a correspondence between statistical questions and questions of optimal recovery and information-based complexity.

Keywords: ADAPTIVE ESTIMATION; BESOV SPACES; DENSITY ESTIMATION; MINIMAX ESTIMATION; NONPARAMETRIC REGRESSION; OPTIMAL RECOVERY; SPATIAL ADAPTATION; WAVELET ORTHONORMAL BASES

1. CLASSICAL MINIMAXITY

Consider the problem of estimating a single normal mean. We have data $Y \sim N(\theta, \sigma^2)$ and we wish to estimate θ. We choose a loss function $l(t)$ and define the risk $R(\hat{\theta}, \theta) = E_\theta l\{\hat{\theta}(Y) - \theta\}$. Then, in a sense described by the minimax theorem (Wolfowitz, 1950), the estimator $\hat{\theta}(Y) = Y$ is optimal: if the loss l is symmetric and bowl shaped,

$$R(Y, \theta) = \inf_{\hat{\theta}} \sup_{\theta} \{R(\hat{\theta}, \theta)\}. \tag{1}$$

This simple and natural result has many familiar implications. For example, the minimax estimator of a mean μ from n samples $X_1, \ldots, X_n$, with $X_i \sim_{\text{IID}} N(\mu, \sigma^2)$ is just the sample mean $\overline{X}$. There are a variety of asymptotic implications, via the

†*Address for correspondence*: Department of Statistics, Stanford University, Stanford, CA 94305, USA.
E-mail: imj@stat.stanford.edu

 0035–9246/95/57301

theory of local asymptotic normality; for example, that in parametric settings the maximum likelihood estimator is locally asymptotically minimax, and that in nonparametric settings the sample median of $X_1, \ldots, X_n$, with $X_i \sim_{\text{IID}} F$, is locally asymptotically minimax for estimating the median med(F).

An important aspect of result (1) is *generality*: the form of the minimax estimator does not depend on the loss function. Hence Y is optimal for a wide variety of purposes, and not just for minimum mean-square estimation.

2. POST-CLASSICAL MINIMAXITY

In recent years, mathematical statisticians have been interested in estimating infinite dimensional parameters—curves, densities, images, ... A paradigmatic example is the problem of *nonparametric regression*,

$$y_i = f(t_i) + \sigma z_i, \qquad i = 1, \ldots, n, \tag{2}$$

where f is the unknown function of interest, the t_i are equispaced points on the unit interval and $z_i \sim_{\text{IID}} N(0, 1)$ is Gaussian white noise. Other problems with similar character are *density estimation*, recovering the density f from $X_1, \ldots, X_n \sim_{\text{IID}} f$, and *spectral density estimation*, recovering f from $X_1, \ldots, X_n$ a segment of a Gaussian zero-mean second-order stationary process with spectral density f.

After extensive study of this setting, mathematical statisticians have achieved some important results, a few of which we describe below. Unfortunately, such results lack the coherence and simplicity of the classical minimax result (1). Instead of a single, natural minimax theorem, there is a whole forest of results, growing in various and sometimes conflicting directions; it requires considerable effort to master and keep abreast of this rapidly developing body of knowledge. Moreover, as we shall see, the literature's very complexity has generated, in the practically minded, a certain degree of scepticism of theory itself.

The literature has reached the current complex demanding state by playing out a series of questions and responses with their own internal logic.

2.1. *No Immediate Infinite Dimensional Analogue of Result (1)*

By the 1960s it was known that it was not possible to derive estimates which work well for *every* function f. Results appeared showing that, for any estimator $\hat{f}$, there was a function f which caused it to misbehave, so that

$$\sup_f \{R_n(\hat{f}, f)\} \not\to 0, \qquad n \to \infty. \tag{3}$$

Compare Farrell (1967) and the more refined negative results of Birgé (1985).

2.2. *Development of a Minimax Paradigm*

To derive a non-void theory—i.e. one containing positive results—it was necessary to look elsewhere than result (1). In the 1970s and 1980s a certain *minimax paradigm* developed, in a long series of work by many researchers worldwide. In this paradigm we seek solutions to minimax problems over bounded parameter spaces. This paradigm has three basic parts.

First, we assume that the function of interest belongs to a specific, known, functional 'ball' $\mathscr{F}(C)$. Standard examples include Hölder balls,

$$\Lambda^{\alpha}(C)=\{f:\ |f(x)-f(y)|\leqslant C|x-y|^{\alpha}\}, \tag{4}$$

if $0<\alpha<1$, with generalizations to $\alpha>1$ (see equation (15)), the L^2 Sobolev balls

$$W_2^m(C)=\left\{f:\ \int_0^1 f^{(m)}(t)^2\,dt\leqslant C^2\right\}, \tag{5}$$

where $f^{(m)}(t)$ denotes the mth derivative of f at t, and the L^p Sobolev classes

$$W_p^m(C)=\left\{f:\ \int_0^1 |f^{(m)}(t)|^p\,dt\leqslant C^p\right\}. \tag{6}$$

Second, we assume a specific risk measure. Standard examples include *risk at a point*,

$$R_n(\hat{f}, f)=E\{\hat{f}(t_0)-f(t_0)\}^2, \tag{7}$$

global squared L^2-norm risk,

$$R_n(\hat{f}, f)=E\|\hat{f}-f\|^2_{L^2[0,1]}, \tag{8}$$

and other measures, such as risk in estimating some derivative at a point, or estimating the function with global L^p-loss or estimating some derivative of the function with global L^p-loss.

Third, we attempt to solve for an estimator which is *minimax* for the class $\mathscr{F}(C)$ and risk R_n,

$$\sup_{\mathscr{F}(C)}\{R_n(\hat{f}, f)\}=\inf_{\hat{f}}\sup_{\mathscr{F}(C)}\{R_n(\hat{f}, f)\}, \tag{9}$$

or, if that proves too difficult, which is *asymptotically minimax*,

$$\sup_{\mathscr{F}(C)}\{R_n(\hat{f}, f)\}\sim\inf_{\hat{f}}\sup_{\mathscr{F}(C)}\{R_n(\hat{f}, f)\},\qquad n\to\infty, \tag{10}$$

or, even if that proves still too difficult, as it usually does, which *attains the minimax rate*

$$\sup_{\mathscr{F}(C)}\{R_n(\hat{f}, f)\}\asymp\inf_{\hat{f}}\sup_{\mathscr{F}(C)}\{R_n(\hat{f}, f)\},\qquad n\to\infty. \tag{11}$$

Particularly in the case where exact or asymptotic optimality obtains, we may think of this three-part paradigm as a process of 'rational estimator design': we obtain an estimator $\hat{f}$ as the solution of a certain optimality problem.

2.3. *Implementation of Minimax Paradigm*

In the 1970s and 1980s, the minimaxity paradigm was developed to fruition. The space of possible results is a four-dimensional factorial design, where one specifies the observation model (regression, density, spectral density), risk R_n type (at a point, globally) and form (L_2, L_1, L_∞, ...), and function class $\mathscr{F}$ (Hölder, Sobolev, ...).

Many combinations of these factors have now been explored, and minimaxity and near minimaxity results have been obtained for a wide variety of cases.

A sampling of these results would go as follows.

(a) Speckman (1979) showed that for estimating a function at a point $f(t_0)$ with squared error loss, with ellipsoidal (L^2-smoothness) class $\mathscr{F}(C)$, the penalized spline estimate is minimax among linear estimates. Actually, it is nearly minimax among all estimates (Ibragimov and Khas'minskii, 1982; Donoho and Liu, 1991; Donoho *et al.*, 1990; Donoho, 1994a).

(b) Sacks and Ylvisaker (1981) showed that for estimating a function at a point, with squared error loss and a quasi-Hölder class $\mathscr{F}(C)$, the linear minimax estimate is a kernel estimate with specially chosen kernel and specially chosen bandwidth; this estimate is within 17% of asymptotically minimax among all procedures (Donoho and Liu, 1991). Donoho and Liu (1991) also showed how to derive optimal kernels for true Hölder classes.

(c) Bretagnolle and Carol-Huber (1979), Stone (1982) and Ibragimov and Khas'minskii (1982) studied problems of estimating the whole object with global loss $\|\hat{f}_n - f\|_{L^p}^p$ and L^p Sobolev *a priori* class $W_m^p(C)$ (same p in both) and found that certain kernel estimates attain the minimax rate—i.e. achieve rate (11).

(d) Pinsker (1980), Efromovich and Pinsker (1981, 1982) and Nussbaum (1985) showed that for estimating the whole object with quadratic global loss $\|\hat{f}_n - f\|_{L^2}^2$, and L^2 Sobolev *a priori* class $W_m^2(C)$, a windowed Fourier estimate is asymptotically minimax—i.e. achieves approximation (10). This was the first infinite dimensional estimation problem in this category in which precise asymptotic minimaxity was achieved.

(e) Korostelev (1993) and Donoho (1994b) showed that for estimating the whole object or its kth derivative with sup-norm global loss $\|\hat{f}_n^{(k)} - f^{(k)}\|_{L^\infty}$, and Hölder *a priori* class $\Lambda^\alpha(C)$, a certain kernel estimate is asymptotically minimax—i.e. achieves approximation (10).

(f) In one of the most surprising developments, Nemirovskii *et al.* (1985) and Nemirovskii (1986) showed that for estimating functions in certain classes (e.g. decreasing functions, Sobolev spaces W_1^m), and certain loss functions (e.g. L^p-loss, $p > 1$), no linear method can achieve the optimal rate. Thus kernel, spline and windowed Fourier methods face problems that they cannot solve, even at the level (11). In principle a certain least squares projection operator, finding the closest object from the class $\mathscr{F}(C)$ to the data, achieves the optimal rate in such cases. For most classes $\mathscr{F}(C)$ this method is non-linear.

(g) Birgé (1983) showed that a certain method based on throwing down ϵ-coverings of the parameter space $\mathscr{F}(C)$ by balls, and then testing between balls, achieves the minimax rate for quite general classes $\mathscr{F}(C)$.

There are many other significant results in this highly developed literature; we list here only the very few that we refer back to later.

2.4. *Practical Indifference*

Despite the impressive array of technical achievements present in the above work, the reaction of the general statistical community has not been uniformly enthusiastic.

For example, a large number of computer packages have appeared over the last 15 years, but the work of the minimax paradigm has had relatively little effect on software. We identify several explanations for this.

2.4.1. *Philosophical common sense*

The minimax paradigm designs estimators on the assumption that certain smoothness conditions hold; yet we never know such smoothness to be the case. (There are even results showing that it is impossible to tell whether or not a function belongs to some W_p^m (Donoho, 1988).) There is therefore a disconnection between the suppositions of the minimax paradigm and the actual situation when we are confronted with real data. This makes the applicability of the results *a priori* doubtful.

This concern would be of little import if the results of working through the paradigm did not much depend on the assumptions; but in fact they do. Different assumptions about $\mathscr{F}(C)$ and R_n lead to markedly incompatible estimators. For example, if we assume that the underlying object is Lipschitz, $|f(x)-f(y)| \leqslant C|x-y|$, with known Lipschitz constant C, and we wish to estimate f at the point t_0 (risk measure (7)), then a minimax kernel estimator has as kernel the solution of a special optimization problem, and a bandwidth $h_n \asymp n^{-1/3}$ attains the minimax rate $n^{-2/3}$. However, if we assume two L^2-derivatives and global L^2-loss, then an estimator with bandwidth $h_n \asymp n^{-1/5}$ attains the minimax rate $n^{-4/5}$. But suppose that we use the method designed under one set of assumptions to solve the problem defined by the other set of assumptions. The outcome will be disappointing in both cases;

(a) the estimator designed for a Lipschitz function attains only the rate $n^{-2/3}$ in case f has two L^2-derivatives, not $n^{-4/5}$;
(b) the estimator designed for two L^2-derivatives may have a risk tending to 0 at rate $n^{-2/5}$ in case f is only Lipschitz, and not $n^{-2/3}$.

But suppose that neither assumption holds, for example that the function is only of bounded variation. Under the assumption of global loss (8) and $\mathscr{F}(C)$ the collection of functions of bounded variation less than or equal to C, the estimator assuming Lipschitz behaviour has a risk tending to 0 like $n^{-1/3}$; the estimator assuming two L^2-derivatives has a risk tending to 0 like $n^{-1/5}$. Both fall far short of the minimax rate, which is $n^{-2/3}$. In this case, moreover, the issue is not just the proper choice of bandwidth; no linear method achieves better than the rate $n^{-1/2}$ uniformly over bounded variation balls, so that any kernel method is unsatisfactory.

2.4.2. *Computational common sense*

Minimaxity results are sometimes held to be uninteresting from a practical point of view. The methods most frequently discussed—kernel methods, spline methods, orthogonal series—were already well known by practitioners before the minimax paradigm was in place. From this point of view, the principal findings of the minimaxity literature—optimal kernels, optimal bandwidths, optimal penalization and so forth—amount to minor variations on these themes, rather than wholesale innovations.

Complementary is the claim that those methods coming out of minimaxity theory which are really new are also impractical. For example, Nemirovskii in a personal communication explained that he had not succeeded in implementing his

least-squares-based method on data sets of realistic size, because it required the solution of a non-linear optimization problem whose running time went up roughly as $O(n^{3.5})$ for n data. The abstract ϵ-covering approach of Birgé is perhaps even more challenging to implement; it requires the implementation of a code for laying down an ϵ-covering on the function space $\mathscr{F}(C)$, and we know of no practical example of such a method in use.

2.4.3. *Spatial common sense*

A third argument for scepticism takes as given that theoretical methods found by the minimax paradigm are, generally, spatially non-adaptive, whereas real functions exhibit a variety of shapes and spatial inhomogeneities. It holds that such spatially variable objects should be addressed by spatially variable methods. Since the minimax paradigm does not seem to give methods with such properties, it argues, minimaxity should be abandoned; it concludes that we should construct methods (heuristically, if necessary) which address the 'real problem'—spatial adaptation.

This point of view has had considerable influence on software development and daily statistical practice, apparently much more than the minimax paradigm. Interesting spatially adaptive methods include classification and regression trees (CART) (Breiman *et al.*, 1983), TURBO (Friedman and Silverman, 1989), multivariate adaptive regression splines (MARS) (Friedman, 1991) and variable bandwidth kernel methods (Breiman *et al.*, 1977; Müller and Stadtmuller, 1987; Terrell and Scott, 1992; Brockmann *et al.*, 1993). Such methods implicitly or explicitly attempt to adapt the fitting method to the form of the function being estimated, by ideas like recursive dyadic partitioning of the space on which the function is defined (CART and MARS), adaptively pruning away knots from a complete fit (TURBO) and adaptively estimating a local bandwidth function (variable kernel methods).

The spatial adaptivity camp is, to date, atheoretical, as opposed to antitheoretical, motivated by the heuristic plausibility of their methods, and pursuing practical improvements rather than conclusive theoretical results which might demonstrate specific quantitative advantages of such methods. But, in our experience, the need to adapt spatially is so compelling that the methods have spread far in the last decade, even though the case for such methods has not been proven rigorously.

2.5. *Recent Developments*

The difficulties enumerated above have been partially addressed by the minimax community in recent years.

The seminal proposal of Wahba and Wold (1975) to choose smoothing parameters adaptively by cross-validation has opened the possibility that we can adapt to the unknown smoothness of an object in a simple, automatic way. Translated into the minimax paradigm, the issue becomes the following: can we design a single method $\hat{f}$ which is *simultaneously asymptotically minimax*, i.e. which attains

$$\sup_{\mathscr{F}(C)}\{R(\hat{f}_n, f)\}=\{1+o(1)\}\inf_{\hat{f}}\sup_{\mathscr{F}(C)}\{R(\hat{f}, f)\} \tag{12}$$

for every ball $\mathscr{F}(C)$ arising in a certain function scale? (Corresponding notions of simultaneously asymptotically rate minimax can be defined in the obvious way.)

The existence of such estimators would go far towards alleviating the philosophical objection listed above, namely that 'you never know $\mathscr{F}(C)$'.

Pioneering work in this direction was by Efromovich and Pinsker (1984), who developed a method which exhibited equation (12) for every L^2 Sobolev ball. The method is based on adaptively constructing a linear orthogonal series estimator by determining optimal damping coefficients from data. Compare also Golubev (1987).

Unfortunately, the idea is based on adapting an underlying linear scheme to the underlying function, so it adapts over only those classes where linear methods attain the minimax rate. For other function classes, such as the class of bounded variation, the method is unable to approach the minimax rate.

Another important development was a theory of spatial adaptivity for the Grenander estimator due to Birgé (1989). The Grenander estimator is a method for estimating a monotone density. It is non-linear and, in general, difficult to analyse. Birgé succeeded in showing that the Grenander estimator came within a factor 2 of a kind of optimally adaptive procedure, namely that histogram estimator with variable width bins which achieves the minimum risk among all histogram estimators.

An extension of such results to a general theory of spatial adaptation would be the next step, e.g. to find a non-linear estimator which achieves essentially the same performance as the best piecewise polynomial fit. However, until now, such an extension has been lacking.

2.6. *Epilogue*

The literature on minimax estimation of curves, densities and spectra has elucidated the behaviour of many different proposals under many different choices of loss and smoothness class. The literature has not converged, however, to a single proposal which is simple, natural and works in an optimal or near optimal way for a wide variety of losses and smoothness classes; even the Efromovich–Pinsker estimator, which seems quite general, fares badly over certain smoothness classes.

Another issue is that, of course, the simple model (2) is not by itself the beginning and end of statistical estimation; it is simply a test-bed which we can use to develop ideas and techniques. It is important that whatever is developed generalizes beyond that model, to handle inverse problems, where we have noisy and indirect data—e.g. inverse problems of tomography, deconvolution and Abel inversion. From this point of view, the atheoretical spatial adaptation proposals have defects—they do not seem to generalize, say in the tomographic case, to adapt spatially to structure in the underlying object.

The net result is that the minimax paradigm has led to complexity, nuance, uncertain generality and, finally, to a psychology of qualification and specialization.

3. WAVELET SHRINKAGE

Recently, a growing and enthusiastic community of applied mathematicians has developed the wavelet transform as a tool for signal decomposition and analysis. The field is growing rapidly, both as a practical, algorithm-oriented enterprise and as a field of mathematical analysis. Daubechies (1992) features an algorithmic viewpoint about the wavelet transform; Meyer (1990) and Frazier *et al.* (1991) feature the

functional space viewpoint. Further references and descriptions may be found in our other papers.

Proper deployment of this tool allows us to avoid many of the difficulties, hesitations, qualifications and limitations in the existing statistical literature.

3.1. *The Method*

For simplicity, we focus on the nonparametric model (2) and a proposal 'VisuShrink' of Donoho and Johnstone (1994a); similar results are possible in the density estimation model (Johnstone *et al.*, 1992). We suppose that we have $n = 2^{J+1}$ data of the form (2) and that σ is known.

(a) Take the n given numbers and apply an empirical wavelet transform W_n^n, obtaining n empirical wavelet coefficients $(w_{j,k})$. This transform is an order $O(n)$ transform, so that it is very fast to compute, in fact faster than the fast Fourier transform. A brief description of one such W_n^n is given in Appendix A.
(b) Set a threshold $t_n = \sqrt{(2 \log n)}\sigma/\sqrt{n}$, and apply the soft threshold non-linearity $\eta_t(w) = \mathrm{sgn}(w)(|w| - t)_+$ with threshold value $t = t_n$, i.e. apply this non-linearity to each of the n empirical wavelet coefficients. (In practice, the coefficients at the very coarsest scales are not shrunk—see later.)
(c) Invert the empirical wavelet transform, obtaining the estimated curve $\hat{f}_n^*(t)$.

We mention now some elaborations of the proposal. First, in practice, we do not shrink the coefficients at the very coarsest scales. In the wavelet transform there is a set of coefficients at $j \leqslant j_0$ measuring 'gross structure'; these correspond to basis functions derived from 'father wavelets'; the remainder derive from 'mother wavelets' and measure detail structure. In practice, we only shrink the detail coefficients. Secondly, when the noise level σ is unknown, we take the median absolute deviation of the wavelet coefficients at the finest scale of resolution and divide by 0.6745 to obtain a crude estimate $\hat{\sigma}$. Finally, we can treat densities, spectral densities, indirect data, non-white noise and non-Gaussian data by various simple elaborations of this proposal; see the discussion.

Figs 1(a) and 1(b) show this method (as elaborated in the last paragraph of this section) applied to some nuclear magnetic resonance data supplied by Chris Raphael. The base-line noise has been largely removed without a blurring of the important peaks. A key point is that in the wavelet domain, shown in Figs 1(c) and 1(d), the noise is spread fairly uniformly among all coefficients, whereas the signal is quite sparse, being concentrated into a small number of coefficients. This is the practical motivation for thresholding: we give some theoretical justification in Section 4.

Two further real data examples appear at the end of Section 5.

Fig. 2 shows four spatially inhomogeneous functions—*Bumps*, *Blocks*, *Heavisine* and *Doppler*. These functions, explicitly defined in Donoho and Johnstone (1994a), are designed to display some features encountered in cross-sections of images, mass spectra, spatially varying frequencies, discontinuities etc. Fig. 3 shows noisy versions, according to model (2). Fig. 4 shows reconstructions.

These reconstructions display two properties of interest. The first is their almost *noise-free* character. There is very little of the random oscillation that one associates with noise. The second property is that *sharp features have stayed sharp* in

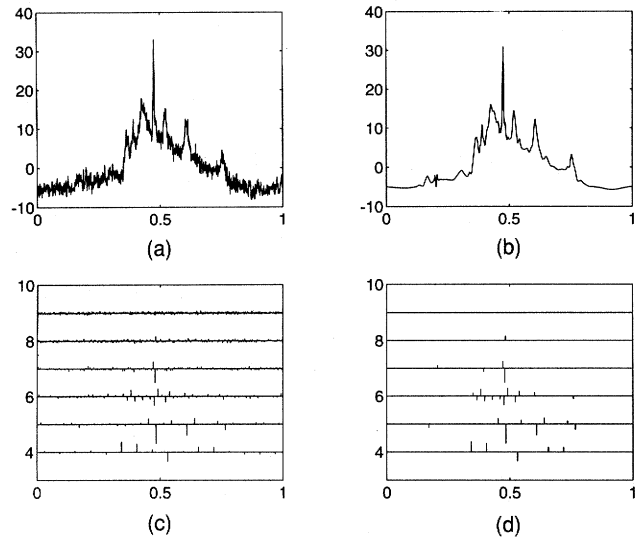

Fig. 1. (a) Nuclear magnetic resonance data from Chris Raphael ($n = 1024$ points); (b) wavelet reconstruction using most nearly symmetric Daubechies wavelets with $N = 6$ and hard thresholding (low frequency cut-off $j_0 = 5$; scale estimated as the median absolute deviation from the median of wavelet coefficients at the highest level); (c) empirical wavelet coefficients; (d) coefficients after thresholding

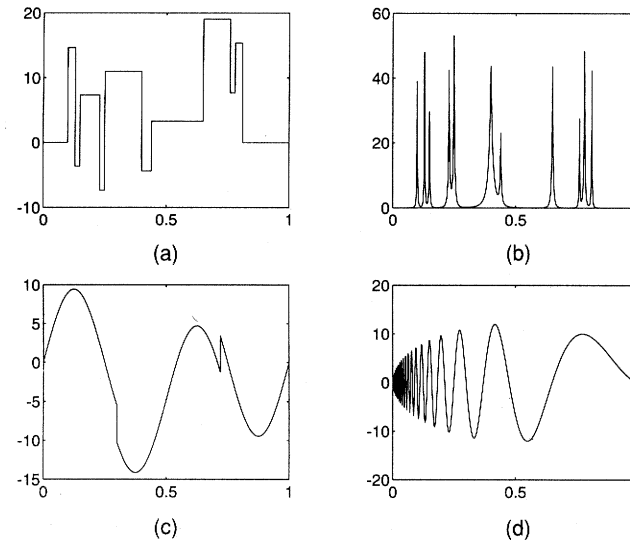

Fig. 2. Four spatially variable functions ($n = 2048$; formulae given in Donoho and Johnstone (1994a)): (a) Blocks; (b) Bumps; (c) Heavisine; (d) Doppler

reconstruction. These two properties are not easy to combine. Linear approaches (such as kernel, spline and windowed Fourier methods) inevitably either blur the sharp features and damp the noise or leave the features intact, but leave the noise intact as well. For comparison, see Figs 5 and 6, which display the results of spline and Fourier series estimates with adaptively chosen penalization and windowing parameters

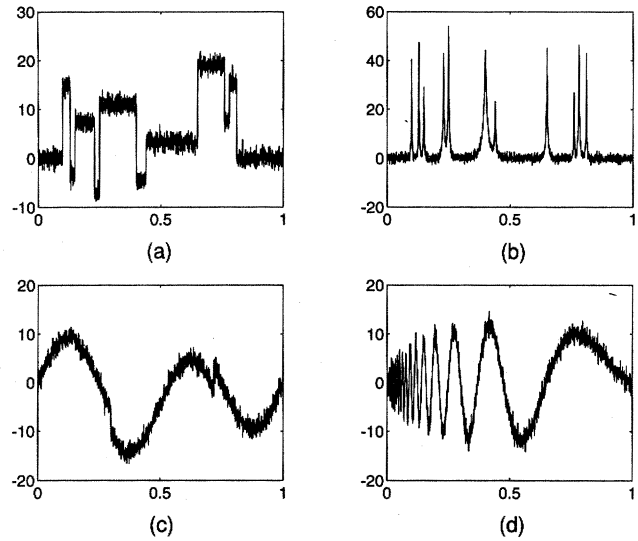

Fig. 3. Four functions with Gaussian white noise, $\sigma = 1$, signal rescaled to have signal-to-noise ratio $\mathrm{SD}(f)/\sigma = 7$: (a) noisy Blocks; (b) noisy Bumps; (c) noisy Heavisine; (d) noisy Doppler

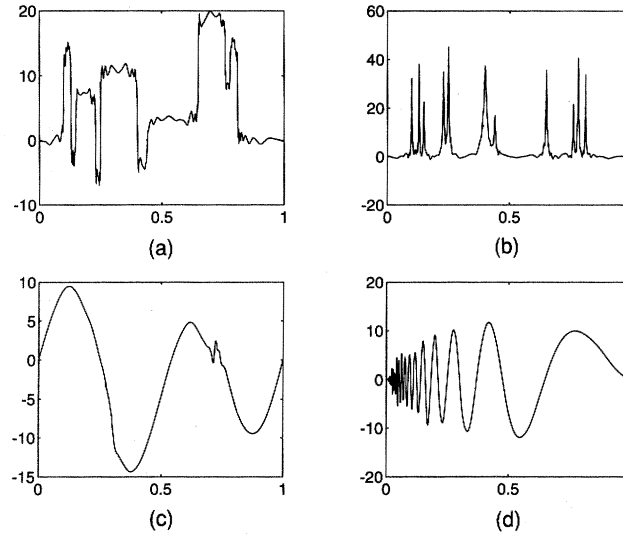

Fig. 4. Reconstructions using soft thresholding and $\lambda = \sqrt{(2 \log n)}$, with most nearly symmetric Daubechies wavelets with $N = 8$ vanishing moments, and low frequency cut-off $j_0 = 6$—notice the 'noise-free' character: (a) noisy Blocks; (b) noisy Bumps; (c) noisy Heavisine; (d) noisy Doppler

respectively, selected to minimize unbiased estimates of mean-squared error. The spline method blurs certain features of the object, such as jumps, while exhibiting certain noise-induced oscillations in areas that ought to be smooth; the windowed Fourier series method is similar: it tends to preserve the features, but again without damping the noise.

These two visual properties of wavelet shrinkage reconstruction prefigure various theoretical benefits to be discussed below.

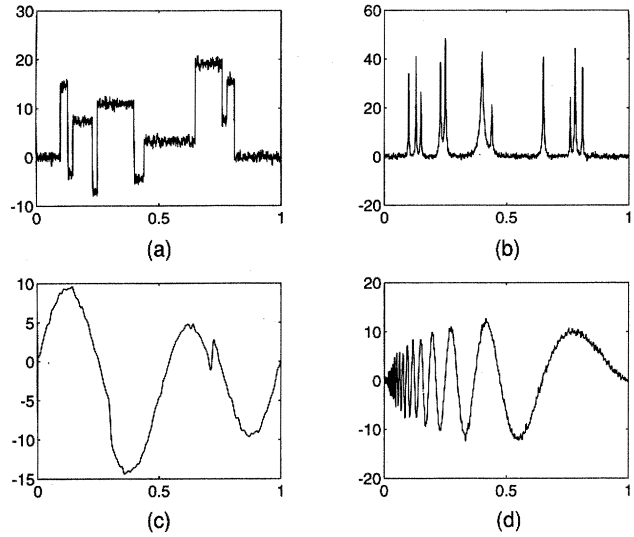

Fig. 5. Reconstructions using spline smoothing with the regularization parameter chosen from the data to minimize the unbiased estimate of mean-squared error: (a) Blocks; (b) Bumps; (c) Heavisine; (d) Doppler

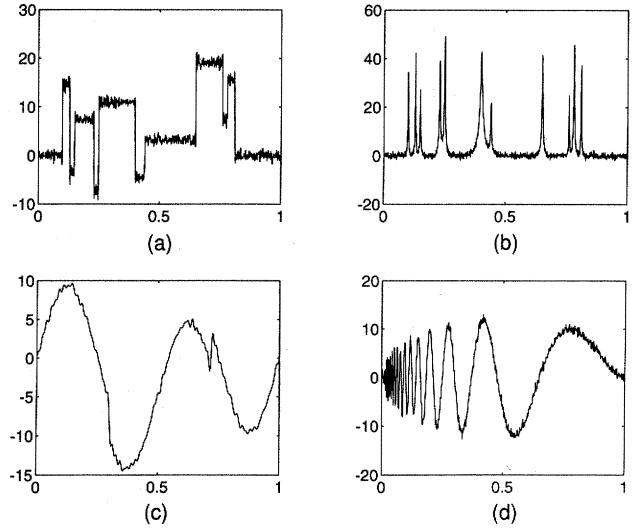

Fig. 6. Reconstructions using a truncated Fourier series, with the number of terms chosen from the data to minimize the unbiased estimate of mean-squared error: (a) Blocks; (b) Bumps; (c) Heavisine; (d) Doppler

Wavelet shrinkage depends in an essential way on the *multiresolution* nature of the wavelet transform. The transform that we use in our examples is either periodic, of the form described briefly in Appendix A, or boundary corrected, as formulated by Cohen *et al.* (1992). (Further details are given in the figure captions, and the figures may be reproduced by interested readers as described in Appendix A.) Our reconstruction takes the form

$$\hat{f}_n^* = \sum_k w_{j_0,k}\phi_{j_0,k} + \sum_{j \geq j_0,k} \hat{\alpha}_{j,k}\psi_{j,k} \tag{13}$$

where the 'scaling functions' $\phi_{j_0,k}$ and 'wavelets' $\psi_{j,k}$ are smooth wiggly functions of 'scale' 2^{-j} and 'position' $k/2^j$. The thresholding gives wavelet coefficients $\hat{\alpha}_{j,k}$, many of which are 0. The result is a sparse reconstruction, with significant contributions from many different scales. By contrast, traditional linear smoothing methods operate in a *monoresolution* fashion, at best with the resolution scale chosen adaptively; the resolution scale is, of course, the bandwidth. Figs 1(c) and 1(d) show the empirical and thresholded wavelet coefficients stratified by scale; contributions of several different scales are present in the display. Thus, although the choice of gross structure scale j_0 plays an important role in practice, the multiscale nature of the non-linear estimator (13) precludes its interpretation as anything other than a 'maximal' bandwidth.

3.2. *Our Claims*

Wavelet shrinkage avoids many of the objections to minimax theory listed in Section 2.4. The method makes no *a priori* assumption that f belongs to any fixed smoothness class; it even accommodates discontinuities, as the figures show. The method is simple and practical, with an algorithm that functions in order $O(n)$ operations. The method is also new, not just a minor variation on something previously in widespread use. The method is spatially adaptive, being able to preserve the spatially inhomogeneous nature of the estimand. Finally, the wavelet shrinkage method also generalizes to high dimensional data, to density estimation and to the treatment of various inverse problems.

While avoiding many common sense objections, the estimator $\hat{f}_n^*$ is nearly optimal for a wide variety of theoretical objectives. It is nearly optimal from the point of view of spatial adaptation. It is nearly optimal from the point of view of estimating an object of unknown smoothness at a point. And it is nearly optimal from the point of view of estimating an object of unknown smoothness in any one of a variety of global loss measures, ranging from L^p-losses to L^p-losses on derivatives, and far beyond.

In brief then, we claim that the wavelet shrinkage method offers all that we might desire of a technique, from optimality to generality, and that it answers by and large the conundrums posed by the current state of minimax theory.

3.3. *Basic Results*

We now state with somewhat more precision the properties of the wavelet shrinkage estimator introduced above. We first mention properties which have been proved elsewhere.

3.3.1. *$\hat{f}_n^*$ is, with high probability, as smooth as the truth*

The empirical wavelet transform is implemented by the pyramidal filtering of Cohen *et al.* (1992); this corresponds to a theoretical wavelet transform which furnishes an orthogonal basis of $L^2[0, 1]$. This basis has elements (wavelets) which are in C^R and have, at high resolutions, D vanishing moments. The fundamental discovery about wavelets that we shall be using is that they provide a 'universal' orthogonal basis: an unconditional basis for a very wide range of smoothness spaces; all the

Besov classes $B^\sigma_{p,q}[0, 1]$ and Triebel classes $F^\sigma_{p,q}[0, 1]$ in a certain range $0 \leqslant \sigma < \min(R, D)$.

Each of these function classes has a norm $\| \|_{B^\sigma_{p,q}}$ or $\| \|_{F^\sigma_{p,q}}$ which measures smoothness. The parameter σ measures the number of derivatives, where the existence of derivatives is required in an L_p-sense (and the parameter q provides a further finer gradation). Special cases include the traditional Hölder (–Zygmund) classes $\Lambda^\alpha = B^\alpha_{\infty,\infty}$ and Sobolev classes $W^m_p = F^m_{p,2}$. In fact, the three parameter classes contain many of the function spaces of analysis (see Triebel (1990)). Furthermore, the Besov and Triebel classes are of statistical and scientific interest because they allow for better models of spatial inhomogeneity. For example, the 'bump algebra' $B^1_{1,1}$ consists of functions f representable as (infinite) linear combinations (with summable coefficients) of arbitrarily located, scaled and signed Gaussian densities. The space of functions of bounded total variation lies sandwiched between $B^1_{1,1}$ and $B^1_{1,\infty}$.

For more about the universal basis property, see Lemarié and Meyer (1986) or Frazier *et al.* (1991) and Meyer (1990). A few details are collected in Appendix A for the reader's convenience.

Definition 1. $\mathscr{L}(R, D)$ is the scale of all spaces $B^\sigma_{p,q}$ and all spaces $F^\sigma_{p,q}$ such that $1/p < \sigma < \min(R, D)$.

The condition $\sigma > 1/p$ ensures that the spaces embed continuously in $C[0, 1]$, i.e. functions f in such spaces are continuous and there is a constant $C = C(\sigma, p, q)$ such that $\sup|f| \leqslant C\|f\|_{B^\sigma_{p,q}}$ (or $\|f\|_{F^\sigma_{p,q}}$). The condition $\sigma < \min(R, D)$ is needed to guarantee that wavelets yield an unconditional basis.

Theorem 1 (Donoho, 1995a). There are universal constants (π_n) with $\pi_n \geqslant 1 - (4\pi \log n)^{-1/2} \to 1$ as $n = 2^{j_1} \to \infty$ and constants $C_1(\mathscr{F}, \psi)$ depending on the function space $\mathscr{F}[0, 1] \in \mathscr{L}(R, D)$ and on the wavelet basis, but not on n or f, so that

$$P\{\|\hat{f}_n^*\|_{\mathscr{F}} \leqslant C_1\|f\|_{\mathscr{F}} \quad \forall\, \mathscr{F} \in \mathscr{L}(R, D)\} \geqslant \pi_n. \tag{14}$$

In words, $\hat{f}_n^*$ is, with increasingly high probability, simultaneously as smooth as f in every smoothness space $\mathscr{F}$ taken from the scale $\mathscr{L}(R, D)$.

Property (14) is a strong way of saying that the reconstruction is noise free. Indeed, as $\|0\|_{\mathscr{F}} = 0$, the theorem requires that if f is the zero function

$$f(t) \equiv 0 \qquad \forall\, t \in [0, 1]$$

then, with probability at least π_n, $\hat{f}_n^*$ is also the zero function. In contrast, traditional methods of reconstruction have the character that, if the true function is 0, the reconstruction is (however slightly) oscillating and bumpy as a consequence of the noise in the observations.

The reader may wish to compare Figs 4, 5 and 6 in the light of this theorem.

3.3.2. *$\hat{f}_n^*$ is near optimal for spatial adaptation*

We first describe a concept of ideal spatial adaptation, as in Donoho and Johnstone (1994a). Suppose that we have a method $T(y, \delta)$ which, given a spatial adaptation parameter δ, produces a curve estimate $\hat{f}$. We are thinking primarily of piecewise polynomial fits, with δ being a vector of break points indicating the boundaries of

the pieces. For a given function f, there is an ideal spatial parameter Δ, satisfying

$$R_n\{T(y, \Delta), f\} = \inf_{\delta} [R_n\{T(y, \delta), f\}];$$

however, since $\Delta = \Delta(f)$, this ideal parameter is not available to us when we have only noisy data. Still, we aim to achieve this ideal and define the *ideal risk*

$$\mathscr{R}_n(T, f) = R_n\{T(y, \Delta), f\}.$$

The ideal risk can be smaller than anything attainable by fixed non-adaptive schemes; to measure this, we fix the risk measure

$$R_n(\hat{f}, f) = n^{-1} \sum_i E\{\hat{f}(t_i) - f(t_i)\}^2.$$

For a generic piecewise constant function with discontinuity, the best risk achievable by linear non-adaptive schemes is of order $n^{-1/2}$, whereas the ideal risk, based on a partition Δ which exactly mimics the underlying piecewise structure of the function, achieves n^{-1}.

Theorem 2 (Donoho and Johnstone, 1994a). With $\mathscr{R}_n(T_{\mathrm{PP}(D)}, f)$ the ideal risk for piecewise polynomial fits by polynomials of degree D, and with the wavelet transform having at least D vanishing moments,

$$R_n(\hat{f}_n^*, f) \leqslant C(\log n)^2 \, \mathscr{R}_n(T_{\mathrm{PP}(D)}, f)$$

for all f and all $n = 2^{J+1}$. Here C depends only on the wavelet transform, and not on f or n.

Hence all the rate advantages of spatial adaptation are reproduced by wavelet shrinkage. (The $(\log n)^2$ bound of this theorem is not sharp for 'most' functions; wavelet shrinkage may perform even better than this indicates.)

In short, we have a theory for spatial adaptation and wavelets are near optimal under that theory.

3.3.3. *$\hat{f}_n^*$ is near optimal for estimating a function at a point*

Fix the risk $R_n(\hat{f}, f) = E\{\hat{f}(t_0) - f(t_0)\}^2$, where t_0 is one of the sample points $t_1, \ldots, t_n$. Suppose that f obeys a Hölder smoothness condition $f \in \Lambda^\alpha(C)$, where, if α is not an integer,

$$\Lambda^\alpha(C) = \{f: \, |f^{(m)}(s) - f^{(m)}(t)| \leqslant C|s - t|^\delta\}, \tag{15}$$

with $m = \lceil \alpha \rceil - 1$ and $\delta = \alpha - m$. (If α is an integer, we use Zygmund's definition (Meyer, 1990).)

Suppose, however, that we are not sure of α and C. If we knew α and C, then we could construct a linear minimax estimator $\hat{f}_n^{(\alpha, C)} = \Sigma_i c_i y_i$ where the (c_i) are the solution of a quadratic programming problem depending on C, α, σ and n (Ibragimov and Khas'minskii, 1982; Donoho and Liu, 1991; Donoho, 1994a). This estimator has worst case risk

$$\sup_{\Lambda^\alpha(C)} [E\{\hat{f}_n^{(\alpha,C)} - f(t_0)\}^2] \sim A(\alpha)(C^2)^{1-r}\left(\frac{\sigma^2}{n}\right)^r, \qquad n \to \infty, \tag{16}$$

where $A(\alpha)$ is the value of a certain quadratic program, and the rate exponent satisfies

$$r = \frac{2\alpha}{2\alpha+1}. \tag{17}$$

This risk behaviour is minimax among linear procedures, and the mean-squared error is within a factor 5/4 of minimax over all measurable procedures.

Unfortunately, if α and C are unknown and we misspecify the degree α of the Hölder condition, the resulting estimator will achieve a worse rate of convergence than the rate which would be optimal for a correctly specified condition.

Can we develop an estimator which does not require knowledge of α and C and yet performs essentially as well as $\hat{f}_n^{(\alpha,C)}$? Lepskii (1991) and Brown and Low (1992) showed that the answer is no, even if we know that the correct Hölder class is one of two specific classes. Hence, for $0<\alpha_0<\alpha_1<\infty$ and $0<C_0, C_1<\infty$,

$$\inf_{\hat{f}_n} \max_{i=0,1} (C_i^{2(r_i-1)} n^{r_i} \sigma^{-2r_i}) \sup_{\Lambda^\alpha(C)} [E\{\hat{f}_n(t_0) - f(t_0)\}^2] \geqslant \text{constant} \times (\log n)^{r_0}. \tag{18}$$

Theorem 3 (Donoho and Johnstone, 1994b). Suppose that we use a wavelet transform with $\min(R, D) > 1$. For each Hölder class $\Lambda^\alpha(C)$ with $0<\alpha<\min(R, D)$, we have

$$\sup_{\Lambda^\alpha(C)} [E\{\hat{f}_n^*(t_0) - f(t_0)\}^2] \leqslant (\log n)^r \, B(\alpha) \, (C^2)^{1-r} \left(\frac{\sigma^2}{n}\right)^r \{1 + o(1)\}, \qquad n \to \infty. \tag{19}$$

Here r is as in equation (17), and $B(\alpha)$ can be calculated in terms of properties of the wavelet transform.

Hence $\hat{f}_n^*(t_0)$ achieves, within a logarithmic factor, the minimax risk for every Hölder class in a broad range. When the Hölder class is unknown, the logarithmic factor cannot be eliminated, because of inequality (18). So the result is optimal in a certain sense.

3.3.4. *$\hat{f}_n^*$ is near optimal for estimating the object in global loss*

Now consider a global loss measure $\| \ \| = \| \ \|_{\sigma',p',q'}$ taken from the $B^\sigma_{p,q}$- or $F^\sigma_{p,q}$-scales, with $\sigma' \geqslant 0$. With $\sigma' = 0$ and p' and q' chosen appropriately, this means that we can consider L^2-loss, L^p-loss, $p>1$, etc. We can also consider losses in estimating the derivatives of some order by picking $\sigma' > 0$. We consider *a priori* classes $\mathscr{F}(C)$ taken from norms in the Besov and Triebel scales with $\sigma > 1/p$—e.g. Sobolev balls.

Theorem 4 (near minimaxity). Pick a loss $\| \ \|$ taken from the Besov or Triebel scales $\sigma' \geqslant 0$ and a ball $\mathscr{F}(C; \sigma, p, q)$ arising from an $\mathscr{F} \in \mathscr{C}(R, D)$, so that $\sigma > 1/p$; suppose that the collection of indices obeys $\sigma > \sigma' + (1/p - 1/p')_+$, so that the object can be consistently estimated in this norm. There is a modulus of continuity $\Omega(\epsilon)$ with the following properties.

(a) The estimator $\hat{f}_n^*$ nearly attains the rate $\Omega(n^{-1/2})$; with constants $C_1\{\mathscr{F}(C), \psi\}$,

$$\sup_{f\in\mathscr{F}(C)}\left(P\left[\|\hat{f}_n^* - f\| \geqslant C_1\,\Omega\left\{\sigma\sqrt{\left(\frac{\log n}{n}\right)}\right\}\right]\right)\to 0, \tag{20}$$

provided that $\sigma' > 1/p'$; if instead $0 \leqslant \sigma' \leqslant 1/p'$, replace C_1 by a logarithmic factor.

(b) No method can exceed the rate $\Omega(n^{-1/2})$: for some other constant $C_2(\|\ \|, \mathscr{F})$

$$\inf_{\hat{f}}\sup_{f\in\mathscr{F}(C)}\left[P\{\|\hat{f} - f\| \geqslant C_2\,\Omega(\sigma/\sqrt{n})\}\right] \to 1; \tag{21}$$

if $(\sigma+\frac{1}{2})p < (\sigma'+\frac{1}{2})p'$, or if $(\sigma+\frac{1}{2})p = (\sigma'+\frac{1}{2})p'$ and we work exclusively in the Besov scale, we may increase $\Omega(\sigma/\sqrt{n})$ to $\Omega[\sigma\sqrt{\{(\log n)/n\}}]$.

In words $\hat{f}_n^*$ is simultaneously within a logarithmic factor of minimax over every Besov and Triebel class in the range indicated, and over a certain subrange it is within a constant factor of minimax.

The modulus of continuity $\Omega(\epsilon)$ depends, of course, on both the loss and the function class. In many cases, $\Omega(\epsilon) \asymp \epsilon^r$ for $r = r(\sigma, p, \sigma', p')$; see inequality (44), later.

By elementary arguments, these results imply similar results for other combinations of loss and *a priori* class. For example, we can reach similar conclusions for L^1-loss, though it is not nominally in the Besov and Triebel scales; we can also reach similar conclusions for the *a priori* class of functions of total variation less than C, also not nominally in $\mathscr{C}(R, D)$. Such variations follow immediately from known inequalities between the desired norms and relevant Besov and Triebel classes.

3.4. *Interpretation*

Theorems 2–4 all have the form that a behaviour which would be attainable by measurable procedures equipped with extra side-information (perhaps a different measurable procedure for different problems) can be obtained, to within logarithmic factors, by the single estimator $\hat{f}_n^*$. Hence, if we are willing to ignore systematically factors of $\log n$ as insignificant, we have a single estimator which is optimal for a wide variety of problems and purposes. Moreover, there is a sense in which, among estimators satisfying theorem 1, these $\log n$ factors are necessary, so if we want the visual advantages of theorem 1 we must accept such logarithmic factors. For results like theorems 2 and 3, logarithmic factors are also unavoidable. Also, the results show that, for a certain range of choices of loss and *a priori* class, the estimator is within a constant factor of optimal.

These results raise an important question: do we want exact optimality for one single decision theoretic purpose or near optimality (within a logarithmic factor) for many purposes simultaneously? The exact optimality approach often leads to very specific procedures for specific problems, defined uniquely as solutions of certain optimization problems, but the procedures so designed might turn out to be unsuitable for other problems. However, the near optimality approach gives us an estimator which is the exact solution of no classical optimization problem, but which *almost* solves many problems simultaneously.

As a simple example, consider the problem of estimating a decreasing function bounded by C in absolute value. The method of least squares gives an estimate which

is decreasing and seems quantitatively quite close to minimax; wavelet shrinkage does not give a decreasing estimate, and so is less well adapted to estimating decreasing objects, yet it is within $(\log n)^{2/3}$ factors of minimax for this class, and continues to work well when the object is not decreasing.

An interesting parallel between the estimators based on wavelet shrinkage and wavelets themselves is the fact that wavelets are the solution of no classical optimization problem; unlike sinusoids and classical orthogonal systems they do not serve as eigenfunctions of a classically important operator, such as differentiation or convolution. Nevertheless, wavelets are 'almost-eigenfunctions' of many operators (Frazier *et al.*, 1991; Meyer, 1990), whereas if they were the exact eigenfunctions of some specific operator (e.g. a convolution operator) they could not continue to be almost-eigenfunctions of many other operators. Here, precise optimality rules out a broad approximate optimality.

There is also a parallel with the theory of robustness. The exact maximum likelihood estimator in certain parametric models has a property of minimum asymptotic variance, but this is accompanied by a non-robustness, an extreme suboptimality at models infinitesimally distant. However, it is possible to find estimators which have almost minimum asymptotic variance but which perform acceptably at a broad range of models close to the original model under consideration. Again exact optimality to one particular set of assumptions rules out a broader approximate optimality.

This interpretation is particularly important in light of the fact that the traditional minimax paradigm makes a rather arbitrary premise: it posits smoothness information that is rarely available. We rarely know that the object of interest has a certain number of derivatives, nor in what space the derivatives ought to be measured (L^p?; L^∞?). Therefore, the expenditure of effort to achieve exact optimality, at the level of constants (10), is particularly difficult to support, except as part of a larger effort to obtain basic understanding.

4. IDEAS UNDERLYING PROOF OF NEAR MINIMAXITY

In this section, we outline some ideas in the proof of theorem 4. A fuller account, including proofs, appears in Donoho *et al.* (1993a).

4.1. *Alternative to Classical Minimaxity for High Dimensions*

At a conceptual level, the wavelet shrinkage method represents a different response to the negative result (3). We obtain an analogue of result (1) that is valid in high dimensions if, instead of trying to do absolutely well uniformly for every f, we try to do nearly as well as the minimax risk for every 'nice' Θ (see Sections 4.2 and 4.4.2).

Informally, the principle that we are exploiting is the following. For estimating an n-dimensional vector θ there is a single shrinkage estimator $\hat{\theta}_n^*$ with the following 'universal near minimax property': for any loss that is in some sense 'bowl shaped and symmetric' and any *a priori* class Θ that is also bowl shaped and symmetric, then

$$\sup_{\Theta}\{R_n(\hat{\theta}_n^*, \theta)\} \leqslant \text{'log } n \text{ factor'} \times \inf_{\hat{\theta}} \sup_{\Theta}\{R_n(\hat{\theta}, \theta)\}. \tag{22}$$

In a sense, this principle has the generality and appeal of result (1): it says that a single estimator is good for a very wide variety of loss functions and purposes.

We put quotation marks around things in inequality (22) to emphasize that we do not prove this principle in this paper. For results like inequality (22), compare Donoho and Johnstone (1994a) and Donoho (1995a) and Donoho and Johnstone (1994b).

4.2. *Translation into Sequence Space*

Consider the following *sequence model*. We start with an index set $\mathscr{I}_n$ of cardinality n, and we observe

$$y_I = \theta_I + \epsilon z_I, \qquad I \in \mathscr{I}_n, \tag{23}$$

where $z_I \sim_{\text{IID}} N(0, 1)$ is Gaussian white noise and ϵ is the noise level. The index set $\mathscr{I}_n$ is the first n elements of a countable index set $\mathscr{I}$. From the n data (23), we wish to estimate the object with countably many co-ordinates $\theta = (\theta_I)_{\mathscr{I}}$ with small loss $\|\hat{\theta} - \theta\|$. The object of interest belongs *a priori* to a class Θ, and we wish to achieve a *minimax risk* of the form

$$\inf_{\hat{\theta}} \sup_{\Theta} (P\{\|\hat{\theta} - \theta\| > \omega\})$$

for a special choice $\omega = \omega(\epsilon)$. About the error norm, we assume that it is *solid* and *orthosymmetric*, namely that

$$|\xi_I| \leqslant |\theta_I| \quad \forall\, I \Rightarrow \|\xi\| \leqslant \|\theta\|. \tag{24}$$

Moreover, we assume that the *a priori* class is also solid and orthosymmetric, so

$$\theta \in \Theta \quad \text{and} \quad |\xi_I| \leqslant |\theta_I| \quad \forall\, I \Rightarrow \xi \in \Theta. \tag{25}$$

Finally, at one specific point, inequality (38) later, we shall assume that the loss measure is either convex, or at least ρ-convex, $0 < \rho \leqslant 1$, in the sense that $\|\theta + \xi\|^\rho \leqslant \|\theta\|^\rho + \|\xi\|^\rho$; 1-convex is just convex.

Results for this model will imply theorem 4 by suitable identifications. Thus we shall ultimately interpret

(a) (θ_I) as wavelet coefficients of f,
(b) $(\hat{\theta}_I)$ as empirical wavelet coefficients of an estimate $\hat{f}_n$ and
(c) $\|\hat{\theta} - \theta\|$ as a norm equivalent to $\|\hat{f} - f\|$.

We shall explain such identifications further in Section 4.5.

4.3. *Solution of Optimal Recovery Model*

Before tackling data from model (23), we consider a simpler abstract model, in which noise is *deterministic* (compare Micchelli (1975), Micchelli and Rivlin (1977) and Traub *et al.* (1988)). The approach of analysing statistical problems by deterministic noise has been applied previously in Donoho (1994a) and Donoho (1994b). Suppose that we have an index set $\mathscr{I}$ (not necessarily finite), an object (θ_I) of interest and observations

$$x_I = \theta_I + \delta u_I, \qquad I \in \mathscr{I}. \tag{26}$$

Here $\delta > 0$ is a known 'noise level' and (u_I) is a nuisance term known only to satisfy $|u_I| \leqslant 1$, $\forall\, I \in \mathscr{I}$. We suppose that the nuisance is chosen by a clever opponent to cause the most damage, and we evaluate performance by the worst case error:

$$E_\delta(\hat{\theta}, \theta) = \sup_{|u_I| \leqslant 1} \|\hat{\theta}(x) - \theta\|. \tag{27}$$

4.3.1. *Optimal recovery—fixed* Θ

The existing theory of optimal recovery focuses on the case where we know that $\theta \in \Theta$, and Θ is a fixed, known *a priori* class. We want to attain the minimax error

$$E_\delta^*(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} \{E_\delta(\hat{\theta}, \theta)\}.$$

Very simple upper and lower bounds are available.

Definition 2. The *modulus of continuity* of the estimation problem is

$$\Omega(\epsilon; \|\ \|, \Theta) = \sup\{\|\theta^0 - \theta^1\|:\ \theta^0, \theta^1 \in \Theta,\ |\theta_I^0 - \theta_I^1| \leqslant \epsilon,\ \forall\, I \in \mathscr{I}\}. \tag{28}$$

Proposition 1.

$$E_\delta^*(\Theta) \geqslant \Omega(\delta)/2. \tag{29}$$

Proof. Suppose that θ^0 and θ^1 attain the modulus. Then under the observation model (26) we could have observations $x = \theta^0$ when the true underlying $\theta = \theta^1$, and vice versa. So whatever we do in reconstructing θ from x must suffer a worst case error of half the distance between θ^1 and θ^0. □

A variety of rules can nearly attain this lower bound.

Definition 3. A rule $\hat{\theta}$ is *feasible* for Θ if, for each $\theta \in \Theta$ and for each observed (x_I) satisfying model (26),

$$\hat{\theta} \in \Theta, \tag{30}$$

$$|\hat{\theta}_I - x_I| \leqslant \delta. \tag{31}$$

Proposition 2. A feasible reconstruction rule has error

$$\|\hat{\theta} - \theta\| \leqslant \Omega(2\delta), \qquad \theta \in \Theta. \tag{32}$$

Proof. Since the estimate is feasible, $|\hat{\theta}_I - \theta_I| \leqslant 2\delta$, $\forall\, I$, and θ, $\hat{\theta} \in \Theta$. The bound follows by the definition (28) of the modulus. □

Comparing inequalities (32) and (29) we see that, quite generally, *any feasible procedure is nearly minimax.*

4.3.2. *Soft thresholding is an adaptive method*

In the case where Θ might be any of a wide variety of sets, we can imagine that it would be difficult to construct a procedure which is near minimax over each of them—e.g. that the requirements of feasibility with respect to many different sets

would be incompatible with each other. Luckily, if the sets in question are all orthosymmetric and solid, a single idea—shrinkage towards the origin—leads to feasibility independently of the details of the set's shape.

Consider a specific shrinker based on the soft threshold non-linearity $\eta_t(y) = \mathrm{sgn}(y)(|y| - t)_+$. Setting the threshold level equal to the noise level $t = \delta$, we define

$$\hat{\theta}_I^{(\delta)}(y) = \eta_t(x_I), \qquad I \in \mathscr{I}. \tag{33}$$

This pulls each noisy coefficient x_I towards 0 by an amount $t = \delta$ and sets $\hat{\theta}_I^{(\delta)} = 0$ if $|x_I| \leqslant \delta$. Because it pulls each coefficient towards the origin by at least the noise level, it satisfies the *uniform shrinkage condition*

$$|\hat{\theta}_I| \leqslant |\theta_I|, \qquad I \in \mathscr{I}. \tag{34}$$

Theorem 5. The soft thresholding esimator $\hat{\theta}^{(\delta)}$ defined by equation (33) is feasible for every Θ which is solid and orthosymmetric.

Proof. $|\hat{\theta}_I^{(\delta)} - x_I| \leqslant \delta$ by definition, whereas inequality (34) and the assumption (25) of solidness and orthosymmetry guarantee that $\theta \in \Theta$ implies $\hat{\theta}^{(\delta)} \in \Theta$. □

This shows that soft thresholding leads to nearly minimax procedures over all combinations of symmetric *a priori* classes and symmetric loss measures. Surprisingly, although the result is both simple and useful, we have been unable to find results of this form in the literature of optimal recovery and information-based complexity.

4.3.3. *Recovery from finite, deterministic noisy data*

The optimal recovery and information-based complexity literature generally posits a *finite* number n of noisy observations, and this is consistent with our model (23). So consider observations

$$x_I = \theta_I + \delta u_I, \qquad I \in \mathscr{I}_n. \tag{35}$$

The minimax error in this setting is

$$E^*_{n,\delta}(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} \|\hat{\theta} - \theta\|.$$

To see how this setting differs from the 'complete data' model (26), we set $\delta = 0$. Then we have the problem of inferring the complete vector $(\theta_I: I \in \mathscr{I})$ from the first n components $(\theta_I: I \in \mathscr{I}_n)$. To study this, we need the following definition.

Definition 4. The *tail-n-width* of Θ in norm $\| \ \|$ is

$$\Delta(n; \| \ \|; \Theta) = \sup\{\|\theta\|: \ \theta \in \Theta, \ \theta_I = 0, \ \forall \ I \in \mathscr{I}_n\}.$$

We have the identity

$$E^*_{n,0}(\Theta) = \Delta(n; \| \ \|; \Theta),$$

which is valid whenever both $\| \ \|$ and Θ are solid and orthosymmetric.

A lower bound for the minimax error is obtainable by combining the $n=\infty$ and the $\delta=0$ extremes:

$$E^*_{n,\delta}(\Theta)\geqslant\max\{\Omega(\delta)/2,\ \Delta(n)\}. \tag{36}$$

Again, soft thresholding comes surprisingly close, under surprisingly general conditions. Consider the rule

$$\hat\theta^{n,\delta}=\begin{cases}\eta_\delta(x_I), & I\in\mathscr{I}_n,\\ 0, & I\in\mathscr{I}\setminus\mathscr{I}_n.\end{cases} \tag{37}$$

Supposing for the moment that the loss measure $\|\ \|$ is convex we have

$$\|\hat\theta^{n,\delta}-\theta\|\leqslant\Omega(2\delta)+\Delta(n),\qquad \theta\in\Theta. \tag{38}$$

(If the loss is not convex, but just ρ-convex, $0<\rho<1$, we can replace the right-hand side by $\{\Omega(2\delta)^\rho+\Delta(n)^\rho\}^{1/\rho}$.)

Comparing inequalities (38) and (36), we again have that soft thresholding is nearly minimax, simultaneously over a wide range of *a priori* classes and choices of loss.

4.4. *Application to Statistical Sequence Model*

We now translate the results on optimal recovery into results on statistical estimation.

4.4.1. *Upper bounds*

The basic idea is the following fact (Leadbetter *et al.*, 1983). Let (z_I) be independently and identically distributed (IID) $N(0,1)$. Define

$$A_n=\{\|(z_I)\|_{l_n^\infty}\leqslant\sqrt{(2\log n)}\};$$

then

$$\pi_n\equiv P\{A_n\}\to 1,\qquad n\to\infty. \tag{39}$$

In words, we have very high confidence that $\|(z_I)\|_{l_n^\infty}\leqslant\sqrt{(2\log n)}$. This motivates us to act as if noisy data (23) were an instance of the deterministic model (35), with noise level $\delta_n=\sqrt{(2\log n)}\epsilon$. Accordingly, we set $t_n=\delta_n$ and define

$$\hat\theta^{(n)}_I=\begin{cases}\eta_{t_n}(y_I), & I\in\mathscr{I}_n,\\ 0, & I\in\mathscr{I}\setminus\mathscr{I}_n.\end{cases} \tag{40}$$

Recall the optimal recovery bound (38) (in the case where the triangle inequality applies). We obtain immediately that whenever $\theta\in\Theta$ and the event A_n holds

$$\|\hat\theta^{(n)}-\theta\|\leqslant\Omega(2\delta_n)+\Delta(n);$$

as this event has probability π_n we obtain the following risk bound.

Theorem 6. If $\|\ \|$ is convex, then for all $\theta\in\Theta$,

$$P\{\|\hat{\theta}^{(n)} - \theta\| \leqslant \Omega(2\delta_n) + \Delta(n)\} \geqslant \pi_n, \tag{41}$$

with a suitable modification if $\| \ \|$ is ρ-convex, $0 < \rho < 1$.

This shows that statistical estimation is not really more difficult than optimal recovery, except by a factor involving $\sqrt{\log n}$.

4.4.2. *Besov and Triebel bodies*

To go further, we specialize our choice of possible losses $\| \ \|$ and *a priori* classes Θ to members of the Besov and Triebel scales of sequence spaces. These are defined as follows. First, we specify that the abstract index set $\mathscr{I}$ is of the standard multiresolution format $I = (j, k)$ where $j \geqslant -1$ is a resolution index and $0 \leqslant k < 2^j$ is a spatial index. We write equally (θ_I) or $(\theta_{j,k})$, and we write $\mathscr{I}^{(j)}$ for the collection of indices $I = (j, k)$ with $0 \leqslant k < 2^j$. We define the Besov sequence norm

$$\|\theta\|_{b^{\sigma}_{p,q}} = \Big[\sum_{j \geqslant -1} \Big\{ 2^{js} \Big(\sum_{\mathscr{I}^{(j)}} |\theta_I|^p \Big)^{1/p} \Big\}^q \Big]^{1/q}, \tag{42}$$

where $s \equiv \sigma + \frac{1}{2} - 1/p$, and the Besov body

$$\Theta^{\sigma}_{p,q}(C) \equiv \{\theta: \ \|\theta\|_{b^{\sigma}_{p,q}} \leqslant C\}.$$

Similarly, the Triebel body $\Phi^{\sigma}_{p,q} = \Phi^{\sigma}_{p,q}(C)$ is defined by the condition $\|\theta\|_{f^{\sigma}_{p,q}} \leqslant C$ where $f^{\sigma}_{p,q}$ refers to the norm

$$\|\theta\|_{f^{\sigma}_{p,q}} = \Big\| \Big(\sum_{I \in \mathscr{I}} 2^{jsq} |\theta_I|^q \chi_I \Big)^{1/q} \Big\|_{L^p[0,1]}, \tag{43}$$

where χ_I stands for the indicator function $\mathbf{1}_{[k/2^j, (k+1)/2^j)}$ and $s \equiv \sigma + \frac{1}{2}$. We remark, as an aside, that Besov and Triebel norms are ρ-convex, with $\rho = \min(1, p, q)$, so that in the usual range $p, q \geqslant 1$ they are convex. Also, for the rest of this section, c_i denote constants depending on (σ, p, q) and (σ', p', q').

Theorem 7 (Besov modulus; Donoho *et al.* (1993a)). Let $\| \ \|$ be a member of the Besov scale, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$, and suppose that $\tilde{\sigma} = \sigma - \sigma' - (1/p - 1/p')_+ > 0$. Then

$$c_0 C^{(1-r)} \delta^r \leqslant \Omega(\delta) \leqslant c_1 C^{(1-r)} \delta^r \qquad 0 < \delta < \delta_1(C), \tag{44}$$

where the rate exponent satisfies

$$r = \min\left(\frac{\sigma - \sigma'}{\sigma + \frac{1}{2}}, \ \frac{\tilde{\sigma}}{\sigma + \frac{1}{2} - 1/p} \right), \qquad \sigma > 1/p, \quad \tilde{\sigma} > 0, \tag{45}$$

except in the critical case where $p' \geqslant p$ and the two terms in the minimum appearing in equation (45) are equal—i.e. $(\sigma + \frac{1}{2})p = (\sigma' + \frac{1}{2})p'$ (for which see Donoho *et al.* (1993a)).

What happens if $\| \ \|$ or Θ, or both, come from the Triebel scales? A norm from the Triebel scale is bracketed by norms from the Besov scales with the same σ and p, but different qs:

$$a_0\|\theta\|_{b^{\sigma}_{p,\max(p,q)}} \leqslant \|\theta\|_{f^{\sigma}_{p,q}} \leqslant a_1\|\theta\|_{b^{\sigma}_{p,\min(p,q)}} \tag{46}$$

(compare Peetre (1975), p. 261, or Triebel (1992), p. 96). Hence, for example,

$$\Theta^{\sigma}_{p,\min(p,q)}(C/a_1) \subset \Phi^{\sigma}_{p,q}(C) \subset \Theta^{\sigma}_{p,\max(p,q)}(C/a_0),$$

and so we can bracket the modulus of continuity in terms of the modulus from the Besov case, but with differing values of q and q'. By inequality (44), the qualitative behaviour for the modulus in the Besov scale, outside the critical case, does not depend on q, q'. The modulus of continuity therefore continues to obey the same general relationships (44) even when the Triebel scale is used for one, or both, of the norm $\|\ \|$ and class Θ.

In addition to concrete information about the modulus, we need concrete information about the tail-n-widths.

Theorem 8 (Donoho *et al.* (1993a)). Let $\|\ \|$ be a member of the Besov or Triebel scales, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$ or a Triebel body $\Phi^{\sigma}_{p,q}(C)$. Then

$$\Delta(n;\ \|\ \|,\ \Theta) \leqslant c_2 n^{-\tilde{\sigma}}, \qquad n = 2^{J+1}.$$

4.4.3. *Lower bound*

With noise levels equated, $\epsilon = \delta$, statistical estimation is not easier than optimal recovery:

$$\inf_{\hat{\theta}} \sup_{\Theta} (P[\,\|\hat{\theta} - \theta\| \geqslant \max\{\Delta(n),\ c\,\Omega(\epsilon)\}\,]) \to 1, \qquad \epsilon = \sigma/\sqrt{n} \to 0. \tag{47}$$

Half of this result is non-statistical; it says that

$$\inf_{\hat{\theta}} \sup_{\Theta} [\,P\{\|\hat{\theta} - \theta\| \geqslant \Delta(n)\}\,] \to 1 \tag{48}$$

and this follows for the reason that (from section (4.3.3)) this holds in the noiseless case. The other half is statistical and requires a generalization of lower bounds developed by decision theorists systematically over the last 15 years—namely the embedding of an appropriate hypercube in the class Θ and using elementary decision theoretic arguments on hypercubes. Compare Samarov (1992), Bretagnolle and Carol-Huber (1979), Ibragimov and Khas'minskii (1982) and Stone (1982).

Theorem 9 (Donoho *et al.* (1993a)). Let $\|\ \|$ come from the Besov scale, with parameter (σ', p', q'). Let Θ be a Besov body $\Theta^{\sigma}_{p,q}(C)$. Then

$$\inf_{\hat{\theta}} \sup_{\Theta} [\,P\{\|\hat{\theta} - \theta\| \geqslant c_3\,\Omega(\epsilon)\}\,] \to 1. \tag{49}$$

Moreover, when $p' > p$ and $(\sigma + \frac{1}{2})p \leqslant (\sigma' + \frac{1}{2})p'$, we obtain an even stronger bound, with $\Omega(\epsilon\sqrt{\log \epsilon^{-1}})$ in place of $\Omega(\epsilon)$.

The proof of theorem 7 constructs a special problem of optimal recovery—recovering a parameter θ known to lie in a certain 2^{j_0}-dimensional l^p-ball ($j_0 = j_0(\epsilon;\ \sigma,\ p,\ \sigma',\ p')$), measuring loss in $l^{p'}$-norm. The construction shows that this finite dimensional

subproblem is essentially as difficult (under model (26)) as the full infinite dimensional problem of optimal recovery of an object in a (σ, p, q)-ball with a (σ', p', q')-loss. The proof of theorem 9 shows that, under the calibration $\epsilon = \delta$, the statistical estimation problem over this particular l^p-ball is at least as difficult as the optimal recovery problem, and sometimes more difficult by an additional logarithmic factor.

4.5. *Translation into Function Space*

The following corollary gives our conclusion from theorems 7–9.

Corollary 1. In the sequence model (23), the single estimator (40) is within a logarithmic factor of minimax over every loss and every *a priori* class chosen from the Besov and Triebel sequence scales. For a certain range of these choices the estimator is within a constant factor of minimax.

Theorem 4 is just the translation of this conclusion back from sequences to functions. We give a sketch of the ideas here, leaving the full argument to Donoho *et al.* (1993a). Fundamental to our approach, in Section 4.2, is the heuristic that observations (2) are essentially equivalent to observations (23). This contains within it three specific subheuristics:

(a) that if we apply an empirical wavelet transform, based on pyramid filtering, to *n noiseless* samples, then we obtain the first *n* coefficients out of the countable sequence of all wavelet coefficients;
(b) that if we apply an empirical wavelet transform, based on pyramid filtering, to *n noisy* samples, then we obtain the first *n* theoretical wavelet coefficients, with white noise added; this noise has standard deviation $\epsilon = \sigma/\sqrt{n}$;
(c) that the Besov and Triebel norms in function space (e.g. L^p- and W_p^m-norms) are equivalent to the corresponding sequence space norms (e.g. $f_{p,2}^0$ and $f_{p,2}^m$).

Using these heuristics, the sequence space model (23) may be viewed as just an equivalent representation of model (2); hence errors in estimation of wavelet coefficients are equivalent to errors in estimation of functions, and rates of convergence in the two problems are identical, when the proper calibration $\epsilon = \sigma/\sqrt{n}$ is made.

These heuristics are just approximations, and several arguments are necessary to derive a full result, covering all cases. Donoho *et al.* (1993a) give a detailed sketch of the connection between the nonparametric and sequence space problems.

5. EXTENSIONS

Wavelet thresholding can, with minor variations, be made to cover other types of problem and data. We mention here some examples.

5.1. *Estimated Scale*

The wavelet shrinkage algorithm, as initially described, assumes that the scale of the errors σ is known and fixed. In our software, we estimate the error scale, as described above, by taking the median absolute deviation of the empirical wavelet coefficients at the finest scale J and dividing by 0.6745. Because $0.6745 < \Phi(1) - \Phi(-1)$, the result is a statistic that, with increasing probability, overestimates σ:

$$\inf_f(P\{\hat{\sigma}>\sigma\})\to 1, \qquad n\to\infty,$$

but not by much. If $\mathscr{F}(C)$ is a ball from $\mathscr{C}(R, D)$,

$$\sup_{f\in\mathscr{F}(C)}(P\{\hat{\sigma}\leqslant 1.01\sigma\})\to 1, \qquad n\to\infty.$$

It is then easy to obtain risk upper bounds paralleling the scale-known case, but involving $\Omega(2.02\sigma\sqrt{(2\log n)/\sqrt{n}}$ in place of $\Omega(2\sigma\sqrt{(2\log n)/\sqrt{n})}$. The conclusions of theorem 4 hold for this estimator.

5.2. *More General Risk Measures*

The results quoted in Section 2.3 typically studied integral L^p-risk measures such as equation (8). Theorem 4 can be extended to such measures. Indeed, Borell's inequality tells us that the noise never exceeds $\sqrt{(2\log n)}$ by very much:

$$P\{\|(z_I)\|_{l_n^\infty}>t+\sqrt{(2\log n)}\}\leqslant\exp(-t^2/2), \qquad t>0.$$

By systematically exploiting this observation, we can obtain bounds on integral risks (8), and we conclude that

$$E\|\hat{f}_n^*-\hat{f}\|^s\leqslant\text{constant}\times[\Omega\{c\sqrt{(\log n)/\sqrt{n}}\}+\Delta(n)]^s, \qquad f\in\mathscr{F}(C),$$

as we would expect; the argument is similar to the way in which conclusions for 0–1 loss are extended to power law losses in Birgé (1983) and Donoho and Liu (1991).

5.3. *Higher Dimensions, Area Samples*

Consider d-dimensional observations indexed by $i=(i_1,\ldots,i_d)$ following one of the models

$$d_i=f(t_i)+\sigma z_i, \tag{50}$$

$$d_i=\text{ave}\{f|Q(i)\}+\sigma z_i. \tag{51}$$

In either case, $0\leqslant i_1,\ldots,i_d<m$, $t_i=(i_1/m,\ldots,i_d/m)$ and the z_i are IID $N(0, 1)$. We set $m=2^{J+1}$ and $n=m^d$. The first case is a straightforward generalization of model (2). In the second case, $Q(i)$ is the cube

$$Q(i)=\{t:\ i_1/m\leqslant t_1<(i_1+1)/m,\ldots,i_d/m\leqslant t_d<(i_d+1)/m\},$$

In the case $d=2$ this may be taken as a model of noisy digital camera charge-coupled device imagery.

For this setting an empirical wavelet transform derives from a d-dimensional pyramid filtering operator $U_{j_0,\,j_1}$ which is based on a tensor product construction; this requires only the repeated application, in various directions, of the one-dimensional filters developed by Cohen *et al.* (1992). To process these observations we follow exactly the three-step prescription described in Section 3.1: empirical wavelet transform, followed by soft thresholding at level $\sqrt{(2\log n)}\sigma/\sqrt{n}$, followed by an inversion of the empirical wavelet transform, giving $\hat{f}_n^*$.

Adaptivity results paralleling theorem 4 are available in these settings. For the point sampling model, the appropriate function space scale $\mathscr{G}(R, D)$ is the collection of Besov and Triebel spaces $B^{\sigma}_{p,q}([0, 1]^d)$ and $F^{\sigma}_{p,q}([0, 1]^d)$ with $\min(R, D) > \sigma > d/p$. For the area sampling model, a *broader* scale $\mathscr{L}(R, D)$ may be used, consisting of all spaces $B^{\sigma}_{p,q}$ and $F^{\sigma}_{p,q}$ which embed in L^1 so that their averages are well defined. The condition now amounts to $\min(R, D) > \sigma > d(1/p - 1)$.

For balls and losses in the scale appropriate to each model, we obtain errors bounded, with overwhelming probability, by $(\log n)^e(\sigma/\sqrt{n})^r$ where the rate exponent satisfies

$$r = \min\left(\frac{\sigma - \sigma'}{\sigma + d/2}, \frac{\tilde{\sigma}}{\sigma + d/2 - d/p}, 2\tilde{\sigma}\right), \tag{52}$$

as long as $\tilde{\sigma} = \sigma - \sigma' - (d/p - d/p')_+ > 0$. In the point sampling case, the logarithmic exponent is $e = e_1 + e_2 + r/2$, where we replace expressions like $1/p$ by d/p etc. throughout. In the area sampling model, $e_1 = 0$. For definitions of e_i, see Donoho *et al.* (1993a).

Moreover, no estimator can do better over any individual ball in this scale than $n^{-r/2}$, so again the wavelet shrinkage estimator $\hat{f}_n^*$ is nearly optimal. The advantage of area sampling is the broader scale of function spaces accommodated and the simplification of the logarithmic terms in the upper bound (i.e. $e_1 \equiv 0$). The proof is parallel to the proof of theorem 4.

5.4. *Alternative Threshold Choices*

This paper concentrates on the simplicity and theoretical and visual benefits of using a single soft threshold $\sigma\sqrt{(2 \log n)}$ applied to all wavelet coefficients above a certain coarse resolution level j_0. However, other choices of threshold may be more appropriate in certain situations. For example, focus on a mean-squared error criterion leads to a threshold somewhat smaller than $\sigma\sqrt{(2 \log n)}$ (Donoho and Johnstone, 1994a).

Perhaps more importantly, *level-dependent* threshold choices arise naturally in certain inverse problems (Donoho, 1995b), and also if we wish to 'tune' wavelet estimators to be nearly exactly minimax over specific function classes (Donoho and Johnstone, 1995a; Johnstone, 1994). Level-dependent thresholds can also be selected adaptively from data by using an unbiased risk criterion—e.g. 'SUREShrink' in Donoho and Johnstone (1995b) (see also Nason (1994)). For such estimators, exact global asymptotic rates of convergence (without logarithmic terms) are possible. Finally, we may wish to choose level- and even position-dependent thresholds subjectively in the context of a particular data set.

In practice, we may also prefer in certain cases to use a 'hard threshold' rule $\eta_t(w) = w\{|w| > t\}$ which does not shrink the retained coefficients. Hard thresholding can lead to better reproduction of peak heights and discontinuities, but at the price of occasional artefacts that can roughen the appearance of the estimate—the uniform shrinkage property (34) does not hold for hard thresholding.

5.5. *Density Estimation*

Wavelet thresholding may be used to obtain near optimal rates in density estimation (Johnstone *et al.*, 1992; Donoho *et al.*, 1993b). Suppose that $X_1, \ldots, X_n$ are IID f, where f is an unknown density supported in $[0, 1]$. Let

$$W_{j,k} = n^{-1} \sum_{i=1}^{n} \psi_{j,k}(X_i)$$

where $\psi_{j,k}$ is again the appropriate wavelet basis function. Define thresholds $t_j = 0$, $j < j_0$, and $t_j = A\sqrt{j}$, where $j_0 \leqslant j \leqslant J$, $J = \log_2 n - \log_2(\log n)$ and A is some constant. The thresholded wavelet series estimator of Johnstone *et al.* (1992) is

$$\hat{f}_n^+ = \sum_{j=-1}^{J} \sum_{k} \eta_{t_j}(W_{j,k})\psi_{j,k}. \tag{53}$$

For this estimator, Johnstone *et al.* (1992) give optimal rate results which are the exact parallel of the results we derived earlier in theorem 4 (although the proofs are entirely different). This is no accident, as the problems are known to be closely connected.

We mention here a slightly weaker, but simple, corollary of the approach of the present paper, obtained in work with Eric Kolaczyk, a doctoral student at Stanford, which makes an interesting comparison with Johnstone *et al.* (1992). Suppose that we let $\mathscr{I}_n$ denote the collection of wavelet coefficients up to level J where $J = \lfloor \log_2 n - 1 \rfloor$. We define a density estimator by

$$\hat{f}_n^* = \sum_{I \in \mathscr{I}_n} \eta_{t_n}(W_{j,k})\psi_{j,k},$$

where $t_n = 2(\log n)C/\sqrt{n}$, with $C = \sup(2^{-j/2}\|\psi_{j,k}\|_\infty)$. This parallels our treatment of regression observations, except that the threshold behaves like $\log n$, not $\sqrt{\log n}$.

Theorem 10 (density estimation). Fix the loss $\| \ \| = \| \ \|_{\sigma',p',q'}$ with $0 \leqslant \sigma \leqslant \min(R, D)$. For each ball $\mathscr{F}(C; \sigma, p, q)$ arising from an $\mathscr{F}$ satisfying $1/p < \sigma < \min(R, D)$, the estimator $\hat{f}_n^*$ attains the rate $\{(\log n)^2/n\}^{r/2}$, where r is as in the earlier results (45).

Estimator (53) has been implemented in Tribouley (1993). An alternate approach, useful when the data are (or can be) binned, is illustrated in the following examples.

5.6. *Two Examples*

Good and Gaskins (1980) give a listing and detailed analysis of a scattering reaction data set consisting of 25752 events aggregated into 172 bins of width 10 MeV. For a simple analysis by wavelet thresholding, the binned counts were padded to a vector (N_i) of length 256 by 'prepending' and appending 42 bins with five counts each. The counts are transformed by the Anscombe (1948) variance stabilizing transformation for Poisson data, $y_i = 2\sqrt{(N_i + 0.375)}$, and then processed exactly as though the data obeyed the white noise model. The results are shown in Fig. 7. Fig. 8 compares the wavelet shrinkage fit with the penalized likelihood fit of Good and Gaskins (1980), Table 1. From the difference plots and the close-ups, we see that the wavelet estimator is remarkably close to the Good–Gaskins estimator (whose regularization parameter β was chosen from the data); the wavelet estimator perhaps smooths the main peak less, whereas the Good–Gaskins estimator locates the second bump more appropriately. Of course the wavelet estimator is much more quickly computed.

Fig. 9 shows a spectrum from electron spectroscopy for chemical analysis (ESCA) based on millions of counts aggregated into 1024 bins (data by courtesy of Jean-Paul Bibérian). To simulate a noisier 'low statistics' situation, a simulated spectrum was

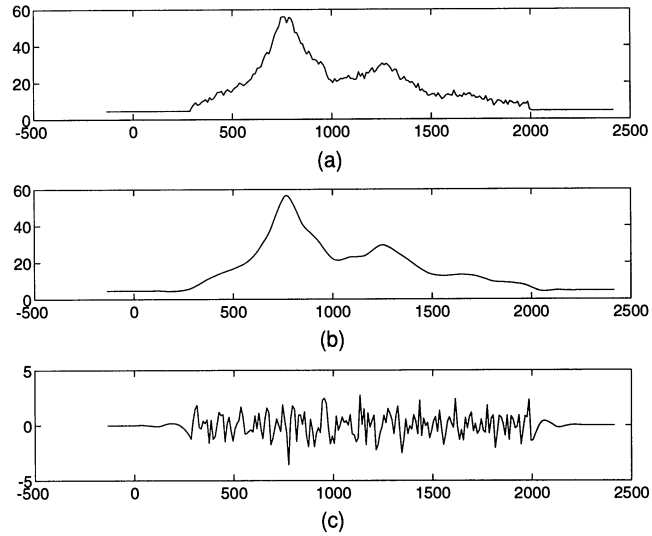

Fig. 7. Good–Gaskins scattering data: (a) $y_i = 2\sqrt{(N_i + 0.375)}$ plotted against energy (256 equally spaced bins of width 10 MeV; (b) wavelet shrinkage estimate using coiflet (see Daubechies (1992), p. 258 ff.) of order 3, coarse resolution cut-off $j_0 = 5$; (c) residuals of (b) from (a)

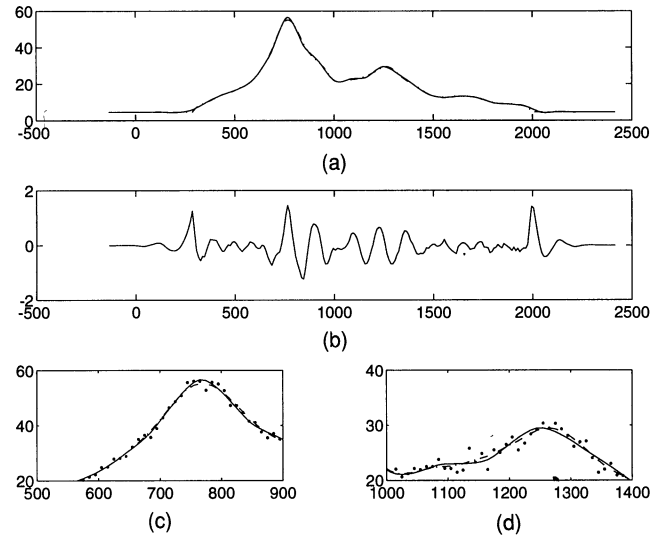

Fig. 8. Good–Gaskins scattering data: (a), (b) superposition of the wavelet and (Anscombe-transformed) Good–Gaskins estimates; (c), (d) close-ups (———, wavelet shrinkage; --------, Good–Gaskins estimates; ·, bin counts)

constructed by using Fig. 9(a) as a 'true' intensity function and drawing Poisson samples with maximum 4000 counts per bin. This simulated spectrum is again transformed by the Anscombe (1948) variance stabilizing transformation given above and is then processed as in the white noise model (although in this case the boundary filters of Cohen *et al.* (1992) are used in performing the wavelet transform). For visual clarity, the figures show only the subset of the reconstructions from bins 400–900.

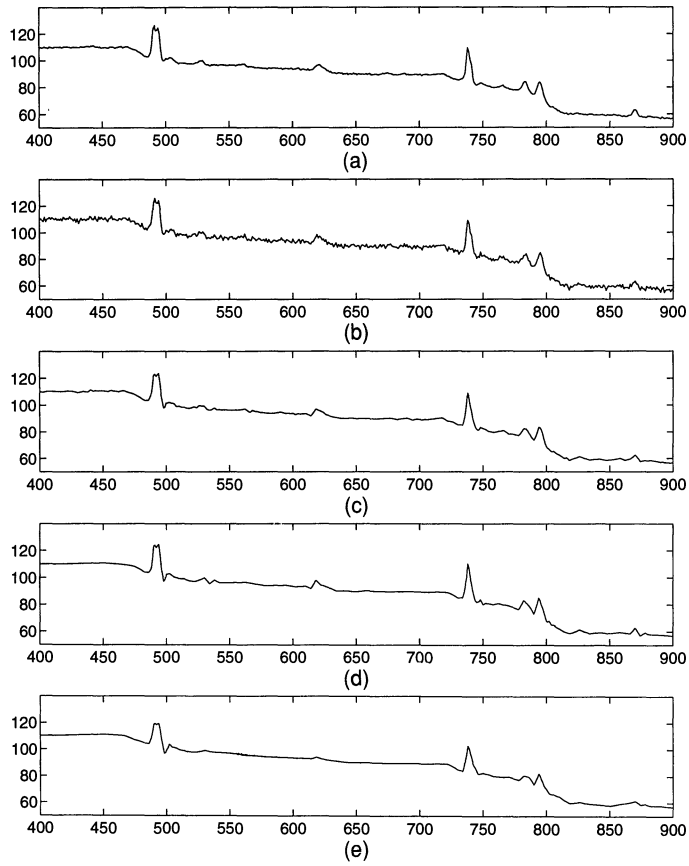

Fig. 9. (a) ESCA spectrum provided by Jean-Paul Bibérian (the maximum cell count is 50011, before normalization, but it is treated as the true intensity spectrum I_i, $i = 1, \ldots, 1024$; only bins 400–900 are shown); (b) simulated spectrum with $N_i \sim \text{Poisson}\{4000 I_i / \max(I_i)\}$; (c)–(e) wavelet transform calculated using Daubechies filters of order 3, combined with the corresponding edge filters of Cohen *et al.* (1992) (thresholding of wavelets was done using (e) soft and (d) hard thresholding at $\sqrt{(2 \log n)}$ and using (c) SUREShrink, in which thresholds are chosen level by level to minimize an unbiased estimate of risk; for a detailed description, see Donoho and Johnstone (1995b) or code listings in the 'TeachWave' package—see Appendix A)

In this case soft thresholding obscures the peak at about bin 620, perhaps owing to excessive shrinkage of wavelet coefficients. One may alternatively use hard thresholding, or even level-dependent thresholds chosen adaptively from the data using an unbiased estimate of risk criterion (SUREShrink)—these both retain the peak at about bin 620, although SUREShrink has some residual high frequency roughness, e.g. in bins 100–450.

6. DISCUSSION AND INSIGHTS

6.1. *Minimaxity and Spatial Adaptivity*

The implicit position of the 'spatial adaptivity community' that minimax theory leads to spatially non-adaptive methods is no longer tenable. Donoho and Johnstone

(1995a) show that minimax estimators can generally be expected to have a spatially adaptive structure, and we see in this paper that a specific nearly minimax estimator exhibits spatially adaptive behaviour—in actual reconstructions. The lack of spatial adaptivity in previous minimax estimators is due to the narrow range of classes $\mathscr{F}(C)$ studied.

6.2. *Need for Non-linearity*

The 'minimax community' has, until now, not fully assimilated the results of Nemirovskii and co-workers on the need for non-linear estimation. Ildar Ibragimov has proposed, privately, that the rate inefficiency of linear estimators in various cases is due to a kind of misstatement of the problem—a mismatch of the norm and function class. Here we have shown that a very simple and natural procedure achieves near optimal performance both over classes where linear estimators behave well and over those where they behave relatively poorly. Moreover, tests on data show that there are evident visual advantages of wavelet shrinkage methods. Now that we have near optimal non-linear methods and can test them out, we see that their advantages are not due to a mathematical pathology, but are intuitive and visual.

6.3. *Modulus of Continuity; Optimal Recovery*

Donoho and Liu (1991) and Donoho (1994a) demonstrated that, for problems of estimating a linear functional of an unknown object in density and regression models, the minimax risk was measured by a geometric object—namely the modulus of continuity of the functional under consideration, over the *a priori* set $\mathscr{F}$. Since that time, it has been natural to inquire whether there was a 'modulus of continuity for the whole object'. Johnstone and Silverman (1990) have proposed lower bounds based on a kind of modulus of continuity. We have shown here that a specific modulus of continuity gives both upper and lower bounds over a broad variety of *a priori* classes $\mathscr{F}$ and losses $\|\ \|$. Essentially, this modulus of continuity works for parameter estimation problems over function classes which are orthosymmetric and solid in some orthogonal basis.

In Donoho (1994a) and Donoho (1994b) in addition, it was shown that quantitative evaluations of minimax risk may be made by exploiting a connection between optimal recovery and statistical estimation. Here similar ideas are used to show that evaluations which are somewhat weaker—i.e. only accurate up to logarithmic terms—carry through in considerable generality. The method appears to have many other applications, e.g. in the estimation of non-linear functionals and in the study of inverse problems (Donoho, 1995b).

6.4. *Importance of Bases*

A key advantage of wavelets is that they lead to unconditional bases of the Besov and Triebel scales of function spaces, as described earlier. An important consequence is that shrinking coefficients in such an expansion cannot introduce large amounts of roughness into the function. We emphasize that the Fourier basis does *not* have this property. For example, de Leeuw *et al.* (1977) showed that given any (perhaps highly irregular) square integrable function f, one may construct a *continuous* function g by judiciously increasing the modulus and changing the phases of certain Fourier

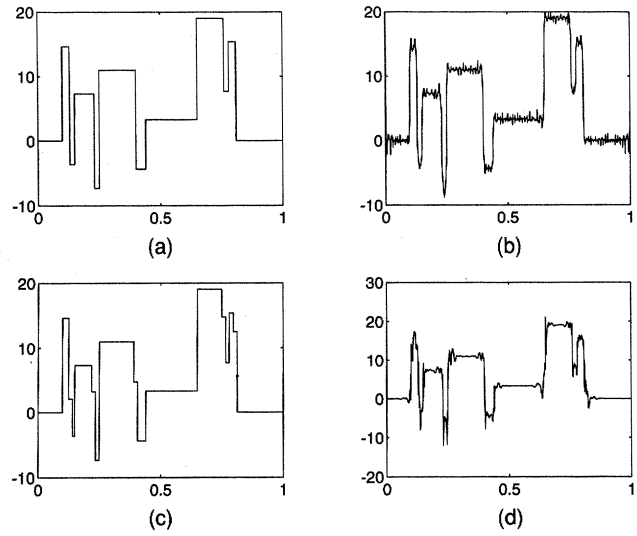

Fig. 10. Effect of mutilation of coefficients in wavelet and Fourier bases: (a) Blocks; (b) zeroing of real and imaginary Fourier coefficients beyond the 32 lowest frequencies; (c), (d) zeroing of negative wavelet coefficients beyond resolution level $5 = \log_2 32$

coefficients. As a graphical illustration, Fig. 10 shows what happens to the basic function blocks if all negative components are set to 0 above resolution level 5, using both the Haar (Fig. 10(c)) and a smoother wavelet. By contrast, when negative real and imaginary components of the Fourier coefficients beyond the lowest 32 are zeroed, the resulting function has much greater local roughness.

The success of wavelet bases has spurred wide interest in alternative time–frequency and time–scale decompositions using 'libraries' of bases built by using wavelet packets and cosine packets (e.g. Coifman *et al.* (1989) and Auscher *et al.* (1992)) and in projection-pursuit-style algorithms for choosing basis functions adapted to particular functions or signals (Mallat and Zhang, 1993). To understand the properties and possibilities of these tools for reconstructing objects measured in noise presents many fascinating problems for future statistical work.

6.5. *Relations to Other Work*

There is at the moment a large amount of work by applied mathematicians and engineers in applying wavelets to practical signal processing problems. Within this activity, there are several groups working on the applications of wavelets to denoise signals: Coifman and collaborators at Yale, Mallat and collaborators at Courant, Healy and collaborators at Dartmouth, DeVore at South Carolina and Lucier at Purdue. These groups have independently found that thresholding of wavelet coefficients works well to denoise signals. They have claimed successes on acoustic signals, and photographic and medical images, which encourages us to believe that our theoretical results describe phenomena observable in the real world.

Of these efforts, the closest to the present in point of view is the work of DeVore and Lucier (1992), who have announced results for estimation in Besov spaces paralleling our own. Obtained from an approximation theoretic point of view, the

parallel is perhaps to be expected, because of the well-known connections between optimal recovery and approximation theory.

In the statistics literature, a series of reports by Hall and Patil (1993a, b) and Fan *et al.* (1993a, b) have looked at properties of fixed threshold estimators (particularly using the $\sqrt{(2 \log n)}$ threshold) via asymptotic expansions of mean-squared error about a fixed function f. References to much of the other current statistics literature appear in the discussion following the paper.

6.6. *On Meaning of 'Asymptopia'*

There are of course many objections that one can make to the opinions expressed here. Certainly we have ignored the significance of logarithm terms, of irregularly spaced data, of non-Gaussian data, of non-translation invariance, of small sample performance and the power-of-two limitation, and we have unduly emphasized the Besov and Triebel spaces rather than real data sets. For the record, many specific improvements to the simple estimator described here can be made to enhance small sample performance, to reduce the prevalence of logarithm terms and to handle irregular data, and we hope to describe these elsewhere.

In this connection, the title word 'asymptopia' is meant to be thought provoking. We can easily envision positive and negative connotations, just as 'utopia' has both kinds of connotation.

In this paper, we have proposed an operational definition of the term. We believe that the ultimate goal of asymptotic minimax theory must be to develop, by rational mathematical criteria, new approaches to estimation problems, with previously unsuspected properties. If we attain this goal, and if the results look promising for certain applications, we are in asymptopia.

ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation grant DMS 92-09130 and National Institutes of Health grant CA 59039-18. The authors would like to thank Paul-Louis Hennequin, who organized the École d'Été de Probabilités at Saint Flour, 1990, where this collaboration began, and to Université de Paris VII (Jussieu) and Université de Paris-Sud (Orsay) for supporting visits of DLD and IMJ. The authors would like to thank Ildar Ibragimov and Arkady Nemirovskii for the personal correspondence cited, and the referees for their suggested improvements in presentation.

Many of the unpublished references by the authors are available by anonymous file transfer protocol from playfair.stanford.edu in the directory /pub/reports.

APPENDIX A

A.1. *Discrete Wavelet Transform*

For the reader's convenience, we give here a short account of a particular form of the empirical wavelet transform W_n^n described in Section 3.1. Our summary is derived from Daubechies (1992), section 5.6, which contains a full account. The original papers by Stephane Mallat are also valuable sources (e.g. Mallat (1989)).

We consider a periodized, orthogonal, discrete transform. Our implementation of this transform (along with boundary-corrected and biorthogonal versions) are available as part

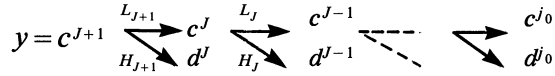

Fig. 11. Cascade structure of the discrete wavelet transform

of a larger collection of MATLAB routines, TeachWave, which may be obtained over Internet from the authors via anonymous file transfer protocol from playfair.stanford.edu in directory /pub/software. The figures in this paper may be reproduced (or modified!) by using the M-files in the directory /Scripts/Asymptopia in the TeachWave release. (Wavelet toolkits in the S language include 'WaveThresh' by G. P. Nason, available from Statlib at lib.stat.cmu.edu, and S+WAVELETS by A. G. Bruce and H.-Y. Gao soon to be available within S-PLUS.)

The forward transform maps data y, of length $n=2^{J+1}$ onto wavelet coefficients $w=(c^{(j_0)}, d^{(j_0)}, d^{(j_0+1)}, \ldots, d^{(J)})$ as diagrammed in Fig. 11. Thus $d^{(j)}$ is a vector of 2^j 'detail' coefficients at resolution level j. (Our convention for indexing j is the reverse of that of Daubechies.) Let $\mathbf{Z}_r=\{0, 1, \ldots, r-1\}$. The operators L_j and H_j map $\mathbf{Z}_{2^j}$ onto $\mathbf{Z}_{2^{j-1}}$ by convolution and downsampling:

$$\begin{aligned} c_k^{(j-1)} &= \sum_s h_{s-2k}\, c_s^{(j)}, \\ d_k^{(j-1)} &= \sum_s g_{s-2k}\, c_s^{(j)}. \end{aligned} \tag{54}$$

The summations run over $\mathbf{Z}_{2^j}$, and subscripts are extended periodically as necessary. The 'low pass' filter (h_s) and 'high pass' filter $(g_s=(-1)^s h_{1-s})$ are finite real-valued sequences subject to certain length, orthogonality and moment constraints associated with the construction of the scaling function ϕ and wavelets ψ: longer filters are required to achieve greater smoothness properties. Daubechies (1992) gives full details, along with tables of some of the celebrated filter families. In the simplest (Haar) case, $h_s \equiv 0$ except for $h_0=h_1=1/\sqrt{2}$, and equations (54) become

$$\begin{aligned} c_k^{(j-1)} &= (c_{2k}^{(j)} + c_{2k+1}^{(j)})/\sqrt{2}, \\ d_k^{(j-1)} &= (c_{2k}^{(j)} - c_{2k+1}^{(j)})/\sqrt{2}. \end{aligned}$$

However, this choice entails no smoothness properties and so is in practice generally replaced with longer filter sequences (h_s).

Regardless of the value of j_0 at which the cascade is stopped, the forward transform W_n^n is an orthogonal transformation. Thus the inverse wavelet transform may be implemented using the adjoint, yielding the equations

$$c_s^{(j+1)} = \sum_k h_{s-2k}\, c_k^{(j)} + g_{s-2k}\, d_k^{(j)}, \qquad j_0 \leqslant j \leqslant J, \tag{55}$$

which in the Haar case reduce to

$$\begin{aligned} c_{2r}^{(j+1)} &= (c_r^{(j)} + d_r^{(j)})/\sqrt{2}, \\ c_{2r+1}^{(j+1)} &= (c_r^{(j)} - d_r^{(j)})/\sqrt{2}. \end{aligned}$$

Since the filter sequences (h_s) and (g_s) appearing in equations (54) and (55) are of (short) finite length, the transform and its inverse involve only $O(n)$ operations.

The 'scaling functions' $\phi_{j_0,k}=(\phi_{j_0,k}(t_i),\ i=1, \ldots, n)$ and 'wavelets' $\psi_{j,k}=(\psi_{j,k}(t_i),\ i=1, \ldots, n)$ appearing in equation (13) are just the rows of W_n^n as constructed above and

so are easily plotted by applying the inverse transform to delta sequences. We use quotation marks since the true wavelet ψ and scaling function ϕ of the mathematical theory are not used explicitly in the algorithms applied to finite data: rather they only appear, for example, as limits of the infinitely repeated cascade. Again Daubechies (1992) has a full account. The existence and properties of ϕ and ψ are, however, crucial to establishing the properties of wavelet shrinkage for noisy data described in Sections 3.1.1–3.1.4.

A.2. *Unconditional Bases and Besov Spaces*

Again, for the reader's convenience, we summarize a few definitions and consequences from the references cited in Section 3.3. A sequence $\{e_n\}$ of elements of a separable Banach space E is called a Schauder basis if for all $v \in E$ there are *unique* $\beta_n \in \mathscr{L}$ such that $\Sigma_1^N \beta_n e_n$ converges to v in the norm of E as $N \to \infty$. A Schauder basis is called *unconditional* if there is a constant C with the following property: for every n, sequence (β_j) and constants (α_j) with $|\alpha_j| \leqslant 1$,

$$\| \sum_1^n \alpha_j \beta_j e_j \| \leqslant C \| \sum_1^n \beta_j e_j \|. \tag{56}$$

Thus, shrinking the coefficients of any element of E relative to an unconditional basis can increase its norm by at most the factor C.

Here is one of the classical definitions of Besov spaces. We follow DeVore and Popov (1988). Let $\Delta_h^{(r)} f(t)$ denote the rth difference

$$\sum_{k=0}^{r} \binom{r}{k} (-1)^k f(t + kh).$$

The rth modulus of smoothness of f in $L^p[0, 1]$ is

$$w_{r,p}(f; t) = \sup_{h \leqslant t} \| \Delta_h^{(r)} f \|_{L^p[0,\, 1-rh]}.$$

The *Besov* seminorm of index (σ, p, q) is defined for $r > \sigma$ by

$$|f|_{B^\sigma_{p,q}} = \left[\int_0^1 \left\{ \frac{w_{r,p}(f; h)}{h^\sigma} \right\}^q \frac{dh}{h} \right]^{1/q}$$

if $q < \infty$, and by

$$|f|_{B^\sigma_{p,\infty}} = \sup_{0<h<1} \left\{ \frac{w_{r,p}(f; h)}{h^\sigma} \right\}$$

if $q = \infty$. The Besov norm $\|f\|_{B^\sigma_{p,q}}$ is then defined as $\|f\|_{L^p[0,\,1]} + |f|_{B^\sigma_{p,q}}$.

An important consequence of the results of Lemarié and Meyer is that this norm is equivalent to the sequence norm (42), i.e. given a wavelet transform of sufficient regularity that associates to f the coefficients $(\theta_I(f))$ there are constants C_1 and C_2, not depending on f, so that

$$C_1 \|f\|_{B^\sigma_{p,q}} \leqslant \|\theta\|_{b^\sigma_{p,q}} \leqslant C_2 \|f\|_{B^\sigma_{p,q}}.$$

(For the original equivalence result on $\mathscr{R}$ see Lemarié and Meyer (1986); for a comprehensive development of the ideas see Frazier *et al.* (1991); for a version applying to $[0, 1]$, see Meyer (1992); for the version adapted to the statistical application in theorem 4, see Donoho (1992).)

The norm equivalence means that we may work with the sequence norms (42) and (43). These are clearly solid and orthosymmetric in the sense (24) (and so the unconditional basis

property (56) for the original norm $\|f\|_{B^\sigma_{p,q}}$, and all equivalent norms, follows). Thus it is the wavelet transform that renders the pleasant properties of soft thresholding in solid, orthosymmetric norms applicable to the Besov and Triebel functions spaces of statistical and scientific interest.

REFERENCES

Anscombe, F. (1948) The transformation of poisson, binomial and negative-binomial data. *Biometrika*, **35**, 246–254.

Auscher, P., Weiss, G. and Wickerhauser, M. (1992) Local sine and cosine bases of Coifman and Meyer and the construction of smooth wavelets. In *Wavelets—a Tutorial in Theory and Applications* (ed. C. K. Chui), pp. 237–256. New York: Academic Press.

Birgé, L. (1983) Approximation dans les éspaces metriques et théorie de l'estimation. *Z. Wahrsch. Ver. Geb.*, **65**, 181–237.

———(1985) Nonasymptotic minimax risk for Hellinger balls. *Probab. Math. Statist.*, **5**, 21–29.

———(1989) The Grenander estimator: a nonasymptotic approach. *Ann. Statist.*, **17**, 1532–1549.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1983) *CART: Classification and Regression Trees.* Belmont: Wadsworth.

Breiman, L., Meisel, W. and Purcell, E. (1977) Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135–144.

Bretagnolle, J. and Carol-Huber, C. (1979) Estimation des densités: risque minimax. *Z. Wahrsch. Ver. Geb.*, **47**, 119–137.

Brockmann, M., Gasser, T. and Herrmann, E. (1993) Locally adaptive bandwidth choice for kernel regression estimators. *J. Am. Statist. Ass.*, **88**, 1302–1309.

Brown, L. D. and Low, M. G. (1992) A constrained risk inequality with applications to nonparametric functional estimation. Unpublished.

Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1992) Multiresolution analysis, wavelets, and fast algorithms on an interval. *Compt. Rend. Acad. Sci. Paris* A, **316**, 417–421.

Coifman, R. R., Meyer, Y., Quake, S. and Wickerhauser, M. V. (1989) Signal processing and compression with wave packets. In *Proc. Int. Conf. Wavelets, Marseille* (ed. Y. Meyer). Paris: Masson.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

DeVore, R. and Lucier, B. (1992) Fast wavelet techniques for near-optimal image processing. In *Proc. IEEE Military Communications Conf.* New York: Institute of Electrical and Electronics Engineers Communications Society.

DeVore, R. A. and Popov, V. A. (1988) Interpolation of Besov spaces. *Trans. Am. Math. Soc.*, **305**, 397–414.

Donoho, D. (1988) One-sided inference about functionals of a density. *Ann. Statist.*, **16**, 1390–1420.

———(1992) Interpolating wavelet transforms. *Technical Report 408*. Department of Statistics, Stanford University, Stanford.

———(1994a) Statistical estimation and optimal recovery. *Ann. Statist.*, **22**, 238–270.

———(1994b) Asymptotic minimax risk for sup-norm loss; solution via optimal recovery. *Probab. Theory Reltd Flds*, **99**, 145–170.

———(1995a) De-noising via soft-thresholding. *IEEE Trans. Inform. Theory*, to be published.

———(1995b) Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harm. Anal.*, to be published.

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

———(1994b) Neo-classical minimax problems, thresholding, and adaptation. *Technical Report*. Department of Statistics, Stanford University, Stanford.

———(1995a) Minimax estimation via wavelet shrinkage. *Ann. Statist.*, to be published.

———(1995b) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, to be published.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1993a) Wavelet shrinkage: asymptopia? *Technical Report 419*. Department of Statistics, Stanford University, Stanford.

———(1993b) Density estimation by wavelet thresholding. *Technical Report 426*. Department of Statistics, Stanford University, Stanford.

Donoho, D. L. and Liu, R. C. (1991) Geometrizing rates of convergence, III. *Ann. Statist.*, **19**, 668–701.

Donoho, D. L., Liu, R. C. and MacGibbon, K. B. (1990) Minimax risk over hyperrectangles, and implications. *Ann. Statist.*, **18**, 1416–1437.

Efromovich, S. and Pinsker, M. (1981) Estimation of square-integrable density on the basis of a sequence of observations. *Prob. Inform. Transmssn*, **17**, 182–195.

———(1982) Estimation of square-integrable probability density of a random variable. *Prob. Inform. Transmssn*, **18**, 175–189.

———(1984) A learning algorithm for nonparametric filtering (in Russian). *Autom. Telem.*, **11**, 58–65.

Fan, J., Hall, P., Martin, M. and Patil, P. (1993a) Adaption to high spatial inhomogeneity based on wavelets and on local linear smoothing. *Technical Report CMA-SR18-93*. Centre for Mathematics and Its Applications, Australian National University, Canberra.

———(1993b) On local smoothing of nonparametric curve estimators. *Technical Report CMA-SR23-93*. Centre for Mathematics and Its Applications, Australian National University, Canberra.

Farrell, R. (1967) On the lack of a uniformly consistent sequence of estimates of a density function in certain cases. *Ann. Math. Statist.*, **38**, 471–474.

Frazier, M., Jawerth, B. and Weiss, G. (1991) *Littlewood–Paley Theory and the Study of Function Spaces*. Providence: American Mathematical Society.

Friedman, J. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–67.

Friedman, J. and Silverman, B. (1989) Flexible parsimonious smoothing and additive modeling. *Technometrics*, **31**, 3–21.

Golubev, G. (1987) Adaptive asymptotically minimax estimates of smooth signals. *Prob. Pered. Inform.*, **23**, 57–67.

Good, I. and Gaskins, R. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data (with discussion). *J. Am. Statist. Ass.*, **75**, 42–73.

Hall, P. and Patil, P. (1993a) Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Technical Report*. Australian National University, Canberra.

———(1993b) On wavelet methods for estimating smooth functions. *Technical Report*. Australian National University, Canberra.

Ibragimov, I. A. and Khas'minskii, R. Z. (1982) Bounds for the risks of non-parametric regression estimates. *Theory Probab. Applic.*, **27**, 84–99.

Johnstone, I. (1994) Minimax bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics, V* (eds S. Gupta and J. Berger), pp. 303–326. New York: Springer.

Johnstone, I., Kerkyacharian, G. and Picard, D. (1992) Estimation d'une densité de probabilité par méthode d'ondelettes. *Compt. Rend. Acad. Sci. Paris* A, **315**, 211–216.

Johnstone, I. and Silverman, B. (1990) Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.*, **18**, 251–280.

Korostelev, A. (1993) Asymptotic minimax estimation of regression function in the uniform norm. *Theory Probab. Applic.*, to be published.

Leadbetter, M. R., Lindgren, G. and Rootzen, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer.

de Leeuw, K., Kahane, J. and Katznelson, Y. (1977) Sur les coefficients de Fourier des fonctions continues. *Compt. Rend. Acad. Sci. Paris* A, **285**, 1001–1003.

Lemarié, P. and Meyer, Y. (1986) Ondelettes et bases hilbertiennes. *Rev. Mat. Iberam.*, **2**, 1–18.

Lepskii, O. (1991) On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Applic.*, **35**, 454–466.

Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.

Mallat, S. and Zhang, Z. (1993) Matching pursuits with time–frequency dictionaries. *IEEE Trans. Signal Process.*, **41**, 3397–3415.

Meyer, Y. (1990) *Ondelettes et Opérateurs*: vol. I, *Ondelettes*; vol. II, *Opérateurs de Calderón–Zygmund*; vol. III, *Opérateurs Multilinéaires*. Paris: Hermann.

———(1992) Ondelettes sur l'intervalle. *Rev. Mat. Iberam.*, **7**, 115–133.

Micchelli, C. A. (1975) Optimal estimation of linear functionals. *Technical Report 5729*. IBM.

Micchelli, C. A. and Rivlin, T. J. (1977) A survey of optimal recovery. In *Optimal Estimation in Approximation Theory* (eds C. A. Micchelli and T. J. Rivlin), pp. 1–54. New York: Plenum.

Müller, H.-G. and Stadtmuller, U. (1987) Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, **15**, 182–201.
Nason, G. (1994) Wavelet regression by cross-validation. *Technical Report 447*. Department of Statistics, Stanford University, Stanford.
Nemirovskii, A. (1986) Nonparametric estimation of smooth regression function. *J. Comput. Syst. Sci.*, **23**, no. 6, 1–11.
Nemirovskii, A., Polyak, B. and Tsybakov, A. (1985) Rate of convergence of nonparametric estimates of maximum-likelihood type. *Prob. Inform. Transmssn*, **21**, 258–272.
Nussbaum, M. (1985) Spline smoothing and asymptotic efficiency in l_2. *Ann. Statist.*, **13**, 984–997.
Peetre, J. (1975) *New Thoughts on Besov Spaces*, vol. 1. Durham: Duke University.
Pinsker, M. (1980) Optimal filtering of square integrable signals in gaussian white noise. *Prob. Inform. Transmssn*, **16**, 120–133.
Sacks, J. and Ylvisaker, D. (1981) Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.*, **9**, 334–346.
Samarov, A. (1992) Lower bound for the integral risk of density function estimates. In *Advances in Soviet Mathematics* (ed. R. Khasminskii), vol. 12, pp. 1–6. Providence: American Mathematical Society.
Speckman, P. (1979) Minimax estimates of linear functionals in a Hilbert space. Unpublished.
Stone, C. (1982) Optimal global rates of convergence for nonparametric estimators. *Ann. Statist.*, **10**, 1040–1053.
Terrell, G. and Scott, D. (1992) Variable kernel density estimation. *Ann. Statist.*, **20**, 1236–1265.
Traub, J., Wasilkowski, G. and Woźniakowski, H. (1988) *Information-based Complexity*. Reading: Addison-Wesley.
Tribouley, K. (1993) Estimation de densité: analyse multidimensionelle et méthodes d'ondelettes. *Thèse de Doctorat*. Université de Paris VII, Paris.
Triebel, H. (1992) *Theory of Function Spaces*, vol. II. Basel: Birkhäuser.
Wahba, G. and Wold, S. (1975) A completely automatic French curve. *Communs Statist.*, **4**, 1–17.
Wolfowitz, J. (1950) Minimax estimation of the mean of a normal distribution with known variance. *Ann. Math. Statist.*, **21**, 218–230.

DISCUSSION OF THE PAPER BY DONOHO, JOHNSTONE, KERKYACHARIAN AND PICARD

Paul L. Speckman (University of Missouri, Columbia): We have seen a review of truly exciting work, and I wish to congratulate the authors on their masterful presentation. We saw many of the main themes in curve estimation including optimal rates, minimax estimation, spatial adaptivity and computational efficiency, topics that are rarely combined. All this was tied together with wavelets, a subject that has had a large effect on much of the rest of the mathematical and scientific community. This review paper may become a landmark in the field.

I have two brief comments to make. A number of years ago when I was beginning work on minimax theory, an experienced colleague remarked that any statistician who believed in minimax estimation must be morose! Is minimaxity really a desirable property? Consider two familiar examples. Minimax estimation agrees with intuition for estimating the mean of a univariate normal distribution, but the minimax estimator of the binomial p under squared error loss has risk asymptotic to $1/4n$ for all p. Is minimax curve estimation comparable with estimating a normal mean or a binomial proportion? At least one of the early minimax estimators, the one attributed to Pinsker (1980), Efromovich and Pinsker (1981, 1982), Nussbaum (1985) and independently derived by Speckman (1985), is an equalizer rule on balls in a Sobolev space. Is this a good bench-mark, or should 'desirable' estimators do better than minimax estimators in most cases?

Secondly, I continue to be disappointed with wavelet examples. I am not certain that I prefer the visual performance of the wavelet estimators to 'standard' methods in many cases.

As an experiment, I used a technique that R. Eubank and I are developing on the 'Blocks' example. Motivated by partial smoothing splines (see Wahba (1984)), one can model a function with a jump at a point τ by

$$\mu(t) = \beta\,\phi(t-\tau) + f(t),$$

where

$$\phi(t-\tau) = \begin{cases} 0, & t < \tau, \\ 1, & t \geqslant \tau, \end{cases}$$

and f is continuous. In the setting of expression (57), suppose that S is a smoother matched to the presumed smoothness of f. Then (see Speckman (1988)) β can be estimated by solving

$$\min_{\beta} \|(I-S)\,(y - X\beta)\|^2, \tag{57}$$

where $X = (\phi(t_1-\tau), \ldots, \phi(t_n-\tau))'$. This idea can easily be modified to include discontinuities in derivatives at τ as well as jumps at other locations. We can detect the locations of jumps by plotting the estimated $\hat{\beta}$ as a function of τ, and this plot can be calibrated to avoid false detections (see Speckman (1993)).

After identifying locations of discontinuities $(\hat{\tau}_1, \ldots, \hat{\tau}_r)$, the coefficients in the model

$$\mu(t) = f(t) + \sum_{k=1}^{\tau} \beta_k\,\phi(t-\hat{\tau}_k)$$

are fitted by weighted least squares as in expression (57) with an appropriate $n \times r$ design matrix X, and the entire function is estimated as

$$\hat{\mu} = S(y - X\hat{\beta}) + X\hat{\beta}.$$

This technique was applied to simulated data modelled on the authors' 'Blocks' example. The model function is shown in Fig. 12(a), the model with added noise is displayed in Fig. 12(b) and a cross-validated Gasser–Müller kernel estimate is shown in Fig. 12(c). This fit displays the characteristic undersmoothing of a non-spatially adaptive method. With the same bandwidth, the detection method of Speckman (1993) identified all 11 jumps. The semiparametric model was then fitted with new bandwidth 0.09 estimated by generalized cross-validation. The corresponding completely automated fit is shown in Fig. 12(d).

This result is almost embarrassingly good. A step function with isolated discontinuities is an ideal target for our method, which of course would not perform well over the diverse kinds of problem considered in the paper. Without doubt, other methods such as the edge detection schemes based on boundary kernels of Müller (1992) or Wu and Chu (1993) would also do well. It is clear that methods tailored to specific problems can outperform the general wavelet techniques espoused here. Thus I am not completely convinced that 'the wavelet shrinkage method offers all that we might desire of a technique . . .' for all problems.

Examples like this do not detract from the work that we have seen in this paper. The authors are to be congratulated on many fronts. They have contributed greatly to a fundamental understanding of curve estimation methods, they have been remarkably innovative in applying wavelet methodology to statistical prolems, their methods are eminently practical to implement and they have achieved a rare synthesis of advances in theory and practice. I have profound respect for the paper, and it gives me great pleasure to propose a vote of thanks.

J. S. Marron (University of North Carolina, Chapel Hill): This work is fundamentally important. It has stimulated many new ideas in smoothing. It contains not one but three major contributions to statistics: first, the important new wavelet bases for use in orthogonal series estimation; second, the beautifully simple thresholding idea; third, powerful new theoretical tools for studying smoothers have been developed. This body of work represents a breakthrough in nonparametric curve estimation. The attempt to make a direct connection between minimax lower bounds and the practice of statistics is interesting and refreshing.

Spatial adaptation is often an important issue, but it is important to keep in mind that there are three motivations for doing different amounts of smoothing in different locations:

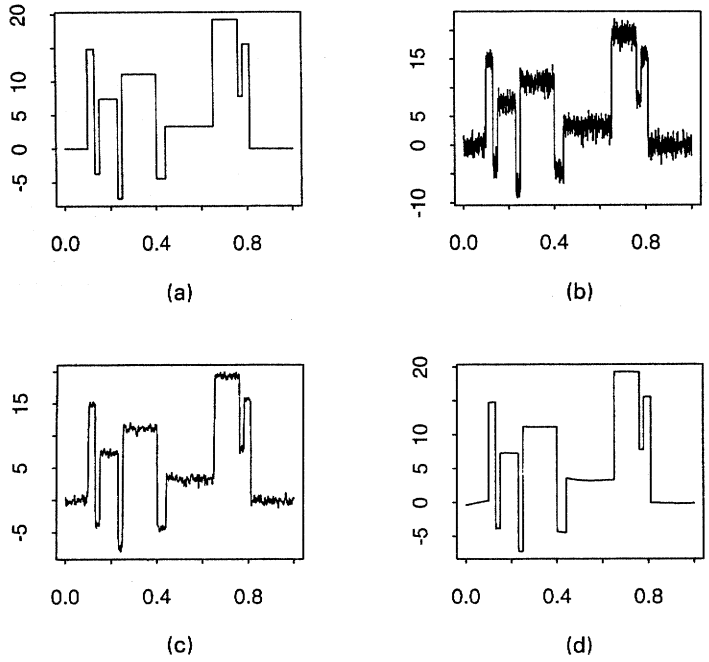

Fig. 12. Simulated example ($n = 2048$) analogous to the 'Block function': (a) ideal function; (b) function with Gaussian white noise, $\sigma = 1$; (c) reconstruction with the cross-validated Gasser–Müller smoother; (d) reconstruction with the cross-validated semiparametric estimator

(a) non-homogeneous 'curvature' in the underlying regression curve θ;
(b) heteroscedasticity in the noise;
(c) an irregular or random design.

Only the first of these is the focus here, but in many cases the others are of more importance. The authors indicate that there is work in progress on the other points, which is anxiously awaited. It appears to be straightforward to construct wavelet estimators, but much more difficult to adapt the thresholding ideas.

Some important (also non-linear) competing technologies should be mentioned. One of these is B-splines with knot deletion, and another is local polynomial estimation with location adaptive bandwidth. A direct comparison of thresholded wavelets with local bandwidth local polynomials, on the same examples as in the present paper, has been done by Fan and Gijbels (1995). The local polynomial method appears to provide a more effective estimate in all these examples.

What can be said about other applications of wavelet methods to statistics? The use of Fourier analysis in smoothing is a minor application, compared with the spectral analysis of time series. Does wavelet analysis of time series similarly have potential to be more important than wavelet smoothing?

Here are some further questions.

(a) When do the asymptotics take effect? Are exact risk calculations, as in Marron and Wand (1992), tractable for wavelets?
(b) How sensitive is the thresholding to the precise distribution $Y \sim N(\theta, \sigma^2)$? Is there any hope to robustify thresholding?

It is my pleasure to second the vote of thanks for this stimulating work.

The vote of thanks was passed by acclamation.

Bernard Silverman (University of Bristol): I would firstly like to refer to some of my current work joint with Guy Nason. The multiresolution analysis provided by wavelet methods of course has diagnostic

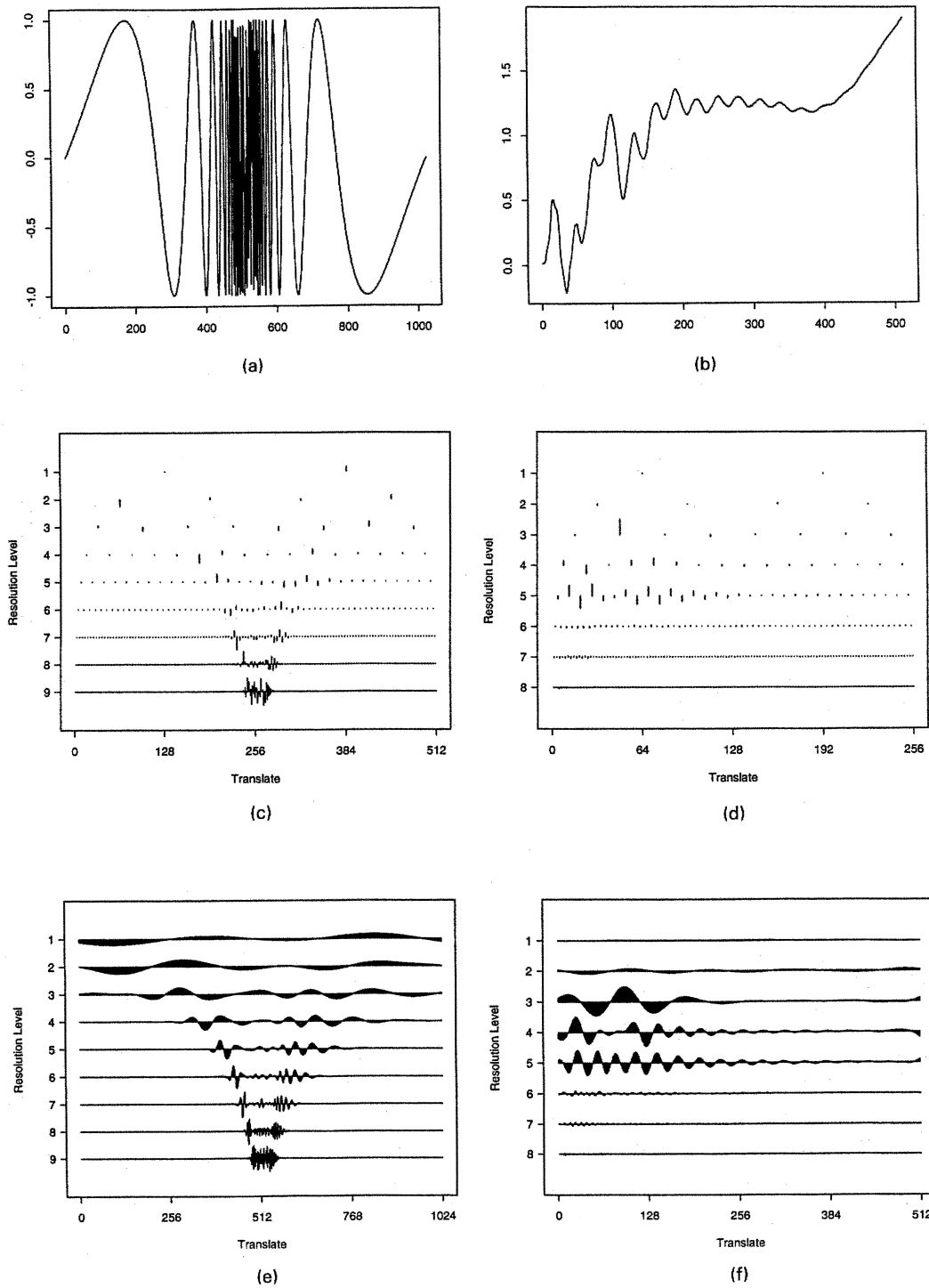

Fig. 13. (a), (b) Signal, (c), (d) wavelet transform and (e), (f) stationary wavelet transforms ((a), (c), (e), a simulated chirp; (b), (d), (f), the force exerted by the forelimb of a horse walking onto a force plate, sampled at 10 kHz)

and exploratory statistical ability beyond its use to construct estimators. The way in which the wavelet coefficients provide information localized both in space and time can be exploited directly to give a form of local spectral analysis. For this it seems preferable (at the cost of a log n factor in computational effort) to use a *stationary* wavelet transform of the data, where wavelets of all scales are placed at all n positions $k/2^n$. This has some connections with the *a trous* algorithm; see Shensa (1992). An example is shown in Fig. 13, where the changing frequency across the data becomes clear by the way that the weight of the wavelet coefficients moves on the frequency scale. The coefficients in the stationary transform are no longer orthogonal, but there is the possibility of using them to find the best positioned ordinary wavelet transform, avoiding the arbitrary choice of origin. We are also investigating ways of interpolating between the dyadic frequency scale, and of smoothing the coefficients at each level by exactly the right amount to remove the (predictable) sinusoidal variation.

Iain Johnstone and I have considered the extension to more general noise models than the white noise considered in the paper. If the noise is stationary it is natural to use *level-dependent* thresholds $t_{n,j}$ when shrinking the empirical wavelet coefficients $w_{j,k}$. Under mild conditions we show that if these are chosen appropriately then a result of the form (22) can be proved, by extension of the techniques of Donoho and Johnstone (1994a). For details see Johnstone and Silverman (1994).

Despite my great enthusiasm for the paper, I still have a little unease about certain aspects of the minimax paradigm. Suppose that you were trying to estimate the derivative of a function f. The results of Section 3.3.4 of the paper show that the derivative of the nearly minimax estimator $\hat{f}^*$ is a nearly minimax estimator of the derivative. So the thing to do is to find a good estimate of f and to differentiate it. (This conclusion has not much to do with wavelets and a similar result for linear smoothers can be obtained by generalizing the results of Speckman (1979).) However, common sense and experience—and other theoretical results—suggest that you should smooth *more* if you intend to differentiate the result. The technical explanation of this paradox is not difficult, but I believe that it raises important questions about the way that minimax results should be viewed in practice.

Guy Nason (University of Bristol): The authors are to be congratulated on a very clear and fascinating paper that unites several important ideas concerning the statistical uses of wavelets.

The authors referred to level-dependent thresholding and its importance for many applications such as inverse problems. The cross-validation approach of Nason (1994a) only provides a single global threshold although in principle it would be possible to modify the procedure to obtain level-dependent thresholds by making the optimization procedure multivariate.

The cross-validation algorithm in Nason (1994a) is similar to twofold cross-validation (Burman, 1989). Rather than using a random split, it splits an $n=2^{J+1}$ data set into two equal halves systematically with one set containing the evenly indexed points and the other containing the odd. The cross-validation score is constructed by building estimates with one half and comparing it with the other and vice versa. The score is numerically minimized and because the score is based on $n/2$ points a correction factor is applied to the minimizing threshold to give the 'n-point' cross-validated threshold. Results from simulation experiments such as those depicted in Fig. 14 are very encouraging and suggest that cross-validation can be made to work well with wavelet methods.

I have recently extended the cross-validation algorithm in Nason (1994b) so that it can handle

(a) two-dimensional functions (images) and
(b) data sets of *any* size using full cross-validation, based on a leaving-one-out method.

The second of these extensions shows great promise. I have compared the performance of twofold and full cross-validation on sets that are a power of 2 in length and they behave similarly (although twofold validation better handles heavy-tailed errors).

Two advantages of using cross-validation are that it does not use any prior knowledge about the error structure and it is simple to use with different norms. Nason (1994b) demonstrates this by attempting to clean up a noisy image by minimizing L^2, L^∞ and first- and second-derivative L^2 global norms.

From a practical point of view the most important aspect that needs to be addressed is the choice of where to begin thresholding (j_0). This choice has almost as much influence over the accuracy of the estimate as the threshold. There is work in progress on choosing j_0 but we would appreciate any practical advice that the authors can give.

Kongming Wang, Burkhardt Seifert and Theo Gasser (University of Zürich): The paper by Dave Donoho and colleagues summarizes an impressive collection of results for wavelets. The fact that near

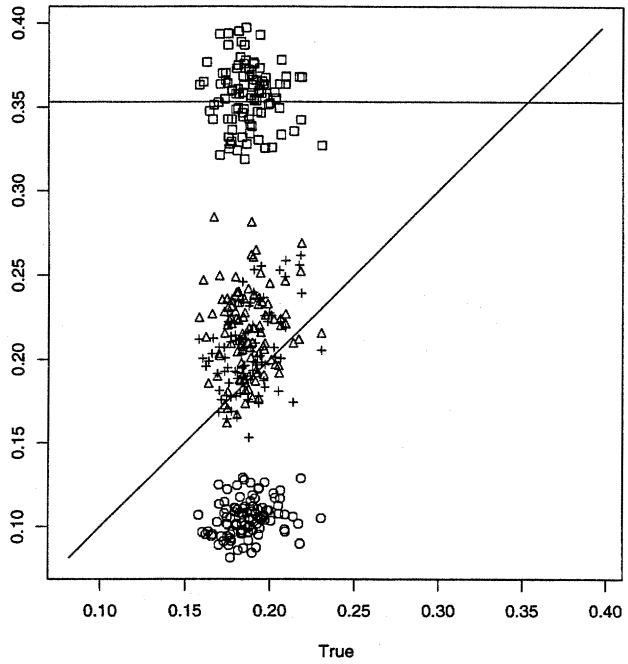

Fig. 14. Thresholds from 100 cross-validation simulations for estimating the piecewise polynomial with discontinuity from Nason and Silverman (1994) in the presence of independent normal additive noise: the true residual sum of squares minimizing threshold is recorded on the x-axis and the thresholds as computed by universal thresholding (□, VisuShrink), GlobalSure (○, SUREShrink with global threshold), twofold cross-validation (△) and full cross-validation (+) are plotted on the y-axis; the horizontal line is $\sqrt{(2 \log n)}$; those points on the diagonal line $y = x$ indicate successful estimation of the optimal threshold for a given simulation (in terms of the specific residual sum of squares)

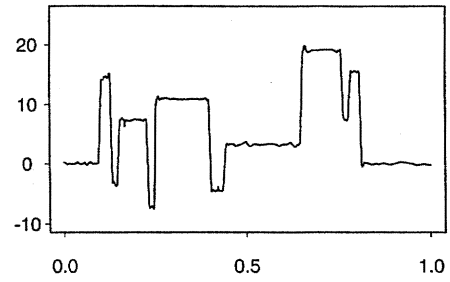

Fig. 15. Kernel smoothing with a local bandwidth: MASE = 1.22

optimality holds for many classes of functions, considered separately so far, has great intellectual appeal. Convolution kernel estimators, with some restricted optimality results, look mediocre in comparison. A local bandwidth selector was proposed in Brockmann *et al.* (1993). Incidentally, it also uses an $O(n)$ algorithm. Fig. 15 shows the results for a comparison of the Blocks function. The wavelet computations were done with the $(2 \log n)^{1/2}$ threshold. It produces a visually nicer fit than the lambda(n) threshold, and was preferred by the authors, despite an increase in mean average squared error (MASE) by a factor of 2. By the same logic we could tolerate a further small increase in MASE to obtain a visually still more pleasing fit by kernel estimation. Part of the difference can be attributed to the different 'kernel order', leading to Gibbs phenomena. The increase in MASE for kernels is due to a loss of steepness for jumps. The kernel fit evidently performs better than the linear fits presented in the paper (probably

TABLE 1

MASE for locally adaptive kernel estimator and wavelets for $r(x) = \exp\{-(x-0.25)^2/0.005\}/\sqrt{(0.005\pi)} + \exp\{-(x-0.5)^2/0.02\}/\sqrt{(0.02\pi)}$

	MASE values for the following values of n and σ:					
	$n=2048$		$n=256$		$n=64$	
	$\sigma=0.2$	$\sigma=1.$	$\sigma=0.2$	$\sigma=1.$	$\sigma=0.2$	$\sigma=1.$
Kernel local	1.1×10^{-3}	1.5×10^{-2}	5.9×10^{-3}	7.8×10^{-2}	1.8×10^{-2}	0.2352
Wavelet shrinkage	2.1×10^{-3}	1.3×10^{-2}	9.1×10^{-3}	7.8×10^{-2}	1.6×10^{-2}	0.2523

because of the method of bandwidth selection). A simple changepoint detector can improve the situation for kernel fits drastically (the possible existence of jumps can often be postulated in the field of application).

We undertook a small scale simulation for a smooth function (100 runs, Table 1), with a high and a low noise level. Since we had difficulty in estimating the residual variance for wavelets, we used the true noise variance, in contrast with kernel estimators. Also a boundary kernel adaptation was used—inflating the variance somewhat—and not a cyclical version. Thus, results are optimistic for wavelets. Considering this, the results for wavelets are good, but not quite competitive. Evidently, these limited results should be treated with great caution. It seems that the theoretical results are so far more impressive than the practical performance. For most applied statisticians there is no hurry to change to wavelets.

Although a local adaptation is heuristically entirely convincing, a price must be paid for estimating a local instead of a global bandwidth: a better mean integrated squared error can be easily achieved asymptotically for a local rule, owing to the larger class of estimators. However, the relative rate of convergence to this better value is slower than for a global rule. For our own method, the convergence of variance decreased from $O_p(n^{-1/2})$ to $O_p(n^{-1/4})$. Thus, some caution is appropriate when increasing the flexibility of estimators by replacing the constant tuning parameter by an infinite dimensional parameter for estimating infinite dimensional objects nonparametrically.

Sam Efromovitch (University of New Mexico, Albuquerque):

Thresholding as an adaptive method

I would like to congratulate the authors on developing a new method of adaptation for orthogonal series estimates based on wavelets. The method is asymptotically 'nearly' optimal over a wide variety of loss functions and estimated curves.

Regarding the adaptation itself, an issue that arises quickly is just how to implement thresholding for small samples, traditional smooth curves and trigonometric bases when the quality of estimation is defined by the mean integrated squared error (MISE).

To analyse the issue, I have restricted my attention to estimating 18 smooth underlying densities and sample sizes from 25 to 1000 observations. A hard threshold adaptation modified for the small sample sizes, based on a preliminary truncated Fourier series, is compared with a linear truncated estimate which is optimal for the setting. Both oracles and adaptive estimates have been compared.

The comparison has shown the following results. For the underlying densities considered and sample sizes the MISE of the threshold oracle is at most 1.6 times greater than the MISE of the linear oracle. The adaptive estimates have been compared via Monte Carlo simulations. The simulations have shown that for all underlying densities except a uniform density the ratio of the MISE of the threshold adaptive estimate to the MISE of the linear adaptive estimate is less than 1.8; for the uniform density it ranges from 2 to 2.4 for various sample sizes. This outcome is not surprising at all because densities which are nearly uniform are the minimax densities (see Efromovich (1985)).

The results, together with asymptotically near minimax properties of threshold adaptation, support the authors' conclusion that thresholding may be considered as an alternative to the known methods of adaptation.

Readers who are interested are referred to Efromovich (1994) where the details of this research may be found.

M. Nussbaum (Institute for Applied Analysis and Stochastics, Berlin): The authors have stressed the analogy of the problem of estimating a single normal mean and of estimating f in the Gaussian nonparametric regression. By Le Cam's (1986) theory, general parametric models can be reduced to the problem of the single normal mean (or of a finite dimensional normal mean). Clearly it is desirable to have an analogue for nonparametric experiments, i.e. to be able to extend the decision theory developed in this paper systematically. We shall describe a step forwards, relating to the problem of density estimation. Le Cam's deficiency distance Δ allows us to define rigorously the statistical closeness of two models; it utilizes Markov kernel transitions between them. For nonparametric models, Brown and Low (1992), treating the Gaussian nonparametric regression and its continuous analogue (the white noise model), have shown how to build an asymptotic decision theory: not via localization and limit experiments as in the parametric case but via an approximation to the original sequence in the sense of Δ, i.e. via *asymptotic equivalence*. Then the two sequences have asymptotically the same risk behaviour, for all problems with bounded loss. For a model assuming independent and identically distributed (IID) observations having density g, we can now state the following. Let the class $F(\alpha, C, \epsilon)$ be the class of all densities g on the unit interval which are in a Hölder ball $\Lambda^{\alpha}(C)$ (see equation (4) in the paper) and which are bounded from below by ϵ. Then if $\epsilon>0$ and $\alpha>\frac{1}{2}$ the IID model with density g is asymptotically equivalent to a regression

$$y_i = g^{1/2}(t_i) + (2n^{1/2})^{-1} z_i, \qquad i = 1, \ldots, n,$$

where z_i are independent standard normal and t_i are equispaced (see Nussbaum (1992)). This partly justifies the 'rootogram' procedure described by the authors. However, it should be stressed that the above equivalence is still an abstract result, which claims the *existence* of appropriate transitions between the two models. To make it fully constructive in the sense of giving recipes for efficient procedures requires further work. The authors' guess of working with binned density data does not seem bad, especially in the light of recent results on the role of the Haar expansion in empirical processes theory (see Koltchinskii (1994)). A common principle might be involved, namely reduction of the data to binomial and using basic normal approximations for these.

The full asymptotic equivalence in the sense of Δ applies not only to the estimation problems treated in the paper. Therefore it is not surprising that large *a priori* classes like total variation are not covered, as is already true for the regression–white noise result.

Yazhen Wang (University of Missouri, Columbia): It was the authors of this paper who pioneered the development of function estimation by using wavelets. Wavelet shrinkage is a breakthrough in function estimation and their work has revolutionized the way that functions are estimated. This very simple estimate is spatially adaptive and theoretically optimal. I very much enjoy reading this well-written and thought-provoking paper. It will undoubtedly stimulate several lines of research activity. I should like to leave the discussion of philosophical points to others and to make two comments on Section 5.4.

For certain inverse problems and function estimation for dependent data, level-dependent thresholds are needed to 'tune' wavelet estimates to be nearly minimax over a wide range of spaces (see Donoho (1993, 1995) and Wang (1994a)). The technique used to study these problems is the wavelet–vaguelette decomposition (see Donoho (1995) and Wang (1994a)). Such a decomposition is also very useful in fractal signal processing (for example see Wornell and Oppenheim (1992)).

Wavelets have a remarkable ability to zoom in on very short-lived frequency phenomena such as transients in signals or singularities in functions and hence provide an ideal tool for jump and sharp cusp detection. Detection by wavelets has several advantages over traditional smoothing methods (see Donoho (1993), Mallat and Hwang (1992) and Wang (1994b)). When a function has jumps, the wavelet estimate exhibits many undesirable spurious oscillations near jump locations. To remedy the drawback, such detection was employed in estimating the function (see Froment and Mallat (1992) and Wang (1994c)).

Rainer von Sachs (Universität Kaiserslautern): My comment is a contribution from quite a practical point of view. I consider wavelet shrinkage for local smoothing in nonparametric curve estimation. Concerning this I partly want to address the reservations which have been around with respect to the finite sample behaviour of these methods, the choice of additional tuning constants, etc. The authors themselves mention that, on the one hand, in this work they have ignored the significance of small sample performance, non-Gaussian data and so on. On the other hand, they are aware that many groups working on applied denoising found that thresholding wavelet coefficients works well.

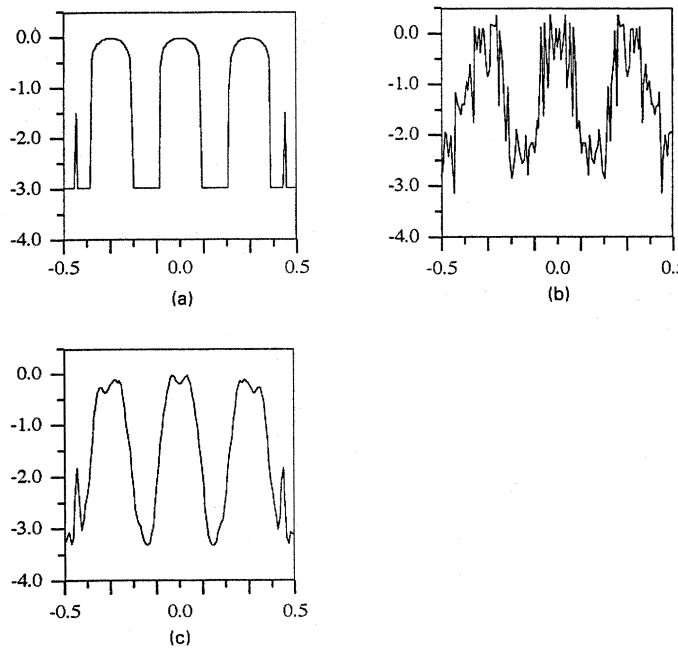

Fig. 16. Frequency cuts (at a fixed time), scaled logarithmically: (a) true evolutionary spectrum; (b) raw periodogram; (c) periodogram smoothed with soft thresholding (128 data points)

As an example which differs considerably from the independent and identically distributed and Gaussian situation, I applied non-linear wavelet techniques to the periodogram of a locally stationary process (as defined in Dahlhaus (1993)) to estimate its evolutionary spectrum (modelling transients, quasi-oscillating behaviour, etc.). Denoising this estimate often calls for local smoothing: von Sachs and Schneider (1994) found that even with small to moderate sample sizes (segments of length of 128–512 observations) we can separate noise and relevant local signal structure—provided that, in advance, the variance is stabilized by a logarithmic transformation (as in Gao (1993a) for the stationary situation). An illustration is given by Fig. 16, which shows cuts in frequency direction of a time–frequency-dependent spectrum, its raw and its locally smoothed periodogram estimate.

Let me give a comparison with 'classical' techniques, e.g. local bandwidth selection rules in kernel estimation: we must estimate the smoothing parameter at each design point, e.g. for a plug-in estimate, the second derivative of the unknown function which calls for another smoothing rule. Though I mention the nice iterative scheme of Brockmann *et al.* (1993) which tries to minimize the tuning effort, it seems to me that switching to the 'space of coefficients' (of really local basis functions) inherently allows a comparatively simple and intuitive choice of the smoothing parameter. This leads to good results, even if you apply the resulting threshold rule uniformly to the set of your wavelet coefficients. As an extension, refinements by introducing scale-dependent rules allow improvements for non-standard situations (as, for example, for the density estimation problem).

I would like to end by asking the authors about their experience with small sample sizes and refined parameter choice in the threshold rule for these non-standard situations.

David R. Brillinger (University of California, Berkeley): So many things are enjoyable about this paper: at times it is so broad in sweep that it appears to relate to all of statistics. The work is driven by theory with practice firmly in mind. In essence: the model is given by expressions (2) and (23), the method is set down in Section 3.1 and the heuristics are laid out in Section 4.5. Shrinkage is basic but has been employed by crystallographers for many years (e.g. Blow and Crick (1959)).

The model involves additive noise of constant level, no weighting and sometimes normality. In the count variate example (Section 5.6), a variance stabilizing transform is employed. In a search for

efficiency and appropriateness we are led to think about non-additive non-Gaussian noise. Well, iteratively reweighted least squares is directly available as a computational technique. Take the count example. A wavelet generalized linear model Poisson technique could involve representing the linear predictor by an orthogonal function expansion, choosing a link function, taking some initial estimates of the coefficients, forming the Poisson weights based on the initial values and then performing the standard generalized linear model iterations, except that the coefficients of the linear predictor are to be shrunk. One iterates to convergence. What is basic is a wavelet model for the mean level function and a variance form depending on the mean level as necessary. Of course, because of the iterations, the number of computations is increased, but unless the data set is extremely large that may not prove seriously difficult.

Brillinger (1994) presents some inferential aspects of the wavelet technique for a deterministic signal in the presence of additive stationary noise. That work was motivated by the problem of whether microtubule movement was diffuse or via jumps. As the authors emphasize, wavelets are convenient for addressing functions with jumps.

Following the notation of equation (13), the model considered is

$$f=\sum_k \beta_{j_0,k}\phi_{j_0,k}+\sum_{j\geqslant j_0,K}\alpha_{j,k}\psi_{j,k}+E$$

with the number of coefficients finite, but unknown, and E stationary mixing noise. Some large sample properties of the least squares estimates of the coefficients, in particular variances, are indicated and then it is shown that for hard limiting the wavelet estimate is asymptotically normal with mean f and variance that may be estimated by

$$\frac{2\pi\hat{f}_2(0)}{n}\left(\sum_k \phi_{j_0,k}^2+\sum_{j,k}\hat{w}_{j,k}^2\psi_{j,k}^2\right).$$

This is pointwise almost everywhere. Here $\hat{w}_{j,k}$ is the wavelet coefficient multiplier and $\hat{f}_2(0)$ is the estimate of the noise power spectrum at frequency 0, formed by averaging the squares of the ordinary least squares type of coefficients at one more than the highest level of resolution, which should be high, but not too high. For example, the asymptotic distribution allows approximate confidence intervals to be constructed.

Michael H. Neumann (Weierstrass Institute for Applied Analysis and Stochastics, Berlin): My comments are directed at possibilities for transferring the method of non-linear wavelet shrinkage to a wide variety of other curve estimation problems. Despite an error structure that is possibly not independently and identically distributed and non-Gaussian, many can be treated like a problem with Gaussian white noise.

Unified approach to curve estimation problems with non-Gaussian noise

Unlike for linear estimators, the risk equivalence between non-linear wavelet estimators in a particular non-Gaussian model and the Gaussian white noise model is not obvious. In particular, a central limit theorem for the empirical wavelet coefficients $w_{j,k}$ would not provide the risk equivalence for unbounded loss functions. However, often we can state asymptotic normality in terms of probabilities of large deviations by showing that

$$\frac{P\{\pm(w_{j,k}-\theta_{j,k})/\sigma_{j,k}\geqslant x\}}{1-\Phi(x)}\to 1 \text{ as } n\to\infty \tag{58}$$

holds uniformly in $(j,k)\in\mathscr{I}_n$, $0\leqslant x\leqslant\Delta_n$, where $\theta_{j,k}$ is the true coefficient $\sigma_{j,k}^2=\text{var}(w_{j,k})$, $\mathscr{I}_n=\{(j,k)\,|\,2^j\leqslant n^{1-\gamma}\}$ for some arbitrarily small $\gamma>0$, and $\Delta_n\asymp n^\delta$ for some positive $\delta=\delta(\gamma)$.

Define the accompanying Gaussian model as

$$\xi_{j,k}=\theta_{j,k}+\epsilon_{j,k},\qquad (j,k)\in\mathscr{I}_n, \tag{59}$$

where $\epsilon_{j,k}\sim N(0,\sigma_{j,k}^2)$.

Proposition. Let $\delta_{j,k}=\delta_{j,k,n}$ be monotone functions with $|\delta_{j,k}(y)| \leqslant |y|$. Then under expressions (58) and (59)

$$\sum_{(j,k)\in\mathcal{I}_n} E\{\delta_{j,k}(w_{j,k})-\theta_{j,k}\}^2 = \sum_{(j,k)\in\mathcal{I}_n} E\{\delta_{j,k}(\xi_{j,k})-\theta_{j,k}\}^2\{1+o(1)\}+O(n^{-1}). \tag{60}$$

We can apply thresholded wavelet estimators to empirical coefficients satisfying expression (58) in *exactly* the same way as in Gaussian white noise models and obtain an equivalent risk. The restriction of this equivalence to levels j bounded away from the finest resolution scale does not matter, because these levels can be neglected under usual smoothness assumptions without loss in asymptotic efficiency.

This approach was used in non-Gaussian regression with heteroscedastic errors and non-uniform design by Neumann and Spokoiny (1994). Neumann (1994) applied it in the framework of spectral density estimation for a stationary, possibly non-Gaussian time series. There, empirical wavelet coefficients were defined as

$$w_{j,k}=\int I(\omega)\,\psi_{j,k}(\omega)\,\delta\omega$$

where $I(\omega)$ is a (possibly tapered) periodogram. Other obvious fields of application are regression and density estimation with dependent observations.

Some additional remarks

An important point for practical applications of any nonparametric estimator is the data-driven choice of the smoothing parameter(s). The slightly suboptimal approach with 'log n thresholds' can be applied with obvious modifications to almost all other curve estimation problems. A levelwise optimization of the thresholds by leave-one-out cross-validation has been investigated in Neumann and Spokoiny (1994) for non-Gaussian regression.

In view of the heteroscedasticity of the accompanying model (59) in many practically relevant cases it makes sense to think about individual thresholds $t_{j,k}(n)$, which depend somehow on the individual variances $\sigma^2_{j,k}$ of the empirical wavelet coefficients.

Should we be satisfied with equivalence results like equation (60) for non-Gaussian models? It has been proved in many instances that estimation in a particular non-Gaussian model is not easier than in some model similar to equation (59). This gives some support for the use of normal theory thresholds.

Trevor Hastie (AT&T Bell Laboratories, Murray Hill) **and Robert Tibshirani** (University of Toronto): We congratulate the authors on a remarkable paper: a rare example of how mathematical power can produce both *elegant* and *useful* tools for data analysis.

Wavelet shrinkage is highly adaptive and well suited to the estimation of rough functions in high signal-to-noise situations. How well does it estimate smooth functions in lower signal-to-noise settings?

To investigate this, we carried out a small simulation study. We chose two functions: the Blocks function (signal-to-noise ratio 7, Fig. 2(a)) and $y=x^2$ with signal-to-noise ratio 2, and 1024 x-values equally spaced in $[-2, 2]$. We compared the authors' wavelet smoother VisuShrink (using Guy Nason's 'wavethresh' S language implementation) and cubic smoothing splines using generalized cross-validation to select the smoothing parameter. The S commands used for VisuShrink were

```
fit<- wd(y, filter.number=8, family="DaubLeAsymm", bc="symmetric")
fit<- wr(threshold(fit, type="soft", levels=6: (fit$nlevels-1),
     dev=function(x)mad(x)^2))
```

Figs 17 and 18 show the average estimates and plus or minus two standard error bands over 50 simulations for the two settings.

In Fig. 17 VisuShrink has lower variance but greater bias than the cubic spline, whereas in Fig. 18 the cubic spline has lower bias and variance. Table 2 shows a summary of the results, for the four function by signal-to-noise ratio scenarios. The Monte Carlo standard error is roughly 0.002. The variance Var is the dominant component of the mean-squared error MSE: VisuShrink wins by a small margin for the Bumps function, whereas the cubic smoothing spline wins easily for the quadratic. Varying the signal-to-noise ratio had little effect on the results.

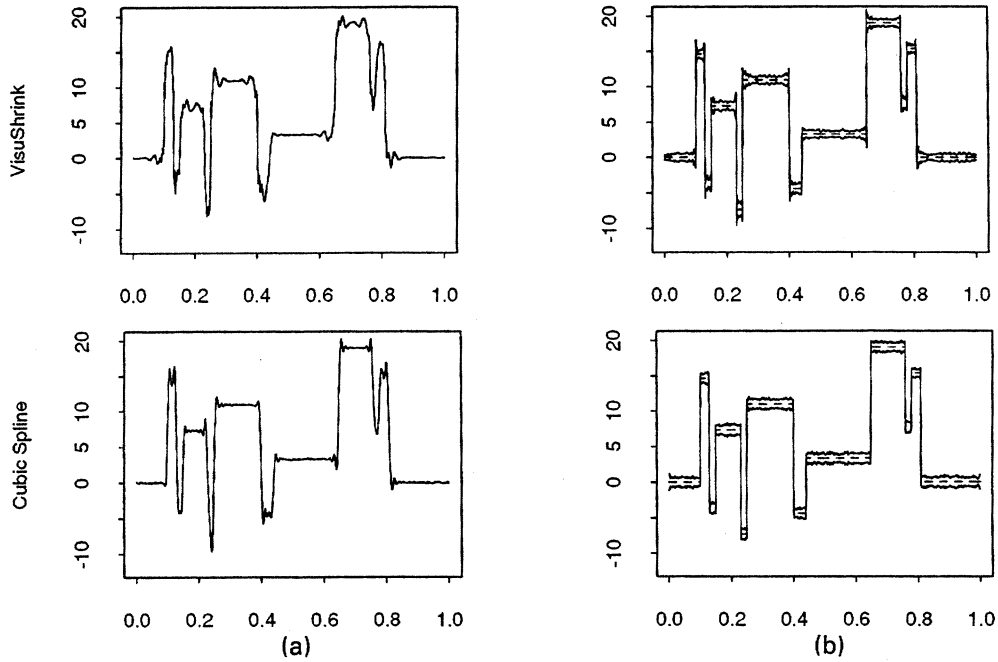

Fig. 17. (a) Average estimate and (b) plus and minus two standard errors for VisuShrink (top) and the cubic smoothing spline (bottom)

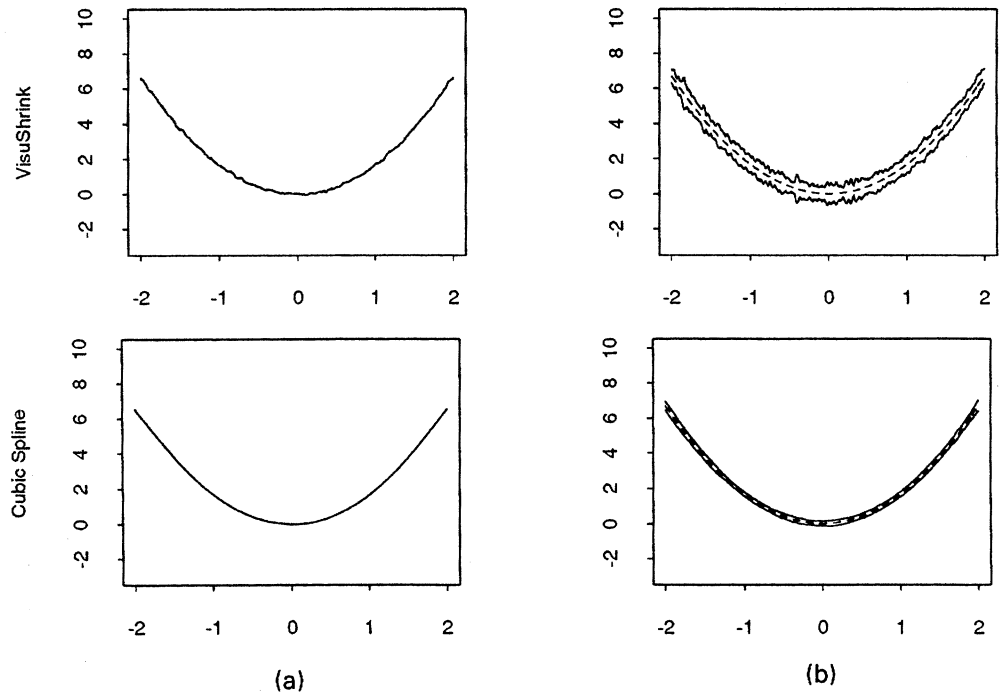

Fig. 18. (a) Average estimate and (b) plus and minus two standard errors for VisuShrink (top) and the cubic smoothing spline (bottom)

TABLE 2

Function	*Signal-to-noise ratio*	*VisuShrink*			*Cubic smoothing spline*		
		Bias^2	*Var*	*MSE*	*Bias*^{2}	*Var*	*MSE*
Bumps	7	0.037	0.067	0.104	0.005	0.113	0.118
	2	0.036	0.069	0.105	0.006	0.114	0.120
Quadratic	7	0.001	0.057	0.058	0.000	0.008	0.008
	2	0.001	0.057	0.058	0.000	0.005	0.005

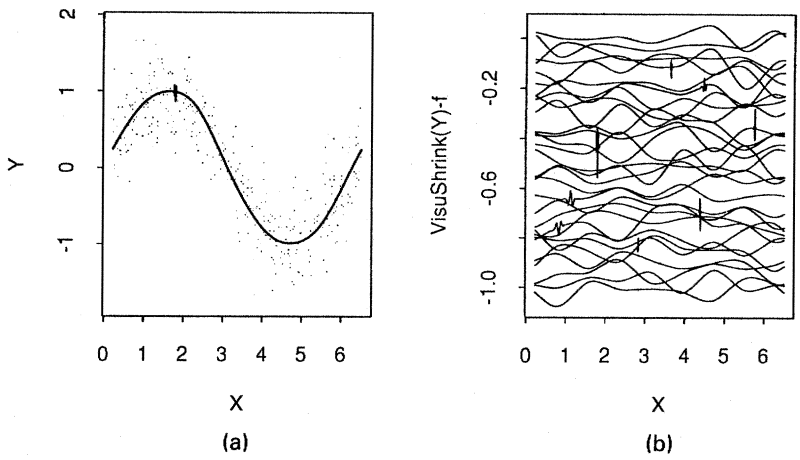

Fig. 19. (a) Data and VisuShrink estimate from a sinusoidal function with Gaussian noise $Y = \sin X + N(0, \frac{1}{3})$; (b) the fit minus the true function for 30 simulations of this model, each offset to avoid overplotting

Even with a moderate amount of noise, VisuShrink still passes some high frequency wiggles. Fig. 19(a) shows an example of such an occurrence using a very smooth function in a simple smoothing problem. Fig. 19(b) shows the fit minus the truth for 30 simulations of this model, each offset to avoid overplotting. Again we are using a smooth basis, here with periodic end effects:

```
fit<- wd(y, filter.number=8, family="DaubLeAsymm", bc="periodic")
fit<- wr(threshold(fit, type="soft", levels=6: (fit$nlevels-1),
    dev=function(x)mad(x)^2))
```

On the basis of this limited experience, it seems that the price paid for near optimal local adaptivity may be too high. We wonder whether the authors can shed light on these results.

Jianqing Fan (University of North Carolina, Chapel Hill): I would like to add that minimax results provide useful guidelines to selecting one class of 'intuitive' estimators among other competitors, particularly when the constant factors are not specified. The constant factor $A(\alpha)$ in expression (16) is explicitly given in Fan *et al.* (1993).

Strengths of wavelets

The major strength of the wavelet thresholding estimators is that they can capture very well local features such as the sharp bumps and discontinuities and global features such as high frequencies alternations. In addition to the authors' examples, I would like to add an application in statistical hypothesis testing, following the Neyman–Pearson paradigm. Fan (1995) showed that traditional nonparametric tests have low power in detecting fine features such as sharp and short aberrant as

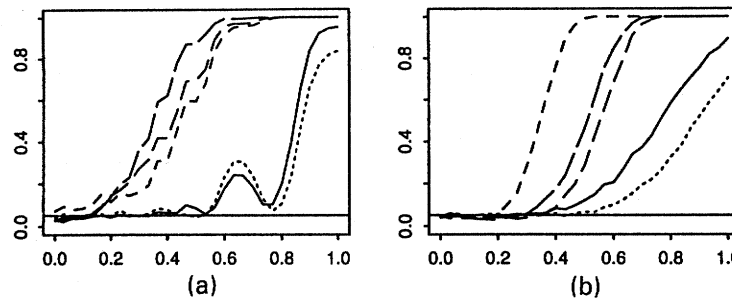

Fig. 20. Power of various nonparametric tests: (a) ———, Kolmogorov–Smirnov; ·····, Cramér–von Mises; --------, adaptive Neyman; – – – – –, wavelet hard thresholding; — — —, wavelet soft thresholding; (b) ———, Wilcoxon rank sum; ·····, Fisher–Yates; --------, adaptive Neyman; – – – – –, wavelet hard thresholding; — — —, wavelet soft thresholding

well as global features such as high frequency components. The drawbacks can be repaired via the wavelet thresholding and the Neyman truncation tests. I include a part of the results here (see Fan (1995)). Fig. 20(a) is used for detecting local features for the one-sample goodness-of-fit problem

$$H_0: F = \text{uniform}(-1, 1) \leftrightarrow H_1: F = F_\mu, \text{ with } F_\mu'(x) = \frac{1}{2} + \frac{1}{2}\sin\left(\frac{2\pi x}{x^2 + 0.1^2}\right) I(|x| < \mu).$$

The power is plotted against μ which indicates the strength of the local character around 0. Fig. 20(b) shows the power of two-sample nonparametric tests for the problem

$$H_0: F_1 = F_2 \leftrightarrow H_1: F_1 \neq F_2.$$

The power, computed at the alternatives $F_1 = N(0, 1)$ and $F_2 = 0.7N(\mu/0.7, 1) + 0.3N(-\mu/0.3, 1)$, is plotted against μ. The horizontal line indicates the 5% significance level and the sample sizes are 200.

Other spatial adaptation methods

There are many references on nonparametric smoothing. Can these methods have a spatial adaptation similar to wavelet thresholding? Fan and Gijbels (1995) proposed a simple data-driven *variable bandwidth* for local polynomial fitting and demonstrated that it works as well as the wavelet thresholding methods. This method trades off the bias and variance at each location and can also be used for non-equispaced designs and for estimation of derivatives. The data-driven local smoothing method is also highly non-linear, which coincides with the authors' claim in Section 6.2. Further, such an approach is applicable to likelihood-based models such as the generalized linear models and the proportional hazards model.

Open problems

(a) Which wavelet bases should be used for a given data set? Can the level-dependent thresholding be chosen to minimize the mean-squared error at each location?
(b) How do we assess the sampling variability to wavelet thresholding estimates?
(c) How does wavelet thresholding behave at boundary regions?
(d) Can the advantages of wavelet transforms carry over to the likelihood-based models, in particular, when it is incorporated with a 'dimensionality reduction principle' such as the generalized partially linear models

$$f(\mathbf{X}) = g\{\eta(\mathbf{X}_1^T\alpha) + \mathbf{X}_2^T\beta\}$$

for some specific link function g, where $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$ is a multidimensional covariate?

The following contributions were received in writing after the meeting.

Anestis Antoniadis (University of Grenoble): In reviewing their work, the authors give a clear and impressive exposition of wavelet shrinkage methods for nonparametric estimation problems. The

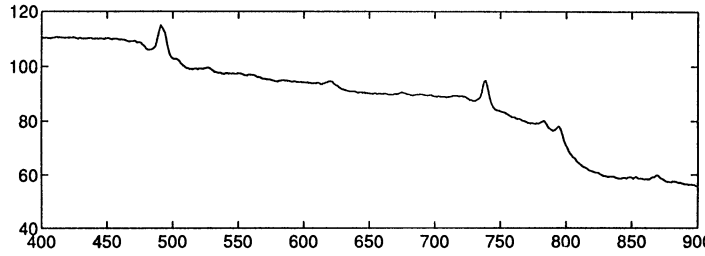

Fig. 21. Smoothly thresholded estimate of the simulated spectrum in Fig. 9(b)

variety of applications outlined shows how useful wavelet shrinkage methods can be. The results make the theory of functional minimax estimation more accessible to the statistical community.

I would like to add some minor comments.

To consider sequence model (23) as essentially an equivalent representation of model (2), we should replace the approximation of the empirial wavelet coefficients of the sampled function f by the theoretical functions. One way to do this is to use coiflets (see Antoniadis (1994a)); another, adopted by the authors, is to use a slightly modified transform which guarantees that the empirical coefficients are precisely the theoretical coefficients of f in this slightly modified transform. Both approaches assume that f is at least in $\mathscr{F} \in \mathscr{C}(R, D)$ (Section 3.3) with $1/p < \sigma < \min(R, D)$ which ensures that $\mathscr{F}$ embeds continuously in $C([0, 1])$. What happens when we consider more unusual function spaces such as $B^{\sigma}_{p,q}$ with $0 < p < 1$ and $\sigma = 1/p$ (spaces of extremely spatial heterogeneity)? Decomposition results for such spaces have recently been developed by Frazier *et al.* (1991) but the sequence and function 'equivalence' no longer holds.

For traditional smoothers, several researchers have concentrated on the question of asymptotic normality of the estimates useful in the determination of confidence bands. When linear wavelet methods (based solely on the scaling function ϕ) are used, we can derive an asymptotic normality of the estimator at dyadic points (Antoniadis *et al.* (1995)). However, at non-dyadic points the asymptotic variance of the estimator oscillates and asymptotic normality cannot be obtained. Do the authors think it possible to develop other approaches for producing confidence bands?

The soft thresholding procedure of Section 4.3.2 is one way to produce an adaptive estimator of the regression function. We may also think about smoothly thresholded estimators of the form

$$\hat{f}_n = \sum_k w_{j_0,k}\, \phi_{j_0,k} + \sum_{j \geqslant j_0} \sum_k \frac{\hat{\alpha}_{j,k}}{1 + 2^{2\beta j}\lambda}\, \psi_{j,k}$$

which can be implemented via a regularization framework (see DeVore and Lucier (1992a) and Antoniadis (1994b)). Fig. 21 shows a reconstruction of the simulated spectrum from electron spectroscopy for chemical analysis given in Fig. 9(b), using such a procedure with $\beta = 0.5$ and a data-based optimal smoothing parameter λ. The peak at about 620 is now retained, without any considerable residual high frequency roughness.

The methodology presented depends heavily on the assumption of independent observations. If wavelet shrinkage is applied to serially correlated data, the estimates can be substantially biased. I would be grateful for any suggestions that the authors may have about how their results could be extended to this case.

Lucien Birgé (Université Paris VI): I agree with most of the comments developed in Section 2 concerning the minimax risk. But even if most results in this regard have little or no practical relevance they were necessary for understanding the very nature of nonparametric problems.

The method proposed by the authors is particularly attractive because of its practical relevance (although it has, from my point of view, very little to do with Utopia) and it is a very interesting contribution towards finding general adaptive methods. Nevertheless, although they put great emphasis on spatial adaptivity, which is clearly the essential practical property of their estimators, from a theoretical point of view, I would consider it as one particular variation on the widely developed theme of adaptivity. The main difference is the introduction of Besov spaces together with non-linear

procedures, which provides the necessary spatial inhomogeneity, instead of the more classical Sobolev or Hölder spaces.

I would rather consider all those problems as special issues of the following problem: we choose a family of estimators depending on some parameters (bandwidth or local bandwidths for kernel estimators, number and location of the knots for adaptive spline estimation and position of the co-ordinates to be estimated in the case of projection estimators) and we try to choose these parameters from the data in some optimal way. This means that we try to find an automatic procedure (what the authors call an 'oracle') to choose the parameters such that the resulting minimax risk is almost as good as if the choice were made with the complete knowledge of the true unknown function to be estimated. Another approach is what is usually called 'model selection': given a family of models (in this paper, the finite dimensional linear spaces generated by some finite subsets of a wavelet basis) and a method of estimation (projection estimators), is it possible to find a model leading to an almost optimal risk without knowing the true function to be estimated? This function need not belong to any of the hypothetical models but we should choose a model that provides an optimal trade-off between bias and variance. The choice of finite wavelet expansions provides the attractive family of models allowing adaptation to spatial inhomogeneity and is particularly well suited to estimation in Besov spaces but many other choices would be possible, leading either to similar or different adaptation properties. If a general 'optimal' or 'close to optimal' model selection scheme were available, it could cope with this and other situations, the resulting properties of the estimators depending only on the choice of the estimation procedure and family of models in hand.

Neil J. Crellin (Stanford University) **and Michael A. Martin** (Australian National University, Canberra): We congratulate the authors on their eloquent and persuasive advocacy of wavelet methods. The paper raises important general questions about the future direction of research into smoothing methods.

Fear, uncertainty, doubt

The authors paint a grim picture of the current state of minimax theory for recovering objects from noisy data. At best, they would describe it as a loose collection of one-trick ponies, each uniquely designed to solve a particular problem. At worst, they would regard it as a muddled morass of competing techniques, difficult to comprehend and apply.

We believe that the authors' pessimism about the state of minimax estimation theory is unjustified. Certainly, the minimax estimation literature is vast and diffuse, but this condition is typical of scientific endeavour, where most advances are small and specialized. Moreover, the search for generality in developing methodology inevitably extracts a price when interest focuses on specific applications for which certain assumptions are justifiable. What cost do wavelet methods incur? The significance of logarithmic terms in the risk bounds associated with wavelet smoothers cannot be ignored. When smoothness assumptions on the target function are justified, methods that are nearly optimal as described here are, simply, non-optimal and can be outperformed by straightforward kernel-based methods with appropriate bandwidth choice. Further, simple wavelet methods will not work without modification for irregularly spaced data sets, or data sets whose sample size is not a power of 2, etc. The modifications necessary to adapt wavelet-based methods to these new circumstances lead to suspicions that the charges of 'complexity, nuance, [and] uncertain generality' levelled at classical methods might apply just as easily to wavelets. The issues of appropriate choice of wavelet bases and the best form of thresholding are also critical questions, research into which will involve a degree of 'qualification and specialization'.

We believe that wavelet methods will augment rather than supplant classical smoothing methods. The near optimality properties of wavelet methods are promising indications of what can be achieved generally, and, for situations where wavelet-based methods *are* optimal, they will attract significant use. Areas in which wavelet methods promise to be especially useful include speech recognition, data compression and discontinuity detection. None-the-less, classical methods that are optimal under smoothness assumptions will continue to gain favour as long as applications exist that warrant these assumptions. Wavelets represent neither the beginning nor the end of our search for good, general smoothing methods, but they are a promising step along the way.

Paul Doukhan (Université de Cergy-Pontoise): Functional estimation is associated with $n^{-r/(1+2r)}$ convergence rates for parameters with regularity r. This optimal rate is generally obtained for adequate values of a smoothing parameter. Adaptation is defined as a data-dependent choice of this smoothing

sequence—e.g. cross-validation and plug-in methods for kernel estimation and smoothing splines. Wavelet shrinkage was defined by the authors to obtain minimax rates for cases where linear estimates do not achieve them. An $\ln n$ loss in the rate yields an 'almost minimax' adaptive procedure when r is unknown; this shows the power of the method. Wavelets were first used as a tool of linear density estimation (Doukhan, 1988; Doukhan and Leon, 1990) to provide a global L^2-deviation result for projection density estimation (Soulier (1991) considered the variance of diffusions). Wavelet analysis—i.e. orthogonal projections on a multiscale analysis—combines the bias advantages of convolution kernels and the variance advantages of orthogonal projections (see Ango-Nze and Doukhan (1993) for dependent data). The asymptotic properties of 'wavelet shrinkage' estimates of conditional expectations are unknown (even for the independent and identically distributed observations case) because such estimates involve denominators that are not non-negative (see equation (53)). Also, there is no result in distribution for non-linear estimators and one wonders about extensions of L^2 global tests of hypothesis in this non-linear frame. Finally, it is possible to consider 'Dirac-like' distributions and parameter spaces of measures around such singular measures (see Donoho *et al.* (1992) and Doukhan and Gamboa (1994) for an analytical framework). Wavelets are known to describe simply any class of distributions; hence, after the present study of classes of regular functions, it would be interesting to investigate this opposite statistical problem.

Joachim Engel (Universität Bonn): In this paper as well as in a series of other publications by the same authors Donoho, Johnstone, Kerkyacharian and Picard demonstrate convincingly that shrinking wavelet coefficients via thresholding offers an appealing method for spatially adaptive nonparametric curve estimation. From a theoretical statistician's point of view wavelet threshold estimators mathematically form a compelling theory leading to minimax results over very broad classes of function spaces. In their generality these results surpass anything known for competing methods by far.

From an applied statistician's perspective wavelets offer some advantages but also raise some questions. Wavelets are prominent because of their computational ease. The empirical wavelet coefficients are fast to compute and the spatially adaptive smoothing rule of soft thresholding is easy to implement. Whereas methods like locally variable kernel estimators estimate a whole auxiliary function, the bandwidth function, wavelet shrinkage achieves its localization by applying 'uniformly' the same threshold to all wavelet coefficients above a coarse resolution level. Problematic for the data analyst are several restrictions that pertain to the minimax theory of wavelet shrinkage estimators: fixed equidistant design; homoscedasticity; powers of 2 sample sizes; normal errors. How sensitive are the estimators to the choice of the low frequency cut-off j_0 and the constant A in the threshold formula for estimating densities?

Although other smoothing methods are characterized by the authors as being atheoretical, these methods are usually based on intuitively quite plausible concepts. A partial answer to some of the above questions as well as some insight of what wavelet estimators are doing to the data may be gained when considering the simplest class of wavelets: the system of Haar functions. Granted that as for the wavelets with zero regularity the nice minimax results do not hold any longer in their stated form, Haar wavelets lead to very intuitive estimators provided that some precaution is observed. In Engel (1993, 1994) a term selection rule for Haar wavelet estimators is developed leading to adaptive histograms and regressograms. This follows from the special structure of the Haar system. Spatial adaptivity is achieved with a hard thresholding type of rule with an additional provision requiring, when selecting a certain wavelet coefficient, inclusion of its 'dyadic predecessors'. The resulting Haar wavelet estimator allows heteroscedasticity and random design data. Therefore it leads to a versatile tool for exploratory data analysis resembling recursive, tree-structured methods like classification and regression trees.

Alexander A. Georgiev (Albemarle Corporation and Louisiana State University, Baton Rouge) **and Hong Liu** (University of Pennsylvania, Philadelphia): We should like to congratulate the authors for the appealing piece of work that draws together a variety of curve estimation problems from disparate areas within the general minimax framework. The paper is an important statistical contribution to the fast growing wavelet transform toolbox for recovering signals and images from noisy data.

We would like to offer an alternative approach for recovering infinite dimensional objects and a brief comparison with the authors' results from Figs 2–6. Instead of using the regression equation (2) as a curve model, we propose (Liu, 1992; Liu and Georgiev, 1993, 1994) a spatially autoregression–regression model

$$y_i = f(y_{i-2}, y_{i-1}, y_{i+1}, y_{i+2}, t_i) + \sigma z_i, \qquad i = 1, \ldots, n, \tag{61}$$

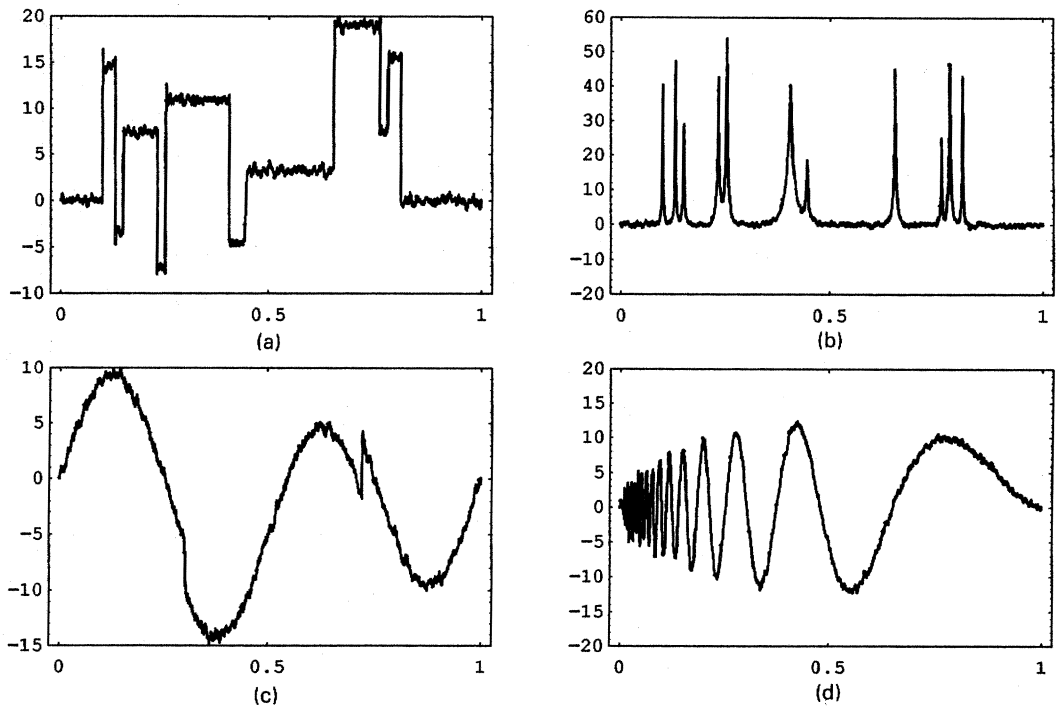

Fig. 22. Reconstructions using the autoregression–regression model and a kernel nonparametric estimate: (a) Blocks; (b) Bumps; (c) Heavisine; (d) Doppler

as a superposition of our previous research (Georgiev, 1984, 1985, 1988) in the area of nonparametric functional estimation. We use the kernel method and the data from Fig. 3 to recover the four functions (Fig. 22).

We also compute the signal-to-noise ratio (SNR) for each reconstruction by using the criteria

$$\mathrm{SNR} = -10 \log_{10} \left\{ \frac{\sum_i (f_i - \hat{f}_i)^2}{\sum_i f_i^2} \right\} \qquad \text{(dB)}$$

where f_i are the noiseless data (Fig. 2), and $\hat{f}_i$ denotes either raw data (Fig. 3) or the reconstructions (Figs 4–6). Table 3 lists SNRs for each method.

We shall finish with a few comments about the choice of model (61). Obviously, there is no unique modelling approach for recovering infinite dimensional objects from noisy data. This makes the

TABLE 3
SNRs

	Blocks	*Bumps*	*Heavisine*	*Doppler*
Data	18.9220	17.9117	17.2323	17.0412
Wavelet	19.8350	17.5714	28.4771	22.1560
Spline	22.3558	21.3147	28.5289	24.0435
Fourier	21.2852	20.4044	27.2747	22.7468
Model (61)	27.8856	21.4224	27.1639	24.1202

comparison of results in the SNR table even more difficult: researchers use a classical model (2) and advance wavelet methods; we use the more sophisticated model (61) and the classical kernel method. We believe that the underlying model is an important assumption in the recovery problem and a factor determining the quality of the reconstruction. Some insight for this methodological paradigm would be welcome.

I. J. Good (Virginia Polytechnic Institute and State University, Blacksburg): I do not understand the wavelet procedure sufficiently well to make a constructive comment, and Appendix A did not help me. (For example, I do not know what 'downsampling' means.) But I do have a question concerning Section 5.6 where a comparison is made between 'wavelet thresholding' and the penalized likelihood fit of Good and Gaskins (1980) as applied to some high energy scattering data. My question is does the wavelet method produce nearly the same 13 bumps as reached by Good and Gaskins? Figs 7 and 8 do not seem to provide an answer to this question, but the text claims that the two estimators were remarkably close. 'Closeness' has five distinct meanings depending on whether one has in mind the probability density as such or, in contrast, the number of bumps or hollows or modes or dips. (Incidentally the authors use 'prepending' in a sense different from the definitions in the *Oxford English Dictionary*.)

Another question is, are the wavelet procedures intended in the spirit of exploratory data analysis? The expression 'soft thresholding' occurs in the caption of Fig. 4 but I could not find a definition. A glossary might have been helpful, especially for the silent majority of readers who are too busy to read every word.

Peter Hall and Prakash Patil (Australian National University, Canberra): Donoho, Johnstone, Kerkyacharian and Picard's paper is particularly noteworthy for its clear introduction to the very powerful technology that they have developed. In the spirit of their challenging contributions we would like to expand the debate by including points to which users of more traditional smoothing methods might like to see reactions from expounders of wavelets. Obviously we have our own responses to those issues, so let us invite a healthy discussion by provocatively making assertions, rather than simply asking questions!

(a) We suggest that the major contribution of wavelet methods to statistics will be to expand significantly the class of problems to which smoothing methods may be applied in a computationally efficient way, rather than to supplant standard methods. The data sets to which the authors apply their techniques are not really typical of many that are traditionally treated by smoothers. Relative to standard settings, the data in the present paper enjoy high signal-to-noise ratios and have signals whose frequency changes very rapidly in space (e.g. at discontinuities). For reasons discussed below, the wavelet methods described by the authors do not necessarily perform well in traditional settings. This is not in any sense a criticism; it is an attempt to place wavelet methods into context.

(b) *Spatial adaptivity*: to a person used to traditional smoothing problems, where the target curve has perhaps a flattish section and one or two not-too-pointed peaks or valleys, spatial adaptivity means the ability to adjust the bandwidth by a relatively small amount along the curve. In the asymptopia of that context, frequency varies by a bounded amount as sample size n increases. The sort of wavelet methods that the authors discuss do not provide this degree of adaptivity. Their methods are adaptive to relatively large changes in frequency, of the order $(\log n)^{1/2}$ or larger. They can be made more adaptive, but then computation becomes a major issue.

(c) If j_0 is kept fixed as n increases—which seems to be the authors' intention—the estimators described in the present paper provide an excessive amount of smoothing when applied to curves that are smooth or piecewise smooth. Their mean-squared errors are asymptotically dominated by bias. They suffer this drawback as a penance for their ready applicability to a very wide range of variable frequency curves, with which more traditional methods do not cope so successfully. The problem can be overcome by choosing j_0 to increase with n at a suitable rate (actually, $n^{1/(2r+1)}$), but then the value of j_0 must be chosen empirically in much the same way that a bandwidth would be.

Eva Herrmann (Technische Hochschule Darmstadt): This discussion paper has set new bench-marks in the theory of nonparametric regression. The proposed method of wavelet shrinkage impressively satisfies asymptotic criteria which go far beyond traditional criteria in their generality.

Theorem 1, especially the nice property of estimating the zero function exactly with increasing probability, cannot be satisfied by traditional methods, even a spatial adaptive method. Unfortunately, this property seems to be lost even for the wavelet shrinkage described in the examples which does not set thresholds at the coarsest scale.

The asymptotic theory of theorems 2–4 is new in its generality. I agree with the authors that this kind of robustness is very attractive but I do not agree with the provocative statement that they 'offer all we might desire of a technique, from optimality to generality'. To justify such a statement something more is needed. However, other methods may exist which are simultaneously better, if not in rate perhaps so in constants. In this context it would be helpful for applications if situations could be described where VisuShrink proposed by the authors really satisfies the asymptotic property (10). Such a statement would give some hints about how to choose the parameter D and other parameters of the wavelet class or if one should use level-dependent thresholds, e.g. those of SUREShrink. However, it is important to know whether a breakdown can happen for VisuShrink in situations which do not satisfy model (2) but are close in some respects. Besides this, simulations show that wavelet shrinkage can compare with traditional spatially adaptive methods even for moderate sample sizes.

Before wavelet shrinkage can become the method of choice in practice it must of course be adapted to some of these situations, especially more general design and heteroscedastic non-Gaussian errors. This can be done for example for the local bandwidth kernel estimator of Brockmann *et al.* (1993), but it is not obvious how to modify VisuShrink. The variance estimator which is proposed in Section 5.1 is only crude, as stated by the authors themselves. It should at least be modified in a way which makes it applicable even for non-Gaussian noise where the asymptotic $\sup_{f \in \mathcal{F}(C)}\{P(\hat{\sigma} \leqslant 1.01\sigma)\} \to 1$ for $n \to \infty$ will fail.

Eric D. Kolaczyk (University of Chicago): I would like to comment briefly on the topic of indirect data. The methods of the authors for estimating f from direct data, $y = f + z$, are especially impressive in that they are simple (transform–threshold–invert transform), fast and theoretically tractable. A generalization of these methods to indirect data, $y = Kf + z$, should have similar characteristics. Donoho (1995) and Kolaczyk (1994) showed that this goal indeed is attainable in many cases.

Donoho (1995) laid out a framework for such a generalization by introducing the wavelet–vaguelette decomposition (WVD). This method is a wavelet analogue of the singular value decomposition, in which the unobserved function f is decomposed with respect to a wavelet basis, $\{\psi_I\}_{I \in \mathcal{I}}$, and the observed function Kf is decomposed with respect to a basis of vaguelettes, $\{\gamma_I\}_{I \in \mathcal{I}}$. The vaguelettes are near wavelet functions whose specific form depends on both K and the ψ_I in such a way that the corresponding vaguelette and wavelet coefficients are equal, i.e. $[\gamma_I, Kf] = \langle \psi_I, f \rangle$. Recovery of the vaguelette coefficients of Kf means recovery of the wavelet coefficients of f. Donoho proposed a method of estimating f based on soft thresholding of these coefficients which retains the simplicity and theoretical tractability of the current authors.

However, to maintain truly the spirit of the methods of the authors, implementation of the WVD must be done with efficient algorithms. Since the vaguelette functions are not actually wavelets and depend on both K and ψ_I, a calculation of the coefficients must be approached in a problem-specific manner. Furthermore, the thresholds used in shrinking the coefficients typically must be level dependent. Kolaczyk (1994) has developed methods for implementing the WVD in the contexts of integration, fractional integration (e.g. Abel transform) and tomography.

To illustrate, in the case of tomography Kolaczyk (1994) used a multiresolution analogue of the traditional filtering of backprojected projections (FBP) algorithm to compute the vaguelette coefficients. The size of these coefficients increases with resolution index j like $2^{j/2}$, so the level-dependent thresholds are designed to increase similarly in size. Thresholding in this manner, before applying the inverse wavelet transform, amounts to a form of regularization which seeks to compensate for the ill-posedness of the problem. Similar to the experience of the authors, in the case of direct data, images reconstructed by using this approach tend to be less noisy than those by using standard FBP approaches, but at the expense of a slight loss of detail.

O. V. Lepskii (Institute for System Analysis, Moscow, and Humboldt-Universität, Berlin), **E. Mammen** (Humboldt-Universität, Berlin) **and V. G. Spokoiny** (Institute for Information Transmission Problems, Moscow, and Institut für Angewandte Analysis und Stochastik, Berlin): First we thank the authors for this excellent presentation. We have really enjoyed this interesting tour through asymptopia. Wavelets have brought new life to curve estimation. The wavelets approach offers new powerful procedures for statistical applications and has motivated much new research.

One of the most interesting features of wavelet estimates is their adaptation to spatially inhomogeneous smoothness. This property has been shown in this paper by demonstrating optimal rates over all

Besov spaces. As noted in the paper, the Besov scale of spaces contains classes of spatially inhomogeneous functions.

Stimulated by the papers of the authors we have considered the following problem (Lepskii *et al.*, 1994). Is there a kernel estimate with locally data adaptive bandwidth that shares with wavelet estimates their good asymptotic performance over Besov classes (including the classes containing functions with spatially inhomogeneous functions)? In particular, the kernel estimate should achieve the minimax rates specified in theorem 9. This would imply that the kernel estimate would adapt well to spatial inhomogeneous smoothness.

We succeeded in constructing a kernel estimate with these asymptotic properties by modifying a general adaptation procedure due to Lepskii (1990). In our set-up the procedure is as follows. For every argument (design point etc.) kernel estimates are calculated for different bandwidths. A criterion is introduced for deciding whether differences between kernel estimates for two different bandwidths are significant. The largest bandwidth is selected for which no significant differences show up in comparisons with smaller bandwidths. We propose to use the kernel estimate with this bandwidth.

There is an important difference between our method and wavelet estimates based on thresholding of empirical wavelet coefficients. The wavelet estimates use localization in the time and in the frequency domain. Our method uses only localization in time as happens also to be the case for other spatially adaptive curve estimates (variable knot splines or other kernel estimates with locally varying bandwidth). We would like to ask the following question: what are the advantages of localization in time and frequency and how does this approach compare with methods based only on localization in the time domain? We have no intuition about the answer of this question and we think that this question is only one example for interesting research problems turning up in asymptopia.

Bradley Lucier (Purdue University, West Lafayette): The method of thresholding wavelet coefficients as described in the paper is related in a strong way to traditional smoothing splines. In smoothing splines, one finds the minimizer $\tilde{f}$ of an expression similar to

$$\min_g (\|f-g\|_{L_2} + \lambda |g|_{W^{m,2}});$$

here f is (related to) the noisy data, g is a candidate for the smoothed approximation and $W^{m,2}$ is the Sobolev space of functions with m weak derivatives in L_2. We can generalize this problem to seeking an approximate minimizer of the problem

$$\|f-g\|_{L_2} + \lambda \|g\|_Y$$

where Y is *any* smoothness space embedded in L_2. If for Y we choose Besov spaces $B_p^\alpha(L_p)$ that have α derivatives in L_p with minimal smoothness to be embedded in L_2, then we find (by a variant of the Sobolev embedding theorem) that

$$\frac{1}{p} = \frac{\alpha}{d} + \frac{1}{2}$$

in d dimensions, and the algorithm for finding an approximate minimizer of this functional is to take the wavelet coefficients of f and to threshold them below a certain value ϵ, which depends on the noise level and which in turn determines λ. Details are given in DeVore and Lucier (1992b). It is important to determine minimax estimation results for these minimally smooth function spaces; this is discussed further in DeVore and Lucier (1992b).

Secondly, I would like to point out that wavelet shrinkage assumes that functions can be characterized completely by their smoothness in various function spaces, and this can have drawbacks for specific applications. For example, none of the function space norms determined by the size of wavelet coefficients assume that large wavelet coefficients at finer scales are located, spatially, near large wavelet coefficients at coarser levels, which is indeed the case when we have a jump discontinuity in a function or an image. If the functions that you wish to approximate have this property, there may be more efficient ways to estimate them. In particular, it is much more informative (and efficient) to model example (a) of

Fig. 2 as a piecewise constant function with finitely many jumps, and with this new model we are likely to achieve better noise removal *for this particular case*. But, if you believe that your underlying set of functions (natural images, for example) has non-trivial regions for which the smoothness space characterization is the most efficient, then I believe that wavelet methods will prove most efficient in noise removal and smoothing.

Peter McCullagh (University of Chicago): In the pre-computer era, when computational efficiency was even more critical than it is now, Yates's algorithm provided an efficient method for the computation by hand of main effects and interactions for factorial designs. Yates's algorithm is a form of fast Fourier transform in which the basis functions are typically tensor products of orthogonal polynomials. The determination of significant effects and the estimation of σ are typically accomplished by using a half-normal plot of absolute standardized contrasts, preferably with main effects omitted. Fitted values are then computed by back-transformation with insignificant effects omitted, an early form of thresholding. It seems plausible that the same technique of half-normal plotting could be used for thresholding wavelet coefficients, at least for independent observations.

Pierre Moulin (Bell Communications Research, Morristown): First I would like to commend the authors for an outstanding paper. In addition to its contributions to minimax theory, the paper fills a void: whereas in recent years engineers have successfully been applying the wavelet transform to signal denoising problems, so far valuable practical achievements have not been matched by comparable theoretical advances.

With the authors providing us with ammunition in Section 6.6, I would like to comment on the small sample problem. The wavelet coefficients of functions in various classes exhibit a distinctive structure, such as decay of fine scale coefficients at a particular rate. When basic wavelet shrinkage is applied to the data, the wavelet coefficients are treated independently of each other, and these dependences are ignored. Unfortunately, this may give rise to rather unpleasant distortions of signal features, in which case the shrinkage technique does not perform nearly as impressively as in the examples given. Correcting for those artefacts would require modelling (Basseville *et al.*, 1992) or adapting to such dependences. Two of the factors affecting the artefacts are the choice of the coarse scale and the choice of the wavelet. The latter should result from a trade-off between properties such as regularity, number of vanishing moments, symmetry and spatial localization (Daubechies, 1993). Whereas the first two properties directly affect the performance of the estimator (measured in a suitable norm), the last two are arguably of visual significance.

Another topic is the estimation of functions corrupted by non-Gaussian noise. Such problems occur in log-spectral density estimation as well as in signal processing applications such as radar and sonar. The wavelet shrinkage approach and the concept of visually noise-free reconstruction may be applied to these problems (Moulin, 1993, 1994; Gao, 1993b). Under mild technical conditions, the distribution of the noise wavelet coefficients converges to a Gaussian distribution at coarse scales, by application of the central limit theorem. At fine scales, however, the distribution of the coefficients may be markedly different. In log-spectral density estimation, large thresholds should be used at fine scales to account for this effect.

Hans-Georg Müller (University of California, Davis): Donoho, Johnstone, Kerkyacharian and Picard are to be congratulated on their interesting theoretical results. It is impressive to see that near minimaxity results for the wavelet shrinkage method for noisy regression data have been obtained. However, the practical behaviour of this method in its current state for the important small to moderate sample size case is shaky and not competitive with 'adaptive methods'. Since we never have enough data to find out meaningfully what the nature of the underlying function space is, it makes empirical sense to consider the spaces $\mathcal{F}_n$, the class of infinitely differentiable functions whose derivatives are all Lipschitz continuous, except at a finite number of $D=D(n)$ discontinuities in function or derivatives. Two-stage adaptive procedures for the case $D>0$, where in the first step the location of discontinuities is determined and in the second step the curve is estimated, adapting to discontinuities as end points, are highly non-linear and lead to curve estimates with the usual asymptotic properties due to the very fast convergence of estimated discontinuity locations (Müller, 1992; Eubank and Speckman, 1993; Wu and Chu, 1993).

These approaches try to model curves explicitly, which is advantageous as it allows us to relate modelling assumptions directly with features in the curve, and can be obtained by modifying classical

smoothing methods, like locally weighted least squares, kernel methods or semiparametric methods. Problems such procedures confront explicitly rather than implicitly are

(a) discriminating between noise in data and discontinuities in underlying curves, i.e. the question whether $D=0$ or rather $D>0$ (Müller and Stadtmüller, 1994) and
(b) if $D>0$, determination of the number of discontinuities D (Yao, 1990).

In these models $\mathcal{F}_n$ consistent estimates of σ based on difference schemes are available as an alternative to the inconsistent estimate described in Section 5.1.

The adaptive smoothing paradigm encompasses practically relevant flexibility besides discontinuity adaptation in a variety of other ways, currently unavailable for wavelets:

(a) no need for normal errors;
(b) adaptation to heteroscedasticity of the data, non-equidistance of the design, and local curvature is possible.

Optimal local bandwidths

$$b^*(x) = c\left\{\frac{g^{(2)}(x)^2}{f(x)}\sigma^2(x)\right\}^{1/5},$$

adapting to local curvature $g^{(2)}(x)$ of the curve to be estimated, to the design density $f(t)$ and to a variance function $\sigma^2(x)$ can be chosen adaptively from the data in models $\mathcal{F}_n$ slightly modifying arguments in Müller and Stadtmüller (1987). Discontinuities in curves or derivatives are an inherent feature of many curve data and the addition of wavelets to the toolkit of statisticians for dealing with such data is a welcome development.

Richard A. Olshen (Stanford University School of Medicine): I congratulate the authors for this paper. They are correct to call attention to spatial adaptivity, to new discoveries that suitably formulated, minimax theory need not lead to non-adaptive methods and to the importance of bases. However, I will argue against some points of view.

One disagreement is with Section 2.4.3, in which the 'spatial adaptivity camp', to which I guess I belong, is described as being 'atheoretical, as opposed to antitheoretical'. I pretend to authority only concerning binary tree-structured methods from among those listed in that section. See Olshen (1994) for a brief survey of tree-structured methods for classification, regression, clustering and survival analysis. Among its references are those to work on asymptotics of these methods (see Breiman *et al.* (1984), chapters 11 and 12, Gordon and Olshen (1984), Nobel (1993) and Nobel and Olshen (1994)). We can do much mathematics to gain understanding of spatially adaptive statistical algorithms that is rather different from asymptotic minimax conclusions concerning functions of a single real variable. As trees proceed to asymptopia, one studies the algorithms applied to 'true' underlying distributions as data accrue (bias—Banach's principle, as in Garsia (1970) is relevant here), as well as the successive discrepancies between application to true and to 'empirical' distributions (variance). The geometry of terminal regions must be 'under control' for good asymptotic behaviour. So 'end cut preference', as follows from the law of the iterated logarithm (Breiman *et al.* (1984), section 11.8), is in tension with the imperative that the 'terminal regions' be in suitable senses a differentiation basis for L^1 of the domain of the predictors, typically though not exclusively an Euclidean space of arbitrary dimension (Gordon and Olshen, 1984). Without successful resolution of this tension there is no first half of the asymptotics for trees. Geometry figures as well in the second half, for the terminal regions must belong to something like a Vapnik–Chervonenkis class so that requisite large deviation results apply. The greedy growing criterion of pruned, tree-structured vector quantization tends to make 'good geometry' (see Nobel and Olshen (1994)), and it is empirically so that tree-structured methods can be suspect when resulting terminal regions are geometrically 'perverse'.

The comment in Section 2.6 concerning the inability of 'atheoretical spatial adaptation proposals' to deal with structure in the underlying object in a tomographic situation seems very wrong to me in view of Heaton (1994).

Finally, there are reasons why tree-structured methods have proven popular in some applied areas. Simply put, they 'work', and sufficiently well that users can be sceptical about looking elsewhere. See, for example, Cosman *et al.* (1994), Goldman *et al.* (1988) and Breiman *et al.* (1984).

A. B. Tsybakov (Université Paris VI): This paper is important from both a mathematical and a practical point of view. Mathematically, it gives a unified treatment of optimal rates of convergence for nonparametric estimation, in a very general set-up. This is achieved by invoking the modern tools of approximation theory, optimal recovery and wavelets. With an elegant thresholding device the statistical minimax problems are reduced to deterministic optimal recovery problems, and wavelet expansions provide an easy representation of sophisticated functional classes in sequence space. A close relationship between nonparametric estimation and approximation theory was emphasized by I. A. Ibragimov and R. Z. Khasminskii as early as in the 1970s. This paper builds a bridge between both theories by applying the optimal recovery approach and demonstrating how powerful it is. I find this approach very fruitful. It will certainly be widely used in the future. Also the paper shows that statistics is no longer satisfied with simply consuming approximation theory results, and it develops this theory for its own purposes (see theorems 7 and 8, which are related to a chain of work on the subject).

Practically, the paper gives universal procedures which work reasonably well in cases where the standard tools, such as splines, kernels or Fourier series, fail. Of course, in a particular problem the standard tools can be modified, by crafted techniques based on intuition, to obtain nice results. The authors suggest that we should not rely too much on intuition and should use an adaptive wavelet shrinkage procedure. This proposal is appealing but it has some problems. For example, the tuning constants D, R and j_0 in the definition of the estimator are not specified. Their choice is a matter of art, as also is the choice of father and mother wavelets. (The choice of a kernel does not influence the performance of kernel estimators much; similar quantitative results would be useful for wavelets.) In many cases, where the underlying functions are smooth, users are quite satisfied with standard nonparametric estimators. Their smoothing parameters can be chosen adaptively. The asymptotic distributions and confidence bands are also available, which is not so for wavelet estimators. However, it is reasonable to prefer wavelet shrinkage when the prior knowledge about the underlying function is poor, or when we suspect that it has a bad form, as in the four examples of the paper.

Grace Wahba (University of Wisconsin, Madison): We thank the authors for an intellectual *tour de force*. Wavelet shrinkage methods may well be the method of choice in applications such as denoising sound and images with certain kinds of noise. The spatial adaptivity of wavelets is important and forces us to consider how it might be incorporated into other nonparametric regression methods.

I give an example of how a modest amount of adaptively chosen spatial variability may be implemented in (quadratically) regularized estimates in a reproducing kernel space (e.g. in an L_2 Sobolev ball). We take the simplest possible case, but generalizations are immediate.

Let

$$\mathscr{H} = W_{2,0}^m = \left\{ f: \int_0^1 f^{(m)}(t)^2 \, dt < \infty, \, f(0) = f'(0) = \ldots = f^{(m-1)}(0) = 0 \right\},$$

with the square norm

$$J_\lambda(f) = \int_0^1 \lambda f^{(m)}(u)^2 \, du.$$

(The initial conditions can be removed.) Given data $y_i = f(t_i) + \sigma z_i$, $i = 1, \ldots, n$, the minimizer f_λ in $\mathscr{H}$ of

$$\frac{1}{n} \sum_{i=1}^n \{y_i - f(t_i)\}^2 + J_\lambda(f)$$

is given by $f_\lambda(\,) = \sum_{i=1}^n c_i R_\lambda(\,, t_i)$ where

$$R_\lambda(s, t) = \int_0^1 \frac{1}{\lambda} G(s, u) \, G(t, u) \, du$$

is the reproducing kernel for $\mathscr{H}$ with the square norm J_λ, here $G(s, u) = (s-u)_+^{m-1}/(m-1)!$ and $c = (c_1, \ldots, c_n)' = (\Sigma_\lambda + I)^{-1} y$ where $y = (y_1, \ldots, y_n)'$ and Σ_λ is the $n \times n$ matrix with ijth entry

$R_\lambda(t_i, t_j)$; see Wahba (1990). Now replace λ by a slowly varying, strictly positive function $\lambda(u)$ (satisfying some regularity conditions). Since, for $f \in \mathscr{H}$,

$$f(t) = \int_0^1 G(t, u)\, f^{(m)}(u)\, du = \int_0^1 \left\{ G(t, u) \frac{1}{\sqrt{\lambda(u)}} \right\} \left\{ f^{(m)}(u) \sqrt{\lambda(u)} \right\} du$$

it is not difficult to show that the reproducing kernel for $\mathscr{H}$ with the (more general) square norm

$$J_\lambda(f) = \int_0^1 \lambda(u) f^{(m)}(u)^2\, du$$

is

$$R_\lambda(s, t) = \int_0^1 G(t, u)\, G(s, u) \frac{du}{\lambda(u)}.$$

Parameterizing $1/\lambda(u)$ by

$$\frac{1}{\lambda(u)} = \sum_{k=1}^{K} \frac{1}{\lambda_k} B_k(u)$$

where the B_k are an appropriate set of shifted hill functions (B-splines!) with the constant functions in their span, we have

$$R_\lambda(s, t) = \sum_{k=1}^{K} \frac{1}{\lambda_k} \int_0^1 G(t, u)\, G(s, u)\, B_k(u)\, du.$$

It is of course desirable to choose the B_k so that the integrals in the formula for R_λ can be obtained analytically. The formula for f_λ minimizing the above penalized least squares problem with J_λ instead of J_λ, which may penalize wiggliness differently in different regions of $[0, 1]$, is obtained by substituting in R_λ for R_λ in the formula for f_λ above. With n large, K appropriately much less than n and the B_k chosen well it should be possible to choose the λ_k by generalized cross-validation or (σ known) by unbiased risk estimation, thus allowing more spatial adaptivity than is obtained with constant λ.

Gilbert G. Walter (University of Wisconsin, Milwaukee): The authors are to be congratulated

(a) for making sense of the confusing array of minimax estimators and
(b) for showing that wavelet estimators are almost as good as the best whatever the class of functions.

Their method is easy to understand and is computationally efficient. It could become a future standard for nonparametric regression function estimation.

Discrete wavelet estimators in general and their estimator in particular have the form

$$\begin{aligned} \hat{f}(t) &= \sum_k \beta_{mk} \varphi_{mk}(t) + \sum_{m \neq j \leq M} \sum_k \alpha_{jk} \psi_{jk}(t) \\ &= \hat{f}_m(t) + \hat{r}_m(t) \end{aligned}$$

where φ_{mk} and ψ_{jk} are respectively the 'scaling functions' and 'wavelets'. The first term $\hat{f}_m(t)$ corresponds to a low pass filter and gives a blurred version of the regression function $f(t)$ whereas the second term $\hat{r}_m(t)$ corresponds to a band pass filter. Any irregularities of f will appear in $\hat{r}_m$ and not in $\hat{f}_m$. Thus the trick is to include these irregularities while excluding the noise. This is done by thresholding: throwing out the small coefficient α_{jk} while shrinking the rest.

The smooth portion $\hat{f}_m$ is given by a projection operator and is a type of kernel estimator. It is recovered automatically if $f(t)$ is smooth since in that case all the coefficients α_{jk} should fall below the threshold. Thus their method includes some kernel estimators.

It also includes interpolation-type estimators since it has been shown by Walter (1992) that wavelet subspaces often contain sampling functions. This requires only that the coefficient β_{mk} be modified.

Much fine tuning remains to be done particularly for small samples. Our own experience (Wu, 1994) seems to show that in this case both hard and soft thresholding sometimes lead to undesirable oscillations. This is not too surprising since we are trying to estimate $f(t)$ but begin with the empirical functional

$$f^*(t) = \sum_{i=1}^{n} y_i \delta(t - t_i),$$

which is wildly oscillating.

Other problems that could be the subject of further study include the existence of Gibbs's phenomenon (Kelly, 1992) and the failure of translation invariance (Walter, 1994). The former causes overshoot at jump discontinuities but may be eliminated by a proper choice of β_{mk}. The latter is potentially more troublesome since a slight shift in t may change the magnitude of the coefficients considerably.

The **authors** replied later, in writing, as follows.

We are very grateful to all contributors for the stimulating comments and questions that they have raised concerning the role of wavelet bases and thresholding in (minimax) nonparametric estimation. We shall not be able to resolve all points in a brief rejoinder—indeed the discussion may be seen as a collective research agenda for the future, along with fair notice that many of the expert contributors are engaged on the problems already. However, we enthusiastically attest to Brillinger's remark that there is much fun to be had in pursuing these ideas, that many questions are still in their infancy and that there are many opportunities for researchers to pursue.

A defining feature of the Royal Statistical Society's discussion papers is that one can indulge in enthusiastic advocacy in the read paper to kindle debate, knowing full well that all possible qualifications will emerge in the discussion. Thus, we can readily accept, along with many discussants, that the wavelet shrinkage method does not offer all that everyone might desire of a technique, but fortunately we did not claim that! And, of course, any specific *instance* of a wavelet shrinkage estimator will not satisfy us in all cases either. However, we do hold that the elements of the wavelet paradigm, namely sparsity, thresholding, fast algorithms, theoretical tractability etc., can be combined to achieve many of the desiderata that one might pose for a nonparametric technique.

Evolution

Thus the algorithm (a)–(c) of Section 3.1 is (of course!) not intended as the last word on wavelets in the regression model (2). Rather, it might be thought of as 'first-generation' wavelet technology, whereas, for example, iterative locally adaptive bandwidth kernel regression methods represent a rather mature 'nth- (10th?) generation' evolution of kernel methods.

Thus we would hope for further development of wavelet shrinkage methods to address many of the challenges raised in the discussion. As one example of a 'second-generation' wavelet shrinkage proposal, we cite the 'translation invariant' denoising illustrated in Fig. 23. Here, the lack of translation invariance of VisuShrink, noted by Walter and others, is remedied by computing VisuShrink on all n translations of the original data, back-translating and averaging. A fast $O(n \log n)$ algorithm exists, and the visual appearance, on Bumps say, is not so much worse than the special purpose technique used by Speckman, or the refined kernel method shown by Gasser. This development arose following ideas of and discussions with Coifman, Nason and Silverman (see also Silverman's contribution).

Criticisms

Many discussants (Marron, Fan, Herrman and Müller) fault wavelet shrinkage methods for not being developed to handle non-uniform designs for predictors and heteroscedasticity for responses. Although these criticisms have merit, they are based on an implicit conception of the type of problems 'traditionally treated by smoothers' (Hall and Patil). Indeed, we would not at present particularly recommend wavelet methods for scattered observational data with small to moderate sample sizes. There are certainly other conceptions, as is suggested by Moulin, who notes that the wavelet transform and thresholding have been widely applied in practice by engineers and signal processors in recent years. Our development is guided partly by the continuing explosion in instrumentally acquired data, as signals and images, directly or indirectly obtained, in which uniform designs, if not homoscedastic additive errors, are the norm.

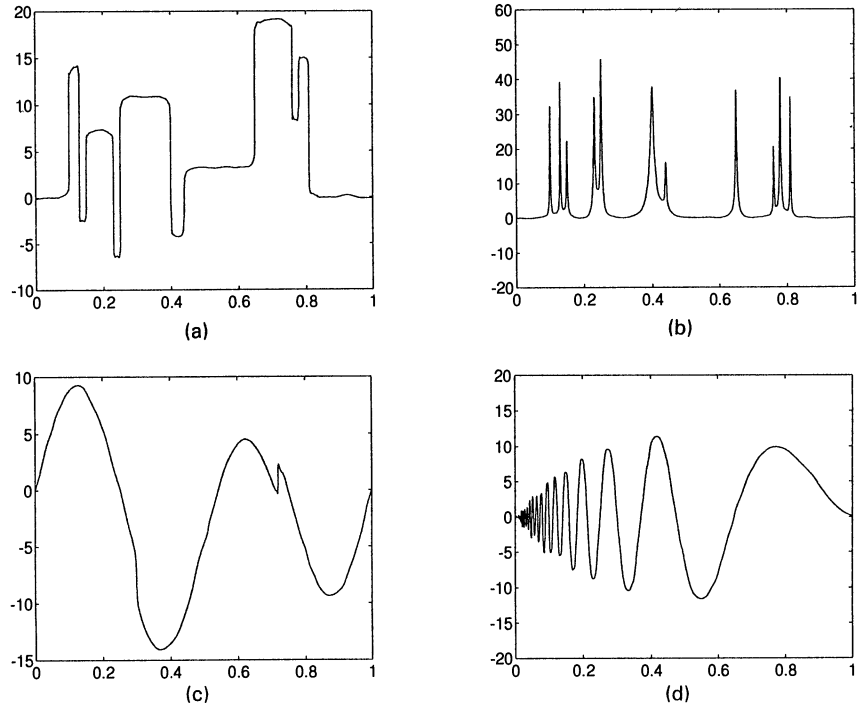

Fig. 23. Translation invariant version of VisuShrink using the Haar wavelet and soft thresholding at $\sqrt{(2 \log n)}$: (a) Blocks; (b) Bumps; (c) Heavisine; (d) Doppler

In this context, it may be useful to contrast the term 'denoising' with 'smoothing'. The word smoothing is usually associated with procedures that downweight high frequencies, on the assumption that they contain more noise than the (smooth) signal. Denoising is a more general notion, that refers to an attempt to remove noise from a signal which may itself be very far from smooth, containing perhaps singularities or transitory high frequency oscillations (such as in music or speech). As is argued in the text, a thresholding approach to denoising is most successful with signals that (in an appropriately chosen basis) are sparsely represented by a relatively small number of coefficients.

We believe that there is a rough equivalence, at least in the white noise model, between many of the popular univariate curve fitting methods, i.e. each of kernel methods with appropriate locally adaptive bandwidth, splines with adaptive knot placement, local polynomial fits and wavelet shrinkage form sufficiently rich classes of methods that they can each be tuned to achieve similar phenomena. Thus, Lepskii, Mammen and Spokoiny describe the construction of a locally adaptive bandwidth kernel estimate that adapts to spatially inhomogeneous smoothness of the kind captured by Besov function classes. Fan and Gijbels (in a tradition that goes back at least to Friedman and Stuetzle (1981), Section 3, and Friedman (1984)) study a data-driven variable bandwith for local polynomial fitting. In this sense, we agree with Crellin and Martin, Gasser, Tsybakov and others that wavelet methods will augment rather than supplant the classical toolkits in the practice of univariate fitting.

As mentioned earlier, any specific instance of a wavelet (or any other) method cannot be satisfactory on all counts. Hall and Patil are right that VisuShrink typically incurs bias rather than variance—this is the price of tuning it to be 'as smooth as the truth' with high probability. To address some of the challenges made by Hall and Patil, it is useful to look at a variant with level-dependent thresholds chosen from the data, such as SUREShrink discussed in Section 5.4 and in Donoho and Johnstone (1995). SUREShrink balances bias and variance contributions to mean-squared error (MSE) to achieve exactly the right rate of convergence over the traditional function classes, without extra logarithmic terms (this might reassure Crellin and Martin also). The figures in Donoho and Johnstone (1995) suggest that SUREShrink can handle a very wide range of variable frequency curves, including those 'traditionally treated by smoothers'. Since, by design, it has better MSE properties than VisuShrink, it may also be

more appropriate for comparison with the MSE results given by Georgiev and Liu. Computation does not become a major issue, since the complexity is $O(n \log n)$. Kerkyacharian *et al.* (1994) further show how this adaptive rate optimality extends to density estimation and more general L_p-losses.

We agree with Hall and Patil that the computational efficiency of wavelet-like algorithms will be a contribution to statistical methods—for example, we can bootstrap $O(n)$ or $O(n \log n)$ algorithms with relative impunity. Perhaps predictably, we cannot go along with the idea that this will be their only contribution—for example wavelet bases bring considerable conceptual clarity to theoretical analysis: they have certainly helped to articulate the key role of sparsity in compression, denoising and estimation.

Exploratory analysis versus modelling

Comments of Speckman, Müller, Lucier, Moulin and others prompt us to agree that there is and should be a continuum of methodologies, ranging from fully nonparametric towards the fully parametric, corresponding to increasingly definite modelling assumptions. After extension to more general distributional structures (see below), wavelet methods fall towards the fully nonparametric end of this spectrum.

Thus, to Good, we would reply that wavelet procedures can certainly be used in the spirit of exploratory data analysis for revealing structure in data. Indeed, the result of an exploratory wavelet-based analysis might be to fit a more specific model such as that described by Speckman. Similarly, Wang, Müller, Walter and others are correct that, although wavelet methods react well to unsuspected singularities, if the singularity structure is known reasonably well, special methods can be designed to do even better. Thus, the class $\mathscr{F}_n$ described by Müller will often be appropriate, but not, for example, for a variable frequency signal like Doppler. Along these lines, we note that we could develop a notion of segmented multiresolution analysis that estimates the location of singularities and performs multiresolution analysis separately on each side, with attendant removal or reduction of Gibbs effects (Donoho, 1994).

The specific distributional assumptions that we imposed were used to aid in the choice of specific thresholds—the utility of transform methods is much broader. In less-heavy-tailed cases of independent and identically distributed (IID) non-Gaussian noise, central limit effects will ensure that coefficients at all but the very top levels will be near Gaussian. Variance estimates that are quite generally consistent can be constructed from wavelet coefficients at the top level, since they are generalized differences. For heavy-tailed error distributions, the forthcoming wavelet toolkit by Andrew Bruce and Hong-Ye Gao from S-PLUS will include resistant median-based multiresolution methods. Log-spectral density estimation requires a level-dependent thresholding strategy, as shown in work of Gao and Moulin.

Questions

We were delighted to note that some discussants provided at least the beginnings of answers to questions raised by other contributors. Thus, Fan and others ask about extensions to likelihood-based models, and Brillinger proposes the use of iteratively reweighted least squares in generalized linear models with linear predictors expressed in wavelet bases. Herrmann asks about robustness to various aspects of model (2) including heteroscedastic non-Gaussian errors, and Neumann and Nussbaum describe work on conditions under which results in the limiting Gaussian situation can be carried over to a variety of observational models. Fan's discussion of goodness-of-fit tests may address Doukhan's query on the topic. Engel describes Haar wavelet estimators allowing heteroscedasticity and random design data. Lepskii, Mammen and Spokoiny ask about the advantages of localization in time and frequency, and von Sachs describes settings with clear frequency structure that changes over time.

Concerning confidence intervals for the wavelet fits, we are happy to note that Brillinger describes work for the stationary error case. In the simpler setting of IID errors, we need not rely explicitly on asymptotics—it would in principle be straightforward (and relatively fast!) to bootstrap fitted residuals. It would certainly be interesting to see how well such intervals reflected the bias inherent in the use of high thresholds such as $\sqrt{(2 \log n)}$. In general, we certainly hope for more work on asymptotic distribution of wavelet-based estimators.

Marron asks how far away asymptopia is, and whether exact risk calculations might be possible. Indeed, they are possible, for example in model (2), by using the formulae for MSE of co-ordinatewise threshold rules developed in Donoho and Johnstone (1994a), and Marron and Johnstone have been following this up. These exact risk calculations will also offer a complement to the interesting simulations reported by Efromovich.

On the specific issue of better dealing with local dependence across scales, we agree with Lucier and Moulin that there is much to be done. We note that for certain combinations of parameters in the critical case $((\sigma+\frac{1}{2})p=(\sigma'+\frac{1}{2})p')$ the optimal rates for Triebel spaces differ from those from Besov spaces

precisely because there is dependence across levels in the least favourable distributions. We hope to include details in a paper that will contain complete proofs of the results of Section 4.

Concerning Good's query on the scattering data, we certainly only meant the term 'close' in an informal, visual sense; however, we did find 13 sign changes in the second differences of our fit that persisted for four or more bins—this is close to, but not identical with, the manner in which Good and Gaskins (1980) counted bumps.

Comments

We certainly echo Birgé's sentiments that the issue of adaptation is basic, and we are following his work with Massart with great interest. The short note Donoho and Johnstone (1994b) looks at adaptation to choice of basis from a huge library of such bases built from a relatively constrained number of individual possible basis vectors.

The results of Neumann and Nussbaum on approximation by the Gaussian (white) noise model are of considerable importance. Nussbaum is right to emphasize that more thought is needed about exactly what kind of approximation is needed and under what conditions. For example, in the case of spectral density estimation for stationary series, Neumann (1994) has Gaussian approximation results, but they do not quite seem to capture the highly level-dependent threshold choices used by Gao (1993b).

We are greatly indebted to McCullagh for noting the possible relevance of Yates's algorithm (see, for example, Yates (1937), pages 15 and 29) for computing factorial effect totals. It turns out that the formal manipulations of Yates's method, as described, for example, in Cochran and Cox (1957) (p. 158 ff.), are *exactly* the same as those used to compute the wavelet packet table (Coifman *et al.*, 1989; Wickerhauser, 1994) in the particular case of the Haar filter. Although the motivation, setting and products of the wavelet and cosine packet methods are quite different, we are amused and stimulated by the universality of clever and powerful computational ideas.

Olshen takes us to task for describing the spatial adaptivity camp as atheoretical. The work by Olshen, Andrew, Lou and the co-authors of CART concerns asymptotic *properties* (end cut preferences, geometry of terminal nodes, Vapnik–Chervonenkis classes of regions) of tree-structured methods that shed light on how it works in practice. That valuable work is differently focused from theoretical *comparisons* with other methods, minimax bench-marks etc., which is the narrow interpretation of what we wrote.

Wahba and Lucier focus on regularization methods: Wahba proposes to achieve some spatial adaptivity within the framework of computationally familiar quadratic problems by replacing the regularization parameter by a spatially varying regularization function. Lucier notes that one can use non-Hilbertian regularization penalties, and that these can lead quite naturally to thresholding methods with attendant spatial adaptivity properties. With L_1-type regularization penalties, we are still within the domain of convex optimization, and so numerical algorithms are now practical even for sample sizes of the order of thousands (see for example Chen and Donoho (1995)).

We thank Olshen for drawing attention to Heaton's interesting tree-structured approach to positron emission tomography, which had not been developed far at the time that the paper was written. Although Heaton's work does now show that tree-structured methods can be adapted for certain indirect data situations, the computational burden of current algorithms is very high.

Practicalities

Engel, Nason, Tsybakov and others note that the (first-generation) WaveShrink depends on several parameters such as the base level j_0 and choice of wavelet filter. Although the qualitative and asymptotic role of these parameters has become quite clear, it is true that on specific data sets they can make a sufficiently large difference that in present practice we would often vary them subjectively depending on the circumstances. For example the fact that high MSE (almost all variance) is observed on the quadratic signal by Hastie and Tibshirani is probably due to the fact that $j_0 = 6$ is too high for soft thresholding on a very smooth signal: many coefficients at levels 3, 4 and 5, say, are left unchanged and are contributing variance but no signal. Obviously some work on automatic choice would be desirable, and this work will be simplified by the fact that there are relatively few sensible choices for j_0, and that the popular filter families depend on an integer parameter. We might also note that the choice of j_0 (in conjunction with a $\sqrt{(2 \log n)}$ threshold) is less critical for hard thresholding than for soft thresholding, since a large signal at low frequencies will be passed through undiminished even if j_0 is very small. For a choice of wavelet basis, we refer also to some preliminary work of Tribouley (1995).

Hastie and Tibshirani also comment on the presence of a high frequency wiggle in about a quarter of the simulations in Fig. 19. This is roughly of the order of magnitude that would be predicted from

the bound (in theorem 1) on the probability that all $n = 1024$ observations on the underlying Gaussian error distribution remain below $\sqrt{(2 \log n)}$ standard deviations. If this is thought objectionable from a visual point of view, then there is little change to the current asymptotic theory and qualitative conclusions in replacing the threshold of $\sqrt{(2 \log n)}$ by, say, $\sqrt{(3 \log n)}$, in which case such wiggles will be much rarer. This counterpoint of robustness of asymptotic conclusions and finite sample sensitivity to thresholds admittedly sounds a dissonant note in an otherwise harmonious asymptopia: some finer analysis at the level of constants may help in resolving the discord.

Historical points

We wish to emphasize the remark made in the discussion that Doukhan (1988) represents, to our knowledge also, the first use of wavelet bases in theoretical statistics (in the context of *linear* estimation methods).

In asserting that the minimax paradigm has had relatively little effect on software, we should have explicitly excepted the pioneering work of Stone (1977, 1980, 1982, 1985) on minimax rates of convergence in general and additive nonparametric models. Stone has of course also been involved in the development of commercial and public domain software ranging from CART (Breiman *et al.*, 1984) to adaptive spline methods for a variety of function estimation problems (e.g. Kooperberg *et al.* (1995a, b)), and the link between minimax theory and algorithms is strong here.

Foundations

Some discussants (Speckman and Silverman) fret about the philosophy of the minimax approach, whereas others surprise us as welcome new converts to the faith (Crellin and Martin). We propose the simultaneous near minimaxity phenomenon studied here as a constructive response to the traditional charge that the minimax method can be derailed by defending against irrelevant worst cases. Here we start with a single estimator and consider how much trouble it can encounter, relative to the minimax method, over a variety of parameter spaces of varying smoothness, shape and size. The corresponding worst cases for these spaces are *very* different (Johnstone, 1994). If we can conclude that the estimator just passes all these hurdles, then we may be somewhat reassured that it will survive at least some other situations. Consider a slightly flippant analogy with market research: the product is test marketed in a range of market sectors before being set loose on the general public.

Conclusion

Taken together, the discussants cover much ground, from tomography (Kolaczyk) to time series (von Sachs and others), from engineering (Moulin) to biomedical data (Silverman) and, of course, from theory to methodology. The wide interest in wavelet and related methods in many disciplines clearly presents exciting opportunities for an expanded trade in ideas and tools into and out of statistics, and we are happy to conclude by reiterating our thanks to all contributors, and to the journal and the Royal Statistical Society for hosting the forum.

REFERENCES IN THE DISCUSSION

Ango-Nze, P. and Doukhan, P. (1993) Estimation fonctionnelle de séries temporelles mélangeantes. *Compt. Rend. Acad. Sci. Paris* A, **317**, 405–408.

Antoniadis, A. (1994a) Smoothing noisy data with coiflets. *Statist. Sin.*, **4**, 651–678.

——— (1994b) Smoothing noisy data with tapered coiflets series. *Technical Report RR 993-M*. University of Grenoble, Grenoble.

Antoniadis, A., Grégoire, G. and McKeague, I. (1995) Wavelet methods for curve estimation. *J. Am. Statist. Ass.*, to be published.

Basseville, M., Benveniste, A., Chou, K. C., Golden, S. A., Nikoukhah, R. and Willsky, A. S. (1992) Modeling and estimation of multiresolution stochastic processes. *IEEE Trans. Inform. Theory*, **38**, 766–784.

Blow, D. M. and Crick, F. H. C. (1959) The treatment of errors in the isomorphous replacement method. *Acta Crystallogr.*, **12**, 794–802.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. New York: Chapman and Hall.

Brillinger, D. R. (1994) Uses of cumulants in wavelet analysis. *Proc. SPIE Adv. Signal Process.*, **2296**, 2–18.

Brockmann, M., Gasser, T. and Herrmann, E. (1993) Locally adaptive bandwidth choice for kernel regression estimators. *J. Am. Statist. Ass.*, **88**, 1302–1309.

Brown, L. D. and Low, M. (1992) Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, to be published.

Burman, P. (1989) A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, **76**, 503–514.
Chen, S. and Donoho, D. L. (1995) Basis pursuit. In *Proc. 28th Asilomar Conf. Signals, Systems and Computers*. Washington DC: Institute of Electrical and Electronics Engineers Computer Society. To be published.
Cochran, W. and Cox, G. (1957) *Experimental Designs*, 2nd edn. New York: Wiley.
Coifmann, R. R., Meyer, Y., Quake, S. and Wickerhauser, M. V. (1989) Signal processing and compression with wave packets. In *Proc. Int. Conf. Wavelets, Marseille* (ed. Y. Meyer). Paris: Masson.
Cosman, P. C., Gray, R. M. and Olshen, R. A. (1994) Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy. *Proc. IEEE*, **82**, 919–932.
Dahlhaus, R. (1993) Fitting time series models to nonstationary processes. Submitted to *Ann. Statist.*
Daubechies, I. (1993) Orthonormal bases of compactly supported wavelets: II, Variations on a theme. *SIAM J. Math. Anal.*, **24**, 499–519.
DeVore, R. and Lucier, B. (1992a) Smoothness spaces and wavelet decomposition. *SPIE J.*, 1830.
———(1992b) Fast wavelet techniques for near-optimal image processing. In *Proc. IEEE Military Communications Conf.* New York: Institute of Electrical and Electronics Engineers Communications Society.
Donoho, D. L. (1993) Nonlinear wavelet methods for recover of signals, densities, and spectra from indirect and noisy data. In *Different Perspectives on Wavelet* (ed. I. Daubechies), pp. 173–205. Providence: American Mathematical Society.
———(1994) On minimum entropy segmentation. In *Wavelets: Theory, Algorithms, Applications* (eds C. K. Chui, L. Montefusco and L. Puccio), pp. 233–269. New York: Academic Press.
———(1995) Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harm. Anal.*, to be published.
Donoho, D. L. and Johnstone, I. M. (1994a) Ideal spatial adaptation via wavelet shrinkge. *Biometrika*, to be published.
———(1994b) Ideal denoising in an orthonormal basis chosen from a library of bases. *Compt. Rend. Acad. Sci. Paris* A, **319**, in the press.
———(1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, to be published.
Donoho, D. L., Johnstone, I. M., Hoch, C. J. and Stern, A. S. (1992) Maximum entropy and the nearly black object (with discussion). *J. R. Statist. Soc.* B, **54**, 41–81.
Doukhan, P. (1988) Formes de Toëplitz associées à une analyse multiéchelle. *Compt. Rend. Acad. Sci. Paris* A, **306**, 663–666.
Doukhan, P. and Gamboa, F. (1994) Prohorov rates in superresolution. *Preprint 94-106*. Université de Paris-Sud, Orsay.
Doukhan, P. and Leon, J. (1990) Déviation quadratique d'estimateurs d'une densité par projection orthogonale. *Compt. Rend. Acad. Sci. Paris* A, **310**, 425–430.
Efromovich, S. (1985) Nonparametric estimation of a density with unknown smoothness. *Theory Probab. Applic.*, **30**, 557–568.
———(1994) On adaptive nonnegative orthogonal series density estimator for small samples. *Technical Report*. University of New Mexico, Albuquerque.
Efromovich, S. and Pinsker, M. (1981) Estimation of square-integrable density on the basis of a sequence of observations. *Prob. Inform. Transmssn*, **17**, 182–195.
———(1982) Estimation of square-integrable probability density of a random variable. *Prob. Inform. Transmssn*, **18**, 175–189.
Engel, J. (1993) Tree structured histogram estimation based on a simple class of wavelets. *Discussion Paper A-415*. University of Bonn, Bonn.
———(1994) A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multiv. Anal.*, **49**, 242–254.
Eubank, R. L. and Speckman, P. (1993) Nonparametric estimation of functions with jump discontinuities. *IMS Lect. Notes Monogr. Ser.*
Fan, J. (1995) Test of significance based on wavelet thresholding and Neyman's truncation. *J. Am. Statist. Ass.*, to be published.
Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1993) Local polynomial fitting: a standard for nonparametric regression. *Discussion Paper 9315*. Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve.
Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Statist. Soc.* B, **57**, 371–394.
Frazier, M., Jawerth, B. and Weiss, G. (1991) *Littlewood-Paley Theory and the Study of Function Spaces*. Providence: American Mathematical Society.
Friedman, J. (1984) A variable span smoother. *Technical Report 5*. Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford.
Friedman, J. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.
Froment, J. and Mallat, S. (1992) Second generation compact image coding with wavelets. In *Wavelets and Their Application* (eds Ruskai *et al.*), pp. 655–677. Cambridge: Jones and Bartlett.
Gao, H.-Y. (1993a) Wavelet estimation of spectral densities in time series analysis. *PhD Dissertation*. University of California, Berkeley.

———(1993b) Choice of thresholds for wavelet estimation of the log spectrum. *Preprint*. Department of Statistics, University of California, Berkeley.
Garsia, A. M. (1970) *Topics in Almost Everywhere Convergence*. Chicago: Markham.
Georgiev, A. A. (1984) Nonparametric system identification by kernel methods. *IEEE Trans. Autom. Control*, **29**, 356–358.
———(1985) Local properties of function fitting estimates with application to system identification. In *Mathematical Statistics and Applications: Proc. 4th Pannonian Symp. Mathematical Statistics* (eds W. Grossmann *et al.*), pp. 141–151. Dordrecht: Reidel.
———(1988) Consistent nonparametric multiple regression: the fixed design case. *J. Multiv. Anal.*, **25**, 100–110.
Goldman, L., Cook, E. F., Brand, D. A., Lee, T. H., Rouan, G. W., Weisberg, M. C., Acampora, D., Stasiulewicz, C., Walshon, J., Terranova, G., Gottlieb, L., Kobernick, M., Goldstein-Wayne, B., Copen, D., Daley, K., Brandt, A. A., Jones, D., Mellors, J. and Jakubowski, R. (1988) A computer protocol to predict myocardial infarction in emergency department patients with chest pain. *New Engl. J. Med.*, **318**, 797–803.
Good, I. and Gaskins, R. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data (with discussion). *J. Am. Statist. Ass.*, **75**, 42–73.
Gordon, L. and Olshen, R. A. (1984) Almost surely consistent nonparametric regression from recursive partitioning schemes. *J. Multiv. Anal.*, **15**, 147–163.
Heaton, A. A. (1994) Tree-structured methods for PET reconstruction. *PhD Dissertation*. Department of Electrical Engineering, Stanford University, Stanford.
Johnstone, I. (1994) Minimax bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics, V* (eds S. Gupta and J. Berger), pp. 303–326. New York: Springer.
Johnstone, I. M. and Silverman, B. W. (1994) Wavelet methods for data with correlated noise. To be published.
Kelly, S. (1992) Pointwise convergence for wavelet expansions. *PhD Dissertation*. Washington University, St Louis.
Kerkyacharian, G., Picard, D. and Tribouley, K. (1994) l_p adaptive density estimation. *Technical Report Mathématiques, URA 1321*. Université de Paris VII, Paris.
Kolaczyk, E. D. (1994) Wavelet methods for the inversion of certain homogeneous linear operators in the presence of noisy data. *PhD Dissertation*. Department of Statistics, Stanford University, Stanford.
Koltchinskii, V. I. (1994) Komlos–Major–Tusnady approximation for the general empirical process and Haar expansions of classes of functions. *J. Theor. Probab.*, **9**, 73–118.
Kooperberg, C., Stone, C. and Truong, Y. (1995a) Hazard regression. *J. Am. Statist. Ass.*, **90**, in the press.
———(1995b) The l_2 rate of convergence for hazard regression. *Scand. J. Statist.*, to be published.
Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*. Berlin: Springer.
Lepskii, O. V. (1990) One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Applic.*, **35**, 459–470.
Lepskii, O. V., Mammen, E. and Spokoiny, V. G. (1994) Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Preprint*.
Liu, H. (1992) A nonparameteric autoregressive-regression model for robust data smoothing and its application to image restorations. *PhD Dissertation*. Medical University of South Carolina, Charleston.
Liu, H. and Georgiev, A. A. (1993) A nonparametric autoregressive-regression procedure for edge preserving smoothing in two-dimensional signal processing. To be published.
———(1994) A nonparametric autoregressive-regression procedure for edge-preserved smoothing: one-dimensional case. To be published.
Mallat, S. and Hwang, W. L. (1992) Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory*, **38**, 617–643.
Marron, J. S. and Wand, M. P. (1992) Exact mean integrated squared error. *Ann. Statist.*, **20**, 712–736.
Moulin, P. (1993) A wavelet regularization method for diffuse radar-target imaging and speckle-noise reduction. *J. Math. Imagng Vis.*, **3**, 123–134.
———(1994) Wavelet thresholding techniques for power spectrum estimation. *IEEE Trans. Signal Process.*, **42**, 3126–3136.
Müller, H. G. (1992) Change-points in nonparametric regression analysis. *Ann. Statist.*, **20**, 737–761.
Müller, H. G. and Stadtmüller, U. (1987) Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610–625.
———(1994) Detecting and estimating jumps in means with simultaneous scale estimation. To be published.
Nason, G. (1994a) Wavelet regression by cross-validation. *Technical Report 447*. Department of Statistics, Stanford University, Stanford.
———(1994b) Wavelet function estimation using cross-validation. To be published.
Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in S. *J. Comput. Graph. Statist.*, **3**, 163–191.
Neumann, M. H. (1994) Spectral density estimation via nonlinear wavelet methods for stationary non-Gaussian time series. *Preprint 99*. Institute for Applied Analysis and Stochastics, Berlin.
Neumann, M. H. and Spokoiny, V. G. (1994) On the efficiency of wavelet estimators under arbitrary error distributions. *Discussion Paper 4*. Humboldt University, Berlin.
Nobel, A. B. (1993) Histogram regression estimation using data-dependent partitions. Submitted to *Ann. Statist.*

Nobel, A. B. and Olshen, R. A. (1994) Termination and continuity of greedy growing for tree structured vector quantizers. Submitted to *IEEE Trans. Inform. Theory*.
Nussbaum, M. (1985) Spline smoothing and asymptotic efficiency in l_2. *Ann. Statist.*, **13**, 984–997.
———(1992) Asymptotic equivalence of density estimation and white noise. *Ann. Statist.*, to be published.
Olshen, R. A. (1994) Binary trees for classification, regression, and clustering. *IEEE Inform. Theory Soc. Newslett.*, **44**, June, 8–10.
Pinsker, M. (1980) Optimal filtering of square integrable signals in gaussian white noise. *Prob. Inform. Transmssn*, **16**, 120–133.
von Sachs, R. and Schneider, K. (1994) Wavelet smoothing of evolutionary spectra by non-linear thresholding. *Preprint*. Universität Kaiserslautern, Kaiserslautern.
Shensa, M. J. (1992) The discrete wavelet transform: wedding the *a trous* and Mallat algorithms. *IEEE Trans. Signal Process.*, **40**, 2464–2482.
Soulier, P. (1991) Déviation quadratique pour des estimateurs de la variance d'une diffusion. *Compt. Rend. Acad. Sci. Paris* A, **313**, 783–786.
Speckman, P. (1979) Minimax estimates of linear functionals in a Hilbert space. Unpublished.
———(1985) Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, **13**, 970–983.
———(1988) Kernel smoothing in partial linear models. *J. R. Statist. Soc.* B, **50**, 413–436.
———(1993) Detection of change-points in nonparametric regression. Unpublished.
———(1994) Fitting curves with features: semiparametric change-point methods. In *Proc. Int. Symp. Interface*. To be published.
Stone, C. (1977) Consistent nonparametric regression (with discussion). *Ann. Statist.*, **5**, 595–645.
———(1980) Optimal rates of convergence of nonparametric estimators. *Ann. Statist.*, **8**, 1348–1360.
———(1982) Optimal global rates of convergence for nonparametric estimators. *Ann. Statist.*, **10**, 1040–1053.
———(1985) Additive regression and other nonparametric models. *Ann. Statist.*, **13**, 689–705.
Tribouley, K. (1995) Practical estimation of multivariate densities using wavelet methods. *Statist. Neerland.*, to be published.
Wahba, G. (1984) Partial spline models for the semiparametric estimation of functions of several variables. In *Analyses for Time Series: Proc. Japan–US Joint Seminar*, pp. 319–329. Tokyo: Institute of Statistical Mathematics.
———(1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
Walter, G. (1992) A sampling theorem for wavelet subspaces. *IEEE Trans. Inform. Theory*, **38**, 881–884.
———(1994) *Wavelets and Other Orthogonal Systems with Applications*. Boca Raton: Chemical Rubber Company.
Wang, Y. (1994a) Function estimation via wavelets for data with long-range dependence. *Technical Report*. Department of Statistics, University of Missouri, Columbia.
———(1994b) Jump and sharp cusp detection by wavelets. Submitted to *Biometrika*.
———(1994c) Estimation of functions with jump via wavelets. *Technical Report*. Department of Statistics, University of Missouri, Columbia.
Wickerhauser, M. (1994) *Adapted Wavelet Analysis from Theory to Software*.
Wornell, G. W. and Oppenheim, A. V. (1992) Estimation of fractal signals from noisy measurements using wavelets. *IEEE Trans. Signal Process.*, **40**, 611–623.
Wu, D. (1994) Probability density estimation with wavelets. *PhD Dissertation*. University of Wisconsin, Milwaukee.
Wu, J. S. and Chu, C. K. (1993) Kernel type estimators of jump points and values of a regression function. *Ann. Statist.*, **21**, 1545–1566.
Yao, Y. C. (1988) Estimating the number of change-points via Schwarz criterion. *Statist. Probab. Lett.*, **6**, 181–189.
Yates, F. (1937) The design and analysis of factorial experiments. *Technical Communication 35*. Imperial Bureau of Soil Science, Harpenden.