THE ROYAL SOCIETY

Wavelets and the theory of non-parametric function estimation

By Iain M. Johnstone

Department of Statistics, Sequoia Hall, Stanford University,
Stanford, CA 94305, USA

Non-parametric function estimation aims to estimate or recover or denoise a function of interest, perhaps a signal, spectrum or image, that is observed in noise and possibly indirectly after some transformation, as in deconvolution. 'Non-parametric' signifies that no *a priori* limit is placed on the number of unknown parameters used to model the signal. Such theories of estimation are necessarily quite different from traditional statistical models with a small number of parameters specified in advance.

Before wavelets, the theory was dominated by linear estimators, and the exploitation of assumed smoothness in the unknown function to describe optimal methods. Wavelets provide a set of tools that make it natural to assert, in plausible theoretical models, that the *sparsity* of representation is a more basic notion than smoothness, and that nonlinear thresholding can be a powerful competitor to traditional linear methods. We survey some of this story, showing how sparsity emerges from an optimality analysis via the game-theoretic notion of a least-favourable distribution.

Keywords: minimax; Pinsker's theorem; sparsity; statistical decision problem; thresholding; unconditional basis

1. Introduction

Within statistics, the first applications of wavelets were to theory. While the potential for statistical application to a variety of problems was also apparent, and is now being realized (as surveyed in other articles in this issue), it was in the theory of non-parametric function estimation that progress was initially fastest. The goal of this article is to survey some of this work retrospectively.

Why should developments in the theory of statistics interest a wider scientific community? Primarily, perhaps, because theory attempts to isolate concepts of broad generality that clarify in what circumstances and under what assumptions particular data analytic methods can be expected to perform well, or not. As a classical example, the most widely used statistical tools—regression, hypothesis tests, confidence intervals—are typically associated with *parametric* models, that is, probability models for observed data that depend on, at most, a (small) finite number of unknown parameters. The true scope, versatility and applicability of these tools was clarified by the development of underlying theoretical notions such as likelihood, sufficiency, unbiasedness, Cramér–Rao bounds, power, and so forth. Many of these concepts have passed into the general toolkit of scientific data analysis.

Phil. Trans. R. Soc. Lond. A (1999) **357**, 2475–2493

A further key point is that theory promotes *portability* of methods between scientific domains: thus, Fisher's analysis of variance was initially developed for agricultural field trials, but he also created theoretical support with such effect that most uses of analysis of variance now have nothing to do with agriculture.

What is meant by the term non-parametric function estimation? The advent of larger, and often instrumentally acquired, datasets and a desire for more flexible models has stimulated the study of *non-parametric* models, in which there is no *a priori* bound on the number of parameters used to describe the observed data. For example, when fitting a curve to time-varying data, instead of an *a priori* restriction to, say, a cubic polynomial, one might allow polynomials of arbitrarily high degree or, more stably, a linear combination of splines of local support.

Despite prolific theoretical study of such infinite-dimensional models in recent decades, the conclusions have not dispersed as widely as those for the parametric theory. While this is partly because non-parametric theory is more recent, it is certainly also partly due to the greater nuance and complexity of its results, and a relative paucity of unifying principles.

The arrival of wavelet bases has improved the situation. Wavelets and related notions have highlighted *sparsity* of representation as an important principle in estimation and testing. Through the *dyadic Gaussian sequence model*, they have bridged parametric and non-parametric statistics and re-invigorated the study of estimation for multivariate Gaussian distributions of finite (but large) dimension. Wavelets have also been the vehicle for an influx of ideas—unconditional bases, fast algorithms, new function spaces—from computational harmonic analysis into statistics, a trend that seems likely to continue to grow in future.

2. A simple model for sparsity

We begin with an apparently naive discussion of sparsity in a 'monoresolution' model. Suppose that we observe an n-dimensional data vector y consisting of an unknown signal θ, which we wish to estimate, contaminated by additive Gaussian white noise of scale σ_n. If the model is represented in terms of its coefficients in a particular orthonormal basis B, we obtain (y_k^B), (θ_k^B), etc., though the dependence on B will usually be suppressed. Thus, in terms of basis coefficients,

$$y_k = \theta_k + \sigma_n z_k, \quad k = 1, \ldots, n, \tag{2.1}$$

and $\{z_k\}$ are independently and identically distributed $N(0,1)$ random variables. Here, we emphasize that $\theta = (\theta_k)$ is, in general, regarded as fixed and unknown. This model might be reasonable, for example, if we were viewing data as Fourier coefficients, and looking in a particular frequency band where the signal and noise spectrum are each about constant.

If, in addition, it is assumed that $\{\theta_k\}$ are random, being drawn from a Gaussian distribution with $\mathrm{Var}(\theta_k) = \tau_n^2$, then the optimal (Wiener) filter, or estimator, would involve *linear shrinkage* by a constant *linear* factor:

$$\hat{\theta}_k = \frac{\rho}{\rho + 1} y_k, \qquad \rho = \frac{\tau_n^2}{\sigma_n^2}. \tag{2.2}$$

The ratio τ_n^2/σ_n^2 (or some function of it) is usually called the *signal-to-noise* ratio.

The two key features of this traditional analysis are

(a) the Gaussian prior distribution leads to linear estimates as optimal; and

(b) the linear shrinkage is invariant to orthogonal changes of coordinates: thus, the same Wiener filter is optimal, regardless of the basis chosen.

Sparsity. In contrast, sparsity has everything to do with the choice of bases. Informally, 'sparsity' conveys the idea that most of the signal strength is concentrated in a few of the coefficients. Thus, a 'spike' signal $\gamma(1, 0, \ldots, 0)$ is much sparser than a 'comb' vector $\gamma(n^{-1/2}, \ldots, n^{-1/2})$, even though both have the same energy, or ℓ_2 norm: indeed these could be representations of the same vector in two different bases. In contrast, noise, almost by definition, is not sparse in any basis, and among representations of signals in various bases, it is the ones that are sparse that will be most easily 'denoised'.

Remark 2.1. Of course, in general terms, sparsity is a familiar notion in statistics and beyond: think of parsimonious model choice, 'Occam's razor', and so forth. It is the motivation for the principal components analysis of Hotelling (1933), suitable for high-dimensional, approximately Gaussian data. However, in the specific domain of non-parametric function estimation, prior to the advent of wavelets, the role of sparsity was perhaps somewhat obscured by the focus on the related, although somewhat more special, notion of smoothness.

Figure 1 shows part of a real signal represented in two different bases: figure 1*a* is a subset of 2^7 wavelet coefficients θ^{W}, while figure 1*b* shows a subset of 2^7 Fourier coefficients θ^{F}. Evidently, θ^{W} has a much sparser representation than does θ^{F}.

The sparsity of the coefficients in a given basis may be quantified using ℓ_p norms,

$$\|\theta\|_p = \left(\sum_1^n |\theta_k|^p \right)^{1/p},$$

for $p < 2$, with smaller p giving more stringent measures. Thus, while the ℓ_2 norms of our two representations are roughly equal,

$$\|\theta^{\mathrm{F}}\|_2 = 25.3 \approx 23.1 = \|\theta^{\mathrm{W}}\|_2,$$

the ℓ_1 norms differ by a factor of 6.5:

$$\|\theta^{\mathrm{F}}\|_1 = 246.5 \gg 37.9 = \|\theta^{\mathrm{W}}\|_1.$$

Figure 2 shows that the sets,

$$\left\{ \theta : \sum_1^n |\theta_k|^p \leqslant C^p \right\},$$

become progressively smaller and more clustered around the coordinate axes as p decreases. Thus, the only way for a signal in an ℓ_p ball to have large energy (i.e. ℓ_2 norm) is for it to consist of a few large components, as opposed to many small components of roughly equal magnitude. Put another way, among all signals with a given energy, the sparse ones are precisely those with small ℓ_p norm.

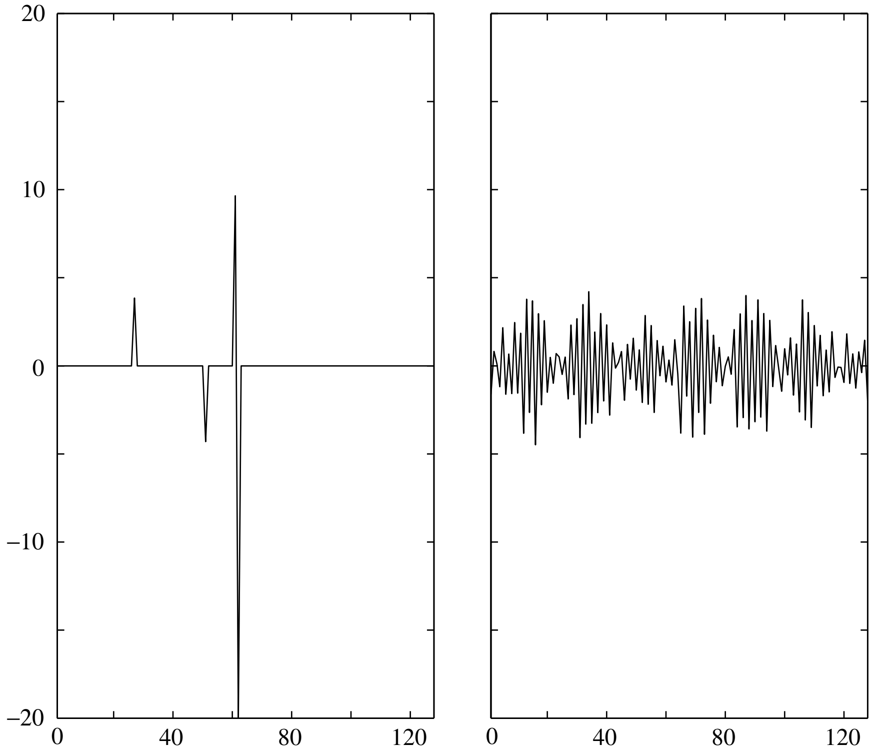

Figure 1. (a) θ_k^{W} = level 7 of estimated NMR reconstruction g of figure 4, while in (b) θ_k^{F} = Fourier coefficients of g at frequencies $65, \dots, 128$, both real and imaginary parts shown. While these do not represent exactly the same projections of f, the two overlap and $\|\theta^{\mathrm{F}}\|_2 = 25.3 \approx 23.1 = \|\theta^{\mathrm{W}}\|_2$.

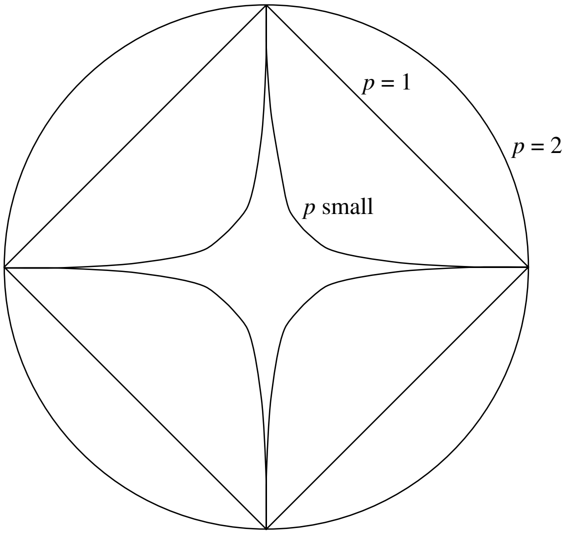

Figure 2. Contours of ℓ_p balls.

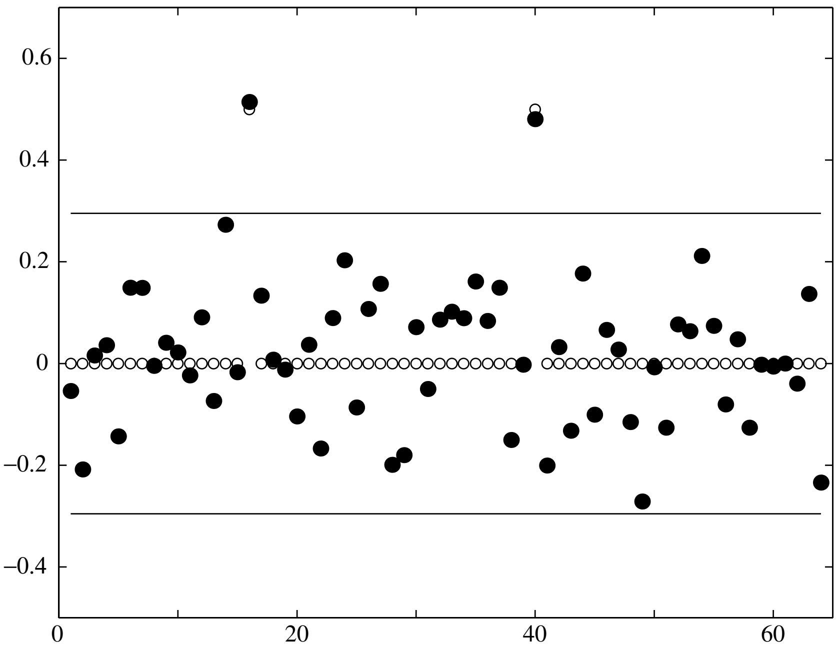

Figure 3. Visualization of model (2.1): open circles are unknown values θ_k, solid circles are observed data y_k, $k = 1, \ldots, n = 64$. Horizontal lines are thresholds at $\lambda_n = \sqrt{\log n/n}$.

Thus, we will use sets $\{\|\theta\|_p \leqslant C\}$ as models for *a priori* constraints that the signal θ has a sparse representation in the given basis. Assume, for simplicity here, that $\sigma_n = 1/\sqrt{n}$ and that $p = C = 1$: it is thus supposed that

$$\sum_1^n |\theta_k| \leqslant 1.$$

Other situations can be handled by developing the theory for general (p, C_n, σ_n) (see Donoho & Johnstone 1994*b*). How to exploit this sparsity information in order to better estimate θ: in other words, can we estimate θ^{W} better than θ^{F}?

Figure 3 shows an idealized case in which all θ_k are zero except for two spikes, each of size 1/2. Two extreme examples of linear estimators are $\hat{\theta}_1(y) \equiv y$, which leaves the data unadjusted, and $\hat{\theta}_0(y) \equiv 0$, which sets every coordinate to zero. The first, a pure 'variance' estimator, has $\mathrm{MSE} = \sigma_n^2 = 1/n$ in each of the n coordinates, for a total $\mathrm{MSE} = 1$. The second, $\hat{\theta}_0$, a pure 'bias' estimator, is exactly correct on all but the two spikes, where it suffers a total $\mathrm{MSE} = 2 \cdot (1/2)^2 = 1/2$. Given the symmetry of the prior knowledge and the statistical independence of the observations, the only other plausible choices for a linear estimator have the form cy, for a constant c, $0 \leqslant c \leqslant 1$. It can be shown that such estimators are effectively a combination of the two extremes, and, in particular, do not have noticeably better MSE performance.

In the situation of figure 3, thresholding is natural. Define the *hard threshold* estimator by its action on coordinates,

$$\hat{\theta}_{\lambda,k}(y) = \begin{cases} y_k, & \text{if } |y_k| \geqslant \lambda\sigma_n, \\ 0, & \text{otherwise,} \end{cases} \tag{2.3}$$

and figure 3 shows a threshold of

$$\lambda_n = \sigma_n \sqrt{\log n} = \sqrt{\log n / n}.$$

For the particular configuration of true means θ_k shown there, the data from the two spikes pass the threshold unchanged, and as such are essentially unbiased estimators. Meanwhile, in all other coordinates, the threshold correctly sets all data to zero except for the small fraction of noise that exceeds the threshold. Thus, it can be directly verified that

$$\mathrm{MSE}(\hat{\theta}_\lambda, \theta) \approx 2\sigma_n^2 + n\sigma_n^2 E\{Z^2, Z^2 > \log n\} \approx 2n^{-1} + 2\sqrt{\log n / n},$$

where Z is a standard Gaussian variate. This mean squared error is of course much better than for any of the linear estimators.

Statistical games and the minimax theorem. The skeptic will object that the configuration of figure 3 was chosen to highlight the advantages of thresholding, and indeed it was! It is precisely to avoid such reasoning from constructed cases that the tools of game theory have been adapted for use in statistics. A sterner and fairer test of an estimator is obtained by creating a statistical two-person zero-sum game, or *statistical decision problem*.

(i) Player I ('the scientist') is allowed to choose any estimator $\hat{\theta}(y)$, linear, threshold or of more complicated type.

(ii) Player II ('nature') may choose $\theta \in \mathbb{R}^n$ *at random*, and may choose a probability distribution π for θ subject only to the sparsity constraint that $E_\pi \|\theta\|_1 \leqslant 1$.

(iii) The pay-off is calculated as the expected mean squared error of $\hat{\theta}(Y)$ when θ is chosen according to π, and then Y satisfies model (2.1): $Y = \theta + \sigma_n z$ for $z \sim N_n(0, I)$. Thus, the pay-off now averages over *both* θ *and* Y:

$$r(\hat{\theta}, \pi) = E_\pi E_{Y|\theta} \|\hat{\theta}(Y) - \theta\|_2^2.$$

Of course, the scientist tries to minimize the pay-off and nature tries to maximize it.

Classical work in statistical decision theory (Wald 1950; Le Cam 1986; see also Johnstone 1998) shows that the minimax theorem of von Neumann can be adapted to apply here, and that the game has a well-defined value, the *minimax risk*:

$$R_n = \inf_{\hat{\theta}} \sup_{\pi} r(\hat{\theta}, \pi) = \sup_{\pi} \inf_{\hat{\theta}} r(\hat{\theta}, \pi). \tag{2.4}$$

An estimator attaining the left-hand infimum in (2.4) is called a *minimax* strategy or *estimator* for player I, while a prior distribution π attaining the right-hand supremum is called *least favourable* and is an optimal strategy for player II. It is the *structure* of these optimal strategies, and their effect on the minimax risk R_n, that is of chief statistical interest.

While these optimal strategies cannot be exactly evaluated for finite n, informative asymptotic approximations are available (Donoho & Johnstone 1994*b*), with the consequence that under our unit norm sparsity constraint,

$$R_n \sim \sqrt{\log n / n},$$

as $n \to \infty$. Indeed, an approximately least-favourable distribution is given by drawing the individual coordinates θ_k independently from a *two-point* distribution with

$$\theta_k = \begin{cases} \sigma_n \sqrt{\log n}, & \text{with probability } \varepsilon_n \doteq 1/\sqrt{n \log n}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.5}$$

This amounts to repeated tossing of a coin highly biased towards zero. Thus, in n draws, we expect to see a relatively small number, namely $\sqrt{n/\log n}$ of non-zero components. The size of these non-zero values is such that they are hard to distinguish from the larger values among the more numerous remaining $n - \sqrt{n/\log n}$ noise observations. Of course, what makes this distribution difficult for player II is that the *locations* of the non-zero components are random as well.

An approximately minimax estimator for this setting is given by the hard thresholding rule described earlier, but with threshold λ_n slightly larger than $\sqrt{\log n}$: for example, $\lambda_n = \sqrt{\log(n \log n)}$ will do. It can also be verified that no linear estimator can achieve a pay-off of better than $1/2$ if nature chooses a suitably uncooperative probability distribution for θ.

It is perhaps the qualitative features of this solution that most deserve comment. Had we worked with simply a signal-to-noise constraint—$E_\pi \|\theta\|_2^2 \leqslant 1$, say—we would have obtained a Gaussian prior distribution as being approximately least favourable, and the linear Wiener filter (2.2) with $\sigma_n^2 = \tau_n^2 = 1/n$ as an approximately minimax estimator. The imposition of a sparsity constraint takes us far away from Gaussian priors and linear estimators.

Sparsity and improved MSE. There is an alternative way to show how sparsity of representation affects the mean squared error of estimation using thresholding. Return to model (2.1), and observe that the MSE of $\hat{\theta}_1(y_i) = y_i$ for estimating θ_i is σ_n^2, while the MSE of $\hat{\theta}_0(y_i) = 0$ is θ_i^2. Given a choice, an omniscient 'oracle' would choose the estimator that yields the smaller of the two MSEs. Repeating this for each coordinate leads to a notion of *ideal risk*:

$$\mathcal{R}(\theta, \sigma) = \sum_i \min(\theta_i^2, \sigma^2). \tag{2.6}$$

Suppose that the coefficients are rearranged in decreasing order:

$$\theta_1^2 \geqslant \theta_2^2 \geqslant \cdots \geqslant \theta_n^2.$$

The notion of 'compressibility' captures the idea that the number of large coefficients, $N_\sigma(\theta) = \#\{\theta_i : |\theta_i| \geqslant \sigma\}$ is small, and also that there is little energy in the tail sums

$$c_k^2 = \sum_{i>k} \theta_i^2.$$

Then, good compressibility is actually equivalent to small ideal risk:

$$\mathcal{R}(\theta, \sigma) = N_\sigma(\theta)\sigma^2 + c_{N_\sigma}^2(\theta).$$

Ideal risk cannot be attained by any estimator, which, as a function of y alone, lacks access to the oracle. However, thresholding comes relatively close to *mimicking* ideal risk: for *soft* thresholding at $\lambda_n = \sqrt{2\log n}$, Donoho & Johnstone (1994*a*), show for all n, and for all $\theta \in \mathbb{R}^n$, that

$$E\|\hat{\theta}_{\mathrm{ST}} - \theta\|^2 \leqslant (2\log n + 1)[\epsilon^2 + \mathcal{R}(\theta, \epsilon)].$$

Thus, sparsity implies good compressibility, which in turn implies the possibility of good estimation, in the sense of relatively small MSE.

Remark 2.2. The scope of model (2.1) is broader than it may at first appear. Suppose that the observed data satisfy

$$Y = A\theta + \epsilon, \tag{2.7}$$

where Y is an $N_n \times 1$ data vector, A is an $N_n \times n$ orthogonal design matrix ($A^{\mathrm{t}}A = mI_n$), and ϵ has independent Gaussian components. Model (2.1) is recovered by premultiplying (2.7) by $m^{-1}A^{\mathrm{t}}$. Thus, A might be a change of basis, and so our analysis covers situations where there is *some* known basis in which the signal is thought to be sparse. Indeed, this is how sparsity in wavelet bases is employed, with A being (part of the inverse of) a wavelet transform.

Remark 2.3. The sparsity analysis, motivated by wavelet methods below, considers a *sequence* of models of increasing dimensionality; indeed, the index n is precisely the number of variables. This is in sharp contrast with traditional parametric statistical theory, in which the number of unknown parameters is held fixed as the sample size increases. In practice, however, larger quantities of data N_n typically permit or even require the estimation of richer models with more parameters. Model (2.7) allows N_n to grow with n. Thus, wavelet considerations promote a style of asymptotics whose importance has long been recognized (Huber 1981, § 7.4).

Remark 2.4. A common criticism of the use of minimax analyses in statistics holds that it is unreasonable to cast 'nature' as a malicious opponent, and that to do so risks throwing up as 'worst cases' parameters or prior configurations that are irrelevant to normal use. This challenge would be most pertinent if one were to propose an estimator on the basis of a single decision problem. Our perspective is different: we analyse *families* of statistical games, hoping to discover the common structure of optimal strategies; both estimators and least-favourable distributions. If an estimator class, such as thresholding, emerges from many such analyses, then it has a certain robustness of validity that a single minimax analysis lacks.

3. The 'signal in Gaussian white-noise' model

The multivariate Gaussian distribution $N_p(\theta, \sigma^2 I)$ with mean θ and p independent coordinates of standard deviation σ is the central model of parametric statistical inference, arising as the large sample limit of other p-parameter models, as well as in its own right.

In non-parametric statistics, the 'signal in Gaussian white-noise' model plays a similar role. The observation process $\{Y(s),\ 0 \leqslant s \leqslant 1\}$ is assumed to satisfy

$$Y(t) = \int_0^t f(s)\,\mathrm{d}s + \sigma W(t), \tag{3.1}$$

where f is a square integrable function on $[0,1]$ and W is a standard Brownian, or Wiener, process starting at $W(0)=0$. In infinitesimal form, this becomes

$$\mathrm{d}Y(t) = f(t)\,\mathrm{d}t + \sigma\,\mathrm{d}W(t),$$

suggesting that the observations are built up from data on $f(t)$ corrupted by independent white-noise increments $\mathrm{d}W(t)$. The unknown parameter is now f, and it is desired to estimate or test f or various functionals of f, such as point values or integrals, on the basis of Y.

As leaders of the Soviet school of non-parametric function estimation, Ibragimov & Khas'minskii (1981) gave a central place to model (3.1), arguing that its challenges are all conceptual and not merely technical. As in the finite-dimensional case, (3.1) arises as an appropriate large-sample or low-noise limit of certain other non-parametric models, such as probability density estimation, regression and spectrum estimation (see, for example, Brown & Low 1996; Nussbaum 1996). It extends to images or other objects if one replaces $t \in [0,1]$ by a multi-parameter index $t \in D \subset \mathbb{R}^d$, and W by a Brownian sheet.

Model (3.1) has an equivalent form in the sequence space of coefficients in an orthonormal basis $\{\psi_I, I \in \mathcal{I}\}$ for $L_2([0,1])$, the square integrable functions on the unit interval. Thus, let

$$y_I = \int \psi_I\,\mathrm{d}Y,$$

and, similarly,

$$\theta_I = \int \psi_I f \quad \text{and} \quad z_I = \int \psi_I\,\mathrm{d}W,$$

the latter being a Wiener–Ito stochastic integral. Then, (3.1) becomes

$$y_I = \theta_I + \sigma z_I, \quad I \in \mathcal{I}, \tag{3.2}$$

where $\theta = (\theta_I) \in \ell_2$, and, by the elementary properties of stochastic integrals, z_I are independent and identically distributed (i.i.d.) standard Gaussian variates.

While (3.2) looks like a straightforward infinite-dimensional extension of the Euclidean $N_p(\theta, \sigma^2 I)$ model, there are significant difficulties. For example, the sequence (y_I) is, with probability one, *not* square summable, because the noise (z_I) is i.i.d. Similarly, there is no probability distribution supported on square summable sequences that is invariant under all orthogonal transformations, or even simply under permutation of the coordinates. If the index set $\mathcal{I}$ is linearly ordered, as for the Fourier basis, one must typically work with sequences of weights, often polynomially decreasing, which lack simplifying invariance properties.

The multi-resolution character of wavelet bases is helpful here. If $\{\psi_I\}$ is now an orthonormal wavelet basis for $L^2[0,1]$, such as those of Cohen *et al.* (1993), then the index $I=(j,k)$ becomes bivariate, corresponding to level j and location $k2^{-j}$ within each level. The index set $\mathcal{I}$ becomes

$$\bigcup_{j\geqslant 0} \mathcal{I}_j \bigcup \mathcal{I}_{-1},$$

with $|\mathcal{I}_j| = 2^j$ counting the possible values of k, and $\mathcal{I}_{-1}$ an exceptional set for the scaling function ϕ_0.

Collect the data coefficients y_{jk} in (3.2) observed at level j into a vector y_j that has a *finite*-dimensional $N_{2^j}(\theta_j, \sigma^2 I)$ distribution. For many theoretical and practical purposes, it is effective to work with each of these level-wise distributions separately. Since they are of finite (but growing!) dimension, it is possible, and often a scientifically reasonable simplification, to give θ_j an orthogonally or permutation-invariant probability distribution. Indeed, the sparsity results of § 2, derived for Gaussian distributions of large *finite* dimension, had precisely this permutation invariance character, and can be applied to each level in the dyadic sequence model.

The full non-parametric estimation conclusions are obtained by combining results across resolution levels. However, it often turns out, especially for minimax analyses, that for a given noise level σ, the 'least-favourable' behaviour occurs at a single resolution level $j = j(\sigma)$, so that conclusions from the $j(\sigma)$th permutation-invariant Gaussian model provide the key to the non-parametric situation. As the noise level σ decreases, the critical level $j(\sigma)$ increases, but in a controllable fashion. Thus, wavelet-inspired dyadic sequence models allow comparatively simple finite-dimensional Gaussian calculations to reveal the essence of non-parametric estimation theory.

In this sense, wavelet bases have rehabilitated the finite-dimensional multivariate Gaussian distribution as a tool for non-parametric theory, establishing in the process a bridge between parametric and non-parametric models.

4. Optimality in the white-noise model

Our discussion of optimality in the white-noise model illustrates the truism that the available tools, conceptual and mathematical, influence the theory that can be created at a given time. Prior to the advent of wavelet bases, formulations emphasizing good properties of linear estimators were the norm; subsequently, theoretical conclusions became possible that were more in accord with recent practical experience with algorithms and data.

As a framework for comparing estimators, we continue to use statistical games and the minimax principle. In the sequence model (3.2), a strategy for player I, the scientist, is a sequence of estimator coefficients $\hat{\theta}(y) = (\hat{\theta}_I)$, which in terms of functions becomes

$$\hat{f}(t) = \sum_I \hat{\theta}_I \psi_I(t).$$

A strategy for player II, nature, is a prior distribution π on θ, subject to a constraint that $\pi \in \mathcal{P}$. In function terms, this corresponds to choosing a random *process* model for $\{f(t),\ 0 \leqslant t \leqslant 1\}$. The pay-off function from the scientist to nature is

$$r(\hat{\theta}, \pi) = E_\pi \|\hat{\theta}(Y) - \theta\|^2 = E_\pi \int (\hat{f} - f)^2.$$

The constraint set $\mathcal{P} = \mathcal{P}(\mathit{\Theta})$ usually requires, in some average sense usually defined by moments, that π concentrates on the set $\mathit{\Theta} = \mathit{\Theta}(\mathcal{F})$.

It is necessary to take $\mathcal{F}$ to be a compact subset of $L_2[0,1]$, because otherwise the minimax risk does not even decrease to zero in the low-noise limit ($\epsilon \to 0$): in other words, even consistency cannot be guaranteed without restricting $\mathcal{F}$. The restrictions usually imposed have been on smoothness, requiring that f have α derivatives with bounded size in some norm. In the 1970s and 1980s, the norms chosen were typically either Hölder, requiring uniform smoothness, or Hilbert–Sobolev, requiring smoothness in a mean square sense.

(*a*) *Linear estimators*

To describe the historical background, we start with linear methods. Estimators that are linear functions of observed data arise in a number of guises in application: they are natural because they are simple to compute and study, and already offer considerable flexibility. In the single time parameter model (3.1)–(3.2), time-shift invariance is also natural, in the absence of specific prior information to the contrary. Thus, in what follows, we switch freely between time domain $\hat{f}$ and Fourier-coefficient domain $\hat{\theta} = (\hat{\theta}_k)$. It then turns out that all shift-invariant estimators have similar structure as follows.

(i) Weighted Fourier series. Using the Fourier series form (3.2) for the data model,

$$\hat{\theta}_k = \hat{\kappa}(hk)y_k, \tag{4.1}$$

where the shrinkage function $\hat{\kappa}$ is decreasing, corresponding to a downweighting of signals at higher frequencies. The 'bandwidth' parameter h controls the actual location of the 'cut-off' frequency band.

(ii) Kernel estimators. In the time domain, the estimator involves convolution with a window function K, scaled to have 'window width' h:

$$\hat{f}(t) = \int \frac{1}{h} K\left(\frac{t-s}{h}\right) \mathrm{d}Y(s). \tag{4.2}$$

The representation (4.1) follows after taking Fourier coefficients.

(iii) Smoothing splines. The estimator $\hat{\theta}$ minimizes

$$\sum (y_k - \theta_k)^2 + \lambda^{2r} \sum k^{2r} \theta_k^2,$$

where the *roughness penalty* term takes the mean square form,

$$c \int_0^1 (D^r f)^2,$$

in the time domain for some positive integer r. In this case, calculus shows that $\hat{\theta}_k$ again has the form (4.1) with

$$\hat{\kappa}(\lambda k) = [1 + (\lambda k)^{2r}]^{-1}.$$

Each of these forms was studied by numerous authors, either in the white-noise model, or in asymptotically similar models—regression, density estimation—usually over Hölder or Hilbert–Sobolev $\mathcal{F}$. A crowning result of Pinsker (1980) showed that linear estimators of the form (4.1) were asymptotically minimax among *all* estimators over ellipsoidal function classes. More specifically, suppose that $\mathcal{F}$ may be represented in the sequence space model (3.2) in terms of an *ellipsoid* with semiaxes determined by a sequence $\{a_k\}$: thus,

$$\mathcal{F} = \left\{ \theta : \sum_k a_k^2 \theta_k^2 \leqslant C^2 \right\}.$$

For example, if $\mathcal{F}$ corresponds to functions with r mean squared derivatives,

$$\int (D^r f)^2 \leqslant L^2,$$

then

$$a_{2k-1} = a_{2k} = (2k)^r \quad \text{and} \quad C^2 = L^2/\pi^{2r}.$$

We denote the resulting space $\mathcal{F}_{r,C}$, and concentrate on these special cases below. Pinsker (1980) constructed a family of *linear* shrinkage estimators $\hat{f}_\epsilon \leftrightarrow \hat{\theta}_\epsilon$ of the form (4.1) with $\hat{\kappa} = \hat{\kappa}_\epsilon$ depending also on (r, C), so that the worse case MSE of $\hat{\theta}_\epsilon(r, C)$ over $\mathcal{F}_{r,C}$ was the best possible in the small-noise limit:

$$\sup_{f \in \mathcal{F}_{r,C}} r(f, \hat{f}_\epsilon) \sim r_\epsilon(\mathcal{F}_{r,C}), \quad \epsilon \to 0.$$

Furthermore, Pinsker (1980) showed that an asymptotically least-favourable sequence of prior distributions could be described by assigning each θ_k an independent Gaussian distribution with mean zero and appropriate scale $\sigma_k^2(\epsilon, r, C)$.

This result would seem to give definitive justification for the use of linear methods: the least-favourable distributions for ellipsoids are approximately Gaussian, and for Gaussian processes, the optimal (Bayes) estimators are linear.

At about this time, however, some cracks began to appear in this pleasant linear/Gaussian picture. In the theoretical domain, Nemirovskii (1985) and Nemirovskii *et al.* (1985) showed, for certain function classes $\mathcal{F}$ in which smoothness was measured in a mean *absolute* error (L_1) sense, that linear estimators were no longer minimax, and indeed had suboptimal *rates* of convergence of error to zero as $\epsilon \to 0$.

Meanwhile, methodological and applied statistical research harnessed computing power to develop smoothing algorithms that used different, and *data-determined*, window widths at differing time points. This is in clear contrast with the fixed-width h implied by the kernel representation (4.2). For example, Cleveland (1979) investigates local smoothing, and Friedman & Stuetzle (1981), in describing the univariate smoother they constructed for projection pursuit regression, say, explicitly,

> the actual bandwidth used for local averaging at a particular value of [the predictor] can be larger or smaller than the average bandwidth. Larger bandwidths are used in regions of high local variability of the response.

This amounts to an implicit rejection of the ellipsoid model. The algorithms of Friedman & Stuetzle and others were iterative, involving multiple passes over the data, and were, thus, beyond theoretical analysis.

(*b*) *Wavelet bases and thresholding*

The appearance of wavelet bases enabled a reconciliation of the Gaussian-linear theory with these divergent trends. Informally, this might be explained by 'Mallat's heuristic', quoted by Donoho (1993):

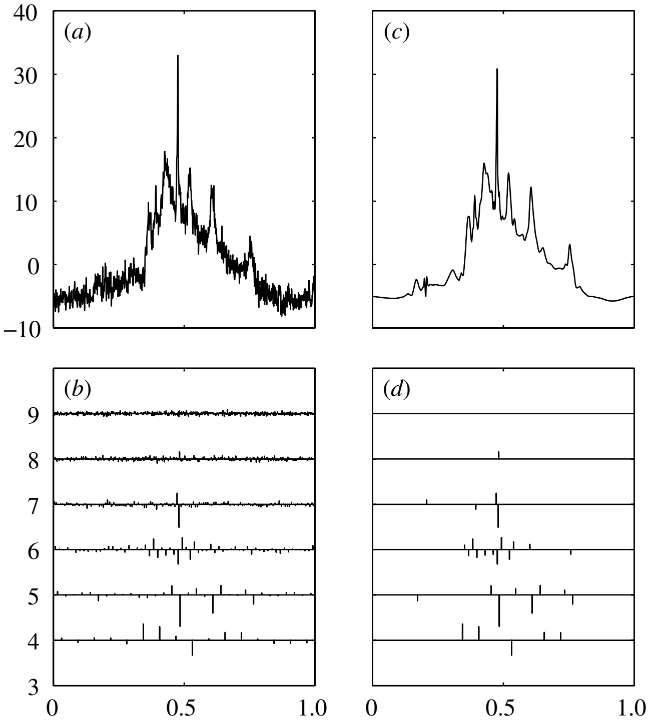

Figure 4. (*a*) Sample NMR spectrum provided by A. Maudsley and C. Raphael; $n = 1024$. (*b*) Empirical wavelet coefficients w_{jk} displayed by nominal location and scale j, computed using a discrete orthogonal wavelet transform: Daubechies near-symmetric filter of order $N = 6$ (Daubechies 1992, ch. 6). (*c*) Reconstruction using inverse discrete wavelet transform of coefficients in (*d*). (*d*) Wavelet coefficients after hard thresholding at $\hat{\sigma}\sqrt{2\log n}$. $\hat{\sigma} = \text{med.abs.dev.}(w_{9k})/0.6745$, a resistant estimate of scale at level 9 (for details, see Donoho *et al.* 1995). (*e*), (*f*) Adaptive (quasi-) linear shrinkage of wavelets coefficients in (*b*) using the James–Stein estimator on the ensemble of coefficients at each level (cf. Donoho & Johnstone 1995), and reconstruction by the discrete wavelet transform.

> Bases of smooth wavelets are the best bases for representing objects composed of singularities, when there may be an arbitrary number of singularities, which may be located in all possible spatial positions.

This captures the notion that a function with spatially varying smoothness (transients at some points, very smooth elsewhere) might be sparsely represented in a smooth wavelet basis and hence well estimated.

Indeed, figure 4*c* illustrates the improvement yielded by wavelet thresholding on a noisy NMR signal in comparison with figure 4*e*, which shows an arguably best near-linear estimator in the spirit of Pinsker's theorem (for further details, see Donoho & Johnstone (1995) and Johnstone (1998)). Clearly, the Pinsker-type estimator fails to adjust the (implied) window width in (4.2) to both capture the sharp peaks *and* to average out noise elsewhere.

We turn now to describe some of the theory that underlies these reconstructions and Mallat's heuristic. More technically, wavelets form an *unconditional basis* simultaneously for a vast menagerie of function spaces, allowing more flexible measures of smoothness than the Hölder and Hilbert–Sobolev spaces hitherto used in statistics. An unconditional basis for a Banach space B with norm $\|\cdot\|$ is defined by a countable family $\{\psi_I\} \subset B$ with two key properties as follows.

(i) Any element $f \in B$ has a *unique representation*:

$$f = \sum_1^\infty \theta_I \psi_I,$$

in terms of coefficients $\theta_I \in \mathbb{C}$.

(ii) *Shrinkage*: there is an absolute constant C such that if $|\theta'_I| \leqslant |\theta_I|$ for all I, then

$$\left\|\sum \theta'_I \psi_I\right\| \leqslant C \left\|\sum \theta_I \psi_I\right\|.$$

The statistical significance of these two properties is firstly that functions $f \in B$ may be described in terms of coefficient *sequences* $\{\theta_I\}$, and secondly that the basic statistical operation of *shrinkage* on these sequences, whether linear or via thresholding, is stable in B, in that the norms cannot be badly inflated. Notably, this property is *not* shared by the Fourier basis (Kahane *et al*. 1977).†

Figure 5 represents the class of Besov spaces schematically in terms of the smoothness index α (equal to the number of derivatives) and the homogeneity index p, plotted as $1/p$ as is customary in harmonic analysis. Each point $(\alpha, 1/p)$ corresponds to a class of Besov spaces. The vertical line $p = 2$ represents the Hilbert–Sobolev smoothness spaces traditionally used in statistics, while points to the right are spaces with $p < 2$, hence having some degree of sparsity of representation in the wavelet domain.

To be more concrete, we consider a single example from the family of Besov spaces, the bump algebra (Meyer 1990, § 6.6).‡ Let

$$g_{\mu,\sigma}(t) = \exp\{-(x-\mu)^2/2\sigma^2\}$$

denote a normalized Gaussian bump with location μ and scale σ. The bump algebra on $\mathbb{R}$ is the collection of all functions f representable as a convergent superposition of signed, scaled and located bumps:

$$f = \sum_1^\infty \alpha_i g_{\mu_i,\sigma_i}, \qquad \sum_1^\infty |\alpha_i| < \infty.$$

This might seem a plausible model for (say) signed spectra with peaks of varying location, width and height (compare with figure 5*b*). As Meyer notes, while the simplicity of this description is perhaps deceptive due to lack of uniqueness in the representation, an equivalent and stable description can be given via a smooth wavelet orthobasis $\{\psi_I\}$. When restricted to $L_2[0,1]$, we may use the index system $\mathcal{I} = \cup_j \mathcal{I}_j$ of § 3. A subset $\mathcal{B}_C$ with norm at most C is defined by those

$$f = \sum \theta_I \psi_I$$

† See Donoho (1993) for a formalization of Mallat's heuristic using the unconditional basis property.

‡ For a discussion in terms of the space of bounded total variation, see Mallat (1998).

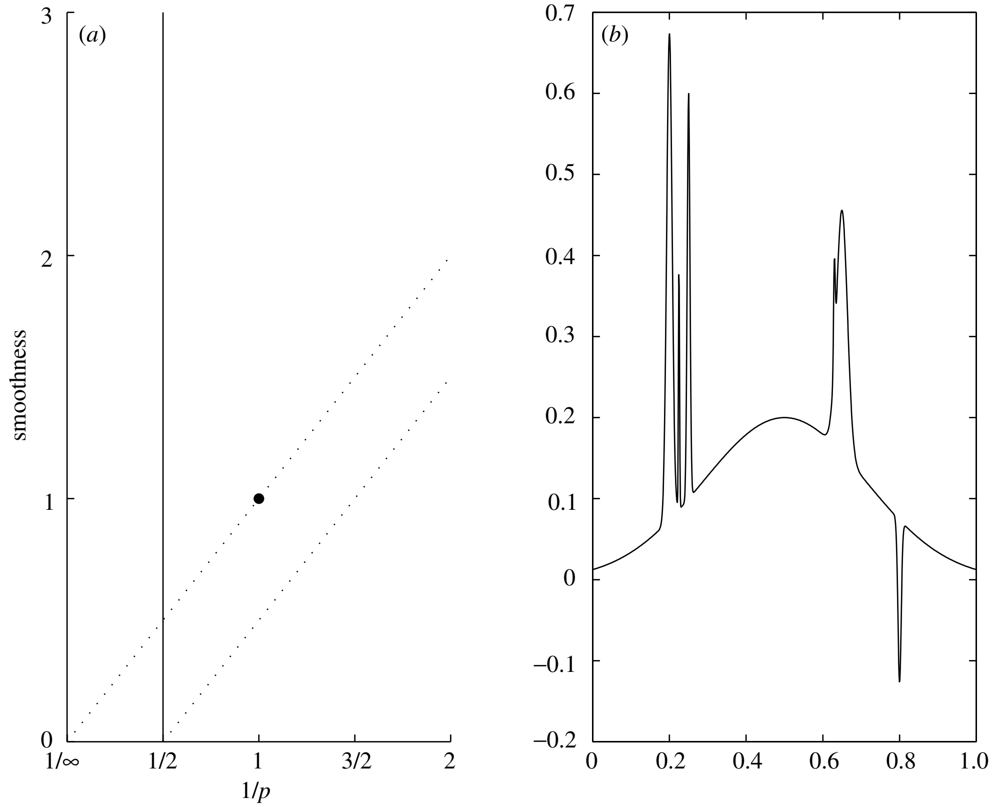

Figure 5. (*a*) Schematic diagram of Besov spaces of varying homogeneity p and smoothness. Spaces above the diagonal line $\alpha = 1/p$ consist of functions that are at least continuous. The point at $(1, 1)$ corresponds to the bump algebra. (*b*) Caricature evoking a function in the bump algebra: a superposition of Gaussians at widely different spatial scales.

for which

$$\sum_{j \geqslant 0} 2^{j/2} \sum_{k \in \mathcal{I}_j} |\theta_{jk}| \leqslant C. \tag{4.3}$$

The condition (4.3) is a scale-weighted combination of the ℓ_1 norms of the coefficients at each level j. The full bump algebra is simply the union of all $\mathcal{B}_C$. This wavelet representation is the key to statistical results.

Now consider a statistical game in the wavelet domain, with the prior constraint family $\mathcal{P}$ consisting of those priors π such that the π expectation of the left-hand side of (4.3) is bounded by C. In view of the analysis in § 2, it is perhaps now not surprising that no linear estimator can achieve optimal rates of convergence. In fact, the minimax risk $R_\epsilon(\mathcal{B}_C)$ decreases like $C^{2/3}\epsilon^{4/3}$, whereas the best rate possible for linear estimators is much slower, namely $O(\epsilon)$.

The sequence space structure provided by the unconditional basis property also implies that optimal estimators in the statistical game are *diagonal*: $\hat{\theta}_I(y) = \delta_I(y_I)$ depends on y_I alone. While these optimal estimators cannot be described explicitly, this separation of variables is an important simplification. For example, the least-favourable distributions are, at least asymptotically, obtained by making the wavelet

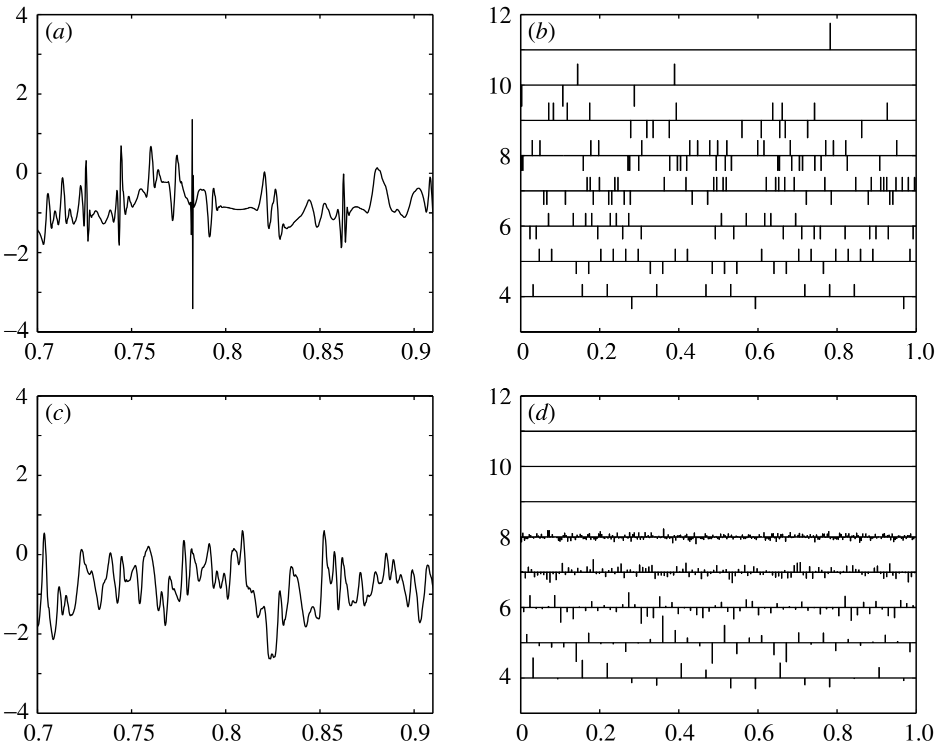

Figure 6. (*a*) A segment of a sample path on $[0,1]$ from a prior distribution on wavelet coefficients that is approximately least favourable for a bump algebra ball of form (4.3). (*b*) The corresponding wavelet coefficients—those at level j—are i.i.d. draws from a three-point distribution $(1-\epsilon_j)\delta_0 + \epsilon_j(\delta_{\mu_j} + \delta_{-\mu_j})$, as described in the text. The wavelets ψ_{jk} are derived from the $N = 8$ instance of the Daubechies (1992, ch. 6) 'closest to linear phase' filter. (*c*) Sample path from the Gaussian process that is the least-favourable distribution for an ellipsoid $\mathcal{F}_{m,C}$ with $m = 1$ square integrable derivatives. (*d*) Corresponding wavelet coefficients with variance σ_j^2 decreasing with j.

coefficients *independent* and, indeed, identically distributed within each level. Furthermore, it turns out that thresholding estimators, as described in § 2 but now with thresholds depending on level, have MSE always within a constant multiple (less than or equal to 2.2) of the optimal value.

Thresholding in a wavelet basis automatically has the spatial adaptivity that previous algorithmic work sought: the effective window width at a given time point is proportional to $2^{-j(t_0)}$, where $j(t_0)$ is the finest level wavelet coefficient that survives thresholding among those wavelets whose support contains t_0.

Sample paths from approximately least-favourable distributions for $\mathcal{B}_C$ are informative. A sample realization of f can be plotted by substituting a sample draw of coefficients, θ_{jk}, into

$$f = \sum \theta_{jk}\psi_{jk}.$$

By considering only threshold rules, which are nearly minimax as just mentioned, it can be shown that a near-least-favourable distribution can be constructed from three point distributions, $(1-\epsilon_j)\delta_0 + \epsilon_j(\delta_{\mu_j} + \delta_{-\mu_j})$, that are quite similar to that given at (2.5). Now, however, the location μ_j and size $\epsilon_j/2$ of the non-zero atom and its reflection depend on the level j, but, within level, the 2^j draws are independent, as they are in (2.5). A numerical optimization (Johnstone 1994) allows evaluation of μ_j and ϵ_j for a given ϵ and $\mathcal{B}_C$. Figure 6*a* shows a representative sample path and figure 6*b* shows the corresponding individual wavelet coefficients for this distribution.

Figure 6*c* shows a corresponding sample path drawn from the Gaussian least-favourable distribution on wavelet coefficients, figure 6*d* corresponding to the ellipsoid $\mathcal{F}_{m,C}$ with $m = 1$ derivatives assumed to be square integrable. Again, the wavelet coefficients are i.i.d. within level, but are now drawn from a Gaussian distribution with variance $\sigma_j^2(m, C, \epsilon)$ determined by Pinsker's solution. The two plots are calibrated to the same indices of smoothness $\alpha = 1$, scale $C = 1$ and noise level $\epsilon = 1/64$.

The qualitative differences between these plots are striking: the Gaussian sample path has a spatially homogeneous irregularity, with the sample wavelet coefficients being 'dense', though decreasing in magnitude with increasing scale or 'frequency octave'. In contrast, the bump algebra sample path has a greater spikiness: the sample wavelet coefficients have increasing sparsity and magnitude with each increasing scale. These differences become even more pronounced if one increases the smoothness α and decreases the homogeneity index p; see, for example, the plots for $\alpha = 2$, $p = 1/2$ in Johnstone (1994).

To summarize: with only the sparsity in mean constraint, and no other restriction on estimators or prior distributions, coordinatewise thresholding and sparse priors emerge as the near optimal strategies for the bump algebra statistical game. Thresholding has better MSE, indeed faster rates of convergence, than any linear estimate over $\mathcal{B}$. This may also be seen visually in the relatively much more noise-free reconstruction using wavelet thresholding, shown in figure 4.

We mention briefly the problem of *adaptation*. The near optimality of thresholding and sparse priors holds in similar fashion for a large class of Besov space constraints described by $(\alpha, 1/p)$ and size parameter C. The optimal threshold estimator in each case will depend on (α, p, C): can one give an estimator with optimal or near-optimal properties without needing to specify (α, p, C)? One very simple possibility, already shown in figure 4 and explored at length in Donoho *et al.* (1995), is to use hard or soft thresholding at threshold $\sqrt{2 \log n}$, where n is the number of observations, or wavelet coefficients. This estimator has a remarkably robust near adaptivity—it nearly achieves, up to logarithmic terms, the minimax rate of convergence simultaneously over a wide range of both functional classes *and* error measures—not solely mean squared error.

5. Concluding remarks

Wavelets have enabled the development of theoretical support in statistics for the important notions of sparsity and thresholding.

In contrast, much work in modern curve fitting and regression treats the imposition of *smoothness* as a guiding principle. Wavelets prompt us to think of smoothness as a particular case of a more general principle, namely sparsity of representation. It is in fact sparsity of representation that determines when good estimation is possible.

Pursuing the sparsity idea for functions of more than one variable leads to systems other than literal wavelets, as discussed by Candès & Donoho (this issue).

We may expect to see more use of ‘dyadic thinking’ in areas of statistics and data analysis that have little to do with wavelets directly. This is likely both in the development of methods, and also as a metaphor in making simpler models for theoretical analysis of other more complex procedures.

Many thanks to Marc Raimondo for his help in the preparation of figure 6. The author gratefully acknowledges financial support from the National Science Foundation (DMS 9505151).

References

Brown, L. D. & Low, M. G. 1996 Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statistics* **3**, 2384–2398.

Cleveland, W. S. 1979 Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.* **74**, 829–836.

Cohen, A., Daubechies, I. & Vial, P. 1993 Wavelets and fast wavelet transform on an interval. *Appl. Comp. Harmonic Analysis* **1**, 54–81.

Daubechies, I. 1992 *Ten lectures on wavelets.* CBMS-NSF Series in Applied Mathematics, no. 61. Philadelphia, PA: SIAM.

Donoho, D. 1993 Unconditional bases are optimal bases for data compression and statistical estimation. *Appl. Comp. Harmonic Analysis* **1**, 100–115.

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

Donoho, D. L. & Johnstone, I. M. 1994*b* Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory Related Fields* **99**, 277–303.

Donoho, D. L. & Johnstone, I. M. 1995 Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. 1995 Wavelet shrinkage: asymptopia? (With discussion.) *J. R. Statist. Soc.* B **57**, 301–369.

Friedman, J. & Stuetzle, W. 1981 Projection pursuit regression. *J. Am. Statist. Assoc.* **76**, 817–823.

Hotelling, H. 1933 Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **24**, 417–441, 498–520.

Huber, P. J. 1981 *Robust statistics.* Wiley.

Ibragimov, I. & Khas’minskii, R. 1981 *Statistical estimation: asymptotic theory.* Springer.

Johnstone, I. M. 1994 Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical decision theory and related topics* (ed. S. Gupta & J. Berger), vol. V, pp. 303–326. Springer.

Johnstone, I. M. 1998 Function estimation: white noise, sparsity and wavelets. Lecture notes.

Kahane, J., de Leeuw, K. & Katznelson, Y. 1977 Sur les coefficients de fourier des fonctions continues. *Comptes Rendus Acad. Sci. Paris* A **285**, 1001–1003.

Le Cam, L. 1986 *Asymptotic methods in statistical decision theory.* Springer.

Mallat, S. 1998 Applied mathematics meets signal processing. In *Proc. of ICM, 18–27 August 1998, Berlin.* (*Doc. Math.* Extra Volume **I**, 319–338.)

Meyer, Y. 1990 *Ondelettes et Opérateurs. I. Ondelettes. II. Opérateurs de Calderón–Zygmund. III. Opérateurs multilinéaires (with R. Coifman).* Paris: Hermann. (English translation published by Cambridge University Press.)

Nemirovskii, A. 1985 Nonparametric estimation of smooth regression function. *Izv. Akad. Nauk. SSR Teckhn. Kibernet.* **3**, 50–60 (in Russian). (Transl. in *J. Comput. Syst. Sci.* **23** (1986), 1–11.)

Nemirovskii, A., Polyak, B. & Tsybakov, A. 1985 Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems Information Transmission* **21**, 258–272.

Nussbaum, M. 1996 Asymptotic equivalence of density estimation and white noise. *Ann. Statist.* **24**, 2399–2430.

Pinsker, M. 1980 Optimal filtering of square integrable signals in Gaussian white noise. *Problems Information Transmission* **16**, 120–133. (Originally published in Russian in *Problemy Peredatsii Informatsii* **16**, 52–68.)

Wald, A. 1950 *Statistical decision functions*. Wiley.