Wavelets and Optimal Nonlinear Function Estimates

by

David L. Donoho
Department of Statistics
University of California, Berkeley

Iain M. Johnstone
Department of Statistics
Stanford University

Technical Report No. 281
December 1990

Department of Statistics
University of California
Berkeley, California 94720

Wavelets and Optimal Nonlinear Function Estimates

David L. Donoho
Department of Statistics
University of California, Berkeley

Iain M. Johnstone
Department of Statistics
Stanford University

November 10, 1990

Abstract

We consider the problem of estimating a smooth function from noisy, sampled data. We use orthonormal bases of compactly supported wavelets to construct nonlinear function estimates which can significantly outperform evey linear method (kernel, smoothing spline, sieve, ...). Our estimates are simple nonlinear functions of the empirical wavelet coefficients and are asymptotically minimax over certain Besov smoothness classes. Our estimates possess the interpretation of *local adaptiveness*: they reconstruct using a kernel which may vary in shape and bandwidth from point to point, depending on the data. Modifications of our estimates based on simple threshold nonlinearities are near minimax and have interesting interpretations as smoothness-penalized least squares estimates or as adaptive depleted-basis spline fits.

Key Words. Minimax Decision theory. Minimax Bayes estimation. Besov Spaces. Nonlinear Estimation. White Noise Model. Nonparametric regression. Hardest Cartesian Subproblems. Renormalization. White Noise Approximation.

Acknowledgements. The first author was supported in part by NSF DMS 88-10192, by NASA Contract NCA2-488, and by a grant from ATT Foundation. The second author was supported in part by NSF grants DMS 84-51753, 86-00235, and NIH PHS grant GM21215-12.

1 Introduction

Suppose we are given n noisy samples of a function f:

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

with $t_i = i/n$, z_i iid $N(0, \sigma^2)$. Our goal is to estimate f with small mean-squared-error, i.e. to find an estimate $\hat{f}$ depending on $y_1, \ldots, y_n$ with small *risk* $R(\hat{f}, f) = E\|\hat{f} - f\|_2^2 = E\int_0^1((\hat{f})(t) - f(t))^2$. In addition, we know a priori that f belongs to a certain class $\mathcal{F}$ of smooth functions, but nothing more. It is reasonable to seek to optimize the *minimax risk* $R(n, \mathcal{F}) = \inf_{\hat{f}} \sup_f R(\hat{f}, f)$. When f is an L_2-Sobolev class or a Hölder class, such problems have been well-studied, with the result that various *linear* algorithms—kernel with fixed bandwidth, or smoothing spline—are known to be near optimal: Stone (1982), Nussbaum(1985).

To make a new point of departure, let us consider function classes $\mathcal{F}$ different from the traditional smoothness classes.

As a first example, consider the *Bump Algebra* (Meyer, 1990, Chapter VI.6, pages 186–189). Let $g_{t,s}(x) = \exp(-(x-t)^2/2s^2)$ denote a Gaussian "bump," normalized to height 1 rather than area 1. The Bump Algebra B is the class of all functions $f : \mathbb{R} \to \mathbb{R}$ which admit the decomposition

$$f(x) = \sum_{i=0}^{\infty} \alpha_i g_{(s_i,t_i)}(x) \tag{2}$$

for some sequence of triplets (α_i, t_i, s_i), $i = 0, 1, 2, \ldots$, which satisfy $\sum_{i=0}^{\infty} |\alpha_i| < \infty$. [Such a representation need not be unique.] The B-norm of such a function is the smallest ℓ_1-norm of the coefficients (α_i) in any such representation:

$$\|f\|_B = \inf \sum |\alpha_i| \qquad \text{such that (2) holds} \tag{3}$$

Under this norm B is a Banach space; in fact, a Banach algebra, since $g_{(t_1,s_1)} \cdot g_{(t_2,s_2)} = \lambda g_{(t_3,s_3)}$, $\lambda < 1$.

This algebra possesses two properties which might spark the interest of readers.

(A) It serves as an interesting caricature of certain function classes arising in scientific signal processing. Functions f obeying (2) with only finitely many nonzero α_i are evidently *polarized spectra* i.e., their graph consists of a set of "spectral lines" located at the (t_i) with "line widths" (s_i), "polarities" $\mathrm{sgn}(\alpha_i)$ and "amplitudes" $|\alpha_i|$. Thus estimating functions in B corresponds to recovery of polarized spectra with unknown locations of the lines, unknown line widths and unknown amplitudes. To make a problem with finite minimax risk, we must have *something* known, so we set $\mathcal{F} = \{ f : \|f\|_B \le C \}$ for a fixed constant C; this corresponds to a constraint on the amplitude of the spectrum to be recovered.

(B) $\mathcal{F}$ contains functions with considerable spatial inhomogeneity. In fact, a single function in $\mathcal{F}$ may be extremely spiky in one part of its domain and extremely flat or smooth in another part of its domain. This would not be possible, for example, in a Holder class, where functions must obey the same local modulus of continuity at

each point. Estimators based in some sense on an idea of spatial homogeneity of the estimand f—such as fixed bandwidth kernel estimates, trigonometric series, and least-squares smoothing splines—will presumably be unable to behave optimally in spatially inhomogeneous settings: either they will oversmooth the spiky part or they will undersmooth the flat part—or both.

These two properties—scientific interest and need for local adaptivity—are by no means unique to the Bump Algebra. For another example, consider the class $\mathcal{F}$ of functions of *Bounded Variation*: $\mathcal{F} = \{ f : TV(f) \leq C \}$. This class possesses the same two properties:

(A) *Scientific Interest.* For example, the key geophysical parameter in the acoustic theory of reflection seismology is the *acoustic impedance*, a function which is necessarily non-smooth, because it has jumps at certain changes in media, but which may be modelled as an object of finite variation.

(B) *Spatial Inhomogeneity.* Functions of bounded variation may have jumps localized to one part of the domain and be very flat elsewhere. Local adaptivity in reconstruction is obviously very desirable.

The story does not end here, either. The Bump Algebra and Total Variation fit into a continuum of examples which exhibit spatial inhomogeneity and may be of scientific interest. We show in this paper, that these examples, which are *Besov-type spaces with index* $p < 2$, exhibit phenomena which are unexpected on the basis of previous theoretical experience with L_2-Sobolev or Hölder classes, in three ways.

Ph. A *Nonlinear estimators can convincingly outperform linear estimators.* Let $R_L(n, \mathcal{F})$ devote the minimax risk for n observations from (1) when estimators are restricted to be *linear* in the data (y_i). Let $R(n, \mathcal{F})$ denote the minimax risk when estimators are unrestricted. We will show that for certain definite rate exponents r_L, r_N, $0 < r_L < r_N < 1$, $R_L(n, \mathcal{F}) \asymp n^{-r_L}$ while $R(n, \mathcal{F}) \asymp n^{-r_N}$. Hence $R_L(n, \mathcal{F})/R(n, \mathcal{F}) \to \infty$. In short, traditional linear methods are unable to compete effectively with nonlinear estimates.

Ph. B *Wavelets allow us to construct near-minimax estimators, which although (necessarily) nonlinear, have a very simple structure.* The theory of *wavelets* Meyer (1990) provides an orthogonal decomposition for L_2 which is an alternative to the usual orthogonal decompositions based on Fourier analysis or orthogonal polynomials. We will show how an empirical version of the wavelet decomposition may be used for nonlinear estimation. Specifically, we estimate the unknown function in a near-optimal way by applying to each empirical wavelet coefficient a special nonlinear transform which optimally rejects noise in a minimax sense.

Thus although one might have expected a minimax nonlinear estimate to be a quite arbitrary nonlinear function of the data, in fact it has a computationally and conceptually convenient form when expressed in the wavelet expansion.

Ph. C *These near-minimax estimates have an intrinsic local adaptivity.* Our wavelet method has a representation as an adaptive kernel estimator which may change locally —in both shape and bandwidth— in response to the data.

We emphasize that these phenomena are general and apply to a continuum of different smoothness classes $\mathcal{F}$.

2 Wavelets

The theory of wavelets has been enthusiatically developed in recent years by a large number of workers. Our point of entry into this literature was the books of Y. Meyer (1990a, b). Synthesizing a large body of superficially different work in fields ranging from Fourier analysis to operator theory to image analysis, Meyer develops fully the idea of multiresolution analysis and its use in the study of various function spaces and various operators. The papers of Daubechies (1988) and Mallat (1989 a,b,c) are also extremely helpful.

Starting from a pair of functions, φ and ψ, which are C^r and of *compact support*, one defines the translated and dilated functions

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \qquad j \in \mathbf{Z}, k \in \mathbf{Z}$$

and, for a fixed ℓ,

$$\varphi_{\ell,k}(t) = 2^{\ell/2}\varphi(2^\ell t - k) \qquad k \in \mathbf{Z}.$$

The functions φ and ψ provide an *r-regular wavelet analysis* if

(1) φ, ψ are C^r and of compact support

(2) $\int \varphi = 1$, $\int \psi t^k = 0$, $0 \le k \le r$,

(3) $\{\psi_{j,k}\}_{j\in\mathbf{Z},k\in\mathbf{Z}}$ make an orthonormal basis of $L_2(\mathbb{R})$

(4) $\{\varphi_{\ell,k}\}_{k\in\mathbf{Z}} \cup \{\psi_{j,k}\}_{j\ge\ell,k\in\mathbf{Z}}$ make an orthonormal basis of $L_2(\mathbb{R})$

Properties (3) and (4) yield immediately the usual decomposition and reconstruction formulas valid for orthonormal bases. However, it is convenient to give them special names. The *wavelet coefficients* of an $f \in L_2$ are

$$\alpha_{j,k} = \int f\psi_{j,k}, \qquad k \in \mathbf{Z}, j \in \mathbf{Z} \tag{4}$$

$$\beta_k = \int f\varphi_{\ell,k}, \qquad k \in \mathbf{Z} \tag{5}$$

The *homogeneous* wavelet reconstruction of f is

$$f = \sum_{j,k} \alpha_{j,k}\psi_{j,k} \tag{6}$$

and the *inhomogeneous* reconstruction formula for f is

$$f = \sum_{k=-\infty}^{\infty} \beta_k\varphi_{\ell,k} + \sum_{j\ge\ell}\sum_{k=-\infty}^{\infty} \alpha_{j,k}\psi_{j,k} \tag{7}$$

The functions ψ and φ are called the mother and father of the wavelets, respectively. The functions $\psi_{j,k}$ are the *wavelets*. To a certain extent, $\psi_{j,k}$ is "localized" at position $x = 2^{-j}k$ and frequency 2^j, while φ_k is "localized" at $2^{-\ell}k$ and occupies the frequency band $[-2^\ell, 2^\ell]$. Thus (β_k) represents the low frequency content of f near k and $\alpha_{j,k}$ represents the content in frequency space near 2^j at spatial position $2^{-j}k$.

In a more formal sense, if V_0 represents the L_2 span of (φ_k) and W_j the L_2 span of $\{\psi_{j,k}, k \in \mathbf{Z}\}$, then the homogeneous reconstruction formula represents a partition

$$L_2(\mathbb{R}) = \cdots \oplus W_{k-1} \oplus W_0 \oplus W_1 \oplus \cdots$$

into different frequency ranges, and the inhomogeneous reconstruction represents a partition

$$L_2(\mathbb{R}) = V_0 \oplus W_0 \oplus W_1 \oplus \cdots$$

into V_0, the low frequencies, and the various high frequency components.

In the definition of wavelet analysis, the property of compact support is not strictly speaking necessary; Meyer does not insist on it (exponential decay at ∞ would be sufficient). However, Daubechies (1988) has proven the existence of compactly supported wavelets of regularity r for each $r > 0$. Compact support is so frequently useful both for practical calculations and for mathematical proofs that we limit ourselves to that case in this paper.

Wavelet analysis has many desirable proerties but we mention two in particular. First, the representations (6)–(7) are true not just in $L_2(\mathbb{R})$ but in many spaces of locally integrable functions. For example the homogeneous reconstruction formula is valid in L_p, $1 < p < \infty$; and the inhomogeneous one is valid in L_p for $1 \leq p \leq \infty$. General orthogonal series decompositions often fail for L_p spaces outside a certain range (Askey and Wainger, 1965); in this sense wavelet reconstruction is much better behaved than general orthogonal series.

Second, the wavelet coefficients can be used to measure quite precisely the smoothness properties of a function. Consider first the local smoothness properties. Suppose we have an r-regular wavelet analysis, $r > 1$. Set $Q(j,k) = supp(\psi_{j,k})$. Jaffard (1989) points out that if f is locally Holderian at x_0, with exponent δ, then $\alpha_{j,k} = O(2^{-(1/2+\delta)j})$ for every sequence (j,k_j) with $j \to \infty$, $x_0 \in Q(j,k_j)$. Meyer (1990) points out that if f is differentiable at x_0 then $\alpha_{j,k} = o(2^{-3/2j})$ for every sequence (j,k_j) with $j \to \infty$, $x_0 \in Q(j,k_j)$. Moreover, both results have near-converses.

The wavelet coefficients also measure global smoothness quite well. Let $\Delta_h^{(r)} f$ denote the r-th difference $\sum_{k=0}^{r} \binom{r}{k} (-1)^k f(t+kh)$. The r-th modulus of smoothness of f in L_p is

$$w_{r,p}(f;h) = ||\Delta_h^{(r)} f||_p.$$

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

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

if $q < \infty$, and by

$$|f|_{B(\sigma,p,\infty)} = \sup_{h>0} \frac{w_{r,p}(f;h)}{h^\sigma}$$

if $q = \infty$.

This measure of smoothness includes, for various settings (σ, p, q), other commonly used measures. For example let C^δ be the set of functions with $|f(s) - f(t)| \leq c|s-t|^\delta$ for some $c > 0$. Then f has for a given $m = 0, 1, \ldots$ a distributional derivative $f^{(m)}$ satisfying

$f^{(m)} \in C^\delta$, $0 < \delta < 1$, if and only if $|f|_{B(m+\delta,\infty,\infty)} < \infty$. Similarly, f has a distributional derivative $f^{(m)}$ satisfying $f^{(m)} \in L_2$ iff $|f|_{B(m,2,2)} < \infty$. Finally, f belongs to B, the bump algebra, iff $|f|_{B(1,1,1)} < \infty$. See for example, Meyer (1990a) Chapter VI. It is a significant fact that the Besov seminorm is essentially a functional of the wavelet coefficients $(\alpha_{j,k})$.

Theorem 1 *Let an r-regular wavelet analysis be given. Define*

$$|\alpha|_{b(s,p,q)} = \left(\sum_{j=-\infty}^{\infty} \left(2^{js} \left(\sum_{k=-\infty}^{\infty} |\alpha_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q}$$

[with the standard interpretation if $q = \infty$]. Then with $\alpha = \alpha(f)$ we have

$$c|f|_{B(\sigma,p,q)} \le |\alpha|_{b(s,p,q)} \le C|f|_{B(\sigma,p,q)}$$

with $s = \sigma + 1/2 - 1/p$, where c, C depend on $(\psi, \varphi, p, q, r, \sigma)$.

See Meyer (1990) Page 197, Proposition 4. In our opinion, it could probably be derived from earlier work in atomic decompositions, in nonlinear approximation theory, and in Besov Spaces, predating the development of wavelets; in particular, Peetre (1976), Pietsch (1981), and especially Frazier and Jawerth(1985) and De Vore and Popov (1988). In some sense these earlier authors were working with expansions in terms of "wigglets", that is to say, wavelet-like expansions without the rigid definition and especially the orthogonality properties of wavelet analysis.

In sum, wavelet analysis gives us a transformation from continuous function space into a sequence space with two fundamental properties

(1) If $\hat{f}$ and f are two functions,

$$\begin{aligned} ||\hat{f} - f||^2 &= \sum_{j,k} (\hat{\alpha}_{j,k} - \alpha_{j,k})^2 \\ ||\hat{f} - f||^2 &= \sum (\hat{\beta}_k - \beta_k)^2 + \sum_{j \ge 0} \sum_{k=-\infty}^{\infty} (\hat{\alpha}_{j,k} - \alpha_{j,k})^2 \end{aligned}$$

so there is an exact *isometry* of the L_2 errors. This, of course, follows from the orthonormality of the wavelet basis.

(2) The function f satisfies $|f|_{B(\sigma,p,q)} \le A$ if $|\alpha|_{b(s,p,q)} \le A/c$, for appropriate c. On the other hand, the function f satisfies $|f|_{B(\sigma,p,q)} \le A$ only if $|\alpha|_{b(s,p,q)} \le CA$. Thus there is an *equivalence* (but not precise isometry) at the level of smoothness measures.

These two properties of the sequence transformation dominate all that follows. We will concern ourselves with recovery of functions from noisy data, and we will find that the sequence transformation enables a complete and natural treatment of the problems.

3 Correspondence between Estimation in Function Space and in Sequence Space

Suppose we observe sequence data

$$y_{j,k} = \theta_{j,k} + z_{j,k} \qquad j = 0, 1, 2, \ldots, \quad k = 0, \ldots, 2^j - 1. \tag{8}$$

where $z_{j,k}$ are i.i.d. $N(0, \epsilon^2)$ and $\theta = (\theta_{j,k})$ is unknown. We wish to estimate θ with small squared error loss $||\hat{\theta} - \theta||_2^2 = \sum(\hat{\theta}_{j,k} - \theta_{j,k})^2$. Although θ is in detail unknown, we do know that $||\theta||_{s,p,q} \leq 1$, where

$$||\theta||_{s,p,q} = \left(\sum_{j \geq 0} \left(2^{js} \left(\sum_{k=0}^{2^j - 1} |\theta_{j,k}|^p \right)^{1/p} \right)^q \right)^{1/q}. \tag{9}$$

Thus we have a problem of estimating θ when it is observed in a Gaussian white noise, and is known *a priori* to lie in a certain convex set $\Theta_{s,p,q}(C) \equiv \{\theta : ||\theta||_{s,p,q} \leq C\}$. We often are interested in the case $C = 1$, and put for short $\Theta_{s,p,q} = \Theta_{s,p,q}(1)$.

The difficulty of estimation in this setting is measured by the *minimax risk*

$$R_N^*(\epsilon; \Theta_{s,p,q}) = \inf_{\hat{\theta}} \sup_{\Theta_{s,p,q}} E||\hat{\theta} - \theta||_2^2 \tag{10}$$

and by the *minimax linear risk*

$$R_L^*(\epsilon, \Theta_{s,p,q}) = \inf_{\substack{\hat{\theta} \\ \text{linear}}} \sup_{\Theta_{s,p,q}} E||\hat{\theta} - \theta||_2^2 \tag{11}$$

where estimates are restricted to be linear.

Because of the wavelet isometry, there is a close connection between minimax estimation in this model and in the regression model (1) with which the paper began.

Theorem 2 Correspondence Theorem. *Let an r-regular wavelet analysis be given, $r \geq 0$. Suppose that the function class $\mathcal{F}$ may be written $\mathcal{F} = \mathcal{P} + \mathcal{H}$, where $\mathcal{P}$ is either $\{0\}$ or else the set of all polynomials of degree $\leq r$, and where $\mathcal{H}$ is the class of locally integrable functions $h(t)$ with wavelet coefficients bounded in the s, p, q seminorm:*

$$\mathcal{H} = \{h : h = \sum \alpha_{j,k} \psi_{j,k}, |\alpha|_{b(s,p,q)} \leq C\}.$$

Suppose in addition that $\mathcal{H} \subset BV(C')$ for some $C' > 0$ (e.g. suppose that $s > 1/2$, or that $p \leq 1$, $q \leq 1$ if $s = 1/2$).

Then if $n \to \infty$ along powers of 2,

$$R_N(n, \mathcal{F}) \sim R_N^*(\frac{\sigma}{\sqrt{n}}, \Theta_{s,p,q}(C)) \tag{12}$$

$$R_L(n, \mathcal{F}) \sim R_L^*(\frac{\sigma}{\sqrt{n}}, \Theta_{s,p,q}(C)) \tag{13}$$

Moreover, estimators attaining $R_N^*(\frac{\sigma}{\sqrt{n}}, \Theta)$ or $R_L^*(\frac{\sigma}{\sqrt{n}}, \Theta)$ in the sequence estimation problem may be used to attain $R_N(n, \mathcal{F})(1 + o(1))$ or $R_L(n, \mathcal{F})(1 + o(1))$ in the sampled-data problem.

The proof, which we defer to sections 7-9, establishes a close connection between the "empirical wavelet coefficients" of a sequence (y_i) obtained via (1) and the sequence data $(y_{j,k})$ obtained from (8), *provided* we make the calibration $\epsilon = \frac{\sigma}{\sqrt{n}}$.

First, however, we use the motivation provided by the Theorem and make, in sections 4-6, a detailed study of the behavior of minimax risks and minimax estimates in the sequence problem.

4 Minimax Estimation in Sequence Space: Rates

The general problem of minimax estimation in a sequence model when the object is known to lie in a compact subset of ℓ_2 has been discussed in Donoho, Liu, and MacGibbon (1990) (hereafter [DLM]). However, their notation and definitions are somewhat different. To use their results in the context of data (8) and prior information (9), we must define sequences (η_i) and (v_i) by lexicographic reordering elements of $(\theta_{j,k})$ and $(y_{j,k})$. Let $\eta_{2^j+k} = \theta_{j,k}$, $v_{2^j+k} = y_{j,k}$; we then have a problem of estimating (η_i) from white noise data (v_i).

The set $\Theta_{s,\infty,\infty}$ corresponds to a certain hyperrectangle $\{\eta : |\eta_i| \le \tau_i\}$, $\tau_{2^j+k} = 2^{-js}$; the set $\Theta_{s,2,2}$ corresponds to an ellipsoid $\sum a_i^2\eta_i^2 \le 1$ with $a_{2^j+k} = 2^{js}$; and more generally $\Theta_{s,p,p}$ corresponds to a set $\sum a_i^p|\eta_i|^p \le 1$, or ℓ_p-body.

From an intuitive point of view, prior information that η lies in a certain ℓ_p body means that for large index i, η_i must be a priori near 0, and hence that a good estimator will not use the raw data v_i as its estimate of η_i, but rather something "shrunk" towards the a priori likely value of 0. Linear shrinkers of the form c_iv_i with $c_i \in [0,1]$ are of course well known. The nonlinear shrinker $sgn(v_i)(|v_i| - \lambda_i)_+$ is less well-known, but has certain advantages.

[DLM] identify the following basic fact about estimation over ℓ_p-bodies: if $p \ge 2$, $R_L^*/R_N^* \le 1.25$; but if $p < 2$, $R_L^*(\epsilon)/R_N^*(\epsilon) \to \infty$, $\epsilon \to 0$. Indeed if $p < 2$, $R_L^*(\epsilon) \asymp (\epsilon^2)^{\frac{2s}{2s+1}}$ while $c(\epsilon^2)^{\frac{s+1/p-1/2}{s+1/p}} \le R_N^*(\epsilon) \le M^2(\epsilon)(\epsilon^2)^{\frac{s+1/p-1/2}{s+1/p}}$ with $M^2(\epsilon) = O(\log^2(\epsilon))$. In words, if $p \ge 2$, linear estimates can be within 25% of minimax among all estimates; if $p < 2$, however, linear estimates can be outperformed by nonlinear estimates even at the level of *rate* as $\epsilon \to 0$.

Underyling these results is an intuitive picture. The case $p < 2$ allows for a situation of *sparsity*: a very few η_i are comparatively large, and the rest of the η_i essentialy zero. Linear shrinkers cope poorly with such sparse situations. Nonlinear shrinkers like $sgn(v_i)(|v_i| - \lambda_i)_+$cope better; speaking informally, they identify the cases y_i which are likely to be large and estimate those coordinates in a relatively conventional way; all other cases are shrunk to zero.

Unfortunately for us, the $\Theta_{s,p,q}$ with $p \ne q$ are not ℓ_p-bodies, and so are not covered by the results of [DLM]. We are able to prove, in section 10 below, closely analogous results.

Theorem 3 *For any $s, p, q \ge 0$,*

$$R_L^*(\epsilon) \asymp (\epsilon^2)^{\frac{2s}{2s+1}}.$$

If p and q both are at least 2

$$\frac{R_L^*}{R_N^*} \le 1.25.$$

Theorem 4 *If $s, q > 0$, $0 < p < 2$.*

$$R_N^*(\epsilon) \asymp (\epsilon^2)^{\frac{s+1/p-1/2}{s+1/p}}$$

Taken together, these results show that for $p < 2$, nonlinear estimates can significantly outperform linear ones in the model (8)-(9).

In combination with the Correspondence Theorem, we get **Ph. A** of the introduction.

Corollary 1 *(Bump Algebra) Let* $\mathcal{F} = \{f : ||f||_B \leq C\}$. *Then for data (1),*

$$R_L(n, \mathcal{F}) \asymp n^{-1/2}$$

$$R_N(n, \mathcal{F}) \asymp n^{-2/3}$$

Indeed, the Bump Algebra coincides with the choice $r > 1$, $\mathcal{P} = \{0\}$, $(s, p, q) = (1/2, 1, 1)$.

Corollary 2 *(Total Variation) Let* $\mathcal{F} = \{f : TV(f) \leq C\}$. *Then for data (1),*

$$R_L(n, \mathcal{F}) \asymp n^{-1/2}$$

$$R_N(n, \mathcal{F}) \asymp n^{-2/3}$$

Indeed, we apply the correspondence theorem with $r = 0$ (the Haar Wavelet Basis), $\mathcal{P} = \{\text{constants}\}$; a simple computation shows that bounded variation is a superset of the case $(s, p, q) = (1/2, 1, 1)$ and is contained in $(s, p, q) = (1/2, 1, \infty)$.

We also see immediately that phenomenon **Ph. A** of the introduction holds more generally: *For Besov Spaces with index* $p < 2$, *nonlinear estimates can be essentially better than linear ones.* Indeed the corresponding fact is true for the sequence model, by Theorems 3 and 4; and the Correspondence Theorem lets us transfer the conclusion to the function estimation problem.

5 Minimax Risk and Minimax Estimator

In this section we study carefully the structure of minimax estimates in the sequence problem. By the Correspondence Theorem this information will carry over to the function estimation model. This will ultimately establish **Ph. B** of the introduction: the idea that the wavelet isometry maps us into a coordinate system where an asymptotically minimax estimator has a very simple structure.

5.1 Minimax Bayes Estimation

Consider the following *Minimax Bayes* estimation problem. We observe data according to the sequence model (8), only now $(\theta_{j,k})$ is a *random variable*, which may be arbitrary except for the single constraint that

$$||\tau||_{s,p,q} \leq 1$$

where

$$\tau_{j,k} = (E|\theta_{j,k}|^p)^{1/p}.$$

In short, we replace the "hard" constraint that $||\theta||_{s,p,q} \leq 1$ by the "in mean" constraint $||\tau||_{s,p,q} \leq 1$. We define the minimax Bayes risk

$$R_B^*(\epsilon; \Theta_{s,p,q}) = \inf_{\hat{\theta}} \sup_{\tau \in \Theta_{s,p,q}} E||\hat{\theta} - \theta||^2.$$

As "hard" constraints are more stringent than "in mean" constraints, $R_B^* \geq R_N^*$.

In this section, we develop two main results. First, we show that minimax estimators for R_B^* are *separable* nonlinearities.

Theorem 5 *Let $s, p, q > 0$, $q \geq p$. The minimax estimator for $R_B^*(\epsilon)$ has the form*

$$\hat{\theta}_{j,k}^* = \delta_j^*(y_{j,k}) \qquad j = 0, 1, \ldots, \quad k = 0, \ldots, 2^j - 1$$

where $\delta_j^*(y)$ is a scalar nonlinear function of the scalar y. In fact there is a 3-parameter family $\delta_{(r,\epsilon,p)}$ of nonlinear functions of y from which the minimax estimator is built:

$$\delta_j^* = \delta_{(r_j^*, \epsilon, p)} \qquad j = 0, 1, \ldots$$

for a sequence $(r_j^*)_{j=0}^{\infty}$ which depends on s, p, q, and ϵ.

Second, we show that R_B^* gives the exact asymptotics of R_N^*.

Theorem 6 *Let $s, p, q > 0$, $q \geq p$, $q \geq 1$, $p < 2$.*

$$R_N^*(\epsilon) = R_B^*(\epsilon)(1 + o(1)) \tag{14}$$

$$R_B^*(\epsilon) \sim \gamma(\epsilon) C^{2(1-r)} \epsilon^{2r} \tag{15}$$

where

$$r = \frac{s + 1/p - 1/2}{s + 1/p}$$

and

$$\gamma(\epsilon) = \gamma(\epsilon; s, p, q)$$

is a continuous periodic function of $\log_2 \epsilon$.

Combining Theorems 6 and 5, the estimator $\hat{\theta}^*$ is *asymptotically* minimax for R_N^* as $\epsilon \to 0$. In short: *a separable nonlinear rule is asymptotically minimax.*

The proof of these results is not primarily a technical matter; instead, it relies on a variety of concepts which we introduce and develop in the subsections below.

5.2 Minimax Bayes Risk for a Mean With Bounded p-th Moment

Consider now a very special problem. We observe

$$v = \eta + z \tag{16}$$

where η is a random variable, and z is independent of η with distribution $N(0, \epsilon^2)$. We do not know the distribution F of η, but we do know that η satisfies $(E_F|\eta|^p)^{1/p} \leq \tau$. We wish to estimate η with small squared-error loss. Define the minimax Bayes risk

$$\rho_p(\tau, \epsilon) = \inf_{\delta} \sup_{(E_F|\eta|^p)^{1/p} \leq \tau} E_F E_\eta (\delta(y) - \eta)^2. \tag{17}$$

This quantity has been analyzed in Donoho and Johnstone (1990), hereafter [DJ]. There we find that ρ_p satisfies the invariance

$$\rho_p(\tau, \epsilon) = \epsilon^2 \rho_p(\tau/\epsilon, 1) \tag{18}$$

and the asymptotic relation

$$\rho_p(\tau, 1) \sim \begin{cases} \tau^2 & p \geq 2 \\ \tau^p (2\log(\tau^{-p}))^{\frac{2-p}{2}} & p < 2 \end{cases}$$

as $\tau \to 0$. The function ρ_p is continuous, is monotone increasing in τ, is concave in τ^p and has $\rho_p(\tau, \epsilon) \to \epsilon^2$ as $\tau/\epsilon \to \infty$.

There exists a rule $\delta_{(\tau,\epsilon,p)}$ which is minimax for $\rho_p(\tau, \epsilon)$; it is odd, monotone, and satisfies the invariance $\delta_{(\tau,\epsilon,p)}(y) = \epsilon \delta_{(\tau/\epsilon,1,p)}(y/\epsilon)$.

[DJ] also consider a vector version of this problem. Suppose we observe n observations according to (16), i.e.,

$$v_i = \eta_i + z_i \qquad i = 1, \ldots, n \tag{19}$$

with z_i iid $N(0, \epsilon^2)$, (η_i) random, with distribution π, (η_i) independent of (z_i), and $(\Sigma E_\pi |\eta_i|^p)^{1/p} \leq n^{1/p} r$ [i.e., $\mathrm{Ave}_{1 \leq i \leq n} E_\pi |\eta_i|^p \leq r^p$]. Let $\Theta_{p,n}(r) = \{\theta : \Sigma_{i=1}^n |\theta_i|^p \leq r^p\}$ denote the n-dimensional l_p ball of radius r; then for the minimax vector Bayes risk

$$R_B^*(\epsilon; \Theta_{p,n}(r)) = \inf_{\hat{\eta}} \sup \{E_\pi E_\eta ||\hat{\eta} - \eta||_2^2 : \; E_\pi \sum |\eta_i|^p \leq n r^p\}.$$

Setting $\tau_i(\pi) = (E_\pi |\eta_i|^p)^{1/p}$ we see that the moment constraint is equivalent to requiring that $\tau(\pi) \in \Theta_{p,n}(r)$. We have the formula (proved in [DJ])

$$R_B^*(\epsilon; \Theta_{p,n}(n^{1/p} r)) = n \rho(r, \epsilon). \tag{20}$$

This is an expression of the fact that the least favorable prior for η makes the η_i i.i.d., in which situation the problem becomes a product of independent problems, and the risk per coordinate becomes that in the scalar problem, i.e., $\rho(r, \epsilon)$.

5.3 Minimax Bayes Risk Over Cartesian Products

Return now to the problem of estimating $(\theta_{j,k})$ from data $y_{j,k} = \theta_{j,k} + z_{j,k}$ $j = 0, \ldots,$ $k = 0, \ldots, 2^j - 1$ containing white noise $(z_{j,k})$. However, instead of the constraint $(\theta_{j,k}) \in \Theta_{s,p,q}$, we now consider the constraint that the moment sequence lies in a *Cartesian product* set. Given a sequence $r = (r_0, r_1, r_2, \ldots)$ of positive numbers, we define such a set by

$$\Theta_{p,\infty}(r) = \left\{ (\theta_{j,k}) : \sum_{k=0}^{2^j-1} |\theta_{j,k}|^p \leq 2^j r_j^p \right\}.$$

Alternatively, letting $\Theta_{p,n}(r)$ denote an n-dimensional l_p ball as before, the set has the representation

$$\begin{aligned} \Theta_{p,\infty}(r) \equiv \Theta_{p,1}(r_0) &\times \Theta_{p,2}(2^{1/p} r_1) \times \cdots \\ &\times \Theta_{p,2^j}(2^{j/p} r_j) \times \cdots \end{aligned}$$

as a product of l_p-balls of increasing dimensionality. Such product sets have two useful properties:

(1) *Additivity of component risks*:

$$R_B^*(\epsilon; \Theta_{p,\infty}(r)) = \sum_{j=0}^{\infty} R_B^*(\epsilon; \Theta_{p,2^j}(2^{j/p} r_j));$$

this is simply because the estimation problems at different levels j are logically independent and become stochastically so as well, under the least favorable prior. Hence we arrive at the formula

$$R_B^*(\epsilon; \Theta_{p,\infty}(r)) = \sum_{j=0}^{\infty} 2^j \rho_p(r_j, \epsilon); \tag{21}$$

in principle, an exact expression for the minimax Bayes risk.

(2) *Separability of minimax rules.* The rule which is minimax Bayes for $\Theta_{p,\infty}(r)$ is

$$\hat{\theta}_{j,k} = \delta_j(y_{j,k}) \qquad j = 0, 1, 2, \ldots \quad k = 0, \ldots, 2^j - 1 \tag{22}$$

where $\delta_j(y) = \delta_{(r_j,\epsilon,p)}(y)$, with $\delta_{(r,\epsilon,p)}$ the minimax rule for the Bounded-Moment problem of Section 5.2 .

As a result of these two properties, we may find ourselves wishing that we could study estimation over sets $\Theta_{p,\infty}(r)$ rather than $\Theta_{s,p,q}$. There is a sense in which we *can*. Suppose that $\Theta_{p,\infty}(r)$ is a subset of a set Θ. The problem of estimating $(\theta_{j,k})$ from data (8) with prior π and $\tau(\pi)$ known to lie in $\Theta_{p,\infty}(r)$ is called a *Cartesian subproblem* of the full problem of estimation (when τ is only known to lie in Θ).

Theorem 7 *Let $s > 0$ and $q \geq p$. The difficulty of the full problem is equal to the difficulty of the hardest Cartesian subproblem. In symbols:*

$$R_B^*(\epsilon; \Theta_{s,p,q}) = \max\{R_B^*(\epsilon; \Theta_{p,\infty}(r)) : \Theta_{p,\infty}(r) \subset \Theta_{s,p,q}\},$$

the maximum being attained by some sequence $(r_0^*, r_1^*, \ldots)$. The estimator

$$\hat{\theta}_{j,k} = \delta_{(r_j^*, \epsilon, p)}(y)$$

which is minimax Bayes for this Cartesian subproblem is minimax Bayes for the full problem.

Note that this Theorem implies Theorem 5. It is proved in section 10.

5.4 Dyadic Renormalization

Theorem 7 reduces the problem of computing $R_B^*(\epsilon; \Theta_{s,p,q})$ to an optimization problem. By formula (21) we have $R_B^*(\epsilon, \Theta_{s,p,q}) = \mathrm{val}(P_{\epsilon,1})$ where $(P_{\epsilon,C})$ denotes the optimization problem

$$(P_{\epsilon,C}) \quad \sup \sum_{j=0}^{\infty} 2^j \rho_p(r_j, \epsilon) \text{ subject to } \sum_{j=0}^{\infty} (2^{sj}(2^j r_j^p)^{1/p})^q \leq C^q$$

(with obvious reformulation if $q = \infty$).

At first glance, solution of this problem would appear to be beyond reach, owing to the fact that we have no closed form expression for $\rho_p(\tau, \epsilon)$ when $p \neq 2$. However, a certain "renormalizability" of the problem provides a tool to get qualitative insights.

Define the following optimization problem $(Q_{\epsilon,C})$ on the space of bilateral sequences $\mathcal{S} = \{(r_j)_{j=-\infty}^{\infty}\}$

$$(Q_{\epsilon,C}) \quad \sup \sum_{j=-\infty}^{\infty} 2^j \rho_p(r_j, \epsilon) \text{ subject to } \sum_{j=-\infty}^{\infty} (2^{sj}(2^j r_j^p)^{1/p})^q \leq C^q$$

This problem is obviously closely related to $(P_{\epsilon,C})$. If the unilateral sequence $(r_j)_{j=0}^{\infty}$ is feasible for the discrete problem $(P_{\epsilon,C})$ then the extension to a bilateral sequence $(\tilde{r}_j)$ defined by setting $\tilde{r}_j = 0$, $j < 0$ and $\tilde{r}_j = r_j$, $j > 0$, is feasible for the bilateral problem $(Q_{\epsilon,C})$. We conclude that

$$\mathrm{val}(P_{\epsilon,C}) \leq \mathrm{val}(Q_{\epsilon,C}) \qquad \forall \epsilon > 0, \ C > 0.$$

On the other hand, if the bilateral sequence (r_j) is feasible for $(Q_{\epsilon,C})$ then the unilateral sequence $\tilde{r}_j$ formed by dropping the $j < 0$ portion from (r_j) is feasible for $(P_{\epsilon,C})$. Moreover, the part of the objective function which is lost in dropping the negative indices is at most ϵ^2, since $\rho_p(r_j, \epsilon) \leq \epsilon^2$ implies $\sum_{j<0} 2^j \rho(r_j, \epsilon) \leq \epsilon^2$. Hence

$$\mathrm{val}(Q_{\epsilon,C}) \leq \mathrm{val}(P_{\epsilon,C}) + \epsilon^2 \qquad \forall \epsilon > 0, \ C > 0.$$

We know of course that a discrepancy of order ϵ^2 between the value of the two problems is asymptotically negligible. Hence $\mathrm{val}(P_{\epsilon,C}) \sim \mathrm{val}(Q_{\epsilon,C})$, as $\epsilon \to 0$.

Here are the asymptotics of $\mathrm{val}(Q_{\epsilon,C})$.

Theorem 8

$$\mathrm{val}(Q_{\epsilon,C}) = \gamma(\epsilon, C) C^{2(1-r)} \epsilon^{2r} \tag{23}$$

where

$$r = \frac{1+\beta-q/2}{1+\beta}$$

and $\gamma(\epsilon; C)$ *is a continuous, periodic function of* $\log_2 \epsilon$.

The immediate implication of this is that $R_N^*(\epsilon) \leq \mathrm{const}\, \epsilon^{2r}$. Indeed,

$$\begin{aligned} R_N^*(\epsilon) \leq R_B^*(\epsilon) &= \mathrm{val}(P_{\epsilon,1}) \\ &\leq \mathrm{val}(Q_{\epsilon,1}) \\ &\leq \mathrm{const}\ \epsilon^{2r}. \end{aligned}$$

The Theorem follows from a certain homogeneity with respect to scaling and translation of the functionals involved. Let $\rho(v) \equiv rho_p(\epsilon, 1)$, and recall the invariance (18). Set $\beta = q(s - 1/p)$. Define

$$\begin{aligned} J_{\rho,\epsilon}(r) &= \epsilon^2 \sum_{-\infty}^{\infty} 2^j \rho(r_j/\epsilon) \\ J_{q,\beta}(r) &= \sum_{i\infty}^{\infty} 2^{j\beta} r_j^q. \end{aligned}$$

Moreover, let $(\mathcal{U}_{a,k} r)j = a r_{j-k}$. Then by a simple change of variables

$$J_{\rho,\epsilon}(\mathcal{U}_{\epsilon,k} r) = \epsilon^2 2^k J_{\rho,1}(r). \tag{24}$$

Also

$$J_{q,\beta}(\mathcal{U}_{a,k} r) = a^q 2^{\beta k} J_{q,\beta}(r). \tag{25}$$

As $(Q_{\epsilon,C})$ is the problem of optimizing $J_{\epsilon,\rho}$ subject to $J_{q,\beta} \leq C^q$, these scaling relations imply at once that if (r_j) is a solution to $(Q_{1,\Gamma})$ then the renormalized function $\tilde{r}_j = \epsilon r_{j-k}$ is a solution to $(Q_{\epsilon,C})$, with

$$\Gamma = \frac{C}{\epsilon 2^{s-1/p}} \tag{26}$$

In turn, this renormalization implies (23).

We now show why. Let $\mathcal{R}_C$ denote the set of sequences feasible for $(Q_{\epsilon,C})$, i.e. the (r_j) with $J_{q,\beta}(r) \leq C^q$. Then, it follows from (25) with $a = \epsilon$ that

$$\mathcal{U}_{\epsilon,k} \mathcal{R}_\Gamma = \mathcal{R}_C$$

and one may also see that

$$\mathcal{U}_{\epsilon^{-1},-k} \mathcal{R}_C = \mathcal{R}_\Gamma.$$

Hence

$$\begin{aligned} \sup\{J_{\rho,\epsilon}(r) : J_{q,\beta}(r) \leq C^q\} &= \sup\{J_{\rho,\epsilon}(\mathcal{U}_{\epsilon,k} r) : \mathcal{U}_{\epsilon,k} r \in \mathcal{R}_C\} \\ &= \sup\{\epsilon^2 2^k J_{\rho,1}(r) : r \in \mathcal{R}_\Gamma\} \\ &= \epsilon^2 2^k \sup\{J_{\rho,1}(r) : r \in \mathcal{R}_\Gamma\} \end{aligned}$$

In other words

$$\mathrm{val}(Q_{\epsilon,C}) = \epsilon^2 2^k \,\mathrm{val}(Q_{1,\Gamma}).$$

In particular, we note that if $x = k + \delta$, k an integer, $\delta \in [0,1)$,

$$\mathrm{val}(Q_{1,2^{s\beta/q}}) = 2^k \,\mathrm{val}(Q_{1,2^{\delta\beta/q}}). \tag{27}$$

Now define

$$k(\epsilon, C) = \lfloor \log_2(C/\epsilon) \cdot \frac{q}{\beta} \rfloor \quad \in \mathbf{Z}$$

$$\delta(\epsilon, C) = \log_2(C/\epsilon) \cdot \frac{q}{\beta} - k(\epsilon, C) \quad \in [0,1)$$

The definition (26) then yields

$$\Gamma = 2^{\delta(\epsilon,C)\beta/q} \quad \in [1, 2^{\beta/q}).$$

Then

$$\begin{aligned} \mathrm{val}(Q_{\epsilon,C}) &= \epsilon^2 2^k \,\mathrm{val}(Q_{1,\Gamma}) \\ &= \epsilon^2 (C/\epsilon)^{q/\beta} 2^{-\delta(\epsilon,C)} \,\mathrm{val}(Q_{1,\Gamma}). \end{aligned}$$

Putting

$$\gamma(\epsilon, C) = 2^{-\delta(\epsilon,C)} \,\mathrm{val}(Q_{1,\Gamma})$$

and noting (27), we see that γ is a function of $\delta(\epsilon, C)$ and hence a periodic function of $\log_2 \epsilon$ for each fixed C. Continuity follows from:

Lemma 1 *The supremum of $\sum_{-\infty}^{\infty} 2^j \rho(r_j)$ over the class R_Γ of nonnegative sequences satisfying $\sum_{-\infty}^{\infty} 2^{j\beta} r_j^q \le \Gamma$ is attained within the subclass D_Γ of decreasing sequences. The class of sequences $\rho(D_\Gamma) = \{(\rho(r_j)) : r \in D_\Gamma\}$ is a compact subset of l_1. Consequently, the maximum $\sum_{-\infty}^{\infty} \rho(r_j)\, dt$ over $r \in R_\Gamma$ is finite, and the maximum is attained by some $r \in D_\Gamma$. $\sup_{0<r<\infty} |\rho((1-\delta)r) - \rho(r)| \to 0$ as $\delta \to 0$; consequently the maximum value of $J_{1,\rho}$ over D_Γ is continuous in Γ.*

The lemma is proved in the appendix.

5.5 Minimax Risk Over ℓ_p-balls

At this point we have completed the proof of the main results announced in Section 5.1, except for the relation (14), i.e. $R_B^*(\epsilon)/R_N^*(\epsilon) \to 1$ as $\epsilon \to 0$. We now establish this asymptotic equivalence.

Consider (yet) another estimation problem. We observe

$$v_i = \eta_i + z_i, \qquad i = 1, \ldots, n$$

where the z_i are i.i.d. $N(0, \epsilon^2)$ and the η_i are no longer random variables, but instead unknown constants, satisfying $(\eta_i) \in \Theta_{n,p}(r)$. We wish to estimate (η_i) in a minimax fashion with respect to squared-error loss. Define

$$R_N^*(\epsilon; \Theta_{n,p}(r)) = \inf_{\hat{\eta}} \sup_{\eta \in \Theta_{n,p}(r)} E_\eta ||\hat{\eta} - \eta||_2^2.$$

This problem has been studied in [DJ]. While evidently $R_N^*(\epsilon; \Theta_{n,p}(r)) \le R_B^*(\epsilon; \Theta_{n,p}(r))$, the two quantities are often not significantly different.

Theorem 9 *[DJ] Let* $r_n = r_0 n^{1/p}$.

$$\frac{R_B^*(1;\Theta_{n,p}(r_n))}{R_N^*(1;\Theta_{n,p}(r_n))} \to 1 \qquad n \to \infty. \tag{28}$$

To apply this fact about ℓ_p-balls to the study of sets $\Theta_{s,p,q}$, suppose we let (ϵ_k) be a sequence of positive numbers tending to zero according to $k(\epsilon_k, C) = k$, $\delta(\epsilon_k, C) = \Gamma$, independent of k. In detail,

$$\epsilon_k = \frac{C}{\Gamma} 2^{-k(\beta/q)}$$

We will show that

$$R_B^*(\epsilon_k)/R_N^*(\epsilon_k) \to 1 \tag{29}$$

as $k \to \infty$.

Let $r^{(0)}$ be a solution to the optimization problem (R_Γ); then $r^{(0)} \in \mathcal{D}_\Gamma$; it has the representation $r_j^{(0)} = (r_j)$ for a certain bilateral sequence $(r_j)_{j=-\infty}^{\infty}$. Then define the unilateral sequence

$$r_j^{(k)} = \epsilon_k r_{j-k}^{(0)} \quad j = 0, 1, 2, \ldots.$$

Consequently

$$\sum_{j=0}^{\infty} 2^j \rho(r_j^{(k)}/\epsilon_k) = \sum_{j \geq 0} 2^j \rho(r_{j-k}^{(0)}),$$

and

$$R_B^*(\epsilon_k; \Theta_{s,p,q}) \sim \epsilon^2 \sum_{j \geq 0} 2^j \rho(r_{j-k}^{(0)}), \qquad k \to \infty.$$

Define

$$\lambda_{j,\ell} = \frac{R_N^*(1;\Theta_{p,2^j}(r_\ell 2^{j/p}))}{R_B^*(1;\Theta_{p,2^j}(r_\ell 2^{j/p}))}.$$

By Theorem 9 above $\lambda_{j,\ell} \to 1$ for ℓ fixed, $j \to \infty$. Now as $\Theta_{\rho,\infty}(r^{(k)}) \subset \Theta_{s,p,q}$,

$$R_N^*(\epsilon_k; \Theta_{s,p,q}) \geq R_N^*(\epsilon_k; \Theta_{p,\infty}(r^{(k)}))$$

but by definition of $\lambda_{j,\ell}$

$$R_N^*(\epsilon_k; \Theta_{p,\infty}(r^{(k)})) = \epsilon_k^2 \sum_{j \geq 0} 2^j \lambda_{j,j-k} \rho(r_{j-k}).$$

Changing variables $\ell = j - k$, $j = k + \ell$, we therefore get that

$$\lim_{k \to \infty} \frac{R_B^*(\epsilon_k; \Theta_{s,p,q})}{R_N^*(\epsilon_k; \Theta_{s,p,q})} \leq \lim_{k \to \infty} \frac{\sum_{\ell=-k}^{\infty} 2^\ell \rho(r_\ell)}{\sum_{\ell=-k}^{\infty} 2^\ell \lambda_{\ell+k,\ell} \rho(r_\ell)}$$

By Fatou's Lemma and Monotone Convergence both the top and bottom of this ratio converge to $\sum_{\ell=-\infty}^{\infty} 2^\ell \rho(r_\ell) = \mathrm{val}(R_\Gamma) \in (0, \infty)$. Hence the indicated limit is 1.

Suppose now that (ϵ_k) is an arbitrary sequence, with associated scaling factors $(\Gamma_k) \subset [1, 2^{\beta/q})$. This sequence has accumulation points. Suppose w.l.o.g. that the sequence actually has a limit, Γ_0, and without essential loss of generality that the limit is strictly in

the interior of $(1, 2^{\beta/q})$. Given $\delta > 0$ we can construct two sequences $(\epsilon_k^{(1)})$ and $(\epsilon_k^{(2)})$ which are subsequences of dyadically scaled sequences with base scales $\Gamma_0^{(1)}$ and $\Gamma_0^{(2)}$ satisfying $\Gamma^{(1)} \leq \Gamma_0 \leq \Gamma^{(2)}$, $|\Gamma^{(1)} - \Gamma^{(2)}| \leq \delta$ and such that for sufficiently large k, $\epsilon_k^{(1)} \leq \epsilon_k \leq \epsilon_k^{(2)}$. By monotonicity of minimax risk in the noise level,

$$\lim_{k\to\infty} \frac{R_B^*(\epsilon_k)}{R_N^*(\epsilon_k)} \leq \lim_{k\to\infty} \frac{R_B^*(\epsilon_k^{(2)})}{R_N^*(\epsilon_k^{(1)})} = \frac{\mathrm{val}(R_{\Gamma^{(2)}})}{\mathrm{val}(R_{\Gamma^{(1)}})}.$$

By continuity of $\mathrm{val}(R_\Gamma)$ (see Lemma 1) this ratio can be made as close to 1 as desired by picking δ small enough. Hence (14) holds.

This completes the proof of Theorems 5 and 6.

6 Near-Minimaxity of Threshold Estimates.

We have derived an asymptotically minimax estimator for $\Theta_{s,p,q}$ built out of coordinatewise nonlinearities from the family $\delta_{(r,\epsilon,p)}$. Unfortunately, these nonlinearities are not available to us in closed form. In this section we show that simple "threshold" nonlinearities provide near-minimax behavior when $q \geq p$. We consider two possibilities: first, the "soft" nonlinearity

$$\delta_\lambda(y) = \mathrm{sgn}(y)(|y| - \lambda)_+$$

which is continuous and Lipschitz; second, the "hard" nonlinearity $\delta_\mu(y) = y1_{\{|y|\geq\mu\}}$ which is discontinuous. [We adopt the convention that δ refers to a scalar nonlinearity whose type depends on the lexicography of the subscript: (ϵ, r, p), λ, and μ referring to different nonlinearities.]

Suppose we are in the Minimax-Bayes model of Section 5.1, so our data are $y_{j,k} = \theta_{j,k} + z_{j,k}$ with $\theta_{j,k}$ random variables satisfying the moment constraint $\tau \in \Theta_{s,p,q}$. Consider the use of separable estimators built out of thresholds, i.e. set $\lambda = (\lambda_{j,k})$ and

$$\hat{\theta}^\lambda_{j,k} = \delta_{\lambda_{j,k}}(y_{j,k}) \qquad j = 0, 1, 2, \ldots; \quad k = 0, \ldots, 2^j - 1.$$

The minimax risk among soft-threshold estimates is defined

$$R^*_\lambda(\epsilon, \Theta) = \inf_{(\lambda_{j,k})} \sup_{\tau\in\Theta} E||\hat{\theta}^\lambda - \theta||_2^2.$$

For hard thresholds $\hat{\theta}^\mu_{j,k} = \delta_{\mu_{j,k}}(y_{j,k})$, the minimax risk $R^*_\mu(\epsilon, \Theta)$ is defined similarly. The minimax risk among all estimates is of course $R^*_B(\epsilon; \Theta)$. In the subsections to follow, we develop results which establish

Theorem 10 *Let $q \geq p > 0$. There are constants $\Lambda(p)$, $M(p)$, both finite, with*

$$\begin{aligned} R^*_\lambda(\epsilon, \Theta_{s,p,q}) &\leq \Lambda(p) R^*_B(\epsilon, \Theta_{s,p,q}) \\ R^*_\mu(\epsilon, \Theta_{s,p,q}) &\leq M(p) R^*_B(\epsilon, \Theta_{s,p,q}) \end{aligned}$$

There exist thresholds which attain these performances; they have the form

$$\lambda_{j,k} = \epsilon \cdot \ell(\epsilon, r^\lambda_j, p) \qquad j = 0, 1, \ldots; \quad k = 0, \ldots, 2^j - 1.$$

and

$$\mu_{j,k} = \epsilon m(\epsilon, r^\mu_j, p) \qquad j = 0, 1, \ldots; \quad k = 0, \ldots, 2^j - 1$$

for certain functions ℓ and m and certain sequences r^λ and r^μ such that $\Theta_{p,\infty}(r^\lambda) \subset \Theta_{s,p,q}$, $\Theta_{p,\infty}(r^\mu) \subset \Theta_{s,p,q}$.

In short, with optimal choice of threshold, we can obtain near-minimax behavior relative to r. We remark that $\Lambda(1) \leq 1.6$, so the near-minimaxity is numerically effective.

Finally, by (14), these estimates are within a factor $\Lambda(p)$ (resp. $M(p)$) of being asymptotically minimax for the frequentist criterion $R^*_N(\epsilon)$.

6.1 Minimax Bayes, Bounded p-th Moment (Encore).

Return briefly to the scalar situation of Section 5.2 ,with $v = \eta + z$, η random, z independent of η and $N(0, \epsilon^2)$. We are interested in estimating η with squared-error loss. The distribution of η is known to satisfy $E|\eta|^p \leq \tau^p$, and the minimax Bayes risk is by definition $\rho_p(\tau, \epsilon)$.

To measure the performance of threshholds in this situation, we define

$$\rho_{\lambda,p}(\tau, \epsilon) = \inf_{\lambda \in [0,\infty]} \sup_{(E|\eta|^p)^{1/p} \leq \tau} E(\delta_\lambda(y) - \eta)^2$$

and

$$\rho_{\mu,p}(\tau, \epsilon) = \inf_{\mu \in [0,\infty]} \sup_{(E|\eta|^p)^{1/p} \leq \tau} E(\delta_\mu(y) - \eta)^2;$$

under our typographical convention, these are worst case risks for soft (λ) and hard (μ) thresholds, respectively.

To compare these performances with the Bayes Minimax estimates we define

$$\Lambda(p) \equiv \sup_{\tau,\epsilon} \frac{\rho_{\lambda,p}(\tau, \epsilon)}{\rho_p(\tau, \epsilon)} < \infty.$$

and

$$M(p) \equiv \sup_{\tau,\epsilon} \frac{\rho_{\mu,p}(\tau, \epsilon)}{\rho_p(\tau, \epsilon)} < \infty.$$

[DJ] show that for $p \in (0, \infty]$, $\Lambda(p) < \infty$ and $M(p) < \infty$, In short, the minimax δ_λ is within a factor $\Lambda(p)$ of minimax, and the minimax δ_μ is within a factor $M(p)$ of minimax.

In fact, $\Lambda(p)$ and $M(p)$ are both smaller than 2.22 for all $p \geq 2$; and computational experiments indicate $\Lambda(1) \leq 1.6$. Quantitatively, $\Lambda(p)$ tends to be somewhat smaller than $M(p)$, which says that "soft" thresholding offers a quantitative superiority. (Compare the conclusions of Bickel (1983) in a different Bayes-minimax problem).

6.2 Hardest Cartesian Subproblems for Thresholds.

Return now to the sequence experiment. The problem of estimating θ when the moment vector τ is known to lie in a Cartesian product $\Theta_{p,\infty}(r) \subset \Theta$, is called a *Cartesian Subproblem* of Θ. For such subproblems we have the formula

$$R^*_\lambda(\epsilon, \Theta_{p,\infty}(r)) = \sum_{j \geq 0}^{\infty} 2^j \rho_{\lambda,p}(r_j, \epsilon)$$

expressing the worst-case risk in the infinite-dimensional problem in terms of one-dimensional worst-case risks, and similarly

$$R^*_\mu(\epsilon, \Theta_{p,\infty}(r)) = \sum_{j \geq 0}^{\infty} 2^j \rho_{\mu,p}(r_j, \epsilon).$$

We have the following analog of Theorem 19.

Theorem 11 *Let $q \geq p$, $s, p > 0$. The difficulty, for soft threshold estimates, of the full problem, is equal to the difficulty, for soft threshold estimates, of the hardest Cartesian subproblem:*

$$R^*_\lambda(\epsilon; \Theta) = \sup\{R^*_\lambda(\epsilon, \Theta_{p,\infty}(r)) : \Theta_{p,\infty}(r) \subset \Theta\}.$$

The supremum is attained by a sequence (r^λ_j), and a soft threshold estimator with thresholds

$$\lambda_{j,k} = \epsilon \cdot \ell(r^\lambda_j, \epsilon, p)$$

is minimax among soft thresholds for a certain function $\ell(r, \epsilon, p)$. Similarly

$$R^*_\mu(\epsilon; \Theta) = R^*_\mu(\epsilon; \Theta_{p,\infty}(r^\mu))$$

for a sequence (r^μ_j), and a hard threshold estimator with thresholds

$$\mu_{j,k} = \epsilon \cdot m(r^\mu_j, \epsilon, p)$$

is minimax among hard thresholds, for some function $m(\tau, \epsilon, p)$.

Hence, the "hardest Cartesian subproblems" heuristic works in this case as well. The proof is given in the appendix. Theorem 10 follows directly:

$$\begin{aligned} R^*_\lambda(\epsilon, \Theta_{s,p,q}) &= R^*_\lambda(\epsilon; \Theta_{p,\infty}(r^\lambda)) \\ &= \sum_{j \geq 0} 2^j \rho_{\lambda,p}(r^\lambda_j, \epsilon) \\ &\leq \Lambda(p) \sum_{j \geq 0} 2^j \rho_p(r^\lambda_j, \epsilon) \\ &= \Lambda(p) R^*_B(\epsilon; \Theta_{p,\infty}(r^\lambda)) \\ &\leq \Lambda(p) R^*_B(\epsilon; \Theta_{s,p,q}) \end{aligned}$$

and similarly for μ-thresholds.

7 Function Estimation in White Noise

At this point, we have a rather complete understanding of minimax and near-minimax estimation in the sequence model. We now turn to the Correspondence Theorem. Once this is established, **Ph. B** of the introduction will follow, and the interpretation of that result will lead to **Ph. C**, completing the three major aims of the paper.

If our attitude towards the results of this paper were simply that they represent asymptotic results of potential practical use, we would relegate the proof of the correspondence theorem to an appendix and bring the body of the paper to a speedy conclusion. However, we have implemented our proposed estimator in computer software; it is algorithmically efficient and gives appealing results. We have found that the issues in the correspondence theorem itself are those which arise in practical implementation of an empirical wavelet transform. Therefore, we spend the next three sections developing the idea that results in the sequence problem may be used for smoothing of noisy data.

There are three main questions which we will address in the process.

(Q1) Membership of a function f in Besov space is determined by the seminorm $|\alpha|_{b(s,p,q)}$, which is bilateral in j and k; membership of a sequence θ in $\Theta_{s,p,q}$ is defined by the norm $||\theta||_{s,p,q}$, which is unilateral in j and finite in k. Why the apparent discrepancy?

(Q2) Wavelets are designed for functions on the whole line; data (1) is restricted to the unit interval $[0,1]$. How do we use the wavelet transform for problems with a boundary?

(Q3) The wavelet transform of a function f requires the calculation of integrals $\int f\psi_{j,k}$. However, the discrete data (1) admit, at best, noisy Riemann sum approximations to such integrals. How do we use the wavelet transform for sampled data?

In this section we consider a different kind of estimation problem which allows us to understand (Q1) and (Q2) fully - the *white noise model.* Suppose we observe the stochastic process $Y(t)$, $t \in [0,1]$ where the process Y is characterized by

$$Y(dt) = f(t)\,dt + \epsilon W(dt) \qquad t \in [0,1] \tag{30}$$

with W a standard Wiener process, and f the function of interest. We wish to estimate f on the basis of these data and the *a priori* information that $f \in \mathcal{F}$ a convex class of functions. We use squared-error loss, and are interested in

$$R_N^*(\epsilon;\mathcal{F}) = \inf_{\hat{f}} \sup_{\mathcal{F}} E||\hat{f} - f||_2^2 \tag{31}$$

as well as

$$R_L^*(\epsilon:\mathcal{F}) = \inf_{\hat{f}\ \text{linear}} \sup_{\mathcal{F}} E||\hat{f} - f||_2^2. \tag{32}$$

This type of problem is called "function estimation in white noise". It can be related to data (1) as follows. Observing data(1) is evidently equivalent to observing

$$\begin{aligned} Y_n(t) &= \frac{1}{n}\sum_{t_i \le t} y_i \\ &= \frac{1}{n}\sum_{t_i \le t} f(t_i) + \frac{1}{n}\sum_{t_i \le t} z_i \end{aligned}$$

With $\epsilon = \frac{\sigma}{\sqrt{n}}$ the process Y_n is visibly a Riemann sum approximation to Y. (Q3) is in some sense about the adequacy of this approximation. By focusing on data Y in this section we implicitly take an affirmative answer to (Q3) for granted and separate the study of questions (Q1) and (Q2) from (Q3). The sampling issues associated with the process (Q3) will be addressed in section 9.

7.1 Functions of Bounded Variation

Suppose that $\mathcal{F}$ is the class of functions f supported in $[0,1]$ and of total variation $TV(f) \leq 1$.

The *Haar basis* is then the appropriate wavelet basis for this case. Let $\varphi = 1_{[0,1]}$, and $\psi(t) = 1_{[1/2,1]} - 1_{[0,1/2]}$. Define $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, $j = 0,1,2,\ldots,$ $k = 0,\ldots,2^j - 1$. Let $f \in L_2[0,1]$ and put

$$\beta_0 = \int \varphi_0 f, \qquad \theta_{j,k} = \int \psi_{j,k} f.$$

Then the inhomogeneous wavelet reconstruction formula gives, in this case $f = \beta_0 + \sum_{j\geq 0}\sum_{k=0}^{2^j-1}\theta_{j,k}\psi_{j,k}$ (convergence in L_2). This is the wavelet representation of f in the Haar basis.

Consider now the data

$$\begin{aligned} b_0 &= \int \varphi_0 Y(dt) \\ y_{j,k} &= \int \psi_{j,k} Y(dt) \qquad j = 0,1,2,\ldots, \quad k = 0,\ldots,2^j - 1. \end{aligned}$$

From properties of the Wiener process,

$$\begin{aligned} b_0 &= \beta_0 + z_0 \\ y_{j,k} &= \theta_{j,k} + z_{j,k} \qquad j = 0,1,2,\ldots, \quad k = 0,\ldots,2^j - 1 \end{aligned}$$

with z_0, $z_{j,k}$ iid $N(0,\epsilon^2)$.

Suppose now that we treat the data $y_{j,k}$ as sequence data, and form empirical estimates $(\hat{\theta}_{jk})$ of the corresponding $(\theta_{j,k})$. Then the series reconstruction $\hat{f}$

$$\hat{f}(t) = b_0 + \sum \hat{\theta}_{j,k}\psi_{j,k}$$

has the loss

$$\|\hat{f} - f\|^2_{L_2[0,1]} = (b_0 - \beta_0)^2 + \sum_{j,k}(\hat{\theta}_{j,k} - \theta_{j,k})^2.$$

In words, there is an exact isometry between estimating error in one domain and in the other. As the isometry goes in both directions, we conclude in an obvious notation that

$$\begin{aligned} R^*_N(\epsilon;\mathcal{F}) &= \epsilon^2 + R^*_N(\epsilon;\Theta) \\ R^*_L(\epsilon;\mathcal{F}) &= \epsilon^2 + R^*_L(\epsilon;\Theta); \end{aligned}$$

here the terms on the left hand side represent minimax risks for the problem in function space (30)–(32) and those on the right for the problem (8)–(10) in sequence space. Evidently, the term ϵ^2 is of negligible importance, compared to the minimax risks.

We get an answer to (Q1) above: for estimating a function on the interval $[0,1]$ the correct correspondence is between the function space $\mathcal{F}$ and the unilateral in j, finite in k, sequence space.

7.2 Functions in the Bump Algebra.

Suppose now that f is known to belong to the Bump Algebra B described in the introduction. To make this quantitative, let us be given φ and ψ, which provide an r-regular wavelet system, $r > 1$, and let $\mathcal{F}$ be the class of all $f : \mathbb{R} \to \mathbb{R}$ satisfying $\sum_{j=-\infty}^{\infty} 2^{j/2} \sum |\alpha_{j,k}| \leq 1$ with $\alpha_{j,k} = \int_{-\infty}^{\infty} f\psi_{j,k}$.

Based on the case with the functions of Total Variation, one might expect an isometry between this function problem and one in sequence space. Actually, this holds only approximately. Because we must use a wavelet with regularity $r > 1$ to express the Bump Algebra, we can no longer use the Haar basis. For other wavelet bases, boundary effects come into play. However, as we will show, the effects are asymptotically negligible.

Theorem 12 *Let $s \geq 1/2$, and let $\mathcal{F}$ be a class of functions on the real line defined by $f = \sum \alpha_{j,k}\psi_{j,k}$, with $|\alpha|_{b(s,p,q)} \leq 1$. Let $\Theta_{s,p,q}$ be the class of sequences defined by the condition $||\theta||_{s,p,q} \leq 1$. Then, using definitions (8)–(10) and (30)–(32) we have*

$$R_L^*(\epsilon; \mathcal{F}) = R_L^*(\epsilon; \Theta_{s,p,q})(1 + o(1)) \qquad \epsilon \to 0 \tag{33}$$

$$R_N^*(\epsilon; \mathcal{F}) = R_N^*(\epsilon; \Theta_{s,p,q})(1 + o(1)) \qquad \epsilon \to 0 \tag{34}$$

Here of course, terms on the left denote in the continuous space white noise problem, and those on the right in the sequence space white-noise problem.

Theorem 12 is proved in two steps. First, a certain nonasymptotic lower bound, of the form $R^*(\epsilon; \mathcal{F}) \geq R^*(\epsilon; \Theta_{s,p,q})$ holds (where $R^* = R_L^*$ or R_N^*) ,showing that the function problem is always harder than the sequence problem. Consider the class $\mathcal{F}_0$ of functions

$$f = \sum_{j \geq 0} \sum_{k=0}^{2^j - 1} \theta_{j,k}\psi_{j,k}.$$

with $\theta \in \Theta_{s,p,q}$. Evidently, $\mathcal{F}_0 \subset \mathcal{F}$. Hence $R^*(\epsilon; \mathcal{F}) \geq R^*(\epsilon; \mathcal{F}_0)$, where $R^* = R_A^*$ or R_N^*. Suppose we could observe $Y(t)$ for all $t \in (-\infty, \infty)$ and not just $[0, 1]$. This is obviously more informative than just observing Y, in the sense of comparison of experiments.

Now as dW is Gaussian white noise, it is not hard to see that the projection of Y on span $[\mathcal{F}_0]$ is sufficient for estimation of $f \in \mathcal{F}_0$. Consequently, the coefficients $y_{j,k} = \int_{-\infty}^{\infty} \psi_{j,k} Y(dt)$ are sufficient for estimating f (and, equivalently, $(\theta_{j,k})$). But the coefficients are exactly of the form (11) in the sequence space problem. Hence, with R denoting either R_L^* or R_N^*,

$$\begin{aligned} R(\epsilon, \mathcal{F}) &\geq R(\epsilon; \mathcal{F}_0) \equiv R(\epsilon, \mathcal{F}_0, \{Y(t) : t \in [0, 1]\}) \\ &\geq R(\epsilon; \mathcal{F}_0, \{Y(t) : t \in (-\infty, \infty)\}) \\ &= R^*(\epsilon; \Theta_{s,p,q}) \end{aligned}$$

Hence for both linear and nonlinear procedures, the sequence problem provides an exact, nonasymptotic lower bound on the difficulty of estimation. Let us now show that it provides an asymptotic upper bound.

Theorem 13 *Fix $\delta > 0$. For every estimator $\hat{\theta}_{j,k}$ in the sequence problem, there is an estimator $\hat{f}$ in the function problem with*

$$\sup_{f \in \mathcal{F}} E||\hat{f} - f||^2 \leq \sup_{\theta \in \Theta_{s,p,q}} E||\hat{\theta} - \theta||^2 + O(\epsilon^{2-\delta}).$$

Here the $O(\epsilon^{2-\delta})$ does not depend on the method $\hat{\theta}$.

Note that for sufficiently small δ, $R^*(\epsilon) \gg \epsilon^{2-\delta}$ whenever $0 < s < \infty$, and so Theorem 13 completes the proof of Theorem 12.

Theorem 13 may at first glance appear purely technical, but the proof exposes a fundamental issue for applications: *boundary behavior of wavelets*. The problem we are faced with is as follows. We have data about $f(t)$ only for $t \in [0,1]$, but the wavelet basis $\psi_{j,k}$ makes no provision for the finiteness of the interval—it is adapted to analysis on the whole real line.

We propose to "solve" this problem as follows. Let $\tilde{Y}(t)$ be an extension of the process $Y(t)$ to all of $\mathbb{R}$ via the rule that for $t < 0$, $\tilde{Y}(t) = W_1(-t)$ where W_1 is a Wiener process started at zero and independent of Y; for $t > 1$, $\tilde{Y}(t) = Y(1) + W_2(t-1)$ with W_2 a Wiener process started at zero independent of Y and W_1.

Now we have observations equivalent to

$$\tilde{Y}(dt) = \tilde{f}(t)\,dt + \epsilon W(dt) \qquad T < \mathbb{R}$$

with $\tilde{f}$ the *mutilation* $f1_{[0,1]}$ of f.

We propose to reconstruct f on $[0,1]$ using simple operations on the *empirical wavelet coefficients* $\int \varphi_{l,k}\tilde{Y}(dt)$ and $\int \psi_{j,k}\tilde{Y}(dt)$ of $\tilde{Y}$.

From the traditional point of view of linear time-invariant methods (or, equivalently Fourier Analysis), replacing the smooth object f by the mutilated one $\tilde{f}$ as an object to be estimated is an exceptionally bad idea. Roughly, the behavior of bias of the estimate in the neighborhood of 0 and 1 would be so bad as to completely dominate the mean-squared error over $[0,1]$.

One of the truly impressive facts about wavelets is that no such effect occurs if they are used in even a crude fasion. In some sense wavelets are robust against even rather brutal operations like mutilation.

Let us give the details about our proposal.

First, an important remark about our basis. We have assumed that φ and ψ are of compact support both contained in $[-S, S]$, with S an integer.

Second, our method requires two positive integer constants, $\ell(\epsilon, S, \delta)$ and $m(\epsilon, S, \delta)$. We will describe how these are chosen later.

We will employ the *inhomogeneous* wavelet algorithm. Let $\varphi_{\ell,k}(t) = 2^{\ell/2}\varphi(2^\ell t - k)$, $k \in \mathbf{Z}$, and let $\psi_{j,k} = 2^{j/2}\psi(2^j t - k)$, $j \geq 0$, $k \in \mathbf{Z}$ as before. Then any $f \in L_2(\mathbb{R})$ has the representation $f = \sum_{k=\infty}^{\infty} \beta_k \varphi_{\ell,k} + \sum_{j>\ell} \sum_{k=-\infty}^{\infty} \alpha_{j,k}\psi_{j,k}$. This is an inhomogeneous reconstruction formula starting at the "base frequency" 2^ℓ. Our reconstruction will have the general form

$$\hat{f} = \sum_{k=-\infty}^{\infty} \hat{\beta}_k \varphi_{\ell,k} + \sum_{j>\ell} \sum_{k=-\infty}^{\infty} \hat{\alpha}_{j,k}\psi_{j,k}$$

However, because of the compact support of φ and ψ, we will have no need for infinite sums. Define $K = \{k : -S \le k \le 2^\ell + S\}$ and, for each $j \ge \ell$, $B(j) = \{k : -S \le k \le S\} \cup \{2^j - S \le k \le 2^j + S\}$, and $I(j) = \{k : S < k < 2^j - S\}$. By the support properties of φ and ψ, we know that unless $k \in K$, $\varphi_{\ell,k}$ vanishes on $[0,1]$ and unless $k \in B(j) \cup I(j)$, $\varphi_{j,k}$ vanishes on $[0,1]$. We note that if $k \in I(j)$, then $supp\psi_{j,k} \subset [0,1]$—these are the $\psi_{j,k}$'s "interior to $[0,1]$"; and that if $k \in B(j)$, then $\psi_{j,k}$ is not interior to $[0,1]$ but may still play a role in the reconstruction of f. We call such terms "Boundary terms."

In detail, our reconstruction is of the form

$$\hat{f} = \sum_{k \in K} \hat{\beta}_k \varphi_{l,k} + \sum_{m \ge j \ge \ell} \sum_{k \in B(j)} \hat{\alpha}_{j,k} \psi_{j,k} + \sum_{j \ge \ell} \sum_{k \in I(j)} \hat{\alpha}_{j,k} \psi_{j,k}.$$

Here the first sum represents smooth low frequency structure, the second represents boundary behavior and the last sum represents high-frequency detail.

We obtain the coefficients in the first two sums using the process $\tilde{Y}$.

$$\begin{aligned} \hat{\beta}_k &= \int_{-\infty}^{\infty} \varphi_{l,k} \tilde{Y}(dt) \qquad k \in K \\ \hat{\alpha}_{j,k} &= \int_{-\infty}^{\infty} \psi_{j,k} \tilde{Y}(dt) \qquad l < j \le m, \quad k \in B(j). \end{aligned}$$

The coefficients for the final sum—and these are the important ones—are obtained from the estimator $\hat{\theta}$.

Define a sequence

$$y_{j,k} = \begin{cases} \int_0^1 \psi_{j,k} \tilde{Y}(dt) & j > \ell, \quad k \in I(j) \\ z_{j,k} & j \ge 0, \quad k \notin I(j) \end{cases}$$

where $z_{j,k}$ are iid $N(0, \epsilon^2)$ independent of $\tilde{Y}$. Then $(y_{j,k})$ is exactly of the form (8) where $\theta_{j,k} = Ey_{j,k}$ obeys

$$\theta_{j,k} = \begin{cases} \alpha_{j,k} & j > \ell, \quad k \in I(j) \\ 0 & j \ge 0, \quad k \notin I(j) \end{cases} \tag{35}$$

Indeed when $k \in I(j)$, the support properties of ψ guarantee that

$$\theta_{j,k} = \int_0^1 \psi_{j,k} f(t)\, dt = \int_{-\infty}^{\infty} \psi_{j,k} f(t)\, dt = \alpha_{j,k}.$$

Apply the estimator $\hat{\theta}$ supplied by the hypothesis of the theorem to the derived data $y_{j,k}$. One obtains estimates for the coefficients for the third term of (7.2) via

$$\hat{\alpha}_{j,k} = \hat{\theta}_{j,k}, \qquad j > \ell, k \in I(j).$$

Let us now analyze the behavior of this estimator. With $\tilde{f} = f1_{[0,1]}$ we have

$$||\hat{f} - f||_{L_2[0,1]} \le ||\hat{f} - \tilde{f}||_{L_2(\mathbb{R})}.$$

Putting $\tilde{\beta}_k = \int \tilde{f}\varphi_k$, $\tilde{\alpha}_{j,k} = \int \tilde{f}\psi_{j,k}$ etc.,

$$\begin{aligned} ||\hat{f} - \tilde{f}||^2_{L_2(\mathbb{R})} &= \sum_{k\in K}(\hat{\beta}_k - \tilde{\beta}_k)^2 \\ &\quad + \sum_{l<j\le m}\sum_{k\in B(j)}(\hat{\alpha}_{j,k} - \tilde{\alpha}_{j,k})^2 \\ &\quad + \sum_{j>m}\sum_{k\in B(j)}\tilde{\alpha}^2_{j,k} \\ &\quad + \sum_{j>l}\sum_{k\in I(j)}(\hat{\alpha}_{j,k} - \alpha_{j,k})^2. \end{aligned} \tag{36}$$

We bound the first two terms as follows. $\hat{\beta}_k \sim N(\tilde{\beta}_k, \epsilon^2)$, $\hat{\alpha}_{j,k} \sim N(\tilde{\alpha}_{j,k}, \epsilon^2)$. Hence

$$E\sum_{k\in K}(\hat{\beta}_k - \tilde{\beta}_k)^2 = Card(K)\cdot\epsilon^2 = (2^l + 2S)\cdot\epsilon^2$$

$$\begin{aligned} E\sum_{l\le j\le m}\sum_{k\in B(j)}(\hat{\alpha}_{j,k} - \tilde{\alpha}_{j,k})^2 &= \epsilon^2\cdot\sum_{l\le j\le m} CardB(j) \\ &= (m-l)(4S+2)\epsilon^2. \end{aligned}$$

Also functions in $\mathcal{F}$ are bounded: $\sup\{||f||_{L_\infty[0,1]} : f\in\mathcal{F}\} = M < \infty$. Hence

$$\begin{aligned} \tilde{\alpha}_{j,k} = \int \tilde{f}\psi_{j,k} = \int f\tilde{\psi}_{j,k} &\le ||f||_\infty||\psi_{j,k}||_1 \\ &\le M||\psi||_1 2^{-j/2}. \end{aligned}$$

Thus

$$\begin{aligned} \sum_{j>m}\sum_{k\in B(j)}\tilde{\alpha}^2_{j,k} &\le (4S+2)M^2||\psi||_1^2\sum_{j>m}2^{-j} \\ &= C2^{-m}, \text{ say.} \end{aligned}$$

Picking m and l appropriately, we get

$$\begin{aligned} 2^l\epsilon^2 &= O(\epsilon^{2-\delta}) \\ (m-l)\epsilon^2 &= O(\epsilon^{2-\delta}) \\ 2^{-m} &= O(\epsilon^{2-\delta}) \end{aligned}$$

simultaneously. Hence

$$E||\hat{f} - \tilde{f}||^2 \le O(\epsilon^{2-\delta}) + E\sum_{j>l}\sum_{k\in I(j)}(\hat{\alpha}_{j,k} - \alpha_{j,k})^2.$$

On the other hand, considering the definitions involved the final term in (36) obeys

$$\sum_{j>l}\sum_{k\in I(j)}(\hat{\alpha}_{j,k} - \alpha_{j,k})^2 \le ||\hat{\theta} - \theta||_2^2.$$

Let Θ be the collection of all sequences generated by prescription (35). As $\theta_{j,k} = \alpha_{j,k} 1_{\{j \geq l, k \in I(j)\}}$, we see immediately that

$$||\theta||_{s,p,q} \leq |\alpha|_{b(s,p,q)} \leq 1$$

Thus $\Theta \subset \Theta_{s,p,q}$. We conclude, as required, that

$$E||\hat{f} - f||^2_{L_2[0,1]} \leq \sup_{\theta \in \Theta_{s,p,q}} E||\hat{\theta} - \theta||^2 + O(\epsilon^{2-\delta}). \qquad \blacksquare$$

This gives an answer to (Q2): by treating the boundary terms in the wavelet expansion slightly differently than the interior terms, we guarantee that they have an asymptotically negligible effect on the mean-squared error.

8 Interpretation of the Procedure.

The development of sections 3–6 above leads to the following proposal for function estimation in the White Noise model (30), for a class $\mathcal{F} = \{f : |\alpha_{j,k}|_{b(s,p,q)} \leq 1\}$.

[1] Form the Empirical Wavelet Coefficients $(\hat{\beta}_k)$, $k \in K$, $(\hat{\alpha}_{j,k})$, $k \in B(j)$, and $y_{j,k}$.

[2] With nonlinearities $(\delta_j)_{j=0}^{\infty}$ chosen optimally from the class $\{\delta_{(r,\epsilon,p)}\}$ or from $\{\delta_\lambda\}$ or from $\{\delta_\mu\}$, apply the formula

$$\hat{\alpha}_{j,k} = \delta_j(y_{j,k}) \qquad j = \ell, \ldots; \quad k \in I(j)$$

to get minimax (resp. near-minimax) estimates of $\alpha_{j,k}$, $k \in I(j)$.

[3] Reconstruct, via

$$\hat{f} = \sum_{k \in K} \hat{\beta}_k \varphi_{\ell,k} + \sum_{\ell \leq j \leq m} \sum_{k \in B(j)} \hat{\alpha}_{j,k} \psi_{j,k} + \sum_{j} \sum_{k \in I(k)} \hat{\alpha}_{j,k} \psi_{j,k}.$$

As we have seen, this procedure yields an $\hat{f}$ which is asymptotically minimax as $\epsilon \to 0$ if the nonlinearities were optimally chosen from the family $\{\delta_{(r,\epsilon,p)}\}$; $\hat{f}$ is within a factor $\Lambda(p)$ [respectively $M(p)$] of asymptotically minimax if the nonlinearities were optimally chosen from $\{\delta_\lambda\}$ (respectively $\{\delta_\mu\}$).

In effect, the three terms in step [3] represent three different aspects of the smoothing problem. Symbolically, we have

$$\hat{f} = \hat{f}_{\text{GROSS}} + \hat{f}_{\text{BOUNDARY}} + \hat{f}_{\text{DETAIL}}$$

where

$$\begin{aligned} \hat{f}_{\text{GROSS}} &= \sum_{k} \hat{\beta}_k \varphi_{\ell,k} \\ \hat{f}_{\text{BOUNDARY}} &= \sum_{l \leq j \leq m} \sum_{k \in B(j)} \hat{\alpha}_{j,k} \psi_{j,k} \\ \hat{f}_{\text{DETAIL}} &= \sum_{j} \sum_{k \in I(j)} \hat{\alpha}_{j,k} \psi_{j,k}. \end{aligned}$$

Let us discuss these three terms in more detail.

$\hat{f}_{\text{GROSS}}$ is a traditional estimate of the orthogonal series type. It involves a reconstruction using the empirical series coefficients corresponding to the low-resolution or smooth terms in a certain series expansion. $\hat{f}_{\text{GROSS}}$ is linear in the data.

$\hat{f}_{\text{BOUNDARY}}$ is a boundary correction of $\hat{f}_{\text{GROSS}}$, again using simple empirical series coefficients, but extending to much higher resolution near the boundary than $\hat{f}_{\text{GROSS}}$ does, to correct for the discontinuity of $\tilde{f}$ at the boundary. $\hat{f}_{\text{BOUNDARY}}$ is linear in the data.

$\hat{f}_{\text{DETAIL}}$ is a detail correction for $\hat{f}_{\text{GROSS}}$ at interior points. It is formed by a nonlinear processing of the high-resolution wavelet coefficients. If the nonlinearities are from the λ or μ threshold family, they can be interpreted as identifying, among those coefficients which $\hat{f}_{\text{GROSS}}$ ignores, those most likely to correspond to signal rather than noise. Indeed, the nonlinearity $\delta_{\mu_j}(y_{j,k})$ sets to zero all those coefficients smaller than $\epsilon \cdot m(\epsilon, r_j^{\mu}, p)$, i.e., all those coefficients where the "empirical signal to noise ratio" is less than m.

8.1 A Locally Adaptive Kernel Estimate.

Note that the "gross structure" and "boundary correction" terms in the wavelet reconstruction are obtained by kernel estimates:

$$\begin{aligned}\hat{f}_{\text{GROSS}}(s) = \sum_{k \in K} \hat{\beta}_k \varphi_{\ell,k}(s) &= \sum \varphi_{\ell,k}(s) \int \varphi_{\ell,k}(t) \tilde{Y}(dt) \\ &= \int \sum \varphi_{\ell,k}(s)\varphi_{\ell,k}(t) \tilde{Y}(dt) \\ &= \int K_G(s,t) \tilde{Y}(dt)\end{aligned}$$

where $K_G(s,t) \equiv \sum_{k \in K} \varphi_{\ell,k}(s)\varphi_{\ell,k}(t)$. Similarly,

$$\hat{f}_{\text{BOUNDARY}}(s) = \sum_{\ell \le j \le m} \sum_{k \in B(j)} \hat{\alpha}_{j,k} \psi_{j,k}(s) = \int K_B(s,t) \tilde{Y}(dt)$$

with $K_B(s,t) \equiv \sum_{\ell \le j \le m} \sum_{k \in B(j)} \psi_{j,k}(s)\psi_{j,k}(t)$.

Turning to "Detail Structure," define $w_j(y)$ so that the identity $\delta_j(y) = y w_j(y)$ holds. Then $\hat{\alpha}_{j,k} = w_j(y_{j,k}) \int \psi_{j,k} Y(dt)$ and

$$\begin{aligned}\hat{f}_{\text{DETAIL}}(s) &= \sum_j \sum_{k \in I(j)} \hat{\alpha}_{j,k} \psi_{j,k}(s) \\ &= \sum_j \sum_{k \in I(j)} w_j(y_{j,k}) \psi_{j,k}(s) \cdot y_{j,k} \\ &= \int \sum_j \sum_{k \in I(j)} w_j(y_{j,k}) \psi_{j,k}(s) \psi_{j,k}(t) Y(dt) \\ &= \int K_D(s,t) Y(dt), \quad \text{say.}\end{aligned}$$

We have symbolically

$$\hat{f} = \int (K_{\text{G}} + K_{\text{B}} + K_{\text{D}})(s,t) \tilde{Y}(dt)$$

where the three "pieces" are orthogonal

$$\int \int K_i(s,t) K_j(s,t)\, ds\, dt = 0 \qquad i \neq j.$$

However K_{D} depends on y, through the $w_j(y_{j,k})$ weights. Consequenty, K_{D} is an *adaptively designed* kernel: it is constructed by adaptively summing kernels $\psi_{j,k}(s)\psi_{j,k}(t)$ of different

bandwidths, using weights based on the apparent need for inclusion of structure at level j and spatial position k.

In detail, put $Q(j,k) = supp\{\psi_{j,k}\} \subset [2^{-j}(k-S), 2^{-j}(k+S)]$, and $W_{j,k}(s,t) = \psi_{j,k}(s)\psi_{j,k}(t)$. Then

$$K_D(s,t) = \sum_{s \in Q(j,k)} w_j(y_{j,k}) W_{j,k}(s,t) :$$

a sum of kernels $W_{j,k}$ with weights. The kernel $W_{j,k}$ is supported in $Q(j,k) \times Q(j,k)$; consequently its bandwidth is $\asymp 2^{-j}$.

Suppose now that δ_j is chosen from the family of λ thresholds. The weights $w_j(y_{j,k})$ are then 0 if $|y_{j,k}| < \lambda_j$; as $|y_{i,k}| \to \infty$, they tend to 1. Hence, a small empirical coefficient $y_{j,k}$ leads to omission of the term $W_{j,k}$ from the detail kernel; a large empirical coefficient leads to inclusion, with full weight 1.

Consequently, if $|y_{j,k}| \gg \lambda_j$, then for $(s,t) \in Q(j,k) \times Q(j,k)$ the kernel $K_D(s,t)$ contains terms of bandwidth $\leq 2^{-j}$. In short, our proposal represents a method of adaptive local selection of bandwidth (and, indeed, kernel shape).

Parallel comments apply when the nonlinearities δ_j are chosen from the other families.

At this point, we have demonstrated **Ph. C** of the introduction – at least for estimation in the white noise model.

8.2 Overfitted Least-Squares with Backwards Deletion

The coefficients $y_{j,k}$ represent the orthogonal projection of Y on the basis functions $\psi_{j,k}$ Thus they represent the "least-squares estimated regression coefficients" in the "linear model"

$$f = \sum \beta_k \varphi_{\ell,k} + \sum_{j \geq \ell} \alpha_{j,k} \psi_{j,k}.$$

However, to build an estimate $\hat{f}$ using all the $\psi_{j,k}$ terms with least-squares coefficients involved serious "overfitting" with the result that the reconstruction is extremely noisy. In fact the "formula"

$$\sum \hat{\beta}_k \varphi_{\ell,k} + \sum_{j \geq \ell} y_{j,k} \psi_{j,k}$$

defines an object so erratic that it can only be interpreted as a distribution, namely $d\tilde{Y}$, not a function.

In traditional statistical modelling one often fits complete models and then removes from consideration those terms with "statistically insignificant" coefficients.

Our method has exactly such an interpretation, if hard thresholds (δ_μ) are employed for the nonlinearity. The standard error of $y_{j,k}$ is ϵ and $\mu_j = m(r_j/\epsilon, 1, p) \cdot \epsilon = m_j \cdot \epsilon$, say, so

$$\hat{\alpha}_{j,k} = \begin{cases} y_{j,k} & |y_{j,k}| \geq m_j \cdot \epsilon \\ 0 & |y_{j,k}| < m_j \cdot \epsilon \end{cases}$$

Hence the reconstruction

$$\hat{f}_{\text{DETAIL}} = \sum \hat{\alpha}_{j,k} \psi_{j,k}$$

includes only those terms $y_{j,k}$ with "z-scores" $y_{j,k}/\epsilon$ exceeding m_j in absolute value. Thus m_j is a "significance threshold."

However, observe that our significance thresholds are determined by a minimax criterion, and not, for example, by some conventional statistical criterion (e.g. $P < .05$). In fact, $m_j \to \infty$ as $j \to \infty$, which means that extreme statistical significance must be attached to a coefficient at high resolution index j before that term is incuded in the reconstruction.

8.3 A Roughness-Penalized Least Squares Estimate.

A popular technique in function smoothing is the use of "penalized least-squares" or "penalized likelihood" methods. Suppose for example that we observe sampled data (y_i) according to (1); a penalized least squares method is

$$\hat{f} = \arg\min_f \sum (f(t_i) - y_i)^2 + \lambda \int_0^1 (f'')^2$$

The resulting $\hat{f}$ is a cubic smoothing spline. The term $\int_0^1 (f'')^2$ is called a roughness penalty.

Our proposal, when used in conjunction with soft thresholds δ_λ, has an interpretation in terms of roughness penalties.

Note the simple identity

$$\delta_\lambda(y) = \arg\min_d (y - d)^2 + \lambda |d| \tag{37}$$

It follows that $\hat{f}_{\mathrm{DETAIL}} = \sum \hat{\alpha}_{j,k} \psi_{j,k}$ solves the problem

$$(\hat{\alpha}_{j,k}) = \arg\min_{(d_{j,k})} \sum (d_{j,k} - y_{j,k})^2 + \sum_{j \geq 0} \lambda_j \sum_{k=0}^{2^j - 1} |d_{j,k}|$$

The first term is a measure of residual sum of squares, or likelihood. The second term is a penalty. We know that $\lambda_j = \epsilon \cdot \ell(r_j^\lambda / \epsilon, 1, p)$ for a certain sequence r_j^λ. In the case $p = q < 2$, from asymptotics for $\ell(r_j^\lambda / \epsilon)$ we know that $\ell_j = \lambda_j / \epsilon$ is asymptotic to $const 2^{j\tilde{s}}$ as $s \to \infty$, with $\tilde{s} = \frac{sp}{2-p}$. It follows that the penalty term is, to within constants, equivalent to

$$\epsilon \cdot c \cdot \sum_{j \geq 0} 2^{j\tilde{s}} \sum_k |\alpha_{j,k}| = \epsilon \cdot c \cdot ||\alpha||_{\tilde{s},1,1}.$$

In short, for each $p = q \in [1, 2)$, the details are estimated with a roughness penalty equivalent to the Besov $(\tilde{\sigma}, 1, 1)$ seminorm, $\tilde{\sigma} = \frac{sp}{2-p} + 1/p - 1/2$.

9 Sampling of Wavelet Series.

We now study issue (Q3): the approximation of wavelet coefficients by sums rather than integrals.

9.1 Sampling Theorem for Wavelets

The clasical sampling theorem says that for an entire function of type π which is in L_2 on the real axis,

$$\sum_{i=-\infty}^{\infty} f^2(i) = \int_{-\infty}^{\infty} f^2(t)dt.$$

This implies, among other things, that certain integrals over the real line may be calculated by sums.

The closest analogous statement for wavelets would be that for a function in V_0, the span of $(\varphi_{0,k})$, the sum of squares of samples taken at the integers is comparable in size to the squared L_2-norm. To guarantee near-equality of sample sums of squares with squared integrals, we have to sample at rate much higher than one per unit time, and normalize the samples by our sampling rate. Let $(s_i)_{i\in\mathbf{Z}}$ be our bilateral sampling mesh, defined by $s_i = (i-1)/n$, at sampling rate n samples per unit time.

Consider a function f in V_0:

$$f = \sum_{k=-\infty}^{\infty} \beta_k \varphi_{0,k}$$

for $\beta_k = \int f\varphi_{0,k}$. Estimating integrals by sums at the sample points $s_i = i/n$ gives

$$\tilde{\beta}_k = n^{-1} \sum_{i\in U} f(s_i)\varphi_{0,k}(s_i).$$

This implicitly defines a linear transformation $U_{0,n} : \ell_2 \to \ell_2$ via $U_{0,n}(\beta) = \tilde{\beta}$.

Lemma 2 *$U_{0,n}$ is a discrete convolution operator. If the wavelets φ are supported inside $[-S, S]$ then the transfer function of the operator is*

$$\hat{u}_n(\lambda) = 1 + \sum_{\substack{k\neq 0 \\ |k|\leq 2S}} c_n(k)e^{ik\theta}$$

where

$$c_n(k) = n^{-1} \sum_{i\in U} \varphi_{0,0}(s_i)\varphi_{0,k}(s_i).$$

If $n \to \infty$ along powers of 2,

$$c_n(k) \to 0 \qquad k \neq 0,\ |k| \leq 2s.$$

It follows immediately from the lemma that the operator norms

$$\begin{aligned} ||U_{0,n}|| &\equiv \sup_{\lambda\in[-\pi,\pi]} |\hat{u}_n(\lambda)| \\ ||U_{0,n}^{-1}|| &\equiv 1/\inf_{\lambda\in[-\pi,\pi]} |\hat{u}_n(\lambda)| \end{aligned}$$

satisfy

$$\begin{aligned} ||U_{0,n}|| &\to 1 \qquad n \to \infty \\ ||U_{0,n}^{-1}|| &\to 1. \end{aligned}$$

In short, the approximation of integrals $(\beta_k)_{k\in\mathbf{Z}}$ by sums $(\tilde{\beta}_k)_{k\in\mathbf{Z}}$ is asymptotically correct, *uniformly* in $(\beta_k) \in \ell_2$. Define the operator $T_n : \ell_2 \to \ell_2$ by

$$T_{0,n}(\beta) = \left(\frac{1}{\sqrt{n}} \sum_k \beta_k \varphi_{m,k}(s_i)\right)_{i\in\mathbf{Z}}.$$

This is the operator that yields normalized samples $(\frac{1}{\sqrt{n}} f(s_i))_{i\in\mathbf{Z}}$ of functions $f \in V_0$. Then we have the crucial identity

$$U_{0,n} = T_{0,n}^* T_{0,n}.$$

As $U_{0,n}$ is almost an isometry, we conclude that $T_{0,n}$ satisfies

$$\begin{aligned} ||T_{0,n}|| &\to 1 \\ ||T_{0,n}^{-1}|| &\to 1 \qquad n \to \infty. \end{aligned}$$

Here we interpret $T_{0,n}^{-1}$ as an operator from Range $(T_{0,n})$ into ℓ_2. These relations imply that $T_{0,n}$ is a near-isometry from functions $f \in V_0$ to samples $\left(\frac{1}{\sqrt{n}} f(s_i)\right)_{i\in\mathbf{Z}}$:

$$\frac{1}{n} \sum_i f^2(s_i) \approx \int_{-\infty}^{\infty} f^2(s) ds$$

for all f in V_0, for large n. This is the "sampling theorem" for wavelet series. It has several implications, such the near-orthogonality, with respect to sums along the grid (s_i), of wavelets $\psi_{j,k}$ and $\psi_{j',k'}$ as long as $j \le 0$.

Analogous relations hold with V_m replacing V_0. Suppose that $n = 2^{m+a}$, for an integer $a > 0$. Then we have

$$\begin{aligned} ||T_{m,n}|| &= ||T_{0,2^a}|| \\ ||T_{m,n}^{-1}|| &= ||T_{0,2^a}^{-1}|| \end{aligned}$$

which extends the sampling theorem to other resolution scales.

9.2 Sampling and Smoothness.

Let $g(t)$ be a function with domain $\mathbb{R}$. Let $V_m : L_2 \to L_2$ be the operator of projection onto the span of $(\psi_{j,k})$ with $j \le m$. Alternatively,

$$(V_m g)(t) = \sum_{k=-\infty}^{\infty} \beta_k^{(m)} \varphi_{m,k}(t) \tag{38}$$

where $\varphi_{m,k}(t) = 2^{m/2} \varphi(2^m t - k)$ and

$$\beta_k^{(m)} = \int \varphi_{m,k}(s) g(s)\, ds \qquad k \in \mathbf{Z}. \tag{39}$$

We now consider the approximation

$$(\tilde{V}_{m,n}g)(t) = \sum_{k=-\infty}^{\infty} b_k^{(m)} \varphi_{m,k}(t) \tag{40}$$

where

$$b_k^{(m)} = n^{-1} \sum_{-\infty}^{\infty} \varphi_{m,k}(s_i) g(s_i) \qquad k \in \mathbf{Z}. \tag{41}$$

In this approximation, the coefficient functionals are approximated by discrete sums.

The following lemmas show that if we restrict attention to pairs (m,n) with $n = 2^{m+a}$ for an integer $a > 0$, the problem of measuring degree of approximation of functions $g \in BV \cap L_2$ by $\tilde{V}_{m,n}g$ has a simple answer.

Lemma 3 *Let* $(P_{m,n,C})$ *denote the optimization problem*

$$(P_{m,n,C}) \quad \sup ||\tilde{V}_{m,n}g - V_m g||_2 \text{ subject to } TV(g) \le C.$$

Then $\mathrm{val}(P_{0,2^a,1}) < \infty$ *for each positive integer* a. *For* $n = 2^{m+a}$

$$\mathrm{val}(P_{m,n,C}) = 2^{-m/2} \cdot C \cdot \mathrm{val}(P_{0,2^a,1}). \tag{42}$$

Let $(Q_{m,n,C})$ *denote the optimization problem*

$$(Q_{m,n,C}) \quad \sup ||V_m g - g||_2 \text{ subject to } \begin{cases} TV(g) \le C \\ \sum_{-\infty}^{\infty} g(s_i) = 0 \end{cases}$$

Then $\mathrm{val}(Q_{0,2^a,1}) < \infty$ *and*

$$\mathrm{val}(Q_{m,n,C}) = 2^{-m/2} \cdot C \cdot \mathrm{val}(Q_{0,2^a,1}). \tag{43}$$

The proof is an application of dyadic renormalization. If g_0 is feasible for $(P_{0,2^a,1})$ (respectively $(Q_{0,2^a,1})$) then $g_{m,C}(t) = C g_0(2^m t)$ is feasible for $(P_{m,n,C})$ (respectively $(Q_{m,n,C})$) and vice versa. As we evidently have $||\tilde{V}_{0,2^a} g_0 - g_0||_2 = 2^{m/2} \cdot ||\tilde{V}_{m,n} g_{m,C} - V_m g_{m,C}||$ and $||V_0 g_0 - g_0|| = 2^{m/2} ||V_m g_{m,C} - g_{m,C}||_2$, the results (42) and (43) are immediate. For them to have meaning, we must prove, however, that $\mathrm{val}(P_{0,2^a,1}) < \infty$ and $\mathrm{val}(Q_{0,2^a,1}) < \infty$. This is done in the appendix.

It follows from the lemma that if $\sum_{-\infty}^{\infty} g(s_i) = 0$, and $g \in BV \cap L_2$, then

$$||\tilde{V}_{m,n}g - g||_2 \le \mathrm{const}\, 2^{-m/2} \tag{44}$$

where the constant depends on $TV(g)$ and on $a = (\log_2 n) - m$. With a fixed, the right-hand side of (44) is of order $1/\sqrt{n}$; this degree of approximation of g by $\tilde{V}_{m,n}g$ is sufficient for our purposes.

We now investigate the smoothness of this approximation. Define

$$\tilde{\alpha}_{j,k} = \begin{cases} \int \psi_{j,k} \tilde{V}_{m,n} g & j \le m,\ k \in \mathbf{Z} \\ 0 & j > m,\ k \in \mathbf{Z} \end{cases}$$

These are the wavelet coefficients of $\tilde{V}_{m,n}g \cdot |\tilde{\alpha}|_{b(s,p,q)}$ is the roughness of $\tilde{V}_{m,n}g$; $|\alpha|_{b(s,p,q)}$ is the roughness of g. We are particularly interested in picking m and n to be sure that $\tilde{V}_{m,n}g$ is not significantly more rough than g.

Lemma 4 *Let* $(R_{m,n})$ *denote the optimization problem*

$$(R_{m,n}) \quad \sup |\tilde{\alpha}|_{b(s,p,q)} \text{ subject to } |\alpha|_{b(s,p,q)} \le 1.$$

Then with $n = 2^{m+a}$, $\mathrm{val}(R_{m,n}) = \mathrm{val}(R_{0,2^a}) < \infty$. *Moreover,*

$$\lim_{a\to\infty} \mathrm{val}(R_{0,2^a}) = 1. \tag{45}$$

Hence for all sufficiently large a, $\mathrm{val}(R_{m,2^{m+a}}) < 1 + \epsilon$. Let a_ϵ denote the smallest such a. Then with $m = n \cdot 2^{-a_\epsilon}$,

$$|\tilde{\alpha}|_{b(s,p,q)} \le (1+\epsilon)|\alpha|_{b(s,p,q)} \tag{46}$$

for all $(\alpha_{j,k})$. In short, $\tilde{V}_{m,n}g$ is nearly as smooth as g itself.

9.3 The Construction.

Given an estimator $\hat{\theta}$ in the sequence experiment we now show how to construct an estimate $\hat{f}$ from noisy samples $y_i = f(t_i) + z_i$ which has an asymptotically equivalent worst-case risk.

The construction has parameters δ, ℓ, m, and a_n. We have $n = 2^{m_n+a_n}$. The sequence (a_n) tends to ∞ in such a way that

$$\mathrm{val}(P_{0,2^{a_n},1}) \le n^\delta \tag{47}$$

$$\mathrm{val}(Q_{0,2^{a_n},1}) \le n^\delta \tag{48}$$

and

$$\mathrm{val}(R_{0,2^{a_n}}) \to 1. \tag{49}$$

$$||T_{m,n}|| \to 1 \tag{50}$$

The construction has 4 steps.

[1] *Removal of Polynomial Trend.* If $\mathcal{P} \neq \{0\}$, let π be a least-squares estimate of f from $\mathcal{P}$

$$\tilde{\pi} = \arg\min \left\{ \sum (\pi(t_i) - y_i)^2 : \pi \in \mathcal{P} \right\}.$$

Define the "trend-adjusted" data

$$\tilde{y}_i = y_i - \tilde{\pi}(t_i) \qquad i = 1, \ldots, n.$$

If $\mathcal{P} = \{0\}$, set $\tilde{y}_i = y_i$, $i = 1, \ldots, n$.

[2] *Calculation of Empirical Wavelet Coefficients.* Set

$$\tilde{b}_k^{(m)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{y}_i \varphi_{m,k}(t_i) \qquad -S \le k \le 2^m + S$$

$$\tilde{f}_{(t)} = \sum \tilde{b}_k^{(m)} \varphi_{m,k}(t)$$

$$\tilde{a}_{j,k} = \int_{-\infty}^{\infty} \psi_{j,k} \tilde{f} \qquad k \in B(j) \cup I(j)$$

$$\tilde{b}_k = \int_{-\infty}^{\infty} \varphi_{\ell,k} \tilde{f} \qquad k \in K$$

[3] *Transfer to Homoscedastic Sequence Experiment.* Define pseudo-data

$$y_{j,k} = \begin{cases} \tilde{a}_{j,k} + z_{j,k} & k \in I(j) \\ z_{j,k} & k \notin I(j) \end{cases}$$

where $z_{j,k} \overset{\text{iid}}{\sim} N(0, \tau_{j,k}^2)$ and the sequence $\tau_{j,k}$ is defined as follows.

Let $\sigma_{j,k}^2 = \text{var}(\tilde{a}_{j,k}) \times n$ and let $\sigma_n^2 = \max_{\ell \le j \le m} \max_{k \in I(j)} \sigma_{j,k}^2$. Then set $\tau_{j,k}^2 = \bar{\sigma}_n^2 - \sigma_{j,k}^2 \ge 0$ if $k \in I(j)$ $\tau_{j,k}^2 = \bar{\sigma}_n^2$, $k \ne I(j)$. The pseudo data have

$$\begin{aligned} Ey_{j,k} &= \begin{cases} \tilde{\alpha}_{j,k} & k \in I(j) \\ 0 & k \in I(j) \end{cases} \\ \text{var}(y_{j,k}) &= \bar{\sigma}^2/n. \end{aligned}$$

Treat these data as if they were from the sequence experiment (8) with $\epsilon^2 = \bar{\sigma}_n^2/n$, and $\Theta = \Theta_{s,p,q}(C(1+\eta_n))$. Here

$$\eta_n \equiv \text{val}(R_{0,2^{a_n}}) - 1 \to 0$$

as $n \to \infty$.

Let $\hat{\theta}_n$ be an estimator for the sequence problem.

[4] *Reconstruction.*

$$\hat{f}(t) = \tilde{\pi}(t) + \sum_{k \in K} \tilde{b}_k \varphi_k + \sum_{\substack{\ell \le j \le m \\ k \in B(j)}} a_{j,k} \psi_{j,k} + \sum_{\substack{\ell \le j \le m \\ k \in I(j)}} \hat{\theta}_{j,k} \psi_{j,k}.$$

This formula works almost as in the case of the White Noise model, with an extra term:

$$\hat{f} = \text{Polynomial Trend} + \text{Gross Structure} + \text{Boundary Terms} + \text{Detail}.$$

We stress once again that the asymptotics of this procedure do not depend in any considerable way on the first three terms: the quality of the method is determined by the quality of the estimator $\hat{\theta}$.

Theorem 14 *With δ, ℓ, m, and a_n as above, $\epsilon = \frac{\bar{\sigma}_n}{\sqrt{n}}$, we have*

$$\sup_{f \in \mathcal{F}} E||\hat{f} - f||^2 \le \sup_{\theta \in \Theta(1+\eta_n)} E||\hat{\theta} - \theta||^2 + O(\epsilon^{2-\delta}).$$

The term on the left refers to the model with n observations; the term on the right refers to the sequence space model. The proof is in the appendix.

9.4 Lower Bound.

To complete the proof of the correspondence theorem, we need lower bounds demonstrating that estimating in function space is not easier than estimating in sequence space.

Theorem 15 *Let $\mathcal{F}_0$ denote the class of functions $f = \sum_{j\geq 0}\sum_{k=0}^{2^j-1}\theta_{j,k}\psi_{j,k}$ with $\theta \in \Theta_{s,p,q}$. Then with $\epsilon = \frac{\sigma}{\sqrt{n}}$*

$$\begin{aligned} R_N(n,\mathcal{F}_0) &\geq R_N^*(\epsilon,\Theta_{s,p,q})(1+o(1)) \\ R_L(n,\mathcal{F}_0) &\geq R_L^*(\epsilon,\Theta_{s,p,q})(1+o(1)) \end{aligned}$$

as $n \to \infty$.

Proof We discuss only the first inequality, as the second follows by entirely parallel arguments. The proof has 3 steps. First, to exhibit a sequence of finite-dimensional cartesion subproblems $\Theta_{p,\infty}(r^{(a)})$ almost as difficult as the full problem $\Theta_{s,p,q}$. The second is to show that there is a near-isometry between θ in the subproblems, and the samples $(f(s_i))_{i\in U}$. The third is to apply the isometry to obtain the lower bound.

As before, we define m_n and a_n by $2^{m_n+a_n} = n$, and we have $a_n \to \infty$, but this time in such a way that $2^{a_n} = O(n^\delta)$ for each $\delta > 0$.

Lemma 5 *Let (r_j) be the sequence defined by*

$$R_B^*(\epsilon,\Theta_{s,p,q}) = R_B^*(\epsilon,\Theta_{p,\infty}(r)),$$

$\Theta_{p,\infty}(r) \subset \Theta_{s,p,q}$. Define $\Theta_n = \Theta_{p,\infty}(r_j^{(n)})$ where $r_j^{(n)} \equiv r_j 1_{\{j\leq m_n\}}$. Then if $2^{a_n} = o(n^\delta)$ for sufficiently small δ

$$R_N^*(\epsilon,\Theta_{s,p,q}) \sim R_N^*(\epsilon,\Theta_n).$$

The finite-dimensional cartesian product defined by this Lemma gives us functions via

$$f_n = \sum_{j=0}^{m_n}\sum_{k=0}^{2^j-1}\theta_{j,k}\psi_{j,k}.$$

Let $\mathcal{F}_n$ denote the class of all such functions. Then $\mathcal{F}_n \subset V_{m_n}$, and so,

$$f_n = \sum_{k=-\infty}^{\infty}\beta_k^{(m)}\varphi_{m,k}$$

for $\beta_k^{(m)} = \int f\varphi_{m,k}$. Estimating integrals by sums gives

$$\tilde{\beta}_k = n^{-1}\sum_{i\in\mathbf{Z}} f(s_i)\varphi_{m,k}(s_i).$$

This implicitly defines a linear transformation $U_{m,n} : \ell_2 \to \ell_2$ via $U_{m,n}(\beta) = \tilde{\beta}$, of the type analyzed in section 9.1. Hence it is a near-isometry.

Lemma 6 *Let T be a nonsingular linear transformation from $\mathbb{R}^d$ into $\mathbb{R}^d$. Let $R(\theta, \sigma, v, \Theta)$ denote the minimax risk, under squared Euclidean norm loss, for estimating θ from data $v_i = \theta_i + z_i$, z_i iid $N(0, \sigma^2)$, θ known to lie in Θ. Let $R(\theta, \sigma, \tilde{v}, \tilde{\Theta})$ denote the minimax risk for estimating θ from data $\tilde{v}_i = T(\theta)_i + \tilde{z}_i$ with $\tilde{z}_i$ iid $N(0, \sigma^2)$, and $\tilde{\theta} = T(\theta)$ known to lie in $\tilde{\Theta} = T(\Theta)$. Then*

$$R(\theta, \sigma, \tilde{y}, \tilde{\Theta}) \geq \frac{||T||^2}{||T^{-1}||^2} R\left(\theta, \frac{\sigma}{||T||}, y, \Theta\right).$$

The lemma applies as follows. Let $\tilde{R}(n, \mathcal{F})$ denote the difficulty of estimating f from observations $y_i = f(s_i) + z_i$, $i \in \mathbf{Z}$. Then

$$\begin{aligned} R(n, \mathcal{F}) &\geq \tilde{R}(n, \mathcal{F}) \\ &\geq \tilde{R}(n, \mathcal{F}_n) \\ &\geq \frac{||T_{m,n}||^2}{||T_{m,n}^{-1}||^2} R_N^*(\epsilon, \Theta_n) \\ &= R_N^*(\epsilon, \Theta_n)(1 + o(1)) \\ &\sim R_N^*(\epsilon, \Theta_{s,p,q}). \end{aligned}$$

Theorem 39 follows.

References

[1] Askey, R. and Wainger, S. (1965). Mean convergence of expansions on Laguerre and Hermite series. *American Journal of Mathematics*, **87**, 695–708.

[2] Bickel, P. J. (1983). Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (M. H. Rizvi, J. S. Rustagi, and D. Siegmund, eds.), Academic Press, New York, 511–528.

[3] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comunications in Pure and Applied Mathematics*, **41**, 909–996.

[4] DeVore, R. and Popov, V. (1988). Interpolation of Besov spaces. *Trans. Am. Math. Soc.*

[5] Donoho, D. L. (1990). Function Estimation and the White Noise model. *Ecole d'Eté de Probilités* (Saint Flour), Springer-Verlag (to appear).

[6] Donoho, D. L. and Johnstone, I. M (1990). Minimax risk over ℓ_p-balls. Technical Report, Department of Statistics, University of California, Berkeley.

[7] Donoho, D. L. and Nussbbaum, M. (1990). Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, **6**, 290–323.

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

[9] Frazier, M. and Jawerth, B. (1985). Decomposition of Besov spaces. *Indiana Univ. Math. J.*, 777–799.

[10] Jaffard, S. (1989). Estimation Hölderiennes Ponctuelle des fonctions au moyen des coefficients d'ondelettes. *Compte Rendus Acad. Sciences Paris* (A).

[11] Mallat, S. (1989a). Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Mat. Soc.*, **315**, 69–87.

[12] Mallat, S. (1989b). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

[13] Mallat, S. (1989c). Multifrequence channel decompositions of images and wavelet models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 2091–2110.

[14] Meyer, Y. (1990a). *Ondelettes.* Paris: Hermann.

[15] Meyer, Y. (1990b). *Operateurs de Calderón et Zygmund.* Paris: Hermann.

[16] Nussbaum, M. (1985). Spline smoothing and asymptotic efficiency in L_2. *Ann. Statist.*, **13**, 984–997.

[17] Peetre, J. (1976). *New Thoughts on Besov Spaces.* Duke Univ. Math. Series. Number 1.

[18] Pietsch, A. (1981). Approximation spaces. *Journal of Approximation Theory*, **32**, 115–134.

[19] Stone, C. (1982). Optimal global rates of convergence for nonparametric estimators. *Ann. Statist.*