Nonlinear Solution of Linear Inverse Problems by Wavelet-Vaguelette Decomposition

David L. Donoho
Department of Statistics
Stanford University

December, 1991
Revised April, 1992

Abstract

We describe the Wavelet-Vaguelette Decomposition (WVD) of a linear inverse problem. It is a substitute for the singular value decomposition (SVD) of an inverse problem, and it exists for a class of special inverse problems of homogeneous type – such as numerical differentiation, inversion of Abel-type transforms, certain convolution transforms, and the Radon Transform.

We propose to solve ill-posed linear inverse problems by nonlinearly "shrinking" the WVD coefficients of the noisy, indirect data. Our approach offers significant advantages over traditional SVD inversion in the case of recovering spatially inhomogeneous objects.

We suppose that observations are contaminated by white noise and that the object is an unknown element of a Besov space. We prove that nonlinear WVD shrinkage can be tuned to attain the minimax rate of convergence, for L^2 loss, over the entire Besov scale. The important case of Besov spaces $B^{\sigma}_{p,q}$, $p < 2$, which model spatial inhomogeneity, is included. In comparison, linear procedures – SVD included – cannot attain optimal rates of convergence over such classes in the case $p < 2$. For example, our methods achieve faster rates of convergence, for objects known to lie in the Bump Algebra or in Bounded Variation, than any linear procedure.

Keywords. Wavelets, Vaguelettes, Singular Value Decomposition, Nonlinear Wavelet Shrinkage, Radon Transform, Abel Transform, Numerical Differentiation, Deconvolution, Besov Spaces, Minimax Estimation, Optimal Rates of Convergence, White Noise Model.

Acknowledgements. The author is on leave from the University of California, Berkeley, and was supported there by NSF grant DMS 88-10192 and by NASA Contract NCA2-488. The author would like to thank Yves Meyer for warm encouragement, and Mario Bertero, and Ingrid Daubechies, for providing copies of unpublished manuscripts.

1 Introduction

Suppose we wish to recover an object $f(t)$ – a function in $L^2(\mathbb{R}^d)$ – but we are able to observe data only about $g(u) = (Kf)(u)$, where K is a linear transformation, such as Radon transform, Convolution transform, or Abel Transform. Such *Linear Inverse Problems* arise in scientific settings ranging from Medical Imaging to Physical Chemistry to Extragalactic Astronomy. Moreover, we assume that the data are noisy, so that we observe $y(u)$ given by

$$y(u) = (Kf)(u) + z(u), \qquad u \in \mathcal{U}$$

where z is a noise (whether stochastic or deterministic, we do not as yet specify). We are interested in recovering f from the data y. For definiteness, we use the $L^2(\mathbb{R}^d)$ norm $||\hat{f} - f||_2$ to measure quality of recovery.

One's first impulse might be to attempt the estimate $\hat{f} = K^{-1}y$. However, in the cases of most interest scientifically, K is not invertible, in the sense that K^{-1} does not exist as a bounded linear operator; such inverse problems are called ill-posed. For reviews of these concepts, see Bertero (1989), O'Sullivan (1986), and Wahba (1989).

1.1 The SVD Paradigm

It is usual to treat such problems by devices such as *quadratic regularization* [Tikhonov, 1962, Phillips 1962, Twomey 1962]: for $\lambda > 0$ one sets

$$\hat{f}_\lambda = (K^*K + \lambda I)^{-1}K^*y;$$

by *iterative damped backprojection* [43]

$$\begin{aligned} \hat{f}_0 &= \mu K^* y \\ \hat{f}_1 &= \hat{f}_0 + \mu K^*(y - K\hat{f}_0) \\ \cdots & \\ \hat{f}_{m+1} &= \hat{f}_m + \mu K^*(y - K\hat{f}_m) \end{aligned}$$

where $\lambda > 0$ is a damping parameter; and by *Singular Value Decomposition* methods, defined as follows. We let $\|\cdot\|_2$ stand equally for the $L^2(dt)$ and $L^2(du)$ norms, and let $\langle , \rangle$ and $[,]$ denote the respective inner products. If K^*K is a compact operator we

let $(e_\nu(t))$ denote its eigenfunctions, k_ν^2 its eigenvalues, and $h_\nu(u)$ the normalized image $h_\nu(u) = (Ke_\nu)(u)/\|Ke_\nu\|$ of these. If no k_ν is zero, we have the reproducing formula

$$f = \sum_\nu k_\nu^{-1}[Kf, h_\nu]e_\nu.$$

A reconstruction rule may be based on this formula, and the idea that the "important" coefficients $\langle f, e_\nu \rangle$ occur early in the series. Then, picking weights w_ν which are near 1 for small ν and near 0 for large ν we get a *Windowed SVD* reconstruction formula:

$$\hat{f}_w = \sum_\nu w_\nu k_\nu^{-1}[y, h_\nu]e_\nu. \tag{1}$$

Weights are chosen so that $(w_\nu/k_\nu) \in \ell^2$. As the eigenvalues of the compact operator K^*K tend to zero this weighting is necessary so that division by near-zero elements does not prevent convergence of the series.

The windowed SVD method, at least theoretically, includes many other approaches to inversion as special cases, simply by suitable choice of the window function w_ν; see Bertero (1989) for example. Thus, if we pick $w_\nu = \frac{k_\nu^2}{k_\nu^2+\lambda}$, we get the method of regularization; and if we pick $w_\nu = (1-(1-\mu k_\nu^2)^m)$ we get the m-th iterative damped backprojection [3, 43].

The singular system approach has been thoroughly studied for applications in a variety of fields; when the singular functions turn out to be of simple functional forms this is particularly so. See Bertero, De Mol, and Pike (1985), and Johnstone and Silverman (1990, 1991) for examples. In fact, the intensive work by many researchers building an extensive edifice of SVD applications qualifies the SVD as a *paradigm* for analyzing and solving linear inverse problems.

Many articles have derived the Singular System of special operators; it is an attractive topic filled with many interesting scientific applications as well as much interesting mathematics.

- Bounded convolution operators on the circle have sinusoids for singular functions.
- The prolate spheroidal functions (Slepian, Landau, and Pollak, 1961) supply the singular system of band and time limiting operators, and of diffraction-limiting operators;
- Spherical Harmonics weighted by special radial functions supply the singular functions for problems of whole-earth geomagnetic inversion Shure, Parker, and Backus (1982);
- Gegenbauer polynomials supply singular functions in problems of tomography (Davisson, 1982). Tchebycheff polynomials weighted by appropriate radial functions supply singular functions in problems of limited angle tomography (Louis, 1986).
- Jacobi Polynomials and Tchebyshev polynomials supply the (e_ν), (h_ν) sequences for singular value decomposition of the Abel transform (Johnstone and Silverman, 1991).
- Bertero and coworkers Pike, De Mol, Boccaci, and others have published a series of elegant *SVD* derivations [4, 5, 7, 8] for operators arising in microscopy, such as time-limited Laplace Transform and the Poisson transform. See also Gori and Guattari for applications in signal processing [27].

Finally, there are optimality results for SVD inversion; we quote an example. Suppose that the singular system of K admits a differential operator D^m of m-th order, with the e_ν as eigenfunctions. Let $\mathcal{F}_m(C)$ denote the class of all functions obeying the smoothness constraint $||D^m f||_2 \leq C$. Suppose that the data y are observed with Gaussian noise z having a covariance that is diagonalized by the e_ν. Then, for an appropriate choice of window (w_ν), the windowed SVD is minimax linear; i.e. it is the linear estimator which minimizes the worst-case risk

$$\sup_{\mathcal{F}_m(C)} E||\hat{f} - f||_2^2.$$

There are even results saying that asymptotically, as the amount of data increases, the best windowed SVD estimator is asymptotically best among *all measurable functions* of the data. See Johnstone and Silverman (1990, 1991) for more information.

1.2 Limitations of SVD

While such results are assuring to users of SVD-based inversion schemes, one must admit that the method has certain limitations. These are rooted in the fact that the basis functions (e_ν), (h_ν) derive from the operator under study, not from the object to be recovered. Thus, if the same operator K occurs in two different fields of scientific inquiry, the basis functions will be the same, even though the type of object to be recovered may be quite different in the two fields. One can easily imagine that in one field of scientific inquiry the f to be recovered could be very efficiently represented in terms of the basis set used; while in the other area, the object is poorly approximated by finite sums of basis terms e_ν even when a fairly large number of terms is used.

Efficient representation of f by singular functions e_ν is essential. Suppose that (for definiteness) the object is observed in white noise, so that the observed singular coefficient obeys

$$[y, h_\nu] = k_\nu \theta_\nu + \epsilon z_\nu$$

where (z_ν) is a Gaussian white noise sequence, ϵ is the noise level, and $\theta_\nu = \langle e_\nu, f \rangle$ is the component of f in the direction e_ν. Then, if we use the best window (w_ν) possible for the function under consideration we would have a mean-squared error within a factor of 2 of

$$\sum_\nu \min(\theta_\nu^2, k_\nu^{-2}\epsilon^2), \tag{2}$$

which we take as a proxy for the difficulty in recovering f. This expression shows that in order to have accurate reconstructions, it is important that there be very few θ_ν which are large, and that those which are large be located at those components ν where k_ν is also large.

In short, even when the SVD window (w_ν) is chosen optimally for the specific function at hand, it is necessary for the coefficients (θ_ν) to have a certain distribution of energy in the singular system basis. Otherwise, the windowed SVD method will not perform well.

In many realistic examples one does not have the desired agreement between the energy distribution of the object and the decay of the singular values. Suppose we are considering an inverse problem involving circular convolution, so that the singular functions are just

sinusoids, and for simplicity, suppose that the singular values are monotone decreasing with increasing frequency $|\nu|$. Suppose that the object to be recovered has a discontinuity. Then its coefficients decay as $|\nu| \to \infty$ only like $1/|\nu|$, which is rather slow; in consequence, the expression $\sum_\nu \min(\theta_\nu^2, k_\nu^{-2}\epsilon^2)$ will tend to zero slowly with ϵ.

This example is repeated at the level of classes of functions. Let $BV(C)$ be the class of objects on the circle which are the integrals of finite signed measures with total mass limited by C. Functions of Bounded Variation may contain jumps, and so their Fourier series decay rather slowly. Results in this paper show that one cannot attain optimal rates of convergence in deconvolution of BV objects by windowed Fourier methods. For any fixed window, the rate at which the reconstruction approaches the truth with increasing sample size/decreasing noise level will be suboptimal.

This is a general phenomenon, and continues outside the special case of circular deconvolution. Moreover, the specific trouble – the spatial variability of objects of bounded variation – occurs in many fields. Objects of bounded variation arise, for example, in seismic inverse problems, where they represent bulk material properties as a function of depth, and so change discontinuously across layer boundaries. They also arise in image processing and medical imaging where they represent the possibility of summable discontinuities in optical or bulk properties of the object.

1.3 Sparse Representations of Objects by Wavelets

Very recently, there has been considerable interest in the use of orthonormal bases of wavelets to represent functions, and many important advantages of wavelet bases have been discovered, particularly as regards sparse representation of objects.

There are many possible wavelet bases [12, 35, 36]. We start with bases of $L^2(I\!R)$. Using the construction of Daubechies (1988), we obtain a function ψ of compact support, having M vanishing moments and M continuous derivatives and unit norm. This function may be described qualitatively as a localized smooth wiggle. We form all dyadic dilations and integer translations of the function, obtaining, for each $\lambda = (j, k)$ with j and k integers, the unit-norm function

$$\psi_\lambda(t) = 2^{j/2}\psi(2^j t - k) \tag{3}$$

The Daubechies construction then insures that $(\psi_\lambda)_\lambda$ is a complete orthonormal system for $L^2(I\!R)$. Hence, if we define the wavelet coefficient

$$\alpha_\lambda = \int \psi_\lambda(t) f(t) dt \tag{4}$$

we have the reconstruction formula

$$f = \sum_\lambda \alpha_\lambda \psi_\lambda, \tag{5}$$

representing f as a sum of localized wiggles at various scales and positions. It is perhaps surprising that one can obtain an orthonormal basis of smooth functions by proceeding in this way from a single function ψ; this possibility was discovered by J.O. Stromberg and independently by Y. Meyer. I. Daubechies showed that it was even possible to take ψ of compact support, a fact we assume below.

The data compression aspects of wavelet transforms are easy to explain. Suppose that $f(t)$ is a piecewise polynomial, with each piece of degree $\leq M$, and a total of P pieces. Then each wavelet coefficient corresponding to a wavelet whose support cube $Q(\lambda)$is contained entirely inside a single piece of the function will vanish (as ψ_λ is orthogonal to polynomials of degree $\leq M$). There are at most $C \cdot P$ wavelets at a single resolution level j which "feel" the boundaries between pieces; hence there are at most $C \cdot P$ nonzero coefficients. In short, the vast majority of the wavelet coefficients α_λ are zero.

We also use wavelet bases of $L^2(I\!R^2)$. We discuss only the tensor-product basis [36], although similar remarks would apply to hexagonal lattice bases [36, 10]. Now the index set is $\lambda = (j, k, \epsilon)$, where j is an integer, k is a member of the integer lattice Z^2, and $\epsilon \in \{1, 2, 3\}$. There are three wavelets $\psi^{[1]}$, $\psi^{[2]}$, $\psi^{[3]}$, corresponding to right-left, up-down, and diagonal directional sensitivity, and one sets

$$\psi_{(j,k,\epsilon)}(t) = 2^j \psi^{[\epsilon]}(2^j t - k). \tag{6}$$

The resulting wavelets (ψ_λ) provide a complete orthonormal system of $L^2(I\!R^2)$, with decomposition and reconstruction formulas as above.

There is considerable empirical evidence that the wavelet transform provides sparse representation of real images. The book of Frazier, Jawerth, and Weiss (1991) gives examples showing that in processing real images one can keep a small percentage – only 1 or 2 percent – of the coefficients in the discrete wavelet transform of real images, and still get a reasonably good reconstruction of the image. (There are of course, even better compression schemes for images containing textures, such as the wavelet packets of Coifman, Meyer, and Wickerhauser).

1.4 Wavelet Representation of Function Classes

Much of the mathematical interest in wavelets concerns the following phenomenon: there is an orthonormal wavelet basis which serves as an unconditional basis for any of the spaces in the Besov and Triebel-Lizorkin scales (Lemarié and Meyer, 1986) (Frazier and Jawerth, 1990) (see also [23, 25, 36]). These scales include all the Hölder, Lipschitz, and Sobolev classes, and many other interesting function spaces as well. Other interesting spaces, such as Bounded Variation, are tightly bracketed between two members of this scale.

The basic result: if the mother wavelet ψ is a little smoother than typical members of one of these function spaces, then the wavelets provide an unconditional basis [36, 25], i.e. if (α_λ) are the wavelet coefficients of an object in the space, then $(\pm_\lambda \alpha_\lambda)$ are wavelet coefficients of another object in the space, for every sequence of signs $(\pm_\lambda)$. So amplitudes alone of the wavelet coefficients, and not any other properties, characterize the members of these function spaces. This property is not shared by classical bases.

We may interpret the unconditional basis property as "diagonalization of the prior information". Suppose we have the prior information that $f \in \mathcal{F}$, where $\mathcal{F}$ possesses an unconditional basis by wavelets; for example suppose $\mathcal{F}$ is a ball of Besov space $B^\sigma_{p,q}$. We can check that $f \in B^\sigma_{p,q}$, the Besov space, by seeing if the norm $\left(\sum_j 2^{jsq}(\sum_k |\alpha_{j,k}|^p)^{q/p}\right)^{1/q} < \infty$ where $s = \sigma + 1/2 - 1/p$ [33, 23, 25, 36]. A coordinatewise test in the wavelet domain reveals membership in this smoothness class.

As a simple example, we consider the Bump Algebra for the real line, defined as follows. (See Meyer's book [36]). Let $g_{(s,t)}(u) = exp(-(u-t)^2/s^2)$ be a Gaussian bump normalized to height one. Then f belongs to the Bump Algebra if it can be expressed as a sum $f = \sum_i a_i g_{(s_i,t_i)}$ where the coefficients (a_i) are ℓ_1-summable. The parameters s_i again have the interpretation as line-widths, the t_i as line locations, the a_i as amplitudes and the $sign(a_i)$ as polarities, so the Bump Algebra on f can again be interpreted as a caricature of scientific spectra. A norm on f is defined by taking the smallest $\sum |a_i|$ in any representation of f. Meyer shows that this norm is equivalent to the norm of the Besov space $B^1_{1,1}$, which is equivalent to the simple functional $\sum_j 2^{j/2} \sum_k |\alpha_{j,k}|$ of wavelet coefficients.

The unconditional basis property represents a kind of optimality of data compression. A measure of the sparsity of representation of an object in a class $\mathcal{F}$ by an orthogonal basis $\mathcal{B}$ is given by the functional

$$J(\epsilon, \mathcal{F}, \mathcal{B}) = \sup_{\mathcal{F}} \sum_i \min(\theta_i^2, \epsilon^2); \tag{7}$$

Consider the rate at which this functional tends to 0 as $\epsilon \to 0$. Suppose that $\mathcal{F}$ is a ball of a functional space for which wavelets furnish an unconiditional basis. Then the rate at which J goes to zero is maximized by the wavelets basis [15]. The wavelet transform has excellent compression capabilities for whole *classes* of smooth functions.

A hint of the significance of this compression relation for our story may be seen by comparing (7) with (2).

1.5 The WVD

We have discussed two types of orthogonal bases for representing functions f. The first type, the SVD, efficiently represents the operator K, offering a kind of diagonalization. The second type, orthonormal wavelet bases, efficiently represent the information that functions obey certain regularity conditions, as expressed by membership in Hölder, Sobolev, or more generally Besov and Triebel spaces. In fact, in a sense, the wavelet bases diagonalize this prior information.

It would be very pleasant to have a single basis which both diagonalizes the operator K and the *a priori* regularity on the object f. Unfortunately, this is in general impossible.

There is, however, a special class of inverse problems in which we can simultaneously quasi-diagonalize both the operator and the a priori information. We will show in sections 2-5 below that there exists a *Wavelet-Vaguelette Decomposition* (WVD) of certain homogeneous inverse problems with the following ingredients.

- Three sets of functions $(\psi_\lambda)_\lambda$ – an orthogonal wavelet basis – and $(u_\lambda)_\lambda$ and $(v_\lambda)_\lambda$ – near-orthogonal sets.

- Quasi-singular value relations

$$K\psi_\lambda = \kappa_j v_\lambda \tag{8}$$

$$K^* u_\lambda = \kappa_j \psi_\lambda. \tag{9}$$

  with quasi-singular values (κ_j), depending on resolution index j but not spatial index k.

- Biorthogonality relations
$$[u_\lambda, v_\mu] = \delta_{\lambda,\mu}. \tag{10}$$
- Near-Orthogonality relations
$$\| \sum_\lambda a_\lambda u_\lambda \|_2 \asymp ||(a_\lambda)||_{\ell^2} \tag{11}$$
$$\| \sum_\lambda a_\lambda v_\lambda \|_2 \asymp ||(a_\lambda)||_{\ell^2}. \tag{12}$$

This decomposition, when it exists, gives rise to the reproducing formula

$$f = \sum_\lambda [Kf, u_\lambda] \kappa_j^{-1} \psi_\lambda$$

which is analogous to the reproducing formula for the SVD, only with wavelets ψ_λ in place of the eigenfunctions e_ν, and u_λ in place of the singular functions h_ν. If we actually had $u_\lambda \equiv v_\lambda$, and $\| \sum_\lambda a_\lambda v_\lambda \|_2 = ||(a_\lambda)||_{\ell^2}$, of course we would have just the SVD. However, the present decomposition is genuinely different. It represents the object in a basis which is effective in representing a wide range of classes of functions; and it represents the operator in a quasi-diagonal form as well. Hence the decomposition achieves the goal of simultaneous quasi-diagonalization of operator K and prior information $\mathcal{F}$ which we aimed for above.

In sections 2-5 below we show that for three dilation-homogeneous operators K – integration, fractional integration, Radon transformation – all sufficiently well-behaved wavelet bases lead to a WVD of K. The common property of these operators is homogeneity with respect to dilation. Let $(D_a f)(t) = f(at)$. Such operators intertwine with D_a via

$$K D_a = a^\alpha D_a K$$

for some exponent α.

1.6 Optimality of WVD Inversion

The WVD decomposition leads to an inversion algorithm: *nonlinear shrinkage of WVD coefficients*. Here we work out the details only for data Y containing measurements contaminated with white Gaussian noise; see section 7. The procedure is very simple:

- Define threshold nonlinearities $\delta_t(y) = sgn(y)(|y| - t)_+$. Choose level-dependent thresholds $t_j \geq 0$.
- Define the reconstructed function
$$\hat{f} = \sum_\lambda \delta_{t_j}([Y, u_\lambda] \kappa_j^{-1}) \psi_\lambda$$

The procedure is nearly as simple as the windowed SVD, with the exception that linear weighting is replaced by nonlinear thresholding. It turns out to have advantages over windowed SVD methods.

- Suppose that $\mathcal{F}$ is a class in the Besov scale for which wavelets offer an unconditional basis. Suppose that K is an operator admitting a WVD, and the thresholds (t_j) are tuned to the class $\mathcal{F}$. Then the procedure attains the minimax rate of convergence for recovery of objects in $\mathcal{F}$ from data Y.

- Suppose the class $\mathcal{F}$ is one of the Besov Spaces $B^{\sigma}_{p,q}$ with $p < 2$ (for example the Bump Algebra). No linear method, in particular the SVD, attains this optimal rate.

As the classes $B^{\sigma}_{p,q}$ with $p < 2$ model spatially inhomogeneous functions, this proves that *nonlinear WVD inversion, when it may be defined, offers significant performance advantages over windowed SVD inversion for recovering spatially inhomogeneous objects.*

1.7 Contents

The paper to follow gives, in sections 2-5, a development of the WVD for homogeneous inverse problems. Section 6 states our main result on the optimality of WVD shrinkage and suboptimality of linear inversion techniques. Sections 7, 8, and 9 give a systematic proof of our results.

2 Weakly Invertible Linear Operators

Let K be a (not necessarily bounded) linear operator from $\mathcal{D}(K)$ to $\mathcal{R}(K)$, where the domain $\mathcal{D}(K) \subset L^2(dt)$ and range $\mathcal{R}(K) \subset L^2(du)$. If K were a bounded linear operator with a bounded linear inverse K^{-1}, the problem of recovering f from noisy measurements Kf would be well-posed in the following sense. From measurements $y = Kf + z$, with noise z having small $L^2(du)$ norm, we obtain an estimate $\hat{f} = K^{-1}y$; this obeys $||\hat{f} - f||_2 \leq ||K^{-1}||_2 ||z||_2$, and so is accurate if the noise level is small. We are interested only in ill-posed situations, which we interpret as saying that K^{-1} does not exist as a bounded linear operator of L^2.

Even when the operator is not strongly invertible it may be possible to get useful information about linear functionals $\langle \psi, f \rangle$ from knowledge of Kf. Such a "Linear Functional Strategy" for inverse problems has been advocated by R.S. Anderssen (1976, 1980, 1986) and M. Golberg (1979). The basic idea is to search for a bounded linear functional $c(\cdot) : L^2(du) \to I\!R$ solving the quadrature problem

$$c(Kf) = \langle \psi, f \rangle \qquad f \in \mathcal{D}(K). \tag{13}$$

If such a linear functional exists, one can stably recover information about $\langle \psi, f \rangle$ from noisy data on Kf. If we observe $y = Kf + z$ where noise z has noise level $||z||_2$ then

$$|c(y) - \langle \psi, f \rangle| \leq ||c|| \cdot ||z||_2 \tag{14}$$

and hence $c(y)$ is a good approximation to $\langle \psi, f \rangle$ if the noise level is small enough.

Which linear functionals can we recover in this way? The articles of Anderssen and Golberg as well as Bertero (1989) give the following answer:

Lemma 1 *Let $\mathcal{D}(K)$ be dense in $L^2(dt)$. The following are equivalent:*

- *There exists a bounded linear functional $c(\cdot)$ of $L^2(du)$ satisfying the quadrature formula (13) for all $f \in \mathcal{D}(K)$.*
- *The inequality*
$$|\langle \psi, f \rangle| \le C||Kf||_2$$
holds for all $f \in \mathcal{D}(K)$.
- $\psi \in \mathcal{R}(K^*)$.

The final criterion is most useful; the second most concrete.

Now suppose we have an orthonormal wavelet basis (ψ_λ) of $L^2(dt)$. We can stably recover the wavelet coefficient $\alpha_\lambda = \langle \psi_\lambda, f \rangle$ from noisy measurements of Kf if, for each λ, we have a bounded linear functional $c_\lambda : L^2(du) \to I\!R$ which satisfies the quadrature relation
$$c_\lambda(Kf) = \langle \psi_\lambda, f \rangle. \tag{15}$$
Lemma 1 says that this is possible provided $\psi_\lambda \in \mathcal{R}(K^*)$ for each λ. On the other hand, it is also convenient to have $\psi_\lambda \in \mathcal{D}(K)$.

Lemma 2 *Suppose that*
$$\psi_\lambda \in \mathcal{D}(K) \cap \mathcal{R}(K^*) \tag{16}$$
for all λ. Then the collection of all functionals $(c_\lambda)_\lambda$ represents the (unbounded) inverse of K in this algebraic sense: for all finite sums $f = \sum_\lambda \alpha_\lambda \psi_\lambda$, we have the reproducing formula
$$f = \sum_\lambda c_\lambda(Kf)\psi_\lambda. \tag{17}$$
We call K weakly invertible.

This reproducing formula is the basis for what follows; to get it, we will always apply criterion (16).

Before analyzing specific operators, we make two remarks.

1. The coefficient functionals have an interpretation in terms of biorthogonality. Denote $\xi_\lambda = K\psi_\lambda$ the image of ψ_λ under K; then
$$c_\lambda(\xi_\mu) = \delta_{\lambda,\mu}$$
where $\delta_{\lambda,\mu}$ denotes the Kronecker delta. As we are working in $L^2(du)$ and c_λ is bounded, there is a Riesz representer γ_λ for c_λ (i.e. $c_\lambda(g) \equiv [\gamma_\lambda, g]$) and so
$$[\gamma_\lambda, \xi_\mu] = \delta_{\lambda,\mu};$$
the (γ_λ) and the (ξ_μ) are biorthogonal with each other.

2. A closely related notion of weak invertibility involves the Gram operator
$$G = K^*K.$$

We say that G is weakly invertible if there is a bounded linear functional s_λ satisfying the quadrature relations

$$s_\lambda(Gf) = \langle \psi_\lambda, f \rangle.$$

Note that G maps a subset $\mathcal{D}(G) \subset L^2(dt)$ into a subset $\mathcal{R}(G) \subset L^2(dt)$, so we continue dealing with objects on the model space where f lives rather than the data space where Kf lives. However, we can translate results into previous terms, formally setting

$$\gamma_\lambda = K\sigma_\lambda,$$

where σ_λ is the Riesz representer of s_λ. Then, formally

$$[\gamma_\lambda, \xi_\mu] = [K\sigma_\lambda, K\psi_\mu] = \langle \sigma_\lambda, K^*K\psi_\mu \rangle = \langle \psi_\lambda, \psi_\mu \rangle = \delta_{\lambda,\mu}.$$

3 Homogeneous Linear Transforms

We now establish the reproducing formula (17) for operators of integration, fractional integration, and Radon Transform.

3.1 Integration

Let $(Kf)(u) = \int_{-\infty}^{u} f(t)dt$. Then from the frequency domain formula $\widehat{(Kf)}(\omega) = (i\omega)^{-1}\hat{f}(\omega)$ we have

$$\mathcal{D}(K) = \{f : \int |\hat{f}(\omega)|^2 |\omega|^{-2} d\omega < \infty\}$$

and

$$\mathcal{R}(K^*) = \{f : \int |\hat{f}(\omega)|^2 |\omega|^2 d\omega < \infty\}.$$

Hence, formally, criterion (16) demands that each ψ_λ have both an integral and a derivative in $L^2(dt)$.

A simple concrete argument says more. Suppose that the mother wavelet ψ is of compact support, integral 0, and C^1 regularity. Then both ψ' and $\psi^{(-1)}(u) = \int_{-\infty}^{u} \psi(t)dt$ are continuous and of compact support. Set $\xi_\mu = K\psi_\mu$ and $\gamma_\lambda = -(\psi_\lambda)'$. Integrating by parts, we get the biorthogonality

$$[\gamma_\lambda, \xi_\mu] = \langle \psi_\lambda, \psi_\mu \rangle = \delta_{\lambda,\mu},$$

which implies the reproducing formula (17).

Note that

$$\gamma_\lambda(u) = -2^j(2^{j/2}\psi'(2^j u - k))$$

so that the representers γ_λ are all scaled and dilated from the one mother representer $\gamma_{(0,0)} = -\psi'$, in the same way as wavelets are all scaled and dilated from ψ – only with an extra factor 2^j inserted. Consequently

$$\|c_\lambda\| = 2^j \cdot \|c_{(0,0)}\|,$$

which indicates the ill-posedness of the problem: by (14) it is increasingly difficult to recover increasingly high resolution components in noise.

3.2 Fractional Integration

Now let $\Omega(\cdot)$ be a not identically vanishing function, homogeneous of degree 0, let $\alpha \in (0,1)$, and set

$$(Kf)(u) = \int_{-\infty}^{\infty} f(t) \frac{\Omega(t-u)}{|t-u|^{1-\alpha}} dt.$$

This is a fractional integration operator which, with Ω = the Heaviside function and $\alpha = 1/2$ can reproduce the Abel transform. We have the frequency domain identity $\widehat{(Kf)}(\omega) = |\omega|^{-\alpha}\hat{\Omega}(\omega)\hat{f}(\omega)$ where $\hat{\Omega}$ is a certain function homogeneous of degree 0, so that (formally)

$$\mathcal{D}(K) = \{f : \int |\hat{f}(\omega)|^2 |\omega|^{-2\alpha} d\omega < \infty\}$$

and

$$\mathcal{R}(K^*) = \{f : \int |\hat{f}(\omega)|^2 |\omega|^{2\alpha} d\omega < \infty\}.$$

The criterion (16) for weak invertibility of K suggests that ψ_λ should have a fractional derivative and a fractional integral in L^2.

Let's work more carefully in the frequency domain. The Plancherel relation $\int f(t)g(t)dt = \frac{1}{2\pi}\int \hat{f}(\omega)\hat{g}(\omega)d\omega$ requires the representer of c_λ to satisfy

$$\hat{\gamma_\lambda}(\omega)|\omega|^{-\alpha}\hat{\Omega}(\omega)\hat{f}(\omega) = \hat{\psi_\lambda}(\omega)\hat{f}(\omega) \qquad \text{a.e. } \omega.$$

The formal solution is

$$\hat{\gamma_\lambda}(\omega) = \hat{\psi_\lambda}(\omega)|\omega|^{\alpha}/\hat{\Omega}(\omega).$$

If the mother ψ has compact support and two continuous derivatives then $\hat{\psi}(\omega) = o(|\omega|^{-2})$ as $\omega \to \infty$; thus $\hat{\gamma_\lambda}(\omega) = o(|\omega|^{-2+\alpha})$. As $0 < \alpha < 1$, $\hat{\gamma_\lambda}$ is an L^2 object. Also, if ψ has two vanishing moments $\int \psi = \int t\psi(t) = 0$ the equation $\xi_\mu = K\psi_\mu$ defines an L^2 object. Indeed, we have the frequency domain formula

$$\hat{\xi}_\mu(\omega) = \hat{\psi}_\mu(\omega)|\omega|^{-\alpha}\hat{\Omega}(\omega);$$

the vanishing moment condition implies that $\hat{\psi}(\omega) = o(|\omega|)$ as $|\omega| \to 0$, and so $\hat{\xi}_\mu(\omega)$ is $o(|\omega|^{1-\alpha})$ as $|\omega| \to 0$; again $0 < \alpha < 1$, and so $\hat{\xi}_\mu(\omega)$ is a continuous function of ω. It decays as $|\omega|^{-2-\alpha}$ as $|\omega| \to \infty$, and hence represents an L^2 object. The biorthogonality relations

$$\begin{aligned}
[\gamma_\lambda, \xi_\mu] &= \frac{1}{2\pi}\int \hat{\gamma_\lambda}(\omega)\hat{\psi_\mu}(\omega)|\omega|^{-\alpha}\hat{\Omega}(\omega)d\omega \\
&= \frac{1}{2\pi}\int \hat{\psi_\lambda}(\omega)\hat{\psi_\mu}(\omega)d\omega \\
&= \langle \psi_\lambda, \psi_\mu \rangle = \delta_{\lambda,\mu}
\end{aligned}$$

now follow, and we get the reproducing formula (17).

Working out explicitly the scaling relationships,

$$\gamma_\lambda(u) = 2^{j\alpha}(2^{j/2}\gamma_{(0,0)}(2^j t - k))$$

so that again all the c_λ derive from a single mother functional under dilation and translation. The norm

$$\|\gamma_\lambda\| = 2^{j\alpha}\|\gamma_{(0,0)}\|;$$

the growth with j expresses the ill-posedness of the deconvolution. Note that different α give rise to different degrees of ill-posedness. $\alpha = 1/2$, which correponds to the Abel transform, presents a lower degree of ill-posedness than $\alpha = 1$, which corresponds to integration.

3.3 Radon Transform

Now we consider objects in $I\!R^2$ and use the 2-d wavelet basis (6). Set $(Kf)(u,\theta) = (P_\theta f)(u)$, where

$$(P_\theta f)(u) = \int_{-\infty}^{\infty} f(\cos(\theta)u + \sin(\theta)v, -\sin(\theta)u + \cos(\theta)v)dv.$$

The usual projection-slice theorem [14] implies the identity

$$\int_0^\pi \int_{-\infty}^{\infty} (P_\theta f)(u)(P_\theta g)^\star(u)dud\theta = \frac{1}{2\pi}\int_{\omega\in I\!R^2} \hat{f}(\omega)\hat{g}(\omega)^\star \frac{1}{r}d\omega,$$

where $r \equiv |\omega|$. We think of the left side as $[Kf, Kg]$; and the right side as a frequency-domain representation of $\langle K^*Kf, g\rangle$. Now the functional s_λ has a representer σ_λ which is characterized by $\langle K^*Kf, \sigma_\lambda\rangle = \langle f, \psi_\lambda\rangle$. In the frequency domain this is

$$\frac{1}{2\pi}\int \hat{f}(\omega)\hat{\sigma_\lambda}^*(\omega)\frac{1}{r}d\omega = \frac{1}{(2\pi)^2}\int \hat{f}(\omega)\hat{\psi_\lambda}^*(\omega)d\omega;$$

for all $f \in L^2$. Hence formally

$$\hat{\sigma_\lambda}(\omega) = \frac{1}{2\pi}\cdot r \cdot \hat{\psi_\lambda}(\omega).$$

If the mother wavelet is of compact support with all partial derivatives of order 2 continuous, then $\hat{\psi_\lambda}(\omega) = o(|\omega|^{-2})$ as $|\omega| \to \infty$ and so this defines a valid L^2 object.

As remarked in section 2, the definition $\gamma_\lambda = K\sigma_\lambda$ then gives the representer of c_λ under sufficient regularity. By the projection-slice theorem this may be written as

$$\gamma_\lambda(u,\theta) = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty} |v|\hat{\psi_\lambda}(v\cos(\theta), v\sin(\theta))e^{ivu}dv.$$

Recalling that we are in $I\!R^2$ and so $\lambda = (j,k,\epsilon)$ where $\epsilon \in \{1,2,3\}$, and $k = (k_x, k_y)$, we may write this as

$$\gamma_{(j,k,\epsilon)}(u,\theta) = 2^j \cdot \gamma_{(0,0,\epsilon)}(2^j u - \cos(\theta)k_x - \sin(\theta)k_y).$$

The γ_λ are all "twisted" dilations of three fixed mother representers. As j increases, they concentrate around certain sine-curves in the (u,θ) plane.

The σ_λ are all dilations and translates of three fixed mother representers; they have norms

$$\|\sigma_\lambda\| = (\frac{1}{(2\pi)^4}\int r^2|\hat{\psi_\lambda}(\omega)|^2 d\omega)^{1/2} = 2^j \cdot \|\sigma_{(0,0,\epsilon)}\|.$$

Similarly, the norm of γ_λ can be derived from the Projection-Slice theorem

$$[\gamma_\lambda, \gamma_\lambda] = [K\sigma_\lambda, K\sigma_\lambda] = \frac{1}{(2\pi)^3} \int |r\hat{\psi}_\lambda(\omega)|^2 \frac{1}{r} d\omega$$

and so

$$||\gamma_\lambda|| = 2^{j/2} \cdot \|\gamma_{(0,0,\epsilon)}\|.$$

Again, the Radon Transform does not have a bounded inverse; but the growth of the coefficient norms is quite moderate.

4 Representers as Vaguelettes

In the cases just studied, we found that the coefficient functionals c_λ have norms growing geometrically in the resolution index, with a certain exponent α:

$$\|c_\lambda\| = 2^{j\alpha} \cdot Const. \tag{18}$$

The functions $u_\lambda = 2^{-j\alpha}\gamma_\lambda$ are thus nearly normalized:

$$\|u_\lambda\| = Const. \tag{19}$$

We would now like to show that under additional assumptions they are nearly orthonormal, in the sense that, for all vectors (α_λ),

$$\|\sum \alpha_\lambda u_\lambda\|_2 \asymp \|(\alpha_\lambda)\|_{\ell^2}. \tag{20}$$

This plays an important role for the theory that follows.

To fill out the picture, we introduce a second set of nearly normalized functions: the (v_μ) defined by

$$v_\mu = 2^{j\alpha} K\psi_\mu.$$

These also have

$$\|v_\mu\| = Const; \tag{21}$$

and they are biorthogonal to the u_λ's:

$$[u_\lambda, v_\mu] = \delta_{\lambda,\mu}.$$

Both systems have much in common with wavelets ψ_λ. They are indexed by the same scheme, and have many qualitative features – localization and cancellation – in common. For example, when K =integration,

$$u_\lambda(t) = 2^{j/2}\psi'(2^j t - k);$$

$$v_\lambda(t) = 2^{j/2}\psi^{(-1)}(2^j t - k);$$

so both systems are formally "wavelet like"; only the mother function does not have all the special properties of a mother wavelet.

There is a specific name for such systems.

Definition 1 *Let (w_λ) be a collection of continuous functions, bounded in $L_2(\mathbb{R}^d)$-norm. Define the Standardization operator*

$$(S_\lambda w)(t) = 2^{-jd/2} w(2^{-j}(t+k)).$$

Suppose that there exist constants C_1, C_2, $\eta > 0$, and $\beta > 0$ so that each standardized function $\tilde{w} = S_\lambda w_\lambda$ satisfies

$$\begin{array}{lrll} (V1) & |\tilde{w}(t)| & \leq C_1(1+|t|)^{-(d+\eta)} & t \in \mathbb{R} \\ (V2) & \int \tilde{w}(t)dt & = 0 & \\ (V3) & |\tilde{w}(t) - \tilde{w}(s)| & \leq C_2|t-s|^\beta & s,t \in \mathbb{R}. \end{array}$$

Then we say, following Yves Meyer, that (w_λ) is a collection of **vaguelettes**.

Assumption (V1) makes the vaguelette w_λ nearly-localized to the cube Q_λ of side 2^{-j} and lower left corner at $2^{-j}k$. Assumption (V2) makes the vaguelette w_λ have mean zero and hence oscillate. Finally, assumption (V3) makes the vaguelette possess at least minimal regularity.

The reason for introducing vaguelettes is the following:

Theorem 1 *(**Y. Meyer, Vol II, page 270**). If (w_λ) is a collection of vaguelettes, then there exists a constant $C = C(C_1, C_2, \eta, \beta)$ so that*

$$||\sum \alpha_\lambda w_\lambda||_2 \leq C||(\alpha_\lambda)||_{\ell^2}.$$

This has the following implication for us:

Theorem 2 *Suppose we have two sets of vaguelettes, (u_λ) and (v_λ) which are biorthogonal for the standard inner product $L^2(dt)$. There are constants c and C so that*

$$c||(\alpha_\lambda)||_{\ell^2} \leq ||\sum \alpha_\lambda u_\lambda||_2 \leq C||(\alpha_\lambda)||_{\ell^2},$$

and

$$c||(\alpha_\lambda)||_{\ell^2} \leq ||\sum \alpha_\lambda v_\lambda||_2 \leq C||(\alpha_\lambda)||_{\ell^2}.$$

This rephrases a standard idea in Hilbert spaces – see Young (1976), Chapter 1, secs. 7 and 8, Chapter 4, sec. 2. The result is used implicitly by Y. Meyer (1990), Volume I, Chapter VI, Page 166.

Our applications of the vaguelette machinery rest on study of homogeneous Fourier multipliers.

Lemma 3 *Suppose that $\psi : \mathbb{R} \to \mathbb{R}$ is of compact support, has M vanishing moments and M continuous derivatives. Let $\hat{\Omega}(\omega)$ be homogeneous of degree 0. Let $|\alpha| + 2 < M$. Define $w : \mathbb{R} \to \mathbb{R}$ by*

$$w(t) = \frac{1}{2\pi} \int e^{it\omega} \hat{\psi}(\omega) \hat{\Omega}(\omega) |\omega|^\alpha d\omega.$$

There exist constants C_1 and C_2 so that

$$\begin{array}{lrll} (R1) & |w(t)| & \leq C_1(1+|t|)^{-2} & t \in \mathbb{R} \\ (R2) & \int w(t)dt & = 0 & \\ (R3) & |w(s) - w(t)| & \leq C_2|s-t| & s,t \in \mathbb{R}. \end{array}$$

We omit the proof of the lemma, which is a lengthy excercise in standard Fourier analysis.

To use this, consider the case K = integration. The family (u_λ) is obtained by dilation and translation from the mother $-\psi'$. The lemma above, with $\alpha = 1$, implies that for $M \geq 3$ this mother satisfies regularity conditions (R1)-(R3), and that therefore the family (u_λ) inherits (V1)-(V3). Similarly, (v_μ) are obtained by translation and dilation from the mother $\psi^{(-1)}$. The lemma above, with $\alpha = -1$, implies that the mother satisfies three regularity conditions (R1)-(R3) and the (v_λ) satisfy (V1), (V2), and (V3) by inheritance. (In this particular case, it is easy to see that the lemma is not sharp: the condition $M \geq 3$ is not necessary in order to get a system of vaguelettes; $M = 2$ would also do).

A similar analysis applies in the case of fractional integration, $0 < \alpha < 1$. If the mother wavelet ψ is of compact support, with $M \geq 3$ vanishing moments, and C^M regularity, $M \geq 3$, one again finds that the systems (u_λ) and (v_μ) are vaguelettes. (Again, the condition $M \geq 3$ is far from necessary).

To study the Radon Transform, first note that the systems (u_λ) and (v_μ) derived as above will not be vaguelettes. The functions are not themselves localized to points, but instead localized near sine-curves in the (u, θ) plane. However, the identity

$$[u_\lambda, u_\mu] = 2^{-j/2-j'/2} \frac{1}{(2\pi)^3} \int \hat{\psi_\lambda}^*(\omega) \hat{\psi}_\mu(\omega)\, r\, d\omega$$

tells us that if we define "partners" $w_\lambda^+(t) = 2^{-j/2} \cdot (2\pi)^{-1/2} \cdot (-\Delta)^{1/4} \psi_\lambda$, then

$$[u_\lambda, u_\mu] = \langle w_\lambda^+, w_\mu^+ \rangle$$

and so the (u_λ) are almost-orthonormal precisely in case their "partners" are. Similarly, the identity

$$[v_\lambda, v_\mu] = 2^{j/2+j'/2} \frac{1}{2\pi} \int \hat{\psi_\lambda}^*(\omega) \hat{\psi}_\mu(\omega) \frac{1}{r} d\omega$$

tells us that the (v_μ) are almost-orthonormal precisely in case their "partners" $w_\lambda^-(t) = 2^{+j/2} \cdot (2\pi)^{1/2} \cdot (-\Delta)^{-1/4} \psi_\lambda$ are almost-orthonormal. Evidently, $\langle w_\lambda^+, w_\mu^- \rangle = \delta_{\lambda,\mu}$.

Establishing the almost-othogonality of the (u_λ) and (v_μ) systems therefore comes down to showing that the (w_λ^+) and (w_λ^-) are vaguelettes. This is again an issue of Fourier multipliers.

Lemma 4 *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be of compact support, with vanishing moments of order $\leq M$ and continuous derivatives of order $\leq M$. Let $|\alpha| + d + 1 < M$. Define $w : \mathbb{R}^d \to \mathbb{R}$ by*

$$w(t) = \frac{1}{(2\pi)^d} \int e^{i\langle t, \omega \rangle} \hat{\psi}(\omega) |\omega|^\alpha d\omega.$$

Then there exist constants C_1 and C_2 so that

$$\begin{array}{lll} (R1) & |w(t)| & \leq C_1/(1+|t|)^{(1+d)}, \qquad t \in \mathbb{R} \\ (R2) & \int w(t)dt & = 0 \\ (R3) & |w(s) - w(t)| & \leq C_2|s-t|, \qquad s, t \in \mathbb{R}. \end{array}$$

We again omit the proof. Compare Y. Meyer (1990) Volume I, Lemme 4, Page 166.

The lemma makes restrictive assumptions; however, it does resolve our application. Suppose that the mother wavelets $\psi^{[i]}$, $i = 1, 2, 3$, are of compact support, with $M \geq 4$ vanishing moments and $M \geq 4$ continuous derivatives; then the (w_λ^+) and (w_λ^+) are vaguelette systems. It follows that the systems (u_λ) and (v_μ) are almost orthonormal.

5 The WVD

The properties of the WVD were described in section 1.5, equations (8)-(12). Now we are in a position to construct a WVD for our three operators. One starts from (ψ_λ), a family of sufficiently regular wavelets of compact support (enough smoothness, enough vanishing moments). One derives the coefficient functionals c_λ; the growth of $||\gamma_\lambda||$ with j gives the κ_j^{-1}. The normalized u_λ and v_λ functions are then defined by $u_\lambda = \kappa_j \gamma_\lambda$ and $v_\lambda = \kappa_j^{-1} \xi_\lambda$; one checks the vaguelette conditions as indicated above, and the almost-orthogonality of the (u_λ) and (v_λ) follows. Hence the development of sections 2-4 has proved:

Theorem 3 *Let K be one of the operators of section 3. Let (ψ_λ) be an orthonormal basis of wavelets of compact support, deriving from a mother with M vanishing moments and M continuous derivatives. If M is sufficiently large, there is a WVD of the operator K, obeying (8)-(12). The quasi-singular values $\kappa_j = 2^{-j\alpha}$ where*

- *Integration: $\alpha = 1$.*
- *α-Fractional Integration: α.*
- *Radon transform: $\alpha = 1/2$.*

The WVD is in some ways similar to the SVD. In fact, if we had $u_\lambda \equiv v_\lambda$ and (u_λ) orthonormal, we would simply be recovering the SVD. Moreover, both decompositions derive from a reproducing formula which gives a kind of diagonal representation of the operator K.

The decompositions seem also to be very different, since singular value decompositions usually have globalized sinusoid-like basis elements, rather than localized wavelet-like functions. Nevertheless, when the WVD exists, wavelets are *nearly* singular functions for K, in the following sense. The basic equations (8)-(9) can be written in words as

$$K(\text{wavelets}) = (\text{multipliers} \cdot \text{vaguelettes})$$

$$K^*(\text{vaguelettes}) = (\text{multipliers} \cdot \text{wavelets}).$$

These say that, although orthonormal wavelet bases are not exactly invariant under K^*K, they are "morally invariant". To the eye, a collection of vaguelettes is the same as a collection of wavelets, so an operator that turns wavelets into vaguelettes has essentially not changed the functions.

5.1 Inhomogeneous Operators

We have so far only shown that three specific types of homogeneous operator possess WVD's. However, inhomogeneous operators can also have WVD's. A typical example in dimension $d = 1$ would be the convolution operator $(Kf)(u) = \int k(u-t)f(t)dt$ where the Kernel k obeys $|\hat{k}(\omega)| \sim |\omega|^{-\alpha}$ as $|\omega| \to \infty$. Suppose that the kernel obeys $\int |k| < \infty$, so that it is a bounded operator of $L^2(dt)$, and the high-frequency regularity

$$\frac{\inf_{|\omega| \leq \Omega} |\hat{k}(\omega)|}{|\hat{k}(\Omega)|} \to 1 \qquad |\Omega| \to \infty.$$

Using wavelet bases of regularity $M > \alpha + 1$, one can obtain the functionals c_λ; they are no longer all dilations and translations of a single mother; instead, at each resolution level j they are translations of $\gamma_{(j,0)}$. The norms $||\gamma_{(j,0)}||$ no longer scale geometrically. Instead, for $j \to -\infty$, they tend to a nonzero constant (in fact $1/|\hat{k}(0)|$). For $j \to +\infty$ they scale geometrically, so that $||\gamma_{(j,0)}|| \sim Const\ 2^{j\alpha}$. Defining the quasi- singular values $\kappa_j = 1$, $j \leq 0$, $= 2^{j\alpha}$, $j > 0$, one can then proceed with defining the WVD and get all the main properties.

To be specific, suppose that $k(x) = 1_{\{x<0\}}e^x$. Then $|\hat{k}(\omega)| \sim |\omega|^{-1}$ as $|\omega| \to \infty$, and so $\alpha = 1$. The formal identity

$$(I - \frac{d}{dt})(Kf)(u) = f(u)$$

– see Widder (1971, page 172) – shows that the functional c_λ has representer

$$\gamma_\lambda = \psi_\lambda - \psi'_\lambda,$$

provided of course that ψ has sufficient regularity. Hence one has

$$||\gamma_\lambda|| \sim 1, \qquad j \to -\infty$$

$$||\gamma_\lambda|| \sim 2^j, \qquad j \to +\infty.$$

This particular convolution is roughly as ill-posed as numerical differentiation. To generate the WVD in this case, one defines $\kappa_j = \min(1, 2^{-j})$ and sets $u_\lambda = \kappa_j(\psi_\lambda - \hat{\psi}'_\lambda)$ and $v_\lambda = \kappa_j^{-1} K\psi_\lambda$. Applying Lemma 3, if the mother ψ has regularity $M \geq 3$, the (u_λ) and (v_λ) are vaguelettes.

As a second example, let $k(x) = \frac{1}{2}e^{-|x|}$. Widder (1971, page 190) gives the identity

$$(I - \frac{d^2}{dt^2})(Kf)(u) = f(u)$$

and so c_λ has representer

$$\gamma_\lambda = \psi_\lambda - \psi''_\lambda$$

for sufficiently smooth mother wavelet ψ. Hence one has

$$||\gamma_\lambda|| \sim 1, \qquad j \to -\infty$$

$$||\gamma_\lambda|| \sim 2^{2j}, \qquad j \to +\infty.$$

Because the growth exponent of the coefficient norms is twice as big as for numerical integration, this deconvolution is "twice as ill-posed" as numerical differentiation. To generate the WVD, define $\kappa_j = \min(1, 2^{-2j})$ and set $u_\lambda = \kappa_j(\psi_\lambda - \hat{\psi}''_\lambda)$ and $v_\lambda = \kappa_j^{-1} K\psi_\lambda$. If the mother ψ has regularity $M \geq 4$, the (u_λ) and (v_λ) are systems of vaguelettes.

Another direction in which a WVD may be defined for inhomogeneous operators concerns convolution operators on the circle. Let $f(\theta)$ be a function on $[0, 2\pi]$ and set

$$(Kf)(t) = \int_0^{2\pi} k(t-\theta) f(\theta) d\theta$$

with $k(t)$ a 2π-periodic function having Fourier series $\hat{k}_\nu \sim |\nu|^{-\alpha}$ as $|\nu| \to \infty$. As in [36, Ch. III, Sec. 11], we can define smooth periodic wavelets (ψ_λ) on the circle, so that $\lambda = (j,k)$ runs through the finite set $(j,0), \ldots, (j, 2^j - 1)$ at level $j > 0$. These periodic wavelets are the same as the wavelets of compact support on the line as soon as j is sufficiently large.

Using this basis of periodic wavelets, one obtains representers γ_λ of the functionals c_λ; the coefficient norms scale as $||c_\lambda|| \sim 2^{j\alpha}$ as $j \to \infty$. One constructs the families $(\tilde{u}_\lambda)$ and $(\tilde{v}_\lambda)$ by the obvious operations, and obtains all the basic relations of the WVD approach.

We will not further pursue such generalizations of the WVD concept to inhomogeneous operators, except to record our conviction that there are numerous applications.

5.2 Inhomogeneous Wavelet Bases

We will find it convenient for later use to adapt the WVD to *inhomogeneous* wavelet bases. In usual treatments of wavelets [12, 35, 36], one introduces, in tandem with the mother wavelet ψ introduced above, a *father wavelet* ϕ. This is a localized, wiggly function of compact support with $\int \phi = 1$, $\int t^k \phi = 0$, $k = 1, \ldots, M$, and $\phi \in C^M$. One forms the dilation and translation

$$\phi_{\ell,k}(t) = 2^{l/2} \phi(2^l t - k);$$

and obtains the *inhomogeneous wavelet expansion*

$$f(t) = \sum_k \langle f, \phi_{\ell,k} \rangle \phi_{\ell,k} + \sum_{j \geq l} \sum_k \langle f, \psi_{j,k} \rangle \psi_{j,k} \tag{22}$$

valid for every $f \in L^2(I\!R)$. This formula consists of the "low frequency" or "gross structure" terms $\sum_k \langle f, \phi_{\ell,k} \rangle \phi_{\ell,k}$ all taken at the single resultion level ℓ, and the "high resolution" piece $(j \geq \ell)$ of the homogeneous wavelet formula $\sum_\lambda \alpha_\lambda \psi_\lambda$. We have given details here for $d = 1$ only, but similar formulae hold for $d > 1$.

The function $\phi_{\ell,k}$ being itself an L^2 function, it has a representation in terms of the homogeneous wavelet expansion

$$\phi_{\ell,k} = \sum_\lambda p_{(\ell,k);\lambda} \psi_\lambda$$

where the coefficients $(p_{(\ell,k);\lambda})$ are of unit norm in ℓ^2, vanish for $j > \ell$, and are of compact support at each resolution level $j \leq \ell$. It follows that we may define the linear functional

$$b_{\ell,k} = \sum_\lambda p_{(\ell,k);\lambda} c_\lambda.$$

If in fact the u_λ are vaguelettes, this series is convergent and defines a bounded linear functional. To check this, one uses the fact that

$$\begin{aligned} \|\sum_{j\le\ell} p_{(\ell,k);\lambda} 2^{j\alpha} u_\lambda\|_2 &\asymp \|(p_{(\ell,k);\lambda} 2^{j\alpha} 1_{\{j\le\ell\}})_\lambda\|_{\ell^2} \\ &\le 2^{\ell\alpha} ||(p_{(\ell,k);\lambda} 1_{\{j\le\ell\}})_\lambda||_{\ell^2} \\ &= 2^{\ell\alpha}. \end{aligned}$$

This leads to the possibility of an *inhomogeneous reproducing formula*

$$f(t) = \sum_k b_{\ell,k}(Kf)\phi_{\ell,k} + \sum_{j\ge l}\sum_k c_{(j,k)}(Kf)\psi_{j,k}. \tag{23}$$

The condition $\int \phi(t)dt = 1$ generally puts ϕ outside of the L^2 domain $\mathcal{D}(K)$. Nevertheless, we can make sense of this expression. We state without proof the following

Lemma 5 *Let K be one of the operators of section 3. Let the father ϕ be of compact support with M continuous derivatives and the mother ψ (mothers $\psi^{[\epsilon]}$ if $d > 1$) be of compact support, have M continuous derivatives and M vanishing moments. If M is sufficiently large, then $K\phi \in C(\mathbb{R}^d)$ and the functionals $b_{\ell,k}$ and c_λ have kernels in $L^1(\mathbb{R}^d)$. Consequently, whenever f is an inhomogeneous wavelet expansion (22) with only finitely many nonzero terms, (23) holds.*

5.3 Functions supported in a Cube

In later sections, we will be interested in expansions for functions supported in a fixed cube Q. We suppose that the gross-structure index ℓ is chosen so that $|supp(\phi_{\ell,k})| \ll |Q|$, i.e. so that $2^{-\ell}$ is small compared with the side length of Q. Because the wavelets ϕ and ψ are of compact support, only a finite number at each level of the expansion have supports intersecting the cube Q. We introduce the set Λ of all λ which may appear with nonzero coefficients for functions supported in Q. $\Lambda = \Lambda_{\ell-1} \cup \Lambda_\ell \cup \ldots \cup \Lambda_j \cup \ldots$. Here $\Lambda_{\ell-1}$ denotes those pairs (ℓ,k) where $|supp(\phi_{\ell,k}) \cap Q| > 0$; and Λ_j denotes, for $j \ge \ell$, those pairs (j,k) for which $|supp(\psi_{j,k}) \cap Q| > 0$. Roughly speaking, $\#\Lambda_j \sim 2^j|Q| + 2S$ where $2S$ is the support width of ψ.

Finally, we define the basis set for $L^2(Q)$, $(\tilde\psi_\lambda)_{\lambda\in\Lambda}$, by setting $\tilde\psi_\lambda = \psi_\lambda$ if $\lambda \in \Lambda_j$, and $\tilde\psi_\lambda = \phi_{\ell,k}$ if $\lambda \in \Lambda_{\ell-1}$. Then for an object supported in Q we have the expansion

$$f = \sum_\Lambda \langle f, \tilde\psi_\lambda\rangle \tilde\psi_\lambda. \tag{24}$$

Similarly, we define $\tilde c_\lambda$ in terms of either c_λ or $b_{l,k}$ depending on the index $\lambda \in \Lambda$, and get the reproducing formula

$$f = \sum_\lambda \tilde c_\lambda (Kf)\tilde\psi_\lambda \tag{25}$$

which is valid for all functions having finite expansions in (24).

5.4 WVD in Inhomogeneous Basis

We now define the WVD as used in our main result. Suppose that the operator K has a WVD in the homogeneous basis, with $||c_\lambda|| \sim 2^{j\alpha}$ as $j \to \infty$ for a certain $\alpha > 0$. We define $\kappa_{\ell-1} = 1$, $\kappa_j = 2^{-j\alpha}$ and let $\tilde{u}_\lambda$ be obtained by scaling the representer of $\tilde{c}_\lambda$ by κ_j: $\tilde{u}_\lambda = \tilde{\gamma_\lambda} \cdot \kappa_j$, $\lambda \in \Lambda$. This gives us the reproducing formula

$$f = \sum_\Lambda [\tilde{u}_\lambda, Kf] \kappa_j^{-1} \tilde{\psi}_\lambda;$$

what is more, the $\tilde{u}_\lambda$ are an almost-orthogonal set:

$$\| \sum_\Lambda \tilde{\alpha}_\lambda \tilde{u}_\lambda \|_2 \asymp ||(\tilde{\alpha}_\lambda)_\Lambda||_{\ell^2}.$$

This last property is the crucial one which makes the WVD worthwhile; we establish it as follows. Given an inhomogeneous representation $u = \sum_\Lambda \tilde{\alpha}_\lambda \tilde{u}_\lambda$, we also have the homogeneous representation $u = \sum \alpha_\lambda u_\lambda$. Hence

$$\| \sum_\Lambda \tilde{\alpha}_\lambda \tilde{u}_\lambda \|_2 = \| \sum_\lambda \alpha_\lambda u_\lambda \|_2 \asymp ||(\alpha_\lambda)||_{\ell^2} = ||(\tilde{\alpha}_\lambda)_\Lambda||_{\ell^2}.$$

The middle step was established for the homogeneous WVD in the last section. The last step follows from the fact that the transformation from coefficients $(\tilde{\alpha}_\lambda)$ to coefficients (α_λ), as a change of basis between two orthonormal bases, is an ℓ^2-isometry.

(In this section we have worked out the inhomogeneous expansion only for $supp(f) \subset Q \subset I\!R$. The same notation and reasoning can yield an inhomogeneous WVD on $I\!R^d$.)

6 Inversion from Noisy Data

We now turn to inversion in the presence of noisy data. We wish to recover f, an object known to be supported in a cube $Q \subset I\!R^d$, where $d = 1$ or 2 for simplicity. We will assume that we have the *white noise* observations

$$Y(du) = (Kf)(u)du + \epsilon W(du) \tag{26}$$

for all $u \in \mathcal{U}$, where W is a Brownian sheet on $\mathcal{U}$ (that is, the integral of a white noise). We know *a priori* that f belongs to a certain class of objects $\mathcal{F}$ which is a Besov ball (more below). Our goal is to recover f as accurately as possible, i.e. to construct an estimator $\hat{f} = \hat{f}_Y$ with worst case risk $\sup_\mathcal{F} E||\hat{f} - f||^2_{L^2(Q)}$ approaching the minimax risk

$$\mathcal{R}^*(\epsilon; \mathcal{F}) = \inf_{\hat{f}} \sup_{\mathcal{F}} E||\hat{f} - f||^2_{L^2(Q)}.$$

6.1 Form of the Estimator

Our proposal is to form the (inhomogeneous) WVD coefficients of the empirical data,

$$y_\lambda = \tilde{c}_\lambda(Y) \qquad \lambda \in \Lambda$$

and to operate on these coefficients nonlinearly, via the form

$$\hat{\alpha}_\lambda = \delta_{t_j}(y_\lambda) \qquad \lambda \in \Lambda$$

where the nonlinearity δ_t is a soft threshold. The reconstruction rule is

$$\hat{f}_{(t_j)} = \sum_{\Lambda} \hat{\alpha}_\lambda \tilde{\psi}_\lambda. \tag{27}$$

Variations on this approach are possible. For example, one may use hard thresholding rather than soft thresholding. One may pick the threshold adaptively from the data, or by minimizing a risk measure.

It might seem that the estimator just described is too simple, and that it can be improved upon by more complicated approaches. Here we show that, on the contrary, the estimator has a minimax optimality.

6.2 Optimality of the Proposal

The object f is known to lie in a ball $\mathcal{F}$ of the Besov space $B^\sigma_{p,q}$. More precisely, [23, 24, 25, 33, 36] the wavelet coefficients of f are known to belong to the set $\Theta^s_{p,q}$ of sequences $(\tilde{\alpha}_\lambda)_\Lambda$ obeying

$$\left(\sum_{j \geq \ell} 2^{jsq} \Big(\sum_{\lambda \in \Lambda_j} |\tilde{\alpha}_\lambda|^p \Big)^{q/p} \right)^{1/q} \leq C. \tag{28}$$

(Fine points: (1) $s = \sigma + d(1/2 - 1/p)$; (2) the sum here omits $\Lambda_{\ell-1}$). We define the *minimax-wavelet threshold risk*:

$$\mathcal{R}^W(\epsilon; \mathcal{F}) = \inf_{(t_j)} \sup_{\mathcal{F}} E||\hat{f}_{(t_j)} - f||^2_{L^2(Q)}.$$

This is the best that can be done, in general, for recovering objects from $\mathcal{F}$ by rules of the form described above, when we choose (t_j) optimally for use with $\mathcal{F}$.

For comparison purposes, we define the minimax linear risk:

$$\mathcal{R}^L(\epsilon; \mathcal{F}) = \inf_{\hat{f} \text{ linear}} \sup_{\mathcal{F}} E||\hat{f} - f||^2_{L^2(Q)}.$$

Theorem 4 *Let K have an inhomogeneous WVD with quasi-singular-values $k_j = 2^{-j\alpha}$ for $j \geq \ell$. Let $\mathcal{F}$ be a Besov ball (28). Suppose that $\sigma > \frac{2-p}{p}\alpha + d(1/p - 1/2)$. The minimax risk tends to zero as $\epsilon \to 0$; in fact*

$$\mathcal{R}^*(\epsilon; \mathcal{F}) \asymp \epsilon^{2r}, \qquad \epsilon \to 0, \tag{29}$$

with rate exponent

$$r = \frac{\sigma}{\sigma + d/2 + \alpha}; \tag{30}$$

Moreover, the minimax wavelet risk tends to zero at the optimal rate:

$$\mathcal{R}^W(\epsilon; \mathcal{F}) \leq Const \cdot \mathcal{R}^*(\epsilon; \mathcal{F}), \qquad \epsilon \to 0. \tag{31}$$

Finally, the minimax linear risk tends to zero

$$\mathcal{R}^L(\epsilon; \mathcal{F}) \asymp \epsilon^{2r'}, \qquad \epsilon \to 0,$$

at the rate

$$r' = \frac{\sigma + d(1/2 - 1/p_-)}{\sigma + d(1 - 1/p_-) + \alpha}, \tag{32}$$

where $p_- = \min(2, p)$*, which is slower than the minimax rate in case* $p < 2$*.*

The proof of this result occupies sections 7, 8 and 9.

6.3 Interpretation: White Noise Model

Some background about the white noise model (26) may help the reader interpret this theorem.

The author is unaware of any real scientific problem where the available data could be described as obeying the white noise model (26). In the Soviet literature of the 1980's, the white noise model has been extensively studied for its own sake, and several elegant and surprising results have been discovered [28, 41].

We study the white noise model because it arises as the large-sample limit of various "real" estimation problems, such as nonparametric regression, density estimation, and time series spectral estimation. A significant trend in the 1980's was the development of asymptotic results in estimation of functions from noisy data by first, solving a problem in the white noise model, and then establishing a correspondence theorem which showed that the solution obtained in the white noise model could be applied in some "real" estimation problems to obtain asymptotically optimal results. Examples of this approach are Efroimovich and Pinsker (1981,1982), Nussbaum (1985), Donoho and Liu (1990), Donoho and Nussbaum (1990), Johnstone and Silverman (1990, 1991), Donoho and Johnstone (1991), Donoho (1991). General results about this approach are given by Brown and Low (1990) and Low(1992).

As a result, one knows from experience that the solution of the white noise model leads to a variety of applications. As the white noise model is mathematically more tractable and homogeneous than the models describing 'real' situations, we study that model here, and leave applications to later work, which presumably would also include reconstructions based on real data.

We briefly describe two results which can be developed on the basis of the white noise calculations given here.

6.3.1 Numerical Differentiation

Let f be an unknown function Let $\mathcal{F}_0$ be the class of functions f supported in $[1/4, 3/4]$,which belong to a Besov ball $\mathcal{F}$ given by (28). Suppose that we observe samples

$$y_i = \int_0^{i/n} f(t)dt + \sigma z_i, \qquad i = 1, \ldots, n, \tag{33}$$

where z_i are i.i.d. N(0,1). Hence we have equispaced samples of the integral of f, observed with sampled white noise. Our goal is to recover f with small L^2 loss. (We assume that we observe data on an interval strictly containing the support of f in order to avoid discussion of boundary effects in this paper.)

We define the process $Y_n(u)$ at the points i/n by

$$Y_n(i/n) = n^{-1} \sum_{k \le i} y_k$$

and we interpolate between the points i/n using Brownian Bridges

$$Y_n(u) = Y_n(i/n) + (t - i/n)y_i + W_0(n(t - i/n))/n$$

for $i/n < u < (i+1)/n$.

Setting $\epsilon = \sigma/\sqrt{n}$, this is an approximation to the process $Y(u)$ defined by (26) for $u \in [0,1]$. For example, for each continuous function $\gamma(\cdot)$ on $[0,1]$, the distribution of $\int \gamma(u) Y(du)$ is $N(\int \gamma(u) g(u) du, \epsilon^2 \int \gamma^2 du)$. On the other hand the distribution of $\int \gamma(u) Y_n(du)$ is $N(n^{-1} \sum_i \gamma(i/n) g(i/n), \epsilon^2 n^{-1} \sum_i \gamma(i/n)^2)$. Hence, for large n, nice linear functionals of Y have almost the same distribution as the same linear functionals of Y_n.

Picking ℓ sufficiently large, all the indices $\lambda \in \Lambda$ satisfy $supp(\tilde{\psi}_\lambda) \in [0,1]$. Hence, every coefficient needed in the expansion $f = \sum_\Lambda \alpha_\lambda \tilde{\psi}_\lambda$ corresponds to a functional c_λ supported in $[0,1]$. Under regularity conditions on our wavelet basis, the functional c_λ has a representer with considerable regularity. We conclude that every functional $c_\lambda(Y_n)$ has almost the same distribution as $c_\lambda(Y)$ for large n.

As a result, one may be convinced of the plausibility of the following result, whose proof we omit.

Theorem 5 *Let $R(n, \mathcal{F}_0)$ denote the minimax risk for estimation of f from the observations (33). Suppose that members of the class $\mathcal{F}_0$ are uniformly bounded: $\sup_{\mathcal{F}_0} ||f||_{L^\infty} \le M$. We have asymptotic equivalence between risk in the sampled-data model (33) and the white-noise models (26)*

$$R(n, \mathcal{F}_0) \sim \mathcal{R}^*(\frac{\sigma}{\sqrt{n}}; \mathcal{F}_0) \qquad n \to \infty.$$

Let $R^L(n, \mathcal{F})$ denote the linear minimax risk for estimation of f from the observations (33). Then

$$R^L(n, \mathcal{F}_0) \sim \mathcal{R}^L(\frac{\sigma}{\sqrt{n}}; \mathcal{F}_0) \qquad n \to \infty.$$

Here is an immediate implication. Let $\mathcal{F}_0$ be the class of objects known to be supported in $[1/4, 3/4]$ and of bounded variation less than or equal to C. Then $\mathcal{F}_0$ is contained in a Besov Ball $B^1_{1,\infty}[0,1]$, and therefore from Theorem 4 with $p = 1$, $\alpha = 1$, and $\sigma = 1$ we get

$$R(n, \mathcal{F}_0) \le Const \cdot n^{-2/5}$$

However, $\mathcal{F}_0$ contains a Besov Ball $B^1_{1,1}[1/4, 3/4]$ of the type in section 7.5 below, and therefore the minimax rate for linear estimates is not better than

$$R^L(n, \mathcal{F}_0) \ge const \cdot n^{-1/4}.$$

Traditional methods of numerical differentiation are linear. All such methods are outperformed in rate of convergence by our proposal, for the class of objects of bounded variation.

6.3.2 Radon Transform

Suppose that events happen at points P_1, P_2, ..., P_n i.i.d. f, where f is a density supported in a cube Q in $I\!R^2$. However, we are not informed of the location of any such point, but only that an event has occurred on a line containing the point; the line is randomly and uniformly oriented, independently of the position of the point. Equivalently, we observe U_1, U_2, ..., U_n i.i.d. g where g is a density on $I\!R \times [0, \pi]$, obtained by Radon transformation of f. Let Y_n be the empirical process on $I\!R \times [0, \pi]$ gotten by summing Dirac Delta measures at the U_i: $Y_n(\cdot) = n^{-1} \sum_{i=1}^{n} \delta_{U_i}(\cdot)$. Let $c(Y) = \int \gamma(u) Y(du)$ be a linear functional with bounded measurable representer γ. Then $Ec(Y_n) = \int \gamma(u) g(u) du$ and

$$n \cdot Var(c(Y_n)) = \int \gamma(u)^2 g(u) du - (\int \gamma(u) g(u) du)^2$$

so, if $||g||_\infty \le M$, $Var(c(Y_n)) \le M/n \int \gamma^2(u) du$. Furthermore, $c(Y_n)$ has an asymptotically Gaussian distribution. In short, $c(Y_n)$ is no worse an estimator of $c(Kf)$ than would be $c(Y)$ in the white noise model at noise level $\epsilon = \sqrt{M}/\sqrt{n}$. (This simple idea is stronger than it sounds at first. Donoho and Liu (1991) have used this principle to develop density estimates within a few percent of asymptotically minimax.)

This suggests the following approach to inversion of Radon data: treat the functionals $\tilde{c}_\lambda(Y_n)$ of the density data as if they arose from the white noise model $\tilde{c}_\lambda(Y)$, at noise level $\epsilon = \sqrt{M/n}$. Consequently, one proposes the inversion formula

$$\hat{f}_n = \sum_{\Lambda} \delta_{t_j}(\tilde{c}_\lambda(Y_n)) \tilde{\psi}_\lambda. \tag{34}$$

Johnstone, Kerkyacharian, and Picard (1991) have analysed the behavior of density estimates by wavelets, and in particular estimates of the form $\hat{f} = \sum \delta_{t_j}(\langle \tilde{\psi}_\lambda, Y_n \rangle) \tilde{\psi}_\lambda$. They do not directly analyze the problem we study here, but their lemmas may be interpreted as support for the idea that nonlinear thresholding of wavelet coefficients behaves as if those coefficients arose from a white noise model at noise level $\sqrt{M/n}$. My own preliminary calculations suggest that our main result for the white noise model carries over to the density model. Thus, it appears:

First, that the estimate $\hat{f}_n$ defined by (34), when tuned for $\mathcal{F}$ the collection of densities supported in Q and belonging to the Besov space $B^\sigma_{p,q}$, with Radon transforms g bounded by $||g||_\infty \le M$,

$$\sup_{\mathcal{F}} E||\hat{f}_n - f||_{L^2(Q)} \le Const \cdot (\frac{M}{n})^r$$

where $r = r(\sigma, p, \alpha)$ is as in our main result. Second, that this rate is optimal among all measurable estimators of the data. Third, that for linear estimators $\hat{f}_{n,L}$, one has

$$\sup_{\mathcal{F}} E||\hat{f}_{n,L} - f||_{L^2(Q)} \ge const \cdot (\frac{M}{n})^{r'}.$$

where r' is as in our white noise result.

These conclusions would imply that, when $p < 2$, linear estimates are suboptimal for Radon inversion from density data and that WVD-based nonlinear estimates are asymptotically within a constant factor of optimal.

Suppose we are trying to recover functions in the 2-dimensional Bump algebra $B^2_{1,1}(I\!R^2)$ known to be supported inside a cube Q. We have Radon density data as indicated above. Then for Theorem 4 we have $\sigma = 2$, $d = 2$, $\alpha = 1/2$. The minimax rate for a set of this form is $n^{-4/7}$, while $p_- = 1$ and so the minimax linear rate is $n^{-2/5}$.

7 Asymptotics of Minimax Risk

In this section we prove the first part of our main result, concerning the asymptotics of the minimax risk. The procedure we have proposed is based on the use of data $y_\lambda = \tilde{c}_\lambda(Y)$, $\lambda \in \Lambda$. Now

$$y_\lambda = \alpha_\lambda + \epsilon\sigma_j z_\lambda \qquad \lambda \in \Lambda \tag{35}$$

where α_λ is the wavelet coefficient $\langle f, \tilde{\psi}_\lambda \rangle$ of the object to be recovered, $\sigma_j = \kappa_j^{-1}$ is (essentially) the norm of $\tilde{c}_\lambda$, and the noise process

$$z_\lambda = [\tilde{u}_\lambda, W] \qquad \lambda \in \Lambda$$

where W is the Wiener sheet ($d > 1$) or Wiener Process ($d = 1$). We are interested in estimating α; because wavelets provide a complete orthonormal system the measure of loss as originally stated, $||\hat{f} - f||^2_{L^2(Q)}$ is essentially the same as $||\hat{\alpha} - \alpha||^2_{\ell^2(\Lambda)}$. Indeed, the function space loss is less than the sequence space loss; but it will follow from section 7.5 below that the two losses yield the same risk asymptotics.

We therefore shift attention to sequence space. We assume that $\alpha \in \Theta^s_{p,q}$, and our measure of performance is the minimax risk

$$\mathcal{R}^*_z(\epsilon) = \inf_{\hat{\alpha}} \sup_{\Theta^s_{p,q}} E||\hat{\alpha}(y) - \alpha||^2_{\ell^2(\Lambda)}.$$

The main fact to emerge below is that the almost-orthogonality of the (u_λ) – which derives from the vaguelettes property – allows us to study the following *discrete white noise* observations

$$x_\lambda = \alpha_\lambda + \epsilon\sigma_j w_\lambda \qquad \lambda \in \Lambda \tag{36}$$

where (w_λ) is a Gaussian white noise. The minimax risk from these observations,

$$\mathcal{R}^*_w(\epsilon) = \inf_{\hat{\alpha}} \sup_{\Theta^s_{p,q}} E||\hat{\alpha}(x) - \alpha||^2_{\ell^2(\Lambda)},$$

will be equivalent, to within constants, to the desired minimax risk $\mathcal{R}^*_z(\epsilon)$.

7.1 Bayes-Minimax Risk

As in Donoho and Johnstone (1990), hereafter **[DJ]**, we develop upper bounds by a Bayes-Minimax approach. We assume we have data (35) where (z_λ) is as before, but now (α_λ) is considered a random field, obeying the constraint that the normalized p-th moments

$$\tau_\lambda = (E|\alpha_\lambda|^p)^{1/p}$$

belong to $\Theta^s_{p,q}$, rather than (α_λ) itself. This is a "softened" or "in-mean" constraint on α rather than a "hard" constraint. We define the minimax Bayes risk

$$\mathcal{B}^*_z(\epsilon) = \inf_{\hat{\alpha}} \sup_{\tau \in \Theta^s_{p,q}} E||\hat{\alpha}(y) - \alpha||^2$$

and similarly for the white noise observations

$$\mathcal{B}^*_w(\epsilon) = \inf_{\hat{\alpha}} \sup_{\tau \in \Theta^s_{p,q}} E||\hat{\alpha}(x) - \alpha||^2.$$

Because every element of $\Theta^s_{p,q}$ may be viewed as a deterministic process whose moments lie in Θ, we have

$$\mathcal{B}^*_z(\epsilon) \geq \mathcal{R}^*_z(\epsilon), \qquad \mathcal{B}^*_w(\epsilon) \geq \mathcal{R}^*_w(\epsilon)$$

relating the Bayes-minimax risks to the minimax risks. We will see in section 7.5 below that, as $\epsilon \to 0$, the Bayes-minimax and the minimax risks become equivalent; but the Bayes-minimax is more convenient to work with because the randomness of α entails a certain separability of the problem.

7.2 Near-Independence of the Noise

In the definition of the W.V.D. and in section 4, we have placed emphasis on the near-orthogonality of the coefficient representers $\tilde{u}_\lambda$. The following explains why.

Lemma 6 *Let Λ be a discrete countable set. Suppose that we define a zero-mean Gaussian random field on Λ by*

$$z_\lambda = [\tilde{u}_\lambda, W] \qquad \lambda \in \Lambda,$$

where $(\tilde{u}_\lambda)$ is a collection of elements of $L^2(du)$ and $W(du)$ is white noise. Suppose that for positive finite constants A_i

$$A_0 \leq ||\tilde{u}_\lambda||_2 \leq A_1 \qquad \lambda \in \Lambda,$$

and that for positive finite constants B_i

$$B_0||(\alpha_\lambda)_\Lambda||_{\ell^2} \leq ||\sum_\Lambda \alpha_\lambda \tilde{u}_\lambda||_2 \leq B_1||(\alpha_\lambda)_\Lambda||_{\ell^2}.$$

Then there exist positive finite constants $0 < \gamma_0 < \gamma_1 < \infty$ so that

$$\gamma_0^2 \leq Var(z_\lambda|(z_\mu)_{\lambda \neq \mu}) \leq Var(z_\lambda) \leq \gamma_1^2, \tag{37}$$

for each $\lambda \in \Lambda$. We may take $\gamma_0 = B_0$ and $\gamma_1 = A_1$.

In the sequel, we call a noise process obeying (37) a *nearly-independent* noise.

Proof. By the L_2-isometry of Abstract Wiener space

$$Var(z_\lambda) \equiv Var([\tilde{u}_\lambda, W]) = ||\tilde{u}_\lambda||^2$$

so $Var(z_\lambda) \le A_1^2 = \gamma_1^2$, say.

On the other hand, as $(z_\lambda)_{\lambda\in\Lambda}$ is a zero-mean Gaussian random field,

$$Var(z_\lambda|(z_\mu)_{\lambda\neq\mu}) = \inf_{(\alpha_\mu)} Var(z_\lambda - \sum_{\lambda\neq\mu} \alpha_\mu z_\mu).$$

Consequently,

$$\begin{aligned} EVar(z_\lambda|(z_\mu)_{\lambda\neq\mu}) &= \inf\{Var(\sum \alpha_\mu z_\mu) : \alpha_\lambda = 1\} \\ &= \inf\{||\sum \alpha_\mu \tilde{u}_\mu||^2 : \alpha_\lambda = 1\} \\ &\ge \inf\{B_0^2 \sum \alpha_\mu^2 : \alpha_\lambda = 1\} \\ &= B_0^2 = \gamma_0^2, \qquad \text{say.} \end{aligned}$$

7.3 Equivalence to White Noise

We now reduce the problem of determining Bayes-minimax risk with noise (z_λ) to the problem of determining the same quantity in the discrete white-noise model (36).

Theorem 6 *Let $(z_\lambda)_{\lambda\in\Lambda}$ be an almost-independent Gaussian noise (37). Then*

$$\mathcal{B}_w^*(\gamma_0\epsilon) \le \mathcal{B}_z^*(\epsilon) \le \mathcal{B}_w^*(\gamma_1\epsilon).$$

Hence, up to constants, asymptotics for the observations (y_λ) derive from those for (x_λ).

To prove this, we introduce more notation. Let π denote the probability distribution (prior distribution) of (α_λ). The Bayes-risk is defined as

$$\mathcal{B}_z(\epsilon,\pi) = \sum_\Lambda E(E\{\alpha_\lambda|y\} - \alpha_\lambda)^2$$

where the expectations refer to the joint distribution on the space of pairs $\{(\alpha_\lambda)_{\lambda\in\Lambda}, (y_\lambda)_{\lambda\in\Lambda}\}$ which derives from the prior π and the noise process (z_λ). $\mathcal{B}_w(\epsilon,\pi)$ is defined similarly in the discrete white noise model.

The individual Bayes risks are interesting because of their relation to minimax Bayes risks. Indeed, applying the Minimax Theorem of statistical decision theory (Le Cam, 1986), we have

$$\mathcal{B}_z(\epsilon;\Theta) = \sup\{\mathcal{B}_z(\epsilon,\pi) : \tau \in \Theta\}$$
$$\mathcal{B}_w(\epsilon;\Theta) = \sup\{\mathcal{B}_w(\epsilon,\pi) : \tau \in \Theta\}.$$

However, there is an additional level of structure in the discrete white noise case.

Lemma 7 *Let π be a prior distribution and let $\tilde{\pi}$ be a prior with the same marginal distribution, but independent coordinates:*

$$\tilde{\pi}\{\alpha_\lambda > t\} = \pi\{\alpha_\lambda > t\} \qquad \textit{for all } t, \lambda.$$

Then $\tilde{\pi}$ is less favorable than π:

$$\mathcal{B}_w(\epsilon,\pi) \le \mathcal{B}_w(\epsilon,\tilde{\pi}).$$

Proof. Let $\tilde{x}_\lambda$ denote the process with discrete white noise and independent coordinates obtained with prior $\tilde{\pi}$.

$$\begin{aligned} Var\{\alpha_\lambda|(x_\lambda)_{\lambda\in\Lambda}\} &\leq Var\{\alpha_\lambda|x_\lambda\} \\ &= Var\{\alpha_\lambda|\tilde{x}_\lambda\} = Var\{\alpha_\lambda|\tilde{x}\}. \end{aligned}$$

Summing across coordinates gives

$$\begin{aligned} B_w(\epsilon,\pi) &= \sum_\Lambda Var\{\alpha_\lambda|(x_\lambda)_{\lambda\in\Lambda}\} \\ &\leq \sum_\Lambda Var\{\alpha_\lambda|\tilde{x}\} = \mathcal{B}_w(\epsilon,\tilde{\pi}). \end{aligned}$$

This lemma means that in searching for least-favorable priors in the discrete white noise model we need only consider independent coordinate priors. Let $\tilde{\Pi}$ denote the collection of such priors. Such priors have $E\{\alpha_\lambda|x\} = E\{\alpha_\lambda|x_\lambda\}$, so that particularly simple estimators – coordinate-by-coordinate nonlinearities $\hat{\alpha_\lambda} = \delta_\lambda(x_\lambda)$ – are optimal. Moreover, one has the formula

$$Var\{\alpha_\lambda|x\} = Var\{\alpha_\lambda|x_\lambda\}, \qquad \tilde{\pi} \in \tilde{\Pi}$$

which leads to

$$\mathcal{B}_w(\epsilon,\tilde{\pi}) = \sum_\Lambda Var\{\alpha_\lambda|x_\lambda\}, \qquad \tilde{\pi} \in \tilde{\Pi}.$$

The theorem rests on the following inequalities for Bayes risks with coordinatewise independent priors.

Lemma 8 *Let (z_λ) be an almost-independent noise (37). Let π be any prior distribution on sequences (α_λ) and let $\tilde{\pi}$ be the prior with the same marginal distributions but independent coordinates.*

$$\mathcal{B}_w(\gamma_0\epsilon,\tilde{\pi}) \leq \mathcal{B}_z(\epsilon,\pi) \leq \mathcal{B}_w(\gamma_1\epsilon,\tilde{\pi})$$

Let us see how these lemmas imply the theorem. First, the lower bound.

$$\begin{aligned} \mathcal{B}_z^*(\epsilon) &= \sup\{\mathcal{B}_z(\epsilon,\pi) : \tau \in \Theta\} \\ &\geq \sup\{\mathcal{B}_z(\epsilon,\tilde{\pi}) : \tau \in \Theta, \tilde{\pi} \in \tilde{\Pi}\} \\ &\geq \sup\{\mathcal{B}_w(\gamma_0\epsilon,\tilde{\pi}) : \tau \in \Theta, \tilde{\pi} \in \tilde{\Pi}\} \\ &= \mathcal{B}_w^*(\gamma_0\epsilon). \end{aligned}$$

Then the upper bound

$$\begin{aligned} \mathcal{B}_z^*(\epsilon) &= \sup\{\mathcal{B}_z(\epsilon,\pi) : \tau \in \Theta\} \\ &\leq \sup\{\mathcal{B}_w(\gamma_1\epsilon,\tilde{\pi}) : \tau \in \Theta, \tilde{\pi} \in \tilde{\Pi}\} \\ &= \mathcal{B}_w^*(\gamma_1\epsilon). \end{aligned}$$

Proof of Lemma 8. Three observations, stated without proof, will be useful.

Lemma 9 *Let (Y_0, Y_1, Y_2) be a Markov Chain. Then*

$$E\,Var\{Y_2|Y_1\} \le E\,Var\{Y_2|Y_0\}.$$

Lemma 10 *Let the reversed process (Y_2, Y_1, Y_0) be an independent-increments process, i.e. let $\Delta_1 = Y_2 - Y_1$, $\Delta_0 = Y_1 - Y_0$ be independent of each other and of Y_2. Then the forward process (Y_0, Y_1, Y_2) is a Markov Chain.*

Lemma 11 *Let (z_λ) be a zero-mean Gaussian random field. Then*

$$z_\lambda = E\{z_\lambda|(z_\mu)_{\lambda\neq\mu}\} + e_\lambda$$

where the random variable e_λ is Gaussian and independent of $(z_\mu)_{\lambda\neq\mu}$, and the random variable

$$\hat{z_\lambda} = E\{z_\lambda|(z_\mu)_{\lambda\neq\mu}\}$$

is also Gaussian and is obtained by a linear combination of the $(z_\mu)_{\lambda\neq\mu}$.

We use these, first to prove the upper bound. By monotonicity of conditional variance in the conditioning set

$$E\,Var(\alpha_\lambda|(y_\lambda)_{\lambda\in\Lambda}) \le E\,Var(\alpha_\lambda|y_\lambda).$$

Let $\tilde{z}_\lambda$ be an independent-coordinate Gaussian process, independent also of (α_λ) and (y_λ), and having variance $Var(\tilde{z}_\lambda) = \gamma_1^2 - Var(z_\lambda) \ge 0$. Define

$$\tilde{y}_\lambda = y_\lambda + \tilde{z}_\lambda \qquad \lambda \in \Lambda.$$

Then setting $Y_2 = \alpha_\lambda$, $Y_1 = y_\lambda$ and $Y_0 = \tilde{y}_\lambda$, the reversed process (Y_2, Y_1, Y_0) is an independent increments process, so the process (Y_0, Y_1, Y_2) is Markov. Hence

$$E\,Var(\alpha_\lambda|y_\lambda) \le E\,Var(\alpha_\lambda|\tilde{y}_\lambda).$$

Summing across coordinates,

$$\mathcal{B}_z(\epsilon, \pi) \le \sum_\Lambda E\,Var(\alpha_\lambda|\tilde{y}_\lambda). \tag{38}$$

Define now the independent process

$$\tilde{x}_\lambda = \tilde{\alpha}_\lambda + \epsilon\gamma_1\sigma_j w_\lambda \qquad \lambda \in \Lambda$$

where $(\tilde{\alpha}_\lambda) \sim \tilde{\pi}$, and $\tilde{\pi}$ has the same marginal distribution as π, but independent coordinates. Then $E\,Var(\tilde{\alpha}_\lambda|\tilde{x}) = E\,Var(\tilde{\alpha}_\lambda|\tilde{x}_\lambda)$ and so, summing across coordinates,

$$\mathcal{B}_w(\gamma_1\epsilon, \tilde{\pi}) = \sum_\Lambda E\,Var(\tilde{\alpha}_\lambda|\tilde{x}_\lambda). \tag{39}$$

Now the construction of $\tilde{y}$ has guaranteed that for each fixed λ, $(\tilde{\alpha}_\lambda, \tilde{x}_\lambda) =_D (\alpha_\lambda, \tilde{y}_\lambda)$. It follows that

$$E\,Var(\alpha_\lambda|\tilde{y}_\lambda) = E\,Var(\tilde{\alpha}_\lambda|\tilde{x}_\lambda)$$

and so, comparing (38) and (39) we obtain the upper bound

$$\mathcal{B}_w(\epsilon, \pi) \le \mathcal{B}_w(\gamma_1 \epsilon, \tilde{\pi}).$$

We now turn to the lower bound. Define $Y_2 = \alpha_\lambda$, $Y_1 = \alpha_\lambda + e_\lambda$, $Y_0 = y_\lambda \equiv \alpha_\lambda + e_\lambda + \hat{z_\lambda}$. Lemma 11 tells us that $\Delta_1 = e_\lambda$ and $\Delta_0 = \hat{z_\lambda}$ are independent of each other; by hypothesis they are independent of α_λ. Hence the reversed process (Y_2, Y_1, Y_0) is an independent increments process, and so the forward process is Markovian. To apply this, Lemma 11 shows the random variables $(z_\mu)_{\lambda \neq \mu}$ can be transformed by a linear isometry into the random variables $(\hat{z}_\lambda, Z_1, Z_2, \ldots)$ where $\hat{z}_\lambda = E\{z_\lambda | (z_\mu)_{\lambda \neq \mu}\}$ and $Z_1, Z_2, \ldots$ are independent of α_λ and of y_λ. Then, omitting the proviso "almost surely" to save ink,

$$\begin{aligned}
Var(\alpha_\lambda | (y_\lambda)_{\lambda \in \Lambda}) &\ge Var(\alpha_\lambda | (y_\lambda)_{\lambda \in \Lambda}, (\alpha_\mu)_{\lambda \neq \mu}) \\
&= Var(\alpha_\lambda | y_\lambda, (z_\mu)_{\lambda \neq \mu}, (\alpha_\mu)_{\lambda \neq \mu}) \\
&= Var(\alpha_\lambda | y_\lambda, \hat{z}_\lambda, Z_1, Z_2, \ldots, (\alpha_\mu)_{\lambda \neq \mu}) \\
&= Var(\alpha_\lambda | y_\lambda, \hat{z}_\lambda) \\
&= Var(\alpha_\lambda | y_\lambda - \hat{z}_\lambda, \hat{z}_\lambda) \\
&= Var(Y_2 | Y_1, Y_0) = Var(Y_2 | Y_1) \\
&= Var(\alpha_\lambda | \alpha_\lambda + e_\lambda).
\end{aligned}$$

This inequality is the heart of the lower bound we seek.

Recall that $Var(e_\lambda) \ge \gamma_0^2$ by hypothesis. Let the Gaussian field $(\tilde{z}_\lambda)$ be defined so that it is independent of (y_λ) and (z_λ), and so that $Var(\tilde{z}_\lambda) = Var(e_\lambda) - \gamma_0^2$. Define

$$x_\lambda = \tilde{\alpha}_\lambda + \epsilon \gamma_0 \sigma_j w_\lambda \qquad \lambda \in \Lambda$$

where $\tilde{\alpha}_\lambda$ has the same marginals as α_λ, and independent coordinates. Set $\tilde{x}_\lambda = x_\lambda + \tilde{z}_\lambda$. Then $(\tilde{\alpha}_\lambda, \tilde{x}_\lambda) =_D (\alpha_\lambda, \alpha_\lambda + e_\lambda)$ for each fixed λ. Hence $E\, Var(\tilde{\alpha}_\lambda | \tilde{x}_\lambda) = E\, Var(\alpha_\lambda | \alpha_\lambda + e_\lambda)$. But the process $Y_0 = \tilde{x}_\lambda$, $Y_1 = x_\lambda$, $Y_2 = \alpha_\lambda$ is again Markovian, so $E\, Var(\tilde{\alpha}_\lambda | \tilde{x}_\lambda) \le E\, Var(\alpha_\lambda | \alpha_\lambda + e_\lambda)$. Summing over coordinates,

$$\begin{aligned}
\mathcal{B}_w(\gamma_0 \epsilon, \tilde{\pi}) &\le \sum_\Lambda E\, Var(\alpha_\lambda | \alpha_\lambda + e_\lambda) \\
&\le \mathcal{B}_z(\epsilon, \tilde{\pi}).
\end{aligned}$$

and the proof is complete.

7.4 Asymptotics via Dyadic Renormalization

We now concentrate exclusively on the discrete white noise observations (36). In this section we show that

$$\mathcal{B}_w^*(\epsilon) \asymp (\epsilon^2)^r \qquad \text{as } \epsilon \to 0$$

with rate exponent

$$r(s, p, \alpha) = \frac{s + d(1/p - 1/2)}{s + d/p + \alpha}.$$

In section 7.6 we will see how this implies our results on minimax risk. We remark that **[DJ]** have thoroughly studied the case of minimax risk from observations $x_\lambda = \alpha_\lambda + \epsilon w_\lambda$, which may be thought of the special case $\sigma_j = 1$ of (36). Our treatment of the more general case follows along the same lines. We assume in this section that $p \le q$ in (28) the other case can be handled much as in **[DJ]**.

First, we reduce evaluation of $\mathcal{B}_w^*(\epsilon)$ to a certain optimization problem in sequence space. Following **[DJ]**, consider the problem of inference from scalar data $x = \theta + \sigma w$, where θ and w are independent random variables, w is $N(0,1)$, and θ has an unknown distribution which is known to obey the constraint $E|\theta|^p \le \tau^p$. The minimax Bayes risk for this problem is

$$\rho_p(\tau, \sigma) = \sup\{E\, Var(\theta|x) : x = \theta + \sigma w, w \sim N(0,1), E|\theta|^p \le \tau^p\}.$$

Now if $(\alpha_\lambda)_{\lambda \in \Lambda}$ are random variables with moments $\tau_\lambda = (E|\alpha_\lambda|^p)^{1/p}$, we have

$$E\, Var(\alpha_\lambda | x_\lambda) \le \rho_p(\tau_\lambda, \epsilon\sigma_j)$$

by definition of $x_\lambda = \alpha_\lambda + \epsilon\sigma_j w_\lambda$ and ρ_p. Hence, if π is a prior making the coordinates independent, the Bayes-risk satisfies

$$\mathcal{B}_w(\epsilon, \pi) \le \sum_\Lambda \rho_p(\tau_\lambda, \epsilon\sigma_j)$$

with equality if the (independent) coordinate priors are chosen carefully, subject to the constraint $E|\alpha_\lambda|^p \le \tau_\lambda^p$. As the least favorable distribution in the white noise model has independent coordinates,

$$\mathcal{B}_w^*(\epsilon) = \sup \sum_\Lambda \rho_p(\tau_\lambda, \epsilon\sigma_j) \text{ subject to } \sum_{j \ge \ell} (2^{js} (\sum_{\lambda \in \Lambda_j} \tau_\lambda^p)^{1/p})^q \le C^q.$$

Donoho and Johnstone (1989) show that ρ_p is concave in τ^p. Set $N_j = Card(\Lambda_j)$. If we hold $r_j^p = N_j^{-1} \sum_{\lambda \in \Lambda_j} \tau_\lambda^p$ fixed, the worst risk is therefore at

$$\tau_\lambda = r_j \qquad \lambda \in \Lambda_j.$$

Using this, the minimax Bayes risk is

$$\sup \sum_{j \ge \ell - 1} N_j \rho(r_j, \epsilon\sigma_j) \text{ subject to } \sum_{j \ge \ell} (2^{js} N_j^{1/p} r_j)^q \le C^q.$$

We get that

$$\mathcal{B}_w^*(\epsilon) = val(P_{\epsilon,C}) + N_{\ell-1}\epsilon^2,$$

where $(P_{\epsilon,C})$ is an optimization problem defined as follows. Applying the invariance $\rho(\tau, \sigma) = \sigma^2 \rho(\tau/\sigma, 1)$ and introducing variables $v_j = r_j/\sigma_j$, set

$$(P_{\epsilon,C}) \qquad \sup \qquad \epsilon^2 \sum_{j \ge \ell} N_j {\sigma_j}^2 \rho(v_j/\epsilon, 1) \text{ subject to } \sum_{j \ge \ell} (2^{js} N_j^{1/p} \sigma_j v_j)^q \le C^q. \tag{40}$$

As in [**DJ**] we introduce an associated problem whose asymptotics follow by renormalization arguments. We suppose that $\sigma_j \sim 2^{j\alpha}$ as $j \to \infty$ and that $Q = [0,1]^d$, so that $N_j \sim 2^{jd}$. Define the problem $(Q_{\epsilon,C})$ on the space of bilateral sequences $(v_j)_{j=-\infty}^{\infty}$ via

$$(Q_{\epsilon,C}) \qquad \sup \qquad \epsilon^2 \sum_{j=-\infty}^{\infty} 2^{j(d+2\alpha)} \rho(v_j, 1) \text{ subject to } \sum_{j=-\infty}^{\infty} (2^{j(s+d/p+\alpha)} v_j)^q \le C^q.$$

Now $(Q_{\epsilon,C})$ differs from $(P_{\epsilon,C})$ by the addition of terms at $j < \ell$, the substitution 2^{jd} for N_j and $2^{j\alpha}$ for σ_j. These substitutions do not change the leading-order asymptotics, and so

$$val(P_{\epsilon,C}) \sim val(Q_{\epsilon,C}) \text{ as } \epsilon \to 0.$$

The problem $(Q_{\epsilon,C})$ possesses a certain dyadic renormalization property which forces $val(Q_{\epsilon,C}) \asymp \epsilon^{2r}$. Let

$$J_{\alpha,\epsilon}(v) = \epsilon^2 \sum_{j=-\infty}^{\infty} 2^{j(d+2\alpha)} \rho(v_j/\epsilon, 1)$$

$$J_{s,p,q}(v) = \sum_{j=-\infty}^{\infty} (2^{j(s+d/p+\alpha)} v_j)^q,$$

so that $val(Q_{\epsilon,C})$ is the maximum of $J_{\alpha,\epsilon}$ subject to $J_{s,p,q}(v) \le C^q$. For positive real a and integer h, introduce the sequence operator $(\mathcal{U}_{a,h} v)_j = a v_{j-h}$. Then we have

$$J_{\alpha,\epsilon}(\mathcal{U}_{\epsilon,h} v) = J_{\alpha,1}(v) \cdot \epsilon^2 \cdot 2^{h(d+2\alpha)}$$

$$J_{s,p,q}(\mathcal{U}_{\epsilon,h} v) = J_{s,p,q}(v) \cdot \epsilon^q \cdot 2^{h(s+d/p+\alpha)}.$$

These two invariances have the following implications. Define Γ via

$$\Gamma(\epsilon, h) = (C/\epsilon)\, 2^{-h(s+d/p+\alpha)}.$$

Then let V_C denote the set of sequences feasible for $(Q_{\epsilon,C})$. The invariances above imply that

$$\mathcal{U}_{\epsilon,h} V_\Gamma = V_C$$
$$\mathcal{U}_{\epsilon^{-1},-h} V_C = V_\Gamma.$$

Hence

$$\begin{aligned} val(Q_{\epsilon,C}) &= \sup\{J_{\alpha,\epsilon}(v) : J_{s,p,q}(v) \le C^q\} \\ &= \sup\{J_{\alpha,\epsilon}(\mathcal{U}_{\epsilon,h} v) : \mathcal{U}_{\epsilon,h} v \in V_C\} \\ &= \sup\{\epsilon^2 2^{h(d+2\alpha)} J_{\alpha,1}(v) : v \in V_\Gamma\} \\ &= \epsilon^2 2^{h(d+2\alpha)} val(Q_{1,\Gamma}). \end{aligned}$$

In particular, if ϵ is of the special form

$$\epsilon_h = 2^{-h(s+d/p+\alpha)}, \qquad h = 1, 2, 3, \ldots$$

then $\Gamma(\epsilon_h, C) = C$ and so

$$val(Q_{\epsilon_h,C}) = (\epsilon_h^2)^{\frac{s+d/p-d/2}{s+d/p+\alpha}} val(Q_{1,C}).$$

Thus, along specially chosen subsequences of ϵ,

$$val(Q_{\epsilon,C}) = Const \cdot \epsilon^{2r}.$$

Moreover the solution of one problem is just $\mathcal{U}_{\epsilon_h,h}$ applied to the solution of the other. This is the dyadic renormalization property.

Now for $\epsilon_{h+1} < \epsilon < \epsilon_h$,

$$val(Q_{\epsilon_{h+1},C}) < val(Q_{\epsilon,C}) < val(Q_{\epsilon_h,C})$$

which establishes that $val(Q_{\epsilon,C}) \asymp \epsilon^{2r}$.

Well, actually, not quite. It is also necessary to show that $val(Q_{\epsilon,C})$ is finite. Results of **[DJ]** on the asymptotic behavior of $\rho_p(v,1)$ as $v \to 0$, can be applied here to show that $val(Q_{\epsilon,C})$ is finite provided that $s > \frac{2-p}{p}\alpha$.

7.5 Lower Bounds on Minimax Risk

To complete our study of the white noise observations (36), we now establish lower bounds on the minimax risk $\mathcal{R}^*_w(\epsilon)$. We will be satisfied here with proving the crude but simple lower bound

$$\mathcal{R}^*_w(\epsilon) \geq const \cdot (\epsilon)^{2r}. \tag{41}$$

We apply the method of hardest rectangular subproblems of Donoho, Liu, and MacGibbon (1990).

First, some notation. While the indices $\lambda \in \Lambda$ all correspond to wavelets $\tilde{\psi}_\lambda$ which intersect the set Q supporting the unknown object f, we now focus attention on the special subset $\Lambda^0 \subset \Lambda$ which indexes wavelets supported entirely in Q. We may stratify this set as $\Lambda^0 = \Lambda^0_{\ell-1} \cup \Lambda^0_\ell \cup \ldots \cup \Lambda_j{}^0 \cup \ldots$ just as with Λ. Then every sum $\sum_{\Lambda^0} \alpha_\lambda \tilde{\psi}_\lambda$ gives a function supported in Q.

Second, an observation. If $\hat{f}$ and f have inhomogeneous wavelet expansions $(\hat{\alpha}_\lambda)_{\lambda\in\Lambda}$ and $(\alpha_\lambda)_{\lambda\in\Lambda}$, we have the lower bound

$$\|\hat{f} - f\|^2_{L^2(Q)} \geq \sum_{\Lambda^0} (\hat{\alpha}_\lambda - \alpha_\lambda)^2.$$

Risk bounds developed for estimating the vector $(\alpha_\lambda)_{\Lambda^0}$ from data $(y_\lambda)_\Lambda$ provide lower bounds not just on the minimax risk in sequence space, but also on the original minimax risk in function space.

Third, some background. Suppose we have a vector $(\tau_\lambda)_{\lambda\in\Lambda^0}$ satisfying $\tau \in \Theta^s_{p,q}$; then

$$\Theta(\tau) = \{\theta : \theta_\lambda = s_\lambda \tau_\lambda, |s_\lambda| \leq 1, \lambda \in \Lambda^0\}$$

is a rectangular subset of $\Theta^s_{p,q}$. The problem of estimating the vector (α_λ) from data (x_λ) when it is known that (α_λ) lies in the subset $\Theta(\tau)$ is called a *rectangular subproblem* of the original estimation problem. The minimax risk of this subproblem is a lower bound on the minimax risk of the full problem. To determine this minimax risk, let $\rho_\infty(\tau,\sigma)$ be the

minimax risk of estimating the scalar θ from scalar data $x = \theta + \sigma w$, where $w \sim N(0,1)$, and $|\theta| \le \tau$. The risk of this problem is precisely

$$\sum_{\lambda \in \Lambda^0} \rho_\infty(\tau_\lambda, \epsilon\sigma_j).$$

There is a prior distribution with independent coordinates which is supported in $\Theta^s_{p,q}$ and has this for its Bayes risk. This prior has for coordinate α_λ the same distribution as the distribution of θ which attains the minimax risk for the bounded normal mean problem described by $\rho_\infty(\tau_\lambda, \epsilon\sigma_j)$. The best lower bound one can get in this manner is obtained by finding the hardest rectangular subproblem of $\Theta^s_{p,q}$, i.e. solving

$$\sup \sum_{\lambda \in \Lambda^0} \rho_\infty(\tau_\lambda, \epsilon\sigma_j) \text{ subject to } \tau \in \Theta^s_{p,q}. \tag{42}$$

A solution of this problem furnishes a prior distribution $\tilde{\pi}_\epsilon$, with independent coordinates, supported in the hardest hyperrectangle, and with Bayes risk equal to the value of this problem.

Finally, we now are in a position to obtain a lower bound on the minimax risk by seeking a hardest rectangular subproblem. (42) may be put in a form similar to that of $(P_{\epsilon,C})$. Define new variables $N_j^0 = Card({\Lambda_j}^0)$ and $r_j = ((N_j^0)^{-1} \sum_{\lambda \in \Lambda_j^0} |\tau_\lambda|^p)^{1/p}$, and rewrite the optimization problem as

$$(P^0_{\epsilon,C}) \qquad \sup \sum_{j \ge \ell - 1} N_j^0 \rho_\infty(r_j, \epsilon\sigma_j) \text{ subject to } \sum_{j \ge \ell} (2^{js}(N_j^0)^{1/p} r_j)^q \le C^q.$$

Making as before the substitutions $N_j^0 \sim 2^{jd}$, $\sigma_j \sim 2^{j\alpha}$, $j \to \infty$, leads to the renormalizable problem

$$(Q^0_{\epsilon,C}) \qquad \sup \epsilon^2 \sum_{j=-\infty}^{\infty} 2^{j(d+2\alpha)} \rho_\infty(v_j, 1) \text{ subject to } \sum_{j=-\infty}^{\infty} (2^{j(s+d/p+\alpha)} v_j)^q \le C^q.$$

The new problem may be shown asymptotically equivalent:

$$val(P^0_{\epsilon,C}) \sim val(Q^0_{\epsilon,C}) \qquad \epsilon \to 0;$$

on the other hand, like $(Q_{\epsilon,C})$, $(Q^0_{\epsilon,C})$ has invariances which yield immediately

$$val(Q^0_{\epsilon,C}) \asymp (\epsilon)^{2r} \qquad \epsilon \to 0.$$

We conclude that the lowerbound (41) holds.

7.6 Conclusion of Risk Asymptotics

Section 7.4 showed that in the discrete white noise model $\mathcal{B}^*_w(\epsilon) \le C \cdot \epsilon^{2r}$ as $\epsilon \to 0$. Moreover, the inequalities of sections 7.1 and 7.3 give $\mathcal{R}^*_z(\epsilon) \le \mathcal{B}^*_z(\epsilon) \le \mathcal{B}_w(\gamma_1 \epsilon)$. We conclude that

$$\mathcal{R}^*_z(\epsilon) \le Const \cdot \epsilon^{2r} \qquad \epsilon \to 0.$$

Section 7.5 showed (41). This means in fact that for each $\epsilon > 0$ there is a prior $\tilde{\pi}_\epsilon$ with independent coordinates such that $\tilde{\pi}_\epsilon\{\alpha \in \Theta^s_{p,q}\} = 1$, and

$$\mathcal{B}_w(\epsilon, \tilde{\pi}_\epsilon) \geq const \cdot (\epsilon)^{2r}, \qquad \epsilon \to 0.$$

Hence

$$\begin{aligned} \mathcal{R}^*_z(\epsilon) &\geq \sup\{\mathcal{B}_z(\epsilon, \tilde{\pi}) : supp(\tilde{\pi}) \in \Theta^s_{p,q},\ \pi \in \tilde{\Pi}\} \\ &\geq \mathcal{B}_z(\epsilon, \tilde{\pi}_{\gamma_0 \epsilon}) \\ &\geq \mathcal{B}_w(\gamma_0 \epsilon, \tilde{\pi}_{\gamma_0 \epsilon}) \\ &\geq const \cdot \epsilon^{2r}. \end{aligned}$$

This completes our derivation of the rate at which the minimax risk goes to zero.

8 Wavelet Shrinkage

We now discuss the performance of our proposed WVD thresholding procedure.

First, background from the theory of the estimating a 1-dimensional normal mean. Consider the scalar observation $x = \theta + \sigma w$ where θ and w are independent scalar random variables and $w \sim N(0,1)$. We wish to estimate θ with small squared error loss, and we know only that $E|\theta|^p \leq \tau^p$. We consider the use of thresholds $\delta_t(x) = sgn(x)(|x| - t)_+$ The minimax-Bayes threshold risk is

$$\rho^s_p(\tau, \sigma) = \inf_t \sup_{E|\theta|^p \leq \tau^p} E(\delta_t(x) - \theta)^2.$$

This has been studied by Donoho and Johnstone (1990), who showed that for a certain constant $C(p) < \infty$ the inequality

$$\rho^s_p(\tau, \sigma) \leq C(p)\rho_p(\tau, \sigma) \tag{43}$$

holds for all $\tau > 0$ and all $\sigma > 0$. The implication is that, although in general the minimax risk ρ_p is attained by a complicated nonlinear function of x, the simple threshold estimator with appropriately chosen threshold t is nearly as good.

Second, introduce the notation

$$r(t, \tau, \sigma) = \sup_{E|\theta|^p \leq \tau^p} E(\delta_t(x) - \theta)^2.$$

This denotes the worst-case risk of using threshold t when the parameter has p-th mean less than τ^p and the noise variance is σ^2. The function $r(t, \tau, \sigma)$ is concave in τ^p for each fixed t and σ.

Finally, we apply this apparatus. Suppose we observe data (35) where the wavelet coefficients (α_λ) are random and obey $(\tau_\lambda) \in \Theta^s_{p,q}$. We will choose thresholds (t_j) for use with the estimator $\hat{\alpha}_\lambda = \delta_{t_j}(y_\lambda)$. Let $s_\lambda = ||\tilde{c}_\lambda||$ be the standard deviation of $(y_\lambda - \alpha_\lambda)/\epsilon$. The risk of such an estimator is

$$\sum_\Lambda E(\delta_{t_j}(y_\lambda) - \alpha_\lambda)^2 = \sum_\Lambda r(t_j, \tau_\lambda, \epsilon s_\lambda).$$

We wish to choose thresholds to minimize the maximum risk, the maximum being taken over all $(\tau_\lambda) \in \Theta^s_{p,q}$. Using the concavity of $r(t,\tau,\sigma)$ in τ^p for each fixed t and σ, **[DJ]** prove a minimax theorem which, after adaptation to the present setting, yields the minimax theorem

$$\inf_{(t_j)} \sup_{\tau \in \Theta^s_{p,q}} \sum_\Lambda r(t_j, \tau_\lambda, \epsilon s_\lambda) = \sup_{\tau \in \Theta^s_{p,q}} \inf_{(t_j)} \sum_\Lambda r(t_j, \tau_\lambda, \epsilon s_\lambda).$$

On the other hand, since by definition

$$\inf_t r(t,\tau,\sigma) = \rho^s_p(\tau,\sigma)$$

the right-hand side of the above minimax identity is equal to

$$\sup_{\tau \in \Theta^s_{p,q}} \sum_\Lambda \rho^s_p(\tau_j, \epsilon s_\lambda).$$

Moreover, $\rho^s_p(\tau,\sigma)$ is monotone increasing in σ for τ fixed, so using $\epsilon s_\lambda \le \epsilon\gamma_1\sigma_j$, we get $\rho^s_p(\tau_j, \epsilon s_\lambda) \le \rho_p(\tau_j, \epsilon\gamma_1\sigma_j)$. Finally, applying (43) we get the upper bound

$$\begin{aligned} \sup_{\tau \in \Theta^s_{p,q}} \sum_\Lambda \rho^s_p(\tau_\lambda, \epsilon\gamma_1\sigma_j) &\le C(p) \sup_{\tau \in \Theta^s_{p,q}} \sum_\Lambda \rho_p(\tau_j, \epsilon\gamma_1\sigma_j) \\ &= C(p)\mathcal{B}^*_w(\gamma_1\epsilon) \\ &\le Const \cdot (\epsilon)^{2r} \qquad \epsilon \to 0. \end{aligned}$$

This shows that

$$\inf_{(t_j)} \sup_{\tau \in \Theta^s_{p,q}} \sum_\Lambda E(\delta_{t_j} - \alpha_\lambda)^2 \le Const \cdot (\epsilon)^{2r} \qquad \epsilon \to 0.$$

We now make the trivial observation that

$$\sup_{\alpha \in \Theta^s_{p,q}} \sum_\Lambda E(\delta_{t_j}(y_\lambda) - \alpha_\lambda)^2 \le \sup_{\tau \in \Theta^s_{p,q}} \sum_\Lambda E(\delta_{t_j}(y_\lambda) - \alpha_\lambda)^2,$$

where α on the left is a constant and α on the right is a random variable. It follows that

$$\begin{aligned} \inf_{(t_j)} \sup_{\alpha \in \Theta^s_{p,q}} \sum_\Lambda E(\delta_{t_j}(y_\lambda) - \alpha_\lambda)^2 &\le Const \cdot (\epsilon)^{2r} \\ &\le Const \cdot \mathcal{R}^*_z(\epsilon) \qquad \epsilon \to 0. \end{aligned}$$

Our threshold estimator is nearly-minimax, and the second part of our main result is established.

9 Minimax Linear Risk

We now complete the proof of our main result. Define the Minimax Linear Risk

$$\mathcal{R}^L_z(\epsilon) = \inf_{\substack{\hat\alpha \\ \text{Linear}}} \sup_{\alpha \in \Theta^s_{p,q}} E||\hat\alpha(y) - \alpha||^2$$

and the corresponding white-noise linear risk

$$\mathcal{R}^L_w(\epsilon) = \inf_{\substack{\hat\alpha \\ \text{Linear}}} \sup_{\alpha \in \Theta^s_{p,q}} E||\hat\alpha(x) - \alpha||^2.$$

9.1 Reduction to White Noise

As in the nonlinear case, we reduce study of asymptotics of $\mathcal{R}_z^L(\epsilon)$ to study of the white noise equivalent $\mathcal{R}_w^L(\epsilon)$.

Theorem 7 *Let Θ be orthosymmetric. Suppose that (z_λ) is an almost-independent noise, with constants γ_0,γ_1. Then*

$$R_w^L(\gamma_0\epsilon) \le R_z^L(\epsilon) \le R_w^L(\gamma_1\epsilon) \qquad \epsilon > 0.$$

The argument is similar to the nonlinear case, but simpler. To give it, we introduce the wide-sense conditional expectation and wide-sense conditional variance. We suppose that Y_i are random variables with finite variance.

$$E^L\{Y_0|Y_1, Y_2, \ldots\} = a_1Y_1 + a_2Y_2 + \ldots$$

where the constants a_1,a_2, ... are chosen to minimize the expression

$$E\{Y_0 - a_1Y_1 + a_2Y_2 + \ldots\}^2.$$

The minimum value of this quadratic expression will be denoted $Var^L(Y_0|Y_1, Y_2, \ldots)$. We remark that these quantities possess properties parelling those of their strict-sense counterparts: (1) Monotonicity under conditioning

$$Var^L(Y_0|Y_1, Y_2) \le Var^L(Y_0|Y_1)$$

(2) Markovian character: if (Y_0, Y_1, Y_2) are Markovian

$$E^L(Y_2|Y_1, Y_0) = E^L(Y_2|Y_1)$$

$$Var^L(Y_2|Y_1, Y_0) = Var^L(Y_2|Y_1)$$

We define, for a zero-mean Θ-valued R.V. α with prior distribution π the Bayes-Linear risk

$$\mathcal{B}_z^L(\epsilon, \pi) = \sum_\Lambda Var^L\{\alpha_\lambda|y\}$$

and

$$\mathcal{B}_w^L(\epsilon, \pi) = \sum_\Lambda Var^L\{\alpha_\lambda|x\}.$$

By the usual linear minimax theorem (Pilz, 1985)

$$\mathcal{R}_z^L(\epsilon) = \sup\{\mathcal{B}_z^L(\epsilon, \pi) : supp(\pi) \in \Theta\}$$

$$\mathcal{R}_w^L(\epsilon) = \sup\{\mathcal{B}_w^L(\epsilon, \pi) : supp(\pi) \in \Theta\}.$$

There are two key steps in the proof of Theorem 7. First, in the white noise model, for any prior π supported in Θ, there is an orthogonal-coordinates prior $\tilde{\pi}$ also supported in Θ with at least as large a linear Bayes risk. Indeed, let α_λ have prior π, and define the R.V.

$$\tilde{\alpha}_\lambda = \pm_\lambda\alpha_\lambda \qquad \lambda \in \Lambda$$

with the signs $\pm_\lambda$ chosen by independent fair coin-tossing. The implied prior $\tilde{\pi}$ is an orthogonal coordinates prior. To see that it is less favorable than π, let $s = (s_\lambda)$ be a (nonrandom) sequence of signs, and let π_s denote the prior of the random variable $(s_\lambda \alpha_\lambda)$. It is easy to see that

$$\mathcal{B}_w^L(\epsilon, \pi_s) = \mathcal{B}_w^L(\epsilon, \pi).$$

Now Bayes-linear risk is concave:

$$\mathcal{B}_w^L(\epsilon, \sum_i \eta_i \pi_i) \geq \sum_i \eta_i \mathcal{B}_w^L(\epsilon, \pi_i).$$

The prior $\tilde{\pi}$ is the average over all sequences of signs $ave_s \pi_s$, which is a (weak-*) limit of finite convex combinations, so

$$\mathcal{B}_w^L(\epsilon, \tilde{\pi}) \geq \mathrm{ave}_s \{\mathcal{B}_w^L(\epsilon, \pi_s)\} = \mathcal{B}_w^L(\epsilon, \pi).$$

This shows that $\tilde{\pi}$ is less favorable than π.

As an immediate application we get the formula

$$\begin{aligned} \mathcal{R}_w^L(\epsilon) &= \sup\{\mathcal{B}_w^L(\epsilon, \tilde{\pi}) : supp(\pi) \in \Theta \\ &\qquad \pi \text{ makes coordinates orthogonal } \} \end{aligned}$$

The quantity on the left is a minimax risk (not Minimax Bayes); no similar formula holds in the nonlinear case.

The second step in the proof is the recognition that if $\tilde{\pi}$ is an orthogonal process

$$\mathcal{B}_w^L(\gamma_0 \epsilon, \tilde{\pi}) \leq \mathcal{B}_z^L(\epsilon, \tilde{\pi}) \leq \mathcal{B}_w^L(\gamma_1 \epsilon, \tilde{\pi}).$$

The proof is as in Lemma 8, only using the Markov properties of wide-sense expectations. The Theorem now follows easily. Let $\tilde{\Pi}$ denote the set of priors with orthogonal coordinates. For the lower bound

$$\begin{aligned} \mathcal{R}_z^L(\epsilon) &= \sup\{\mathcal{B}_z^L(\epsilon, \pi) : supp(\pi) \subset \Theta\} \\ &\geq \sup\{\mathcal{B}_z^L(\epsilon, \tilde{\pi}) : supp(\tilde{\pi}) \subset \Theta, \tilde{\pi} \in \tilde{\Pi}\} \\ &\geq \sup\{\mathcal{B}_w^L(\gamma_0 \epsilon, \pi) : supp(\tilde{\pi}) \subset \Theta, \tilde{\pi} \in \tilde{\Pi}\} \\ &= \mathcal{R}_w^L(\gamma_0 \epsilon). \end{aligned}$$

On the other hand, from

$$\begin{aligned} Var^L\{\alpha_\lambda | (y_\lambda)_{\lambda \in \Lambda}\} &\leq Var^L\{\alpha_\lambda | y_\lambda\} \\ &\leq Var^L\{\alpha_\lambda | \alpha_\lambda + \epsilon \gamma_1 \sigma_j w_\lambda\} \end{aligned}$$

and from the formula

$$\mathcal{B}_w^L(\gamma_1 \epsilon, \tilde{\pi}) = \sum_\Lambda Var^L\{\alpha_\lambda | \alpha_\lambda + \epsilon \gamma_1 w_\lambda\}$$

valid whenever $\tilde{\pi}$ has orthogonal coordinates, we have

$$\begin{aligned} \mathcal{R}_z^L(\epsilon) &= \sup\{\mathcal{B}_z^L(\epsilon, \pi) : supp(\pi) \in \Theta\} \\ &\leq \sup\{\mathcal{B}_w^L(\gamma_0 \epsilon, \pi) : supp(\tilde{\pi}) \subset \Theta, \tilde{\pi} \in \tilde{\Pi}\} \\ &= \mathcal{R}_w^L(\gamma_1 \epsilon). \end{aligned}$$

9.2 Minimax Linear Risk in White Noise

We now study the linear minimax risk in the heteroscedastic white noise model (36). We need the notion of quadratic hull introduced in Donoho, Liu, and MacGibbon (1990).

Definition 2 *Let Θ be a set of sequences. Let Θ_+^2 be the set of sequences $\theta^2 \equiv (\theta_\lambda^2)_{\lambda\in\Lambda}$ arising from $\theta \in \Theta$. Then*

$$QHull(\Theta) = \{\theta : \theta^2 \in Hull(\Theta_+^2)\}.$$

We note the following relationships (Donoho and Johnstone, 1991)

$$QHull(\Theta_{p,q}^s) = \begin{cases} \Theta_{p,q}^s & p,q \geq 2 \\ \Theta_{2,q}^s & p < 2 \leq q \\ \Theta_{2,2}^s & p,q < 2 \end{cases} \tag{44}$$

In particular, if $p < 2$, then $\Theta_{p,q}^s \neq QHull(\Theta_{p,q}^s)$. This has relevance because of the following.

Theorem 8 *Let Θ be a set. Let observations be given by the heteroscedastic white-noise model (36). Then*

$$\mathcal{R}_w^L(\epsilon;\Theta) = \mathcal{R}_w^L(\epsilon;QHull(\Theta)).$$

Hence, for linear estimates, the set Θ is equally difficult as the quadratic hull. For nonlinear estimators this would not be the case.

Proof. As we have seen, $\mathcal{R}_w^L(\epsilon)$ is the Bayes risk of a prior having orthogonal coordinates. Every such prior has for its wide-sense expectation a diagonal linear rule

$$\hat{\alpha}_\lambda = s_\lambda x_\lambda \qquad \lambda \in \Lambda.$$

Therefore the minimax linear procedure is a diagonal rule. (A different argument which does not depend on the minimax theorem may be given.)

For diagonal linear estimates, we have

$$E\|\hat{\alpha}_\lambda - \alpha\|^2 = \sum_\Lambda (s_\lambda - 1)^2\alpha_\lambda^2 + \epsilon^2 \sum s_\lambda^2.$$

This is a linear functional of (α_λ^2) and attains the same extremum over Θ_+^2 as over $Hull(\Theta_+^2)$. Hence the minimax risk over $QHull(\Theta)$ is no larger than over Θ.

9.3 Suboptimality when $p < 2$

We are now in a position to show that linear methods can only attain suboptimal rates of convergence when $p < 2$. Suppose that $p \leq q < 2$. Then we have

$$\begin{aligned} \mathcal{R}_w^L(\epsilon, \Theta_{p,q}^s) &= \mathcal{R}_w^L(\epsilon, QHull(\Theta_{p,q}^s)) \\ &= \mathcal{R}_w^L(\epsilon, \Theta_{2,2}^s) \\ &\geq \mathcal{R}_w^*(\epsilon, \Theta_{2,2}^s) \\ &\geq const\ \epsilon^{2r'} \qquad \epsilon \to 0. \end{aligned}$$

Here $r' = r'(s,p,\alpha) = r(s,2,\alpha)$. As $r(s,2,\alpha) < r(s,p,\alpha)$ for $p < 2$, linear estimators, in particular those based on windowed S.V.D., cannot attain the optimal rate of convergence of the minimax risk to zero. The argument for $q \leq p < 2$ is similar.

10 Discussion

10.1 Refinements and Extensions

It is possible to give much more precise information about the numerical size of minimax risks in the discrete white noise model. Also, it is possible to show that the threshold estimators come reasonably close numerically to the minimax risk. See [**DJ**].

Minimaxity results can be developed in the Lizorkin-Triebel scale of spaces, with parallel conclusions about rates of convergence for the minimax risk and for the minimax linear risk. The details follow from a combination of arguments given above and arguments in [**DJ**].

It is possible to build a WVD starting from a biorthogonal wavelet basis of the type discussed in Cohen, Daubechies, and Feaveau (1990). Such bases have two families of non-orthogonal wavelets $(\psi_\lambda)_\lambda$ and $(\tilde{\psi}_\lambda)_\lambda$, each family generated by dilating and translating a mother wavelet (ψ or $\tilde{\psi}$). These families are defined to yield the biorthogonality $\langle \psi_\lambda, \tilde{\psi}_\mu \rangle = \delta_{\lambda,\mu}$, or equivalently, the L^2-reproducing formula

$$f = \sum_\lambda \langle f, \tilde{\psi}_\lambda \rangle \psi_\lambda .$$

With some slight regularity of the wavelet families, we may define sets (u_λ) and (v_λ) via

$$K\psi_\lambda = \kappa_j v_\lambda$$

$$K^* u_\lambda = \kappa_j \tilde{\psi}_\lambda .$$

Formally,

$$[u_\lambda, v_\mu] = [u_\lambda, \kappa_j^{-1} K \psi_\mu] = \langle \kappa_j^{-1} K^* u_\lambda, \psi_\mu \rangle = \langle \tilde{\psi}_\lambda, \psi_\mu \rangle = \delta_{\lambda,\mu}$$

so the two sets are biorthogonal; with enough regularity of K, ψ, and $\tilde{\psi}$ the two systems will be nearly-orthogonal. As we never really use the orthogonality of the (ψ_λ) in sections 7, 8, 9, Theorem 4 holds for each Besov space in which the biorthogonal set has the unconditional basis property.

10.2 Limitations

The approach developed here in essence is adapted to scale-invariant operators. The reader should not suppose that we claim to say anything about inverse problems with a sharp scale preference. Examples of operators we are not claiming to discuss here include: (1) convolution with a boxcar $k(h) = 1_{\{|h| \le 1\}}$ (has a preferred spatial scale); (2) Fourier multiplication by a boxcar $\hat{k}(\omega) = 1_{\{|\omega| \le 1\}}$ (has a preferred frequency scale); and related quantities (convolution with a Gaussian has a preferred spatial and frequency scale). We are, in essence, only discussing problems which are renormalizable in the sense that Donoho and Low (1992) use the term.

10.3 Simultaneous Diagonalization

The slogan given here – that dilation-homogeneous operators admit of sufficienty regular wavelets as quasi-singular functions – has several precedents.

In the monograph of Frazier, Jawerth, and Weiss (1991) one finds, on page 101,

> *"This ... provides us with an opportunity to discuss an aspect of the general philosphy behind the ϕ and wavelet transforms. A convolution operator T with multiplier m satisfies $T(e^{ix\xi}) = m(\xi)e^{ix\xi}$, if $e^{ix\xi}$ is in the domain of T ... Thus the characters $e^{ix\xi}$ are simultaneous eigenfunctions of all translation invariant operators. For the ψ and wavelet transforms, and an appropriate class of Calderón-Zygmund Operators ... the matrix $\{\langle T\psi_P, \phi_Q\rangle\}$ decays rapidly away from the diagonal. From the ϕ-transform identity $T\psi_P = \sum_Q \langle T\psi_P, \psi_Q\rangle \psi_Q$, this says that ψ_P is an "almost-eigenfunction" of T. Thus the ϕ and wavelet transforms simultaneously "almost diagonalize" a large class of operators, not restricted to convolution type. The idea behind these transforms is to give up precise diagonalization and precise eigenfunctions (which would necessarily be Fourier characters, localized to a point on the Frequency side). Instead, the ψ_Q's have a somewhat, but not completely, localized frequency spectrum. Thus the ψ_Q's are almost eigenfunctions, which gives us almost diagonalization even in certain nonconvolution cases.*
>
> *Another advantage, as we have stressed before, is that the traditional function spaces, like L^p, $1 < p < \infty$, are characterized precisely via the ϕ and wavelet transforms, unlike the Fourier case (except when $p = 2$). Thus these transforms provide a precise tool for studying a large variety of function spaces without losing the essential aspect of (near) diagonalization."*

In the Introduction to Volume II, page vii, of Meyer (1990) one finds the following

> *Lorsqu'on dispose d'une base orthonormée remarquable $\mathbf{e}_j$, $j \in J$, de l'espace de Hilbert de réfèrence $H = L^2(R^n)$, il est impossible de résister à la tentation d'étudier les opérateurs $T : H \to H$ qui sont diagonaux ou presque-diagonaux dans cette base particulière. On peut alors espérer que ces opérateurs soient déjà connus par ailleurs, ce qui confiera à tout l'ensemble une cohérence dont les scientifiques sont friands. Malheureusement, cette attente a, jusqu'aujourd'hui, toujours été déçue. Les opérateurs diagonaux dans les bases orthonormées d'usage courant sont génèriquement pathologiques et ne présentent aucun intérêt. Par example, les opérateurs diagonaux dans le système trigonometrique, c'est à dire vérifiant $T(e^{ikx}) = m_k e^{ikx}$ où m_k est une suite bornée sont, en général, pathologiques ...*
>
> *Les bases orthonormées d'ondelettes constituent le premier, et à notre connaissance, l'unique exemple d'une base orthonormée ayant la propriété que les opérateurs diagonaux (ou presque diagonaux) dans cette base soient intéressants et connus par ailleurs, sous le nom d'opérateurs de Calderón-Zygmund... Ce fait remarquable explique les succès remportés par les séries d'ondelettes, comparées aux autres séries orthogonales. Une altération des coefficients d'ondelettes, du*

type $a_k \rightarrow m_k a_k$, n'a pas de conséquences incôntrolables sur la somme de la série d'ondelettes, lorsque m_k est une suite bornée... Les décompositions en séries d'ondelettes sont donc robustes et cette robustesse provient des relations entre ondelettes et opérateurs que nous venons de decrire."

Since Calderón-Zygmund Operators are defined by certain dilation-homogeneous inequalities, both passages convey the idea of simultaneous diagonalization of dilation-homogeneous operators and of function classes. The first quote says this explicitly; the second, implicitly (recall the slogan of section 1.4 that the unconditional basis property represents a kind of diagonalization of a function class).

The point of view in this paper aligns nicely with these slogans formulated by the Founding Fathers of wavelets. We trace here the consequences of these slogans in the field of linear inverse problems, where the operators need not be bounded or CZO's, but still which possess a certain dilation homogeneity. The simultaneous diagonalization property of the WVD allows us to develop new methods for reconstruction from indirect, noisy measurements – nonlinear methods which can outperform traditional, paradigmatic linear methods.

References

[1] R.S. Anderssen (1980) On the use of linear functionals for Abel-Type integral equations in Applications. in *The Application and Numerical Solution of Integral Equations*, R.S. Anderssen, F. De Hoog, and M. A. Lukas, Eds. Sijthoff and Noordhof International Publishers, The Netherlands.

[2] R.S. Anderssen (1986) The linear Functional Strategy for Improperly posed Problems. *Inverse Problems* J.R. Cannon and U. Hornung, eds. Birkhauser, Basel.

[3] M. Bertero (1989) Linear Inverse and Ill-Posed Problems. in *Advances in Electronics and Electron Physics* . Academic Press: NY.

[4] M. Bertero, P. Bocacci, and E.R. Pike.(1982) On the recovery and resolution of exponential relaxation rates from experimental data: a Singular Value analysis of the Laplace Transform inversion in the presence of noise. *Proc. Roy. Soc. Lond. A* **383** 15-29.

[5] M. Bertero and P. Bocacci. (1989) Computation of the Singular System for a class of integral operators related to data inversion in confocal microscopy. *Inverse Problems* **5** 935-957.

[6] M. Bertero, C. De Mol, and E.R. Pike. (1985) Linear Inverse Problems with Discrete Data I: General Formulation and Singular System Analysis. *Inverse Problems* **1** 301-330.

[7] M. Bertero, C. De Mol, and E.R. Pike. (1986) Particle Size Distributions from Spectral Turbidity: A Singular System Analysis. *Inverse Problems* **2** 301-330.

[8] M. Bertero and E.R Pike. (1986) Intensity Fluctuation distributions from Photon Counting Distributions: A Singular-System analysis of Poisson Transform Inversion. *Inverse Problems* **2**, 259-269.

[9] L.D. Brown and M.G. Low (1990) Asymptotic equivalence of nonparametric regression and white noise. Manuscript.

[10] A. Cohen and I. Daubechies (1990) Orthonormal wavelet bases with Hexagonal Symmetry. Manuscript.

[11] A. Cohen, I. Daubechies, J.C. Feauveau (1990) Biorthogonal bases of compactly supported wavelets. Manuscript.

[12] I. Daubechies. (1988) Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics* **41**, Nov. 1988, pp. 909-996.

[13] M.E. Davisson (1981) A Singular Value Decomposition for the Radon Transform in n-dimensional space. *Numerical Functional Analysis and Optimization* **3** 321-.

[14] S.R. Deans (1983) *The Radon Transform and Some of its Applications.* J. Wiley. New York.

[15] D.L. Donoho (1992) Unconditional bases are optimal bases for data compression and statistical estimation. Manuscript.

[16] D.L. Donoho and I.M. Johnstone (1992) Minimax Estimation via Wavelet Shrinkage. Technical Report, Department of Statistics, Stanford University.

[17] D.L. Donoho and I.M. Johnstone (1990) Minimax Risk over ℓ_p-balls. Technical Report, Department of Statistics, U.C. Berkeley.

[18] D.L. Donoho, R.C. Liu, and K.B. MacGibbon (1990) Minimax Risk over Hyperrectangles, and Implications. *Ann. Statist.* **18** 1416-1437.

[19] Donoho, D.L. and Nussbaum, M. (1990) Minimax Quadratic Estimation of a Quadratic Functional. *J. Complexity* **6** June 1990, 290-323.

[20] D.L. Donoho and M.G. Low (1992) Renormalization Exponents and Optimal pointwise rates of convergence. To Appear, *Ann. Statist*

[21] S.Y. Efroimovich and M.S. Pinsker (1981) Estimation of square-integrable [spectral] density based on a sequence of observations. *Problemy Peredatsii Informatsii* **17** 50-68 (in Russian); *Problems of Information Transmission* (1982) 182-196 (in English).

[22] S.Y. Efroimovich and M.S. Pinsker (1982) Estimation of square-integrable probability density of a random variable. *Problemy Peredatsii Informatsii* **18** 19-38 (in Russian); *Problems of Information Transmission* (1983) 175-189 (in English).

[23] M. Frazier and B. Jawerth (1985) Decomposition of Besov Spaces. *Indiana Univ. Math J.* **34** (1985) 777-799.

[24] M. Frazier and B. Jawerth (1990) A discrete Transform and Decomposition of Distribution Spaces. *Journal of Functional Analysis* **93** 34-170.

[25] M. Frazier, B. Jawerth, and G. Weiss (1991) *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, **79**. American Math. Soc.: Providence, RI.

[26] M. Golberg (1979) A method of Adjoints for solving some ill-posed equations of the first kind, *Applied Mathematics and Computation* **5** 123-130.

[27] F. Gori and G. Guattari (1985) Signal restoration for linear systems with weighted impulse. Singular value analysis for two cases of low-pass filtering. *Inverse Problems* **1** 67-.

[28] I.A. Ibragimov and R.Z. Has'minskii (1984) Nonparametric estimation of the value of a linear functional in a Gaussian white noise. *Theory of Probability and its Applications* **29**, 1-17.

[29] I.M. Johnstone, G.Kerkyacharian, and D. Picard. (1991) Density Estimation using wavelets: comparison between linear and nonlinear methods. Preprint, Department of Mathematics, Université de Nancy I.

[30] I.M. Johnstone and B.W. Silverman. (1990) Speeds of Estimation in Positron Emission Tomography. *Ann. Statist.* **18** 251-280.

[31] I.M. Johnstone and B.W. Silverman. (1991) Discretization effects in Statistical Inverse Problems. *J. Complexity* **7**, 1–34.

[32] H.J. Landau and H.O. Pollak (1961) Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty II. *Bell System Technical Journal* **40** 65-.

[33] P.G. Lemarié and Y. Meyer. (1986) Ondelettes et bases hilbertiennes. *Rev. Mat. Iberoamericana* **2**, 1-18.

[34] A.K. Louis (1986) Incomplete data problems in X-ray computerized tomography I. Singular Value Decomposition of the limited-angle transform. *Numer. Math,* **48**, 251-.

[35] S.G. Mallat (1989) Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Trans. Amer. Math. Soc.* **315** 1, 69-87.

[36] Y. Meyer. (1990) *Ondelettes et Opérateurs: I. Ondelettes* Hermann et Cie, Paris.

[37] Y. Meyer. (1990) *Ondelettes et Opérateurs: II. Opérateurs de Calderón Zygmund* Hermann et Cie, Paris.

[38] M. Nussbaum (1985) Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13**, 984-997.

[39] F. O'Sullivan (1986) A statistical perspective on ill-posed inverse problems. *Statistical Science* **1** 502-527.

[40] J. Pilz (1986) Minimax linear regression estimation with symmetric parameter restrictions. *J. Statist. Plan. Inf.* **13** 297-318.

[41] M.S. Pinsker (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problemy Peredatsii Informatsii* **16**, 2, 52-68 (in Russian). *Problems of Info. Trans.* (1980) 120-133 (in English).

[42] L. Shure, R.L. Parker, and G.E. Backus (1982) Harmonic splines for geomagnetic modeling, *Phys. Earth Plan. Interiors* **28** 215-229.

[43] O.N. Strand (1973) Theory and methods related to the singular function expansion and Landweber's iteration for integral equations of the first kind. *SIAM J. Num. Anal.* **5** 287-.

[44] A.N. Tikhonov (1963) Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Doklady* **4** 1035-.

[45] G. Wahba (1990) *Spline Methods for Observational data* NSF-CBMS Regional Conference Series in Mathematics, SIAM: Philadelphia, PA.

[46] D.V. Widder (1971) *An Introduction to Transform Theory.* Academic Press:New York.

[47] R.M. Young (1976) *An introduction to NonHarmonic Fourier Series.* Academic Press: New York.