PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Ridgelets: a key to higher-dimensional intermittency?

Emmanuel J. Candès and David L. Donoho

Phil. Trans. R. Soc. Lond. A 1999 **357**, 2495-2509
doi: 10.1098/rsta.1999.0444

THE ROYAL SOCIETY

Ridgelets: a key to higher-dimensional intermittency?

By Emmanuel J. Candès and David L. Donoho

Department of Statistics, Stanford University, Stanford, CA 94305–4065, USA

In dimensions two and higher, wavelets can efficiently represent only a small range of the full diversity of interesting behaviour. In effect, wavelets are well adapted for point-like phenomena, whereas in dimensions greater than one, interesting phenomena can be organized along lines, hyperplanes and other non-point-like structures, for which wavelets are poorly adapted.

We discuss in this paper a new subject, ridgelet analysis, which can effectively deal with line-like phenomena in dimension 2, plane-like phenomena in dimension 3 and so on. It encompasses a collection of tools which all begin from the idea of analysis by ridge functions $\psi(u_1x_1 + \cdots + u_nx_n)$ whose ridge profiles ψ are wavelets, or alternatively from performing a wavelet analysis in the Radon domain.

The paper reviews recent work on the continuous ridgelet transform (CRT), ridgelet frames, ridgelet orthonormal bases, ridgelets and edges and describes a new notion of smoothness naturally attached to this new representation.

Keywords: Ridge functions; wavelets; singularities; edges; Radon transform; nonlinear approximation

1. Introduction

This paper is part of a series around the theme—*wavelets: a key to intermittent information?*. The title itself raises a fundamental question; we shall argue that the answer is both *no* and *yes*. We say *no* because wavelets *per se* only address a portion of the intermittency challenge; we intend to make clear how much larger the question is than just the portion which wavelets can face effectively. Roughly speaking, wavelets deal efficiently only with one type of intermittency—singularities at points—and in higher dimensions there are many other kinds of intermittency—singularities along lines, along hyperplanes, etc.—which wavelets do not deal with efficiently. But we also say *yes*, because by using wavelets in a novel way, we have been able to build new systems of representations—ridgelets—which are efficient at many of the tasks where wavelets fail.

In this expository paper, we will primarily focus on the study of objects defined in two-dimensional space since, on one hand, this case already exhibits the main concepts underlying the ridgelet analysis and, on the other hand, it is a very practical setting because of the connection with image analysis. However, we will refer to extensions to higher dimensions wherever it is conceptually straightforward to do so.

(a) *Wavelets and point singularities*

To begin, we call the reader's attention to one of the really remarkable facts about wavelet bases. Suppose that we have a function $f(t)$ of a single real variable $t \in [0,1]$ and that f is smooth apart from a discontinuity at a single point t_0. For example, let $f(t) = t - 1_{\{t>t_0\}}$. In some sense this is a very simple object, and we would like to find an expansion that reveals its simplicity. However, in traditional types of expansions, the representation of this object is quite complicated, involving contributions from many terms. This is so of the Fourier representation; viewing $[0,1]$ as the circle, we can calculate the appropriate Fourier series on $[0,1]$; the number of the Fourier coefficients of f exceeding $1/N$ in absolute value exceeds $c \cdot N$ as $N \to \infty$, for some positive constant c. It is true of traditional orthogonal series estimates; an expansion of f in Legendre polynomials has at least $c \cdot N$ coefficients exceeding $1/N$. In stark contrast, in a nice wavelet orthonormal basis (Daubechies 1988), such as the Lemarié–Meyer inhomogeneous periodized wavelet basis, the number of coefficients exceeding $1/N$ in amplitude grows more slowly that N^ρ for any positive ρ. In effect, the singularity at t_0 causes widespread effects throughout the Fourier and Legendre representations; but the singularity causes highly localized or concentrated effects to the wavelet representation. Alternately, we can say that *in analysing an object exhibiting punctuated smoothness, the wavelet coefficients are sparse, while the coefficients of classical transforms are not sparse.*

The potential for sparsity of wavelet representations has had a wide impact, both in theory and in practice. It has a well-understood meaning for nonlinear approximation and for data compression of objects exhibiting punctuated smoothness (Donoho 1993): since the energy associated with the singularity is mostly concentrated in just a few big coefficients, partial reconstruction using a relatively small number of wavelet terms (the terms associated with the biggest wavelets coefficients) can give excellent approximations. The recognition that wavelets deal successfully with functions which are smooth away from singularities has led to a great deal of interest in their applications in image coding, where a great deal of the important structure consists of singularities—namely, edges. Wavelet-based coders have found wide application in various 'niche' data-compression applications, and are now being considered for inclusion in the JPEG-2000 still-picture data-compression standard.

(b) *Singularities along lines*

Unfortunately some claims for wavelets have been overstated, and wavelets are sometimes being used for applications well outside their actual domain of expertise. To understand this point requires a more careful look at the notion of *singularity*. A function $f(x)$ of n variables may have singularities of any integer dimension d in the range $0, \ldots, n-1$. A zero-dimensional singularity is a point of bad behaviour. A one-dimensional singularity is a curve of bad behaviour. An $(n-1)$-dimensional singularity is a hypersurface of bad behaviour. *Wavelets are fully efficient at dealing with zero-dimensional singularities* only. Unfortunately, in higher dimensions, other kinds of singularities can be present, or even dominant: in typical images, the edges represent one-dimensional singularities, and there are no zero-dimensional singularities to speak of.

To be more concrete, consider the function g supported in the unit square

$$g(x_1, x_2) = 1_{\{x_1+x_2>1/2\}}\ w(x_1, x_2), \quad x \in \mathbb{R}^2, \tag{1.1}$$

where $w(x_1, x_2)$ is a smooth function tending to zero together with its derivatives at the boundary of the unit square. This simple object has a singularity along the line $x_1 + x_2 = \frac{1}{2}$. Such an object poses a difficult problem of approximation both for two-dimensional Fourier analysis and for two-dimensional wavelet analysis. Although the object is very simple, its wavelet transform does not decay rapidly: as $N \to \infty$, there are greater than or equal to $c \cdot N$ orthonormal wavelet coefficients exceeding $1/N$ in size. Its bivariate Fourier series does not decay rapidly either: as $N \to \infty$, there are $\geqslant c \cdot N$ Fourier coefficients exceeding $1/N$ in size. Neither wavelets nor Fourier methods perform really well here. For example, if we used either approach as the basis of transform coders (Donoho 1996), we would have, as a direct corollary of the fact that at least $c \cdot N$ coefficients of g have amplitude $\geqslant 1/N$, that the number of bits one must retain to achieve a distortion less than or equal to ϵ for wavelet transform coding grows as $\epsilon \to 0$ at least as rapidly as $c \cdot \epsilon^{-1}$, and the number of bits one must retain to achieve a distortion ϵ for Fourier transform coding grows as $\epsilon \to 0$ at least as rapidly as $c \cdot \epsilon^{-1}$.

In effect, wavelets are being used in image data compression although their theoretical properties are not nearly as favourable as one might have imagined, given the degree of attention they have received.

The concept of intermittency does not have a universal acceptance. We now take the liberty of identifying this concept as a situation where objects of interest are typically smooth apart from occasional singularities on, say, a set of measure zero. From this point of view we can say that wavelets have a role to play in dealing with a particular kind of intermittency—unusual behaviour at one point (or occasional points)—but not with every kind of intermittency; in dimension two they already fail when asked to deal efficiently with unusual behaviour on a line.

We are entitled here to say that wavelets 'fail' because we know of representing systems which, in a precise sense, can succeed in dealing with unusual behaviour on a line.

(*c*) *Ridgelet analysis*

In this paper we describe a recently developed approach to problems of functional representation—*ridgelet analysis*. Ridgelet analysis makes available representations of functions by superpositions of *ridge functions* or by simple elements that are in some way related to ridge functions $r(a_1x_1 + \cdots + a_nx_n)$; these are functions of n variables, constant along hyperplanes $a_1x_1 + \cdots + a_nx_n = c$; the graph of such a function in dimension two looks like a 'ridge'. The terminology 'ridge function' arose first in tomography (Logan & Shepp 1975), and ridgelet analysis makes use of a key tomographic concept, the Radon transform.

However, multiscale ideas, as found in the work of Littlewood & Paley or Calderòn (Meyer 1990) and culminating in wavelet theory, also appear as a crucial tool in the story. From wavelet theory, ridgelet analysis borrows the localization idea: fine-scale ridgelets are concentrated near hyperplanes at all possible locations and orientations.

As an example of what this family of ideas can do, consider the function g of (1.1). It will turn out that there are ridgelet expansions—by frames and even by orthonormal sets—having the property that the number of coefficients exceeding $1/N$ in

amplitude grows more slowly that N^ρ for any positive ρ. In effect, the singularity in g across the line $x_1 + x_2 = \frac{1}{2}$ has widespread effects in the Fourier and wavelet representation, but the singularity causes highly concentrated effects in the ridgelet representation. Moreover, a ridgelet transform coding method, based on scalar quantization and run-length coding, can code such objects with a bit length that grows more slowly as $\epsilon \to 0$ than any fractional power of ϵ^{-1}. Hence ridgelets do for linear singularities in dimension two what wavelets did for point singularities in dimension one—they provide an extremely sparse representation; neither wavelets nor Fourier can manage a similar feat in representing linear singularities in dimension two.

(*d*) *Ridgelets and ridge functions*

The ability of ridgelets to give a sparse analysis of singularities is just one point of entry into our topic. Another interesting entry point is provided by the connection of ridgelet analysis with the theory of approximation by superpositions of ridge functions. Since the 1970s, it has been proposed that superpositions of ridge functions could offer interesting alternatives to standard methods of multivariate approximation. Friedman & Stuetzle (1981) introduced into statistics the topic of 'projection pursuit regression', specifically suggesting that by such means one might perhaps evade the curse of dimensionality as suffered by then-typical methods of function approximation. Approximation by superpositions of ridge functions acquired further interest in the late 1980s under the guise of approximation by single-hidden-layer feedforward neural nets. In such neural nets, one considers the m-term approximation

$$f(x_1, \ldots, x_n) \approx \sum_{i=1}^{m} c_i \sigma(a_{i,1}x_1 + \cdots + a_{i,n}x_n).$$

Celebrated results in the neural-nets literature include Cybenko's (1989) result that every nice function of n-variables can be approximated arbitrarily well in a suitable norm by a sequence of such m-term approximations, and results of Barron (1993) and Jones (1992) that describe function classes and algorithms under which such m-term approximations converge at given rates, including specific situations in which the rates do not worsen with increasing dimension.

Ridgelet analysis provides an alternate approach to obtaining approximations by superpositions of ridge functions; one which is quantitative, constructive and stable. Roughly speaking, the earlier theory of m-term ridge-function approximations assures us only of the *existence* of superpositions with prescribed features; the theory of ridgelet analysis, growing as it does out of wavelets and computational harmonic analysis, goes to a new level, and gives a particular way to build an approximation which is both constructive and stable. It also gives theoretical insights, previously unavailable, about those objects which can be well represented by ridge functions.

2. The continuous ridgelet transform

The (continuous) ridgelet transform in $\mathbb{R}^2$ can be defined as follows (Candès 1999). Pick a smooth univariate function $\psi : \mathbb{R} \to \mathbb{R}$ with sufficient decay and vanishing mean, $\int \psi(t)\,\mathrm{d}t = 0$. For each $a > 0$, each $b \in \mathbb{R}$ and each $\theta \in [0, 2\pi)$, define the

bivariate function $\psi_{a,b,\theta} : \mathbb{R}^2 \to \mathbb{R}^2$ by

$$\psi_{a,b,\theta}(x) = a^{-1/2} \cdot \psi((\cos\theta x_1 + \sin\theta x_2 - b)/a).$$

This function is constant along 'ridges' $\cos\theta x_1 + \sin\theta x_2 = \text{const}$. Transverse to these ridges it is a wavelet; hence the name *ridgelet*. Given an integrable bivariate function $f(x)$, define its ridgelet coefficients

$$\mathcal{R}_f(a,b,\theta) = \int \bar{\psi}_{a,b,\theta}(x) f(x)\,\mathrm{d}x.$$

Our hypotheses on ψ guarantee that $\int |\hat{\psi}(\lambda)|^2 \lambda^{-2}\,\mathrm{d}\lambda < \infty$, and we suppose further that ψ is normalized so that

$$\int |\hat{\psi}(\lambda)|^2 \lambda^{-2}\,\mathrm{d}\lambda = 1.$$

Candès (1999) proves the exact reconstruction formula

$$f(x) = \int_0^{2\pi} \int_{-\infty}^{\infty} \int_0^{\infty} \mathcal{R}_f(a,b,\theta)\psi_{a,b,\theta}(x)\,\frac{\mathrm{d}a}{a^3}\,\mathrm{d}b\,\frac{\mathrm{d}\theta}{4\pi}$$

valid a.e. for functions which are both integrable and square integrable. This shows that 'any' function may be written as a superposition of 'ridge' functions. Such integral representations have been independently discovered by Murata (1996). In addition, our representation is stable, as we have a Parseval relation:

$$\int |f(x)|^2\,\mathrm{d}x = \int_0^{2\pi} \int_{-\infty}^{\infty} \int_0^{\infty} |\mathcal{R}_f(a,b,\theta)|^2\,\frac{\mathrm{d}a}{a^3}\,\mathrm{d}b\,\frac{\mathrm{d}\theta}{4\pi}.$$

(This relation is, however, absent from Murata's papers.) This approach generalizes to any dimension. Given a ψ obeying

$$\int |\hat{\psi}(\lambda)|^2 \lambda^{-n}\,\mathrm{d}\lambda = 1,$$

define $\psi_{a,b,u}(x) = \psi((u'x - b)/a)/\sqrt{a}$ and $\mathcal{R}_f(a,b,\theta) = \langle f, \psi_{a,b,u}\rangle$. Then there is an n-dimensional reconstruction formula

$$f = c_n \iiint \mathcal{R}_f(a,b,u)\psi_{a,b,u}(x)\,\frac{\mathrm{d}a}{a^{n+1}}\,\mathrm{d}b\,\mathrm{d}u,$$

with $\mathrm{d}u$ the uniform measure on the sphere; and a Parseval relation

$$\|f\|^2_{L^2(\mathbb{R}^n)} = c_n \iiint |\mathcal{R}_f(a,b,\theta)|^2\,\frac{\mathrm{d}a}{a^{n+1}}\,\mathrm{d}b\,\mathrm{d}u.$$

(*a*) *Relation to Radon transform*

The continuous ridgelet transform is intimately connected with the Radon transformation (an excellent reference for the Radon transform is Helgason (1986)). If we put

$$Rf(u,t) = \int f(x)\delta(u'x - t)\,\mathrm{d}x$$

for the integral of f over the hyperplane $u'x = t$, then $\mathcal{R}_f(a, b, u) = \langle \psi_{a,b}, Rf(u, \cdot)\rangle$, where $\psi_{a,b}(t) = \psi((t-b)/a)/\sqrt{a}$ is a one-dimensional wavelet. Hence the Ridgelet transform is precisely the application of a one-dimensional wavelet transform to the slices of the Radon transform where u is constant and t is varying.

(*b*) *An example*

Let g be the mutilated Gaussian

$$g(x_1, x_2) = 1_{\{x_2>0\}} \mathrm{e}^{-x_1^2 - x_2^2}, \quad x \in \mathbb{R}^2. \tag{2.1}$$

This is discontinuous along the line $x_2 = 0$ and smooth away from that line. One can calculate immediately the Radon transform of such a function; it is

$$(Rg)(t, \theta) = \mathrm{e}^{-t^2} \bar{\Phi}(-t \sin\theta / |\cos\theta|) \quad t \in \mathbb{R}, \quad \theta \in [0, 2\pi], \tag{2.2}$$

where

$$\bar{\Phi}(v) \equiv \int_v^\infty \mathrm{e}^{-u^2}\, \mathrm{d}u.$$

We can get immediate insight into the form of the CRT from this formula. Remember that the wavelet transform $\langle \psi_{a,b}, \mathrm{e}^{-t^2} \cdot \bar{\Phi}(-t \sin\theta/|\cos\theta|)\rangle$ needs to be computed. Effectively, the Gaussian window e^{-t^2} makes little difference; it is smooth and of rapid decay, so it does little of interest; in effect the object of real interest to us is $\langle \psi_{a,b}, \bar{\Phi}(-s(\theta)t)\rangle$, where $s(\theta) = \sin\theta/|\cos\theta|$. Define then $W(a, b) = \langle \psi_{a,b}, \bar{\Phi}(-t)\rangle$; this is the wavelet transform of a smooth sigmoidal function. By the scale-invariance of the wavelet transform,

$$\langle \psi_{a,b}, \bar{\Phi}(-s(\theta)t)\rangle = W(s(\theta)a, s(\theta)b) \cdot |s(\theta)|^{-1/2}, \quad \text{for } \theta \in (0, \pi)$$

and, of course, a similar relationship holds for $(\pi, 2\pi)$. In short, for a caricature of $R_f(a, b, \theta)$, we have, for each fixed θ a function of a and b which is a simple rescaling of the wavelet transform of $\bar{\Phi}$ as function of θ. This rescaling is smooth and gentle away from $\theta = \frac{1}{2}\pi$ and $\theta = \frac{3}{2}\pi$, where it has singularities.

We remark that in a certain sense the CRT of g is sparse; if we use a sufficiently nice wavelet, such as a Meyer wavelet, the CRT belongs to $L^p((\mathrm{d}a/a^3)\,\mathrm{d}b\,\mathrm{d}\theta)$ for every $p > 0$. This is a fancy way of saying that the CRT decays rapidly as one moves either spatially away from $b = 0$ or $\theta \in \{\frac{1}{2}\pi, \frac{3}{2}\pi\}$ as one goes to fine scales $a \to 0$.

3. Discrete ridgelet transform: frames

It is important for applications that one obtains a discrete representation using ridgelets. Typical discrete representations include expansions in orthonormal bases. Here we describe an expansion in two dimensions by frames (see also Candès (1999), where the case for all dimensions $n \geqslant 2$ is treated).

We now develop a formula for the CRT of f using the Fourier domain. Obviously, with $\hat{f}$ denoting Fourier transform,

$$\mathcal{R}_f(a, b, \theta) = \frac{1}{2\pi} \int \bar{\hat{\psi}}_{a,b,\theta}(\xi) \hat{f}(\xi)\, \mathrm{d}\xi,$$

where $\hat{\psi}_{a,b,\theta}(\xi)$ is interpreted as a distribution supported on the radial line in the frequency plane. Letting $\xi(\lambda,\theta) = (\lambda\cos(\theta), \lambda\sin(\theta))$ we may write

$$\mathcal{R}_f(a,b,\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} a^{1/2}\bar{\hat{\psi}}(a\lambda)\mathrm{e}^{-\mathrm{i}\lambda b}\hat{f}(\xi(\lambda,\theta))\,\mathrm{d}\lambda. \tag{3.1}$$

This says that the CRT is obtainable by integrating the weighted Fourier transform $w_{a,b}(\xi)\hat{f}(\xi)$ along a radial line in the frequency domain, with weight $w_{a,b}(\xi)$ given by

$$a^{1/2}\bar{\hat{\psi}}(a|\xi|)$$

times a complex exponential in $\mathrm{e}^{-\mathrm{i}\lambda b}$. Alternatively, we can see that the function of b (with a and θ considered fixed), $\rho_{a,\theta}(b) = \mathcal{R}_f(a,b,\theta)$, satisfies

$$\rho_{a,\theta}(b) = \mathcal{F}_1^{-1}\{\hat{\rho}_{a,\theta}(\lambda)\},$$

where $\mathcal{F}_1$ stands for the one-dimensional Fourier transform, and

$$\hat{\rho}_{a,\theta}(\lambda) = a^{1/2}\bar{\hat{\psi}}(a\lambda)\hat{f}(\xi(\lambda,\theta)), \quad -\infty < \lambda < \infty$$

is the restriction of $w_{a,0}(\xi)\hat{f}(\xi)$ to the radial line. Hence, conceptually, the CRT at a certain scale a and angle θ can be obtained by the following steps:

1. two-dimensional Fourier transform, obtaining $\hat{f}(\xi)$;
2. radial windowing, obtaining $w_{a,0}(\xi)\hat{f}(\xi)$, say; and
3. one-dimensional inverse Fourier transform along radial lines, obtaining $\rho_{a,\theta}(b)$, for all $b \in \mathbb{R}$.

We are interested in finding a method for sampling $(a_j, b_{j,k}, \theta_{j,\ell})$ so that we obtain frame bounds, i.e. so we have equivalence:

$$\sum_{j,k,\ell} |\mathcal{R}_f(a_j, b_k, \theta_{j,\ell})|^2 \asymp \iiint |\mathcal{R}_f(a,b,\theta)|^2 \frac{\mathrm{d}a}{a^3}\,\mathrm{d}b\,\mathrm{d}\theta. \tag{3.2}$$

To simplify our exposition, we will suppose that $\hat{\psi}(\lambda) = 1_{\{1\leqslant|\xi|\leqslant 2\}}$ although the frame result holds for a large class of ψ as exposed in Candès (1999). Guided by the Littlewood–Paley and the wavelet theories, the scale a and location parameter b are discretized dyadically, as $a_j = a_0 2^j$ and $b_{j,k} = 2\pi k 2^{-j}$. Following (3.1) the ridgelet coefficients may be written as

$$\mathcal{R}_f(a_j, b_{j,k}, \theta) = \frac{1}{2\pi}2^{-j/2}\int_{2^j\leqslant|\lambda|\leqslant 2^{j+1}} \mathrm{e}^{-\mathrm{i}\lambda 2\pi 2^{-j}}\hat{f}(\xi(\lambda,\theta))\,\mathrm{d}\lambda,$$

and hence, the Plancherel theorem gives

$$\sum_k |\mathcal{R}_f(a_j, b_{j,k}, \theta)|^2 = \frac{1}{\sqrt{2\pi}}\int_{2^j\leqslant|\lambda|\leqslant 2^{j+1}} |w_{2^j,0}|^2|\hat{f}(\xi(\lambda,\theta))|^2\,\mathrm{d}\lambda.$$

In short, at a fixed scale and angular location, the sum of squares of ridgelet coefficients across a varying spatial location amounts to integrating the square of the Fourier transform along a dyadic segment.

Discretizing the angular variable θ amounts to performing a sampling of such segment-integrals from which the integral of $|\hat{f}(\xi)|^2$ over the whole frequency domain needs to be inferred. This is not possible without support constraints on f, as functions f can always be constructed with $f(x)$ having slow decay as $|x| \to \infty$ so that $\hat{f}$ will vanish on a collection of disjoint segments without being identically zero. However, under a support restriction, so that f is supported inside the unit disc (or any other compact set), the integrals over the segments can provide sufficient information to infer $\int |\hat{f}(\xi)|^2 \, \mathrm{d}\xi$.

Indeed, under a support constraint, the Fourier transform $\hat{f}(\xi)$ is a band-limited function, and over 'cells' of appropriate size can only display very banal behaviour. If we sample once per cell, we will capture enough of the behaviour of this object to be in a position to infer the size of the function from those samples. The solution found by Candès (1999) is to sample something like the following with increasing angular resolution at increasingly fine scales:

$$\theta_{j,\ell} = 2\pi\ell 2^{-j}.$$

This strategy gives the equivalence (3.2). It then follows that the collection

$$\{\, 2^{j/2}\psi(2^j(x_1\cos(\theta_{j,\ell}) + x_2\sin(\theta_{j,\ell}) - 2\pi k 2^{-j}))\,\}_{(j \geqslant j_0, \ell, k)}$$

is a frame for the unit disc; for any f supported in the disk with finite L^2 norm,

$$\sum_{j,k,l} |\langle \psi_{a_j, b_{j,k}, \theta_{j,l}}, f\rangle|^2 \asymp \|f\|^2.$$

The construction generalizes to any dimension n; in two dimensions, the discretization involves the sampling of angles from the circle and in n dimensions the sampling of angles from the unit sphere. The angular variable is also sampled at increasing resolution so that at scale j the discretized set is a net of nearly equispaced points at a distance of order 2^{-j} (see Candès (1999) for details).

The existence of frame bounds implies, by soft analysis, that there are 'dual ridgelets' $\tilde{\psi}_{j,k,\ell}$ so that

$$f = \sum_{j,k,\ell} \langle f, \tilde{\psi}_{j,k,\ell}\rangle \psi_{j,k,\ell} \quad \text{and} \quad f = \sum_{j,k,\ell} \langle f, \psi_{j,k,\ell}\rangle \tilde{\psi}_{j,k,\ell},$$

with equality in a an L^2 sense, and so that

$$\sum_{j,k,\ell} |\langle f, \tilde{\psi}_{j,k,\ell}\rangle|^2 \asymp \sum_{j,k,\ell} |\langle f, \psi_{j,k,\ell}\rangle|^2 \asymp \|f\|_{L^2}^2.$$

At the moment, only qualitative properties of the dual ridgelets $\tilde{\psi}_{j,k,\ell}$ are known; for example there are no closed-form expressions for their structure.

4. Orthonormal ridgelets in dimension 2

Donoho (1998) had the idea to broaden somewhat the notion of a ridgelet, to allow the possibility of systems obeying certain frequency/angle localization properties, and showed that if we allow this broader notion, then it becomes possible to have orthonormal ridgelets whose elements can be specified in closed form. Such a system

can be defined as follows: let $(\psi_{j,k}(t) : j \in \mathbb{Z},\ k \in \mathbb{Z})$ be an orthonormal basis of Meyer wavelets for $L^2(\mathbb{R})$ (Lemarié & Meyer 1986) and let

$$(w^0_{i_0,\ell}(\theta),\ \ell = 0, \dots, 2^{i_0} - 1;\ w^1_{i,\ell}(\theta),\ i \geqslant i_0,\ \ell = 0, \dots, 2^i - 1)$$

be an orthonormal basis for $L^2[0, 2\pi)$ made of periodized Lemarié scaling functions $w^0_{i_0,\ell}$ at level i_0 and periodized Meyer wavelets $w^1_{i,\ell}$ at levels $i \geqslant i_0$. (We suppose a particular normalization of these functions). Let $\hat{\psi}_{j,k}(\omega)$ denote the Fourier transform of $\psi_{j,k}(t)$, and define ridgelets $\rho_\lambda(x)$, $\lambda = (j, k; i, \ell, \varepsilon)$ as functions of $x \in \mathbb{R}^2$ using the frequency-domain definition

$$\hat{\rho}_\lambda(\xi) = \tfrac{1}{2}|\xi|^{-1/2}(\hat{\psi}_{j,k}(|\xi|)w^\varepsilon_{i,\ell}(\theta) + \hat{\psi}_{j,k}(-|\xi|)w^\varepsilon_{i,\ell}(\theta + \pi)). \tag{4.1}$$

Here the indices run as follows: $j, k \in \mathbb{Z}$, $\ell = 0, \dots, 2^{i-1} - 1$; $i \geqslant i_0$, $i \geqslant j$. Notice the restrictions on the range of ℓ and on i. Let Λ denote the set of all such indices λ. It turns out that $(\rho_\lambda)_{\lambda\in\Lambda}$ is a complete orthonormal system for $L^2(\mathbb{R}^2)$.

In the present form the system is not visibly related to ridgelets as defined earlier, but two connections can be exhibited. First, define a fractionally differentiated Meyer wavelet:

$$\psi^+_{j,k}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} |\omega|^{1/2}\hat{\psi}_{j,k}(\omega)\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega.$$

Then for $x = (x_1, x_2) \in \mathbb{R}^2$,

$$\rho_\lambda(x) = \frac{1}{4\pi}\int_0^{2\pi} \psi^+_{j,k}(x_1\cos\theta + x_2\sin\theta)w^\varepsilon_{i,\ell}(\theta)\,\mathrm{d}\theta. \tag{4.2}$$

Each $\psi^+_{j,k}(x_1\cos\theta + x_2\sin\theta)$ is a *ridge* function of $x \in \mathbb{R}^2$, i.e. a function of the form $r(x_1\cos\theta + x_2\sin\theta)$. Therefore ρ_λ is obtained by 'averaging' ridge functions with ridge angles θ localized near $\theta_{i,\ell} = 2\pi\ell/2^i$. A second connection comes by considering the sampling scheme underlying ridgelet frames as described in § 3. This scheme says that one should sample behaviour along line segments and that those segments should be spaced in the angular variable proportional to the scale 2^{-j} of the wavelet index. The orthonormal ridgelet system consists of elements which are organized angularly in just such a fashion; the elements $\hat{\rho}_\lambda$ are localized 'near' such line segments because the wavelets $w^\epsilon_{i,\ell}(\theta)$ are localized 'near' specific points $\theta_{i,\ell}$.

Orthonormal ridgelet analysis can be viewed as a kind of wavelet analysis in the Radon domain; if we let $Rf(\theta, t)$ denote the Radon transform and if we let $\tau_\lambda(t, \theta)$ denote the function $\frac{1}{2}(\psi^+_{j,k}(t)w^\varepsilon_{i,l}(\theta) + \psi^+_{j,k}(-t)w^\varepsilon_{i,l}(\theta + \pi))$, the $(\tau_\lambda : \lambda \in \Lambda)$ give a system of antipodally symmetrized non-orthogonal tensor wavelets. The ridgelet coefficients α_λ are given by analysis of the Radon transform via $\alpha_\lambda = [Rf, \tau_\lambda]$. This means that the ridgelet coefficients contain within them information about the smoothness in t and θ of the Radon transform. In particular, if the Radon transform exhibits a certain degree of smoothness, we can immediately see that the ridgelet coefficients exhibit a corresponding rate of decay.

5. Ridgelet synthesis of linear singularities

Consider again the Gaussian-windowed half-space (2.1). The CRT of this object is sparse, which suggests that a discrete ridgelet series can be made which gives a sparse representation of g. This can be seen in two ways.

(a) *Using dual frames*

It can be shown that there exist constructive and simple approximations using dual frames (which are not pure ridge functions) which achieve any desired rate of approximation on compact sets (Candès 1998, ch. 5). Indeed, let $\boldsymbol{A}$ be compact and ψ_i be a ridgelet frame for $L_2(\boldsymbol{A})$. Out of the exact series

$$g = \sum_i \langle g, \psi_i \rangle \tilde{\psi}_i, \tag{5.1}$$

extract the m-term approximation $\tilde{g}_m$ where one only keeps the dual-ridgelet terms corresponding to the m largest ridgelet coefficients $\langle g, \psi_i \rangle$; then the approximant $\tilde{g}_m$ achieves the rate

$$\|g - \tilde{g}_m\|_{L_2(\boldsymbol{A})} \leqslant C_r m^{-r} \quad \text{for any } r > 0,$$

provided, say, ψ is a nice function whose Fourier transform is supported away from 0 (like the Meyer wavelet). The result generalizes to any dimension n and is not limited to the Gaussian window. The argument behind this fact is the sparsity of the ridgelet coefficient sequence; each ridgelet coefficient $\langle \psi_{j,k}, Rg(\theta_{j,\ell}, \cdot) \rangle$ being the one-dimensional wavelet coefficient of the Radon transform $Rg(\theta_{j,\ell}, \cdot)$—for fixed θ. From the relation $Rg(\theta, t) = \mathrm{e}^{-t^2} \bar{\Phi}(-t \cdot \sin\theta / |\cos\theta|)$, it is easy to see that the coefficients $\langle f, \psi_{a,\theta,b} \rangle$ decay rapidly as θ and/or b move away from the singularities

$$(\theta = \tfrac{1}{2}\pi,\ t = 0) \quad \text{and} \quad (\theta = \tfrac{3}{2}\pi,\ t = 0).$$

(b) *Using orthonormal ridgelets*

Donoho (1998) shows that the orthonormal ridgelet coefficients of g belong to ℓ^p for every $p > 0$. This means that if we form an m-term approximation by selecting the m terms with the m-largest coefficients, the reconstruction $f_m = \sum_{i=1}^{m} \alpha_{\lambda_i} \rho_{\lambda_i}$ has any desired rate of approximation.

The argument for the orthonormal ridgelet approximation goes as follows. Because orthonormal ridgelet expansion amounts to a special wavelet expansion in the Radon domain, the question reduces to considering the sparsity of the wavelet coefficients of the Radon transform of g. Now, the Radon transform of g, as indicated above, will have singularities of order 0 (discontinuities) at $(t = 0,\ \theta = \frac{1}{2}\pi)$ and at $(t = 0,\ \theta = \frac{3}{2}\pi)$. Away from these points the Radon transform is infinitely differentiable, uniformly so, outside any neighbourhood of the singularities. If we 'zoom in' to fine scales on one of the singularities and make a smooth change of coordinates, the picture we see is that of a function $S(u, v) = |v|^{-1/2} \sigma(u/|v|)$ for a certain smooth bounded function $\sigma(\cdot)$. The wavelet coefficients of such an object are sparse.

6. Ridgelet analysis of ridge functions

Although ridge functions are not in L^2, the continuous ridgelet transform of a ridge function $f = r(x_1 \cos\theta_0 + x_2 \sin\theta_0)$ makes sense; if the ridge profile r is bounded, the transform can be obtained in a distributional sense and obeys

$$(\mathcal{R}_f)(a, b, \theta) = \delta(\theta - \theta_0) \cdot (Wr(a, b)). \tag{6.1}$$

Thus, the transform is *perfectly localized* to the slice $\theta = \theta_0$ of the precise ridge direction and it amounts to the one-dimensional wavelet transform of the profile function there. This exceptional degree of concentration suggests that ridge functions ought to have very sparse representations by discrete ridgelet systems and that a high rate of approximation can be obtained via m-term ridgelet approximations to such ridge functions using simple thresholding. This can be verified in two ways.

(*a*) *Using dual ridgelets*

Suppose that the ridge profile r is supported in the interval $[-1, 1]$ and obeys a sparsity condition on the wavelet coefficients in a nice wavelet basis: the coefficient sequence $\beta \in w\ell_p$ $(p < 2)$. Then the best m-term one-dimensional wavelet approximation to r has an $L_2[-1, 1]$ convergence rate of order $m^{-(1/p-1/2)}$. There exist approximations by superpositions of m dual ridgelets (which are not pure ridge functions) which achieve the $L^2(\boldsymbol{A})$ rate of approximation $m^{-(1/p-1/2)}$, where $\boldsymbol{A}$ is now the unit disc (Candès 1998, ch. 5 and 7). Such approximants can be constructed by selecting the m terms out of the series (5.1) corresponding to the m largest coefficients.

(*b*) *Using orthonormal ridgelets*

A key point about orthonormal ridgelets is that they are not only in $L^2(R^2)$, but also in $L^1(R^2)$; hence the integral defining orthonormal ridgelet coefficients makes sense for every bounded ridge function. Let the ridge profile $r(t)$ belong to the homogeneous Besov space $\dot{B}^s_{p,p}(\mathbb{R})$, where $s = 1/p$. This means that the best one-dimensional m-term wavelet approximation to r has an $L^\infty(\mathbb{R})$ convergence rate of $m^{-(s-1/p)}$.

Consider now the rate of convergence of thresholded ridgelet expansions. Let $\bar{\eta}_\delta(y, x) = y1_{\{y \cdot x > \delta\}}$ be a thresholding function with a second 'scaling' argument allowing for adjustment of the threshold. For a bounded function f, with

$$\bar{m}(\delta) = \sum_{\Lambda} 1_{\{|\langle f, \rho_\lambda\rangle| > \delta / \|\rho_\lambda\|_{L^\infty(D)}\}}$$

finite, set

$$\tilde{f}_\delta = \sum_{\Lambda} \eta^{(2)}_\delta(\langle f, \rho_\lambda\rangle, \|\rho_\lambda\|_{L^\infty(D)})\rho_\lambda.$$

In effect, thresholding is driven by the interaction between the size of a coefficient and the 'effect' of the corresponding basis function inside the unit disc.

Let $r_\theta(x)$ denote the corresponding ridge function of $x \in \mathbb{R}^2$. Let $\bar{f}_{m(\delta)}$ be the $\bar{m}(\delta)$-term orthonormal ridgelet approximation to the ridge function f. Then

$$\|f - \bar{f}_m\|_{L^\infty(D)} \leqslant C \cdot m^{-(s-1/p)}, \quad m \to \infty. \tag{6.2}$$

In effect, this result is ideal, as it gives the same rate $m^{-(s-1/p)}$ we could hope to obtain by knowing that the underlying approximand was a ridge function in a specific direction and exploiting that information fully—even though the ridgelet thresholding does not 'know' or 'exploit' such information.

These results suggest that dual ridgelet frames and orthonormal ridgelets, although *not* ridge functions, can play the same role in approximation as pure ridge functions. More precisely, suppose an arbitrary function f is well-approximated by a sequence of m-term superpositions of ridge functions; it seems that f should also be well approximated by m-term superpositions from discrete ridgelet systems.

7. Ridge spaces

An important fact about wavelets is their relationship to two special families of functional spaces—the Besov spaces and the Triebel spaces. Taken together, these families of spaces include an important collection of classical functional spaces, such as L^2 spaces, L^p spaces, Sobolev spaces, Hölder spaces and so on. Wavelets provide a special basis for such spaces (an unconditional basis) (Meyer 1990) and provide near-optimal approximations to elements taken from functional balls of such spaces.

With the existence of a new family of transforms, we have the possibility to ask: what are the spaces that these transforms are most naturally associated to? Candès (1998) defines a family of spaces $R^s_{p,q}$—'ridge spaces'—which consist of functions f with ridgelet coefficients obeying certain constraints:

$$\|f\|_{\dot{R}^s_{p,q}} = \left(\int \left[\int |\mathcal{R}_f(a,\theta,b)|^p \, \mathrm{d}b \, \mathrm{d}\theta \right]^{q/p} \frac{\mathrm{d}a}{a^{q(s+1)+1}} \right)^{1/q}$$

and similarly for higher dimensions where $\mathrm{d}\theta$ is replaced by the uniform measure on the sphere and the scale factor $a^{q(s+1)+1}$ by $a^{q(s+n/2)+1}$. (The above display corresponds to the homogeneous ridge spaces (see Candès (1998) for a corresponding inhomogeneous version).) Although the definition looks rather internal, it is possible to give an external characterization of such spaces because of the intimate relationship between the ridgelet analysis and the wavelet analysis of the Radon transform $Rf(u,t)$. In fact, letting $p = q$, one can check that

$$\|f\|^p_{\dot{R}^s_{p,p}} \asymp \mathrm{Ave}_u \|Rf(u,\cdot)\|^p_{\dot{B}^{s+(n-1)/2}_{p,p}},$$

where the notation $\dot{B}^{s+(n-1)/2}_{p,p}$ stands for the usual one-dimensional homogeneous Besov norm. From this characterization, it is clear that s is a smoothness parameter and that both parameters p, q serve to measure smoothness. Here, smoothness has to be understood in a non-classical way; we are not talking about the local behaviour of a function but rather about its behaviour near lines (or if one is in dimension $n > 2$, near hyperplanes).

To capture the essence of such spaces, let us return to our original mutilated Gaussian example, (2.1), generalized to dimension n:

$$g(x_1, \ldots, x_n) = 1_{\{x_n > 0\}} \mathrm{e}^{-(x_1^2 + \cdots + x_n^2)}.$$

From a classical point of view, in any dimension, this object has barely one derivative (in an L_1 sense) meaning that its first derivative is a singular measure, the singularity being supported on the plane $\{x_n = 0\}$. However, under our new definition, this same object is quite smooth and in fact its regularity increases as the dimension increases, as explained in Candès (1998). What do typical elements of these new spaces look like? The mutilated Gaussian is a typical element of $\dot{R}^s_{1,\infty}$ for $s \leqslant 1 + \frac{1}{2}(n-1)$.

For classical Besov spaces, Meyer (1990) tells us that typical elements of $\dot{B}^1_{1,1}$, for instance, are bumps of various scales and at various locations and that the latter space is nothing else than the collection of convex combinations of those bumps (bump algebra). An analogous observation can be made for the ridge spaces (Candès 1998, ch. 4). On the real line, a normalized *point singularity* σ of degree zero, say, is a smooth function away from the origin that may or may not have a pathological behaviour at the origin: that is, we want $|\sigma(t)| \leqslant 1$ and for a few derivatives $|\mathrm{d}^m\sigma(t)/\mathrm{d}t^m| \leqslant |t|^{-m}$ for $t \neq 0$ and $m \leqslant M$. As an example we have the Heaviside $1_{\{x>0\}}$, or a smoothly windowed version of the Heaviside. Next, out of a one-dimensional point singularity σ, we create a *ridge singularity* $\sigma(u'x - b)$, where u is a unit vector and b a scalar, and consider the set of functions arising as convex combinations of such ridge singularities:

$$\mathcal{S} = \left\{ f(x) = \sum_i a_i \sigma_i(u_i' x - b_i), \ \ \sum_i |a_i| \leqslant 1 \right\}.$$

Then, if we look at objects restricted to the unit ball, the membership of an object in $\mathcal{S}$ is essentially equivalent to a statement about the norm of this object in the norm $\dot{R}^s_{p,q}$ for appropriate (s, p, q). More precisely, we have the following double inclusion:

$$R^{1+(n-1)/2}_{1,1}(C_1) \subset \mathcal{S} \subset R^{1+(n-1)/2}_{1,\infty}(C_2), \tag{7.1}$$

saying that compactly supported objects with $R^{1+(n-1)/2}_{1,1}$ norm not exceeding C_1 are convex combinations of ridge singularities, and that every such convex combination has a bounded $\dot{R}^{1+(n-1)/2}_{1,\infty}$ norm.

It follows from this characterization that ridge spaces model very special conormal objects: objects that are singular across a collection of hyperplanes and smooth elsewhere, where there might be an arbitrary number of hyperplanes in all possible spatial orientations and/or locations.

Earlier, we claimed that ridgelets were naturally associated with the representation of ridge spaces. In fact ridgelets provide near-optimal approximations to elements of these spaces, in much the same way that wavelets provide near-optimal approximations to elements of Besov spaces. For instance, we know that the L_2 error of approximation to a mutilated Gaussian by an m-term linear combination of dual-ridgelets decays more rapidly than m^{-r} for any $r > 0$; the space $R^{1+(n-1)/2}_{1,\infty}$ being more or less the convex hull of such mutilated smooth objects, it is natural to guess that ridgelets provide the right dictionary to use for approximating these spaces.

We can make this more precise. Suppose we are given a dictionary $\mathcal{D} = \{g_\lambda, \lambda \in \Lambda\}$ and that we are interested in the L_2 approximation of a generic class of functions $\mathcal{F}$ out of finite linear combinations of elements of $\mathcal{D}$. For a function f and dictionary $\mathcal{D}$, we define its m-term approximation error by

$$d_m(f, \mathcal{D}) \equiv \inf_{(\alpha_i)_{i=1}^m} \inf_{(\lambda_i)_{i=1}^m} \left\| f - \sum_{i=1}^m \alpha_i g_{\lambda_i} \right\|,$$

and measure the quality of approximation of the class $\mathcal{F}$ using m selected elements of $\mathcal{D}$ by

$$d_m(\mathcal{F}, \mathcal{D}) \equiv \sup_{f \in \mathcal{F}} d_m(f, \mathcal{D})$$

(the worst case error over $\mathcal{F}$). Then, let us consider the class $\mathcal{F}$ of functions whose $R^s_{p,q}$-norm is bounded by some constant C (that will be denoted $R^s_{p,q}(C)$), to be approximated in the metric of $L_2(\boldsymbol{A})$ for some compact set $\boldsymbol{A}$. We impose the additional restriction $s > n(1/p - \frac{1}{2})_+$ to guarantee that our class belongs to L_2 also. Then, Candès (1998, ch. 5) shows that no reasonable dictionary would give a better rate of approximation than $m^{-s/d}$: that is, for any reasonable dictionary,

$$d_m(R^s_{p,q}(C), \mathcal{D}) \geqslant Km^{-s/d}.$$

On the other hand, thresholding the ridgelet expansion gives the optimal rate of approximation. Namely, if $|\alpha|_{(m)}$ denotes the mth largest amplitude among the $(|\alpha_i|)$, the m-term series

$$\tilde{f}_m = \sum_i \alpha_i 1_{\{|\alpha_i| \geqslant |\alpha|_{(m)}\}} \tilde{\psi}_i$$

produced by thresholding at $|\alpha|_{(m)}$ achieves the optimal rate

$$\sup_{f \in R^s_{p,q}(C)} \|f - \tilde{f}_m\|_{L_2(\boldsymbol{A})} \leqslant K' m^{-s/d},$$

for some constant $K' = K'(\boldsymbol{A}, C, s, p, q)$.

The result says that we have an asymptotically near-optimal procedure for binary encoding elements of $R^s_{p,q}(C)$: let $L(\epsilon, R^s_{p,q}(C))$ be the minimum number of bits necessary to store in a lossy encoding–decoding system in order to be sure that the decoded reconstruction of every $f \in R^s_{p,q}(C)$ will be accurate to within ϵ (in an L_2 sense). Then, a coder–decoder based on simple uniform quantization (depending on ϵ) of the coefficients α_i followed by simple run-length coding achieves both a distortion smaller than ϵ and a code length that is optimal up to multiplicative factors like $\log(\epsilon^{-1})$ (Donoho 1996).

8. Ridgelets and curves

As we have said earlier, wavelets are in some sense adapted to zero-dimensional singularities, whereas ridgelets are adapted to higher-dimensional singularities; or more precisely, singularities on curves in dimension two, singularities on surfaces in dimension three, and singularities on $(n-1)$-dimensional hypersurfaces in dimension n. Unfortunately, the task that ridgelets must face is somewhat more difficult than the task which wavelets must face, since zero-dimensional singularities are inherently simpler objects than higher-dimensional singularities. In effect, zero-dimensional singularities are all the same—points—while a one-dimensional singularity—lying along a one-dimensional set—can be curved or straight. *Ridgelets are specially adapted only to straight singularities.*

One way to see this is to look at the CRT of a curved singularity. Again in dimension $n = 2$, consider the object $g' = \mathrm{e}^{-x_1^2 - x_2^2} \cdot 1_{\{x_2 > x_1^2\}}$. Qualitatively, it is not hard to see that the Radon transform of such an object has a singularity along a curve, and not just at a point: that is, in the Radon domain, there is a smooth curve $t_0(\theta)$ so that in a neighbourhood of $(t_0(\theta), \theta)$, we have $Rg(t, \theta) \sim w(\theta)(t - t_0(\theta))_+^{1/2}$ for some smooth function w. When we take the wavelet transform in t along each fixed value of θ, we will find that the transform is not nearly as sparse as it was with g.

One can adapt to this situation by the method of localization, which has been frequently used, for example, in time-frequency analysis. We divide the domain in question into squares, and smoothly localize the function into smooth pieces supported on or near those squares either by partition of unity or by smooth orthonormal windowing. We then apply ridgelet methods to each piece. The idea is that, at sufficiently fine scale, a curving singularity looks straight, and so ridgelet analysis—appropriately localized—works well in such cases.

9. Discussion

Because of space limitation, the situation in higher dimensions and the structure of fast ridgelet transform algorithms for lower dimensions, for example, have not been mentioned in this paper. Information on these and related topics can be found in the references below.

References

Barron, A. R. 1993 Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39**, 930–945.

Candès, E. J. 1998 Ridgelets: theory and applications. PhD thesis, Department of Statistics, Stanford University.

Candès, E. J. 1999 Harmonic analysis of neural networks. *Appl. Comput. Harmon. Analysis* **6**(2), 197–218.

Cybenko, G. 1989 Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2**, 303–314.

Daubechies, I. 1988 Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.* **41**, 909–996.

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

Donoho, D. L. 1996 Unconditional bases and bit-level compression. *Appl. Comput. Harmon. Analysis* **3**, 388–392.

Donoho, D. L. 1998 Orthonormal ridgelets and linear singularities. Report no. 1998-19, Department of Statistics, Stanford University.

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

Helgason, S. 1986 *Groups and geometric analysis*. New York: Academic.

Jones, L. K. 1992 A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.* **20**, 608–613.

Lemarié, P. G. & Meyer, Y. 1986 Ondelettes et bases Hilbertiennes. *Rev. Mat. Iberoamericana* **2**, 1–18.

Logan, B. F. & Shepp, L. A. 1975 Optimal reconstruction of a function from its projections. *Duke Math. Jl* **42**, 645–659.

Meyer, Y. 1990 *Ondelettes et opérateurs*. Paris: Hermann.

Murata, N. 1996 An integral representation of functions using three-layered networks and their approximation bounds. *Neural Networks* **9**, 947–956.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES