NATURE VS. MATH: INTERPRETING INDEPENDENT COMPONENT ANALYSIS IN LIGHT OF COMPUTATIONAL HARMONIC ANALYSIS

David L. Donoho

Statistics Department
Stanford University
donoho@stat.stanford.edu

ABSTRACT

ICA has recently been applied to naturally-occurring data to uncover hidden fundamental components. Harmonic analysis has long been used to uncover hidden fundamental components of mathematically-defined objects. In my talk I will explore some recent parallelisms between ICA and CHA – fascinating similarities in the "hidden components" that the two subjects are uncovering. I will suggest that recent work on the independent components of images can best be interepreted in the light of recent constructions in harmonic analysis – such as ridgelets and curvelets. These mathematical objects exhibit surprising similarities to the results of ICA on certain naturally-occurring data.

1. INTRODUCTION

In this article we describe a circle of ideas involving an appealing blend of interactions between computational neurscience, visual physiology, and statistical analysis. Our goal is to propose that recent innovations in harmonic analysis may contribute to the discussion.

1.1. Stimulating Computer Experiments

A convenient starting place is with the article of Olshausen and Field in Nature [30]. This reported the analysis of a database built from natural images by sampling 16 by 16 image patches from the images. Such image patches may be viewed as arrays of numbers, the p-th patch taking the form $X^p = (X^p(i_1, i_2) : 1 \le i_1, i_2 \le n)$, where here n is the image patch extent; in Olshausen and Field's case, $n = 16$. The database $\mathcal{X} = \{X^p : p = 1, \ldots, P\}$ of such image patches was subjected to statistical modelling; Olshausen and Field tried to model each patch X^p as a linear combination of some underlying basis elements $\{\phi_\mu\}$ according to

$$X^p \approx \sum_\mu \theta^p_\mu \phi_\mu.$$

Here the key point is that the basis is *not fixed in advance* but is to be *learned* from the data.

Olshausen and Field proposed that a basis could be learned by searching for a basis that optimized the sparsity of the coefficients (θ^p_μ). The specific objective they proposed was to minimize, over bases Φ for the space of n by n arrays, the functional

$$S(\Phi) = \sum_{p=1}^{P} \left\{ \min_{\theta^p_\mu} \|X^p - \sum_\mu \theta^p_\mu \phi_\mu\|_2^2 + \lambda \sum_\mu \log(1 + (\theta^p_\mu)^2) \right\}$$

subject to the appropriate scale normalization of Φ. The functional S prefers bases which allow sparse representations. Subject to a fixed budget of coefficient 'energy' $\sum_\mu (\theta^p_\mu)^2$, the form $\sum_\mu \log(1 + (\theta^p_\mu)^2)$ is smallest if all but one of the θ^p_μ can be made 0, as can be seen from concavity of $\log(1 + x)$ on $x > 0$. Of course, because of the concavity, true global optimization of $S(\Phi)$ is a doubtful project; however, adapting simple "hill-climbing" ideas to improving an initial guess by following a descent direction provides a way to discover interesting local optima.

The result published by Olshausen and Field is depicted in Figure 1. This shows a collection of basis elements obtained by local optimization of $S(\Phi)$. The elements exhibit a range of orientations and locations. As each basis element ϕ_μ is a 16 by 16 array or image, the elements might be called 'imagelets' by those in a light mood.

(Incidentally, the traditional way of 'learning a basis from data' – Principal Components Analysis (PCA) – would not give this type of result. Several groups have observed that the Principal Components of image

This research was supported by National Science Foundation grants DMS 98–72890 (KDI) and DMS 95–05151; and by AFOSR MURI-95-P49620-96-1-0028.

patch data resemble unlocalized sinusoids, rather than localized oriented features; see [24]. So the search for sparsity seems an important factor in finding oriented, localized features).

Experiments of this general kind have by now been performed by several groups, each one seeking to locally optimize a different objective [4, 20, 21, 23, 22, 26, 27]. Bell and Sejnowksi [4] emphasized, instead of the sparsity of the coefficients θ_μ^p, the closeness to *independence* of those coefficients. Viewing p as selected uniformly at random, the representation coefficient θ_μ^p becomes a random variable θ_μ and we can seek a basis which minimizes an empirical measure of the mutual dependence of the collection $(\theta_\mu)_\mu$. The work of Bell and Sejnowski brought into the picture the notion of *Independent Components Analysis*, a very active area of research which involves workers from many fields, notably digital communications, where ICA has been put to good use in essential ways. By now, many different ICA-style algorithms have been tried on image-patch data, with the typical conclusion that ICA-style analysis produces basis elements displaying in many cases quite pronounced orientational preferences, and ranging over a variety of orientations and positions and scales.

One of the most stimulating extensions was an analysis of video sequences by van Hateren and Ruderman (1997). van Hateren and Ruderman compiled a database of video clips; each clip is an n by n by T array

$$V^c = \{V^c(i_1, i_2, t) : 1 \le i_1, i_2 \le n, 1 \le t \le T\},$$

with $n = T = 12$, so each video clip is composed of $12^3 = 1728$ numerical entries. The complete video database $\mathcal{V} = \{V^c : 1 \le c \le C\}$ can then be statistically analyzed, with the intent to develop a basis Φ allowing to represent the clips via

$$V^c = \sum_\mu \theta_\mu^c \phi_\mu,$$

where $\#\mu = n^2 \cdot T = 1728$, and the θ_μ^c represent the coefficients of V^c in the expansion by elements ϕ_μ. van Hateren and Ruderman used an ICA method oriented towards maximizing the kurtosis of the coefficients θ_μ^c; the algorithm was based on work of Hyvärinen (1996).

The result of analysing a database of video clips in this way is a basis for the space of clips; the resulting "cliplets" or "movielets" might be thought to be intrinsically associated with representing video content.

The results of van Hateren and Ruderman are best presented as videos, since the basis elements themselves are video clips. In this article we can only describe them. Roughly what one sees in examining one such clip is a moving line segment or 'bar', with a certain orientation, direction of propagation, and speed of propagation. Vertical and horizontal alignments are noticeably present; some line segments do not move at all; in general, however, typically one sees movement, over a range of directions and speeds.

1.2. What do you call these things?

Apparently we belong to an era fascinated with the creation of new bases for signals and images. In the last ten years, wavelets, wavelet packets, cosine packets, brushlets, noiselets, and Wilson bases have all been invented [28]; we now have available many systems of representation for signals and images. Where do "imagelets" and "movielets" fit in?

The answer, at least from reading the existing literature, seems to be that there is no clear answer. These images and clips, which are so visually striking, seem to carry different messages to different viewers.

Olshausen and Field initially proposed the idea that their components were *wavelets*; the prepublication version of their paper was titled "*Wavelet-like* receptive fields emerge from a network that learns sparse codes for natural images" (emphasis added). However, the final version of the paper removed any reference to wavelets from the title, so apparently they retreated from that position.

Bell and Sejnowski boldly titled their paper "*Edges* are the independent components of images" (emphasis added); since "edges" don't make a known system of representation (but compare the edgelets of [14]) they seem to be claiming that the system has no counterpart among known systems.

Others, notably Lewicki and Olshausen (1997) and van Hateren and Ruderman (1997), side with the position that these are *Gabor* functions in some sense, perhaps Gabor functions with highly anisotropic Gaussian windows.

1.3. What's at Stake?

This body of work has emerged from a community of researchers engaged in statistical analysis of natural scenes; these workers are motivated by the idea that natural images are what living organisms see as they develop, and that perhaps living organisms develop as they do *because* of the statistical properties of the images they are exposed to [2, 18].

In his review article [32], D. Ruderman points out that

> The development of the mammalian visual system is strongly dependent on visual

> stimulation ... In particular it is known that mammals raised in unusual image environments end up with functionally modified visual systems as adults. This suggests that the visual system's development is influenced by the statistics of its environment, through some as-yet unknown algorithm. Contained in the final "wiring" of the visual system is a set of statistics about the creature's past visual experience. Knowing wich statistics those are might provide great insight toward the nature of visual processing."

The "as-yet unknown algorithm" referred to by Ruderman has been proposed by some to derive from attempts to respond to the statistics of natural scenes to learn how to efficiently represent the scene information, Here the model is taken to be Shannon's information theory, which shows in principle how to adapt to the statistics of messages, such as letter frequencies, learning how to recode them to obtain the most efficient transmission of such images.

In his review article [1], J. Atick says

> "... efficiency of information representation in the nervous system potentially has evolutionary advantages ... much of the processsing in the early levels of sensory pathways might be geared towards building efficient representations of sensory stimuli in an animal's environment.
>
> The ... efficiency principle, formulated as an optimization problem, can be used as a *design* principle to predict neural processing. Starting with the representation of environmental signals as sampled by the array of sensory cells, one can try to find recodings needed to improve efficiency subject to identifiable biological hardware constraints. The several stages of processing required to cast incoming data into the optimal form can then be compared to the stages of neural processing observed in sensory pathways. This principle has been shown to successfully predict retinal processing in space-time and color ... and there are encouraging signs that it could be equally successful in predicting some of the cortical computation strategies."

Underlying the idea that the structure of the visual system might derive from an efficiency seeking principle is a striking fact: the visual system actually *performs* a wonderfully efficient job, reducing roughly 10^7 bits/s of visual data to fit the roughly 40-50 bits/s of perceptual capacity of the deep visual pathway [2].

Now *if* the visual system developed by an efficiency-seeking principle, it *might* be constructed so as to find independent components. From several point of view, an independent components representation of images is a very efficient representation. On the one hand, information theory shows that in certain settings optimal compression results from a transformation to independent random variables, followed by a recoding of the independent data [5]; on the other hand, Barlow [3] has pointed out that computation of joint probabilities is by far easiest when individual components are independent, so a learning mechanism which transformed to independent factors would lead to highly efficient algorithms for processing such information.

1.4. Comparisons with Physiology

Continuing the chain of reasoning of Section 1.3, it is extremely natural that the searchers for independent components in natural images wasted no time in trying to relate the numerical findings to physiological evidence on the structure of simple cells in the V1 area of the higher mammals. Indeed, there is a tradition to do this; see for example Watson, Barlow, and Robson (1983), which attempted to compare human visual reponse to Gabor functions, and the extensive work of Field (1987, 1993), which has energetically advanced the case that statistical properties of natural scenes are reflected in the physiological responses observed in studies of mammals.

The paper of Olshausen and Field, as published, already contained, in its title, the implicit assertion that the components they discovered "were" simple cell profiles. Of course, in that brief paper, no careful documentation of such assertion would have been possible. Later work, particularly of van Hateren and Ruderman (1997) and Lewicki and Olshausen (1998), made careful compilations of the properties of the basis elements they obtained in their experiments, and van Hateren and van der Schaaf (1997), particularly, made comparisons of their findings with physiological evidence on macaque primates, for example as compiled by De Valois and collaborators.

The evidence from these studies seems suggestive but not decisive. In certain respects, the orientation selectivity and spatial bandwidth particularly, the numerical components agree with physiological evidence. In other respects, particularly the number of components within a given range of spatial localization, there is an apparent mismatch between computed components and physiology; see van Hateren and van der Schaaf (1997).

1.5. Unresolved Questions

We now have introduced two unresolved issues:

[U1] How do the computed components compare with basis elements in existing systems of harmonic analysis?

[U2] How do the computed components compare with linear receptive fields in the V1 region of mammals?

It may in fact be premature to attempt such comparisons at this stage. It unclear that numerical experiments with the kinds of image patch data studied to date can accurately reveal much about the structure of the 'true independent components' of continuum scenes with large field of view (supposing that there really are such independent components). There are two significant problems:

- *Aliasing.* One does not observe the full diversity of image components in a 16 by 16 image patch; any coherent feature exerting significant structure over more than a 16 by 16 window (e.g. a long edge) will necessarily be aliased onto a feature fitting in the window.

- *Boundary effects.* Scientists who deal extensively with computation on grids find that boundary effects are often important, often exerting disproportionate effects on results of computations with small problem sizes. But in a 12 by 12 image, a fraction 11/36 of the pixels are boundary pixels, and 20/36 are at most one pixel away from the boundary. In three dimensions, at 12 by 12 by 12, a fraction 728/1728 of the pixels lies on the boundary, and 1216/1728 lie within one pixel. In short, the existence of the boundary might be a significant factor at such small patch/clip sizes.

Many of the problems standing in the way of resolving [U1] and [U2] could be directly addressed by computing independent components on databases compiled from large-sized image patches. This would alleviate the influence of boundaries and of scale aliasing, and give us greater confidence in the stability properties of the obtained solution.

1.6. Curse of Dimensionality

Unfortunately, it does not seem likely (to this author) that reliable results will be obtained for dramatically larger image patches any time soon. For example, obtaining a complete set of estimated independent components for a database with 128-by-128 image patches seems out of reach with existing algorithms, as does the processing of databases of 32 by 32 by 32 video clips. This is because of two effects:

[1] *Computational complexity.* It is algorithmically a rather slow process to obtain independent components. van Hateren and Ruderman report that in computing components for the video clip setting, where they used an algorithm that extracted one component at a time, they were only able to obtain components at the rate of 15/day on a Cray supercomputer. Moreover (see Section 3 below), the algorithms for ICA typically scale with the image size according to very unfavorable exponents, so that as image extents increase, the rate of component extraction will likely slow down dramatically.

[2] *Inferential validity.* Independent components analysis is by its very nature about the properties of a high-dimensional joint distribution of random variables. But one of the cornerstones of modern statistical thought is that estimation of such properties is necessarily highly unstable; the extensive literature on functional data analysis [31] begins from this premise, and posits that special regularization of such problems is necessary to get reliable results. The author suspects that an algorithmic breakthrough which allowed ICA to work with much greater image extents, would simply expose this statistical issue in its full force. In effect, the small image extents which could be attacked to date acted to keep down the dimensionality of the problem and so acted as implicit regularizing factors; if this regularization went away, we would have to face up to the fact that accuractely estimating a basis in high dimensions from empirical data is typically a very poorly-posed problem, suffering from a very high level of statistical variability.

1.7. More Thinking, not More Computing!

It seems promising at this point to simply *think* about what the results of independent components analysis *might* be with much larger n, using mathematical analysis to derive some qualitative predictions. Indeed, recent developments in Computational Harmonic Analysis (CHA) allow us to make interesting proposals about what we might expect to see for image model data containing edges.

CHA is a rapidly developing field, which focuses on developing new representations of signals and images which are rapidly computable – examples include wavelets, wavelet packets, cosine packets, brushlets,

and so on. For surveys see [28, 34]. Of interest for this paper is that fact that some of the results in this field exhibit orthogonal bases – constructed using the tools of mathematical analysis, rather than by data analysis – which combine a certain kind of near-optimality for certain purposes along with a clear interpretatibility.

In my talk I will look at the problem of ICA of images from a CHA perspective, hoping to gain insight into the structure of the independent components of certain synthetic image models. The plan is to study ICA in one of its guises – the one where the goal is to diagonalize cumulant tensors – and, true to the CHA viewpoint, to seek only an approximate diagonalization, but one which is interpretable. To make this a problem in mathematical analysis rather than data analysis, I will forego the analysis of natural images in favor of synthetic models in which a scene is a random superposition of a few objects chosen at random from a library. I will discuss the ability of recent constructions in CHA – *ridgelet analysis* and *curvelet analysis* – to almost diagonalize the cumulant tensor of an appropriate random process model based on a library of edge-dominated objects. I will predict that the results of high-resolution ICA might turn out to be more similar to ridgelets than either wavelets or Gabor functions.

In this paper I will sketch some key parts of my talk, giving a kind of *aide-memoire* for the conference attendee, and providing pointers to more substantial articles. In Section 2 I will describe an important motivating example: the fact that ICA on 1-dimensional signals with singularities can "discover" wavelets. In Section 3 I will describe a recent development in harmonic analysis of two-dimensional objects – orthonormal ridgelets. In Section 4, I will describe an adaptation – tight frames of Curvelets. Section 5 will describe the role that these transforms play in *sparsifying* images with edges. Section 6 will describe the connection between sparsity and Kurtosis.

2. A ONE-DIMENSIONAL MODEL PROBLEM: RAMP

We now consider cumulant-based ICA in a specific 1-dimensional model problem, which we will return to frequently in later sections.

2.1. Yves Meyer's Ramp Process

To get started, we change gears and consider a stochastic process $X(t)$ defined on the index set $T \equiv [0,1]$ as follows. Let ω be chosen from the uniform distribution on $[0,1]$ and define

$$X(t;\omega) = \begin{cases} t & t < \omega \\ t-1 & t \geq \omega \end{cases}$$

This non-Gaussian process, which we call *Ramp*, has very simple behavior: it jumps down by 1 at the uniformly distributed jump time ω; otherwise it just increases at unit slope. The author became interested in this process through a presentation by Yves Meyer [29] who proposed it as an example of a non-Gaussian process in which all the interesting information is carried by the singularity at $t = \omega$, and which is isospectral to Brownian Bridge.

One can check that $EX(t) = 0$ and that

$$Cov(X(t), X(s)) = min(t,s) - ts.$$

This last observation is rather striking; it says that X has the same covariance as the Brownian Bridge, i.e. that the Gaussian process $B(t)$ defined by $B(t) = W(t) - tW(1)$, where W is the standard Wiener process, has the same second-order properties as *Ramp*; this is a striking illustration of the weakness of second order equivalence, since *Ramp*'s behavior is dominated by the singularity, and is otherwise very regular, while the Brownian Bridge is irregular everywhere.

Since X and B are second-order equivalent, they share the same principal components analysis; or to be terminologically correct – since we are in a settng of function-valued realizations rather than finite-dimensional vectors – they share the same Karhunen-Loève Decomposition. The eigenfunctions of the covariance of $B(t)$ are $\phi_k(t) = \sqrt{2}sin(\pi kt)$, for $k = 1, 2, 3, \ldots$, while the eigenvalues

$$\lambda_k = \frac{1}{4\pi k^2}, k = 1, 2, 3, \ldots.$$

This shows that B has a representation

$$B(t) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} Z_k \phi_k(t)$$

where the Z_k are independent standard normal random variables; this is the independent components representation of B. Decades of probability research have proven that B has a very complex and interesting behavior; the reflection of this in the present setting is that the λ_k decay slowly, and it requires many components to closely approximate B. The relation between the number N of terms in a partial reconstruction $B^{(N)} = \sum_{k=1}^{N} \langle B, \phi_k \rangle \phi_k$ and the error of reconstruction is

$$E||B - B^{(N)}||^2 \quad = \quad \sum_{N+1} \lambda_k \sim \frac{1}{4\pi} N^{-1}, \quad N \to \infty.$$

So three-digit relative accuracy requires roughly $10^6/4\pi$ terms. This seems to make intutive sense, because Brownian motion has very rich and interesting behavior, defying many naive expectations.

Because of second-order equivalence between X and B, similar relations hold with the Ramp Process X:

$$X = \sum_{k=1}^{\infty} \sqrt{\lambda_k} U_k \phi_k(t)$$

Here the random variables U_k are mutually orthogonal but not independent. The relation between the number N of terms in a partial reconstruction $X^{(N)} = \sum_{k=1}^{N} \langle B, \phi_k \rangle \phi_k$ and the error of reconstruction is still

$$E\|X - X^{(N)}\|^2 = \sum_{N+1} \lambda_k \sim \frac{1}{4\pi} N^{-1}, \quad N \to \infty. \quad (1)$$

So *Ramp* also appears hard to approximate, in the second-order viewpoint.

The appearance of similarity in these expansions is misleading. Note that this is a principal components analysis of X, but as the U_k are not independent, it is not an independent components analysis of X.

One sign of this distinction is that there are far better ways of building N-term approximations to X than to use $X^{(N)}$. One such approximation was pointed out by Yves Meyer: it is provided by taking the N-largest terms in a wavelet orthonormal basis. Use Daubechies orthonormal wavelets with compact support, adapted to the interval $[0, 1]$. Then the rearranged wavelet coefficients in such a basis decay exponentially fast, so there are at most $C_1 + C_2 \cdot j$ coefficients larger than any threshold $2^{-j/2}$, $j = 1, 2, \ldots$. It follows that three-digit relative accuracy in reconstruction is possible using only about 30 terms chosen adaptively on the basis of the particular realization X, i.e. on the value of ω. In this sense, the second-order equivalence between X and B is entirely misleading; while B is intrinsically hard to approximate, X is not hard to approximate.

We can interpret these observations against the background of ICA. In the case of X, the abandonment of the Karhunen-Loeve basis and linear approximation in favor of a type of nonlinear approximation in another basis works because sinusoids are nothing like independent components for X. On the other hand, they are the independent components for B and so there should be no better scheme of N-term approximation to B, in any other basis, by linear or nonlinear means; the best approximation should be built from these independent components.

2.2. Computational Experiment

Consider a simple numerical experiment. With $n = 32$ we define vectors Y_i, $i = 1, ..., n$, each one a simple digitization of *Ramp*.

$$Y_i(t) = X(t/n, i/n), \quad 1 \le t \le n; \quad 1 \le i \le n.$$

The database $\mathcal{Y}$ thus consists of 32 signal patches, and we use this database as input to the JADE procedure. That is, we calculate from this data the empirical $(32)^4$ cumulant tensor, and we use JADE to attempt a diagonalization of this tensor. The result will be an orthonormal basis with 32 elements depicted in Figure 2. The structure of the basis is rather remarkable; it has many of the features of a wavelet basis.

- *Dyadic Scales.* The elements seem visually to posses a variety of scales; they can be arranged in a dyadic pyramid, with 16 elements at the finest scale, 8 elements at the the next finest scale, and so on.
- *Translations.* Within one scale, the elements seem to be approximately translates of each other, so that (at fine scales particularly) there are elements located roughly at positions $t_{j,k} = k/2^j$.
- *Cancellation.* The elements at fine scales seem to be oscillatory, with two vanishing moments.

Figure 3 shows a Daubechies nearly-symmetric basis with 6 vanishing moments. Figure 4 gives a few side-by-side comparisons between these "JADElets" and certain Daubechies nearly-symmetric wavelets.

2.3. Central Question

So a method motivated by independent components analysis has uncovered wavelet-like elements "hiding" behind the *Ramp* database. We wonder if the same thing is true of images: will a *known system of harmonic analysis* turn out to closely match the results of *ICA analysis of a database*?

In the coming sections we introduce ridgelets and curvelets and provide some evidence in this direction.

3. RIDGELETS

The theory of ridgelets was developed in the Ph.D. Thesis of Emmanuel Candès (1998). In that work, Candès showed that one could develop a system of analysis based on ridge functions

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

He introduced a continuous ridgelet transform $R_f(a,b,\theta) = \langle \psi_{a,b,\theta}(x), f\rangle$ with a reproducing formula and a Parseval relation. He also constructed frames, giving stable series expansions in terms of a special discrete collection of ridge functions. The approach was general, and gave ridgelet frames for functions in $L^2[0,1]^d$ in all dimensions $d \geq 2$ – For further developments, see [8].

Donoho [15] showed that in two dimensions, by heeding the sampling pattern underlying the ridgelet frame, one could develop an orthonormal set for $L^2(\mathbf{R}^2)$ having the same applications as the original ridgelets. The ortho ridgelets are indexed using $\lambda = (j,k,i,\ell,\epsilon)$, where j indexes the ridge scale, k the ridge location, i the angular scale, and ℓ the angular location; ϵ is a gender token. Roughly speaking, the ortho ridgelets look like pieces of ridge functions (2) which are windowed to lie in discs of radius about 2^i; $\theta_{i,\ell} = \ell/2^i$ is roughly the orientation parameter, and 2^{-j} is roughly the thickness. See Figure 6.

A formula for ortho ridgelets can be given in the frequency domain

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

Here the $\psi_{j,k}$ are Meyer wavelets for $\mathbf{R}$, $w^\epsilon_{i,\ell}$ are periodic wavelets for $[-\pi,\pi)$, and indices run as follows: $j,k \in \mathbf{Z}$, $\ell = 0,\ldots,2^{i-1}-1$; $i \geq 1$, and, if $\epsilon = 0$, $i = \max(1,j)$, while if $\epsilon = 1$, $i \geq \max(1,j)$. We let Λ be the set of such λ.

The formula is an operationalization of the *ridgelet sampling principle*:

- Divide the frequency domain in dyadic coronae $|\xi| \in [2^j, 2^{j+1}]$.
- In the angular direction, sample the j-th corona at least 2^j times.

This is depicted in Figure 5.

The sampling principle can be motivated by the behavior of Fourier transforms of functions with singularities along lines. Such functions have Fourier transforms which decay slowly along associated lines through the origin in the frequency domain. As one traverses a constant radius arc in Fourier space, one encounters a 'Fourier ridge' when crossing the line of slow decay. The ridgelet sampling scheme tries to represent such Fourier transforms by using wavelets in the angular direction, so that the 'Fourier ridge' is captured neatly by one or a few wavelets. In the radial direction, the Fourier ridge is actually oscillatory, and this is captured by local cosines.

4. CURVELETS

The curvelet tight frame for $L^2(\mathbf{R}^2)$ is a collection of analyzing elements $\gamma_\mu = \gamma_\mu(x_1,x_2)$ indexed by tuples $\mu \in M'$ to be described below. It has been defined in [9] and has the following key properties:

- Transform Definition:
$$\alpha_\mu \equiv \langle f, \gamma_\mu\rangle, \qquad \mu \in M'.$$
- Parseval Relation:
$$\|f\|_2^2 = \sum_{\mu\in M'} |\alpha_\mu|^2.$$
- L^2 Reconstruction Formula:
$$f = \sum_{\mu\in M'} \langle f, \gamma_\mu\rangle \gamma_\mu.$$

These formal properties are very similar to those one expects from an orthonormal basis, and reflect an underlying stability of representation.

4.1. Analysis

There is a procedural definition of the transform.

- *Subband Decomposition.* We define a bank of subband filters P_0, $(\Delta_s, s \geq 0)$. The object f is filtered into subbands:
$$f \mapsto (P_0 f, \Delta_1 f, \Delta_2 f, \ldots).$$
The different subbands $\Delta_s f$ contain details about 2^{-2s} wide.
- *Smooth Partitioning.* We define a collection of smooth windows $w_Q(x_1,x_2)$ localized around dyadic squares
$$Q = [k_1/2^s, (k_1+1)/2^s) \times [k_2/2^s, (k_2+1)/2^s)$$
Multiplying a function by the corresponding window function w_Q produces a result localized near Q. Doing this for all Q at a certain scale, i.e. for all $Q = Q(s,k_1,k_2)$ with k_1 and k_2 varying but s fixed, produces a smooth dissection of the function into 'squares'. In this stage of the procedure, we apply this windowing dissection to each of the subbands isolated in the previous stage of the algorithm.
$$\Delta_s f \mapsto (w_Q \Delta_s f)_{Q\in\mathcal{Q}_s}.$$

- *Renormalization.* For a dyadic square Q, let

$$(T_Q f)(x_1, x_2) = 2^s f(2^s x_1 - k_1, 2^s x_2 - k_2)$$

denote the operator which transports and renormalizes f so that the part of the input supported near Q becomes the part of the output supported near $[0,1]^2$.

In this stage of the procedure, each 'square' resulting in the previous stage is renormalized to unit scale

$$g_Q = (T_Q)^{-1}(w_Q \Delta_s f), \qquad Q \in \mathcal{Q}_s.$$

- *Ridgelet Analysis.* Each 'square' is analyzed in the orthonormal ridgelet system. This is a system of basis elements ρ_λ making an orthobasis for $L^2(\mathbf{R}^2)$:

$$\alpha_\mu = \langle g_Q, \rho_\lambda \rangle, \qquad \mu = (Q, \lambda).$$

What do curvelets look like? Basically, like windowed ridgelets. Indeed, in the above notation, we have the formula

$$\gamma_\mu = \Delta_s w_Q T_Q \rho_\lambda;$$

in short, we take a ridgelet, translate, window, filter, and renormalize, producing a result which looks like a ridgelet localized to an essentially arbitrary location at an arbitrary scale.

For an understanding of why the procedure might be organized as it is, consider Figure 7. Suppose that we have an object f which exhibits an edge. Upon subband filtering, each resulting fine-scale subband output $\Delta_s f$ will contain a map of the edge in f, thickened out to a width 2^{-2s} according to the scale of the subband filter operator. This gives the subband the appearance of a collection of smooth ridges. When we smoothly partition each subband into 'squares', we see either an 'empty square' – if the square does not intersect the edge – or a ridge fragment. Moreover, the ridge fragments are nearly straight at fine scales, because the edge is nearly straight at fine scales. Such nearly straight ridge fragments are precisely the desired input for the ridgelet transform.

5. SPARSIFYING EDGES

The ridgelet transform and curvelet transforms were developed for the purpose of sparsifying objects with edges.

Consider an object with discontinuity along the line $x_1 a + x_2 b = c$, but which is otherwise smooth. Think, for example, of the mutilated Gaussian:

$$g(x_1, x_2; a, b, c) = 1_{\{x_1 a + x_2 b > c\}} e^{-x_1^2 - x_2^2}. \qquad (3)$$

This case is ideal for representation by ridgelets as compared to other systems of representation. The ortho-ridgelet coefficients are very sparse, and the approximation to g by a superposition of the N ortho-ridgelets with the N largest amplitudes converges to g at any desired rate: $\|g - g_N\|_2 \le C_\beta \cdot N^{-\beta}$ for all $\beta > 0$, $n = 1, 2, 3, \ldots$. In contrast N-term wavelet approximations converge only at the rate $N^{-1/2}$ and N-term Fourier approximations converge only at the rate $N^{-1/4}$. One can get a high quality approximation to an object with a (perfectly) straight edge using many fewer ridgelets than wavelets or sinusoids.

If the edge is curved, we should instead use curvelets. This system is better than other systems such as wavelet, Fourier, and ridgelets. For analyzing an object f with C^2 smoothness except for a discontinuity along a C^2 curve, the curvelet approximation to f by a superposition of the N curvelet terms having coefficients with the N largest amplitudes converges to f at essentially the N^{-1} rate: $\|f - f_N\|_2 \le C_\beta \cdot N^{-\beta}$ for all $\beta > 1$, and $N = 1, 2, 3, \ldots$. In contrast N-term wavelet approximations converge only at the rate $N^{-1/2}$ and N-term Fourier approximations converge only at the rate $N^{-1/4}$. Roughly speaking the rate of convergence by curvelets is double the rate for wavelets and 4 times the rate for sinusoids. One can get a high quality approximation to an object with a curved edge using substantially fewer curvelets than wavelets or sinusoids.

Figure 8 helps illustrate a key point about the quantitative performance of the curvelet procedure. The procedure extracts a ridge fragment from subband s with aspect ratio 2^{-s} by 2^{-2s}, and renormalizes, obtaining an object which, in the frequency domain, has support localized to the frequency band $|\xi| \approx 2^s$, and lives in a region of width ≈ 1. In short, the Fourier Transform of a ridge fragment is a ridge fragment. By the very construction of the ridgelet transform, one sees that one is expecting to encounters an object with a Fourier transform looking like such a ridge fragment, and that the ridgelet tranform is defined so that a very few ridgelet coefficients will be needed to represent such an object.

We can see why it would not be very helpful to use classical transforms for such ridge fragments. The Fourier transform uses sinusoids, which correspond to points in the frequency domain. A ridge fragment's Fourier transform is again a ridge of dimensions 2^s by 1. Hence order 2^s coefficients are needed to represent a single ridge fragment using sinusoids. The Wavelet transform has elements which correspond to annular rings in the frequency domain, multiplied by sinusoids; their angular support is very large, effectively constant, indpendent of scale. The ridge fragment is supported

in a band of angular resolution $O(2^{-s})$. Hence it also takes order 2^s coefficients to represent a single ridge fragment. Only the ridgelet basis has the required angular localization to mimic the ridge fragment signatures. For rigorous analysis, see the references, for example, [15, 7].

6. KURTOSIS AND SPARSITY

Empirically, transforms with sparse outputs often yield coefficients with high normalized kurtosis. Here is a heuristic explanation. Suppose that we are considering a transform with the property that that coefficients of a typical object are naturally grouped in subbands, and that, in each subband, the individual coefficient amplitudes either vanish or else take a certain common nonzero value v_s. Assume the coefficients are equally like to be positive and negative. Let ε_s be the fraction of nonzero items in the subband. Then defining the normalized empirical kurtosis at subband s by $K_{4,s} = Ave|X_i|^4/(Ave|X_i|^2)^2 - 3$, and assuming small ε_s, we have

$$K_{4,s} \approx \frac{\varepsilon_s v_s^4}{(\varepsilon_s v_s^2)^2} \approx \varepsilon_s^{-1}.$$

Hence, under this model, the sparser the subband, the higher the kurtosis. Note that the amplitude of the nonzero coefficients in the subband does not matter.

For the model of edges along curves given above, and for wavelets, Fourier and curvelets, the above heuristic is reasonable. Wavelets at subband j have about 2^j nonzero coefficients. Fourier at Corona j has at least 2^j nonzero coefficients, and possibly many more if edges span a wide range of directions. But, for curvelets chosen from a matching frequency range $|\xi| \in [2^{j-1}, 2^{j+2}]$ (so that $s = j/2$), there are roughly $2^{j/2}$ nonzero coefficients (i.e. a few per curvelet subband cell). These yield the predictions that in analysing smooth objects with discontinuities along curves,

- Kurtosis in the Fourier basis scales with subband index j at best like 2^j, and probably much worse.
- Kurtosis in the Wavelet basis scales with subband index j like 2^j.
- Kurtosis in the Curvelet basis scales with subband index j like $2^{3j/2}$.

In our experience there is rough agreement between this rule of thumb and what we see on real images. Although I will present more evidence on this at the conference, I can say here that, for example, we have the following result on a 256×256 version of the Barbara image: at subband $j = 6$, the wavelet basis gave empirical Kurtosis 19.560 while the curvelets frame gave empirical Kurtosis 53.8995.

Now we know that when there is a truly independent components representation, the kurtosis will be larger in that representation than in any alternative representation [13, 12]. Hence qualitatively curvelets seem 'close' to an IC representation, closer than anything else we know.

Now our analysis of kurtosis scaling at fine scales derives from a sparsity argument. Also, we know that curvelets give optimally sparse decompositions of images which are smooth functions away from edges – optimal among all bases, and even among all overcomplete representations.

This suggests that curvelets might do, in an appropriate sense *an optimal job of making the kurtosis large*. In fact, there is considerable mathematical evidence of such optimality, to be discussed at the workshop. There is also visual evidence, for example, comparing Figures 1 and 6 and also displays to be presented at the workshop.

One is tempted to claim that 'curvelets are the independent components of images'; we leave this issue for discussion at the ICA 2000 Workshop.

7. REFERENCES

[1] Atick, J.J. (1992) Could information theory provide an ecological theory of sensory processing? *Network* **3**, 213-251.

[2] Barlow, H.B. (1958) Sensory Mechanisms, the Reduction of Redundancy, and Intelligence. Proc. Symp. Mechanization of Thought Processes, National Physical laboratory, Teddington Middlesex.

[3] Barlow, H.B. (1989) Unsupervised Learning. *Neural Computation* **1** 295-311.

[4] Bell, A.J. and Sejnowski, T.J. (1995) An information-maximization approach to blind separation and blind deconvolution. *Neural Computation* **7** 1129-1159.

[5] Berger, T. (1971) *Rate-Distortion theory: a mathematical basis for data compression.* Englewood Cliffs, NJ: Prentice-Hall.

[6] Candès, E. (1999). Harmonic analysis of neural networks. *Appl. Comput. Harmon. Anal.* **6** 197–218.

[7] Candès, E. (1998) *Ridgelets: Theory and Applications.* Ph.D. Thesis, Department of Statistics, Stanford University.

[8] Candès, E. and Donoho, D. (1999). *Ridgelets: the key to high-dimensional intermittency?. Phil. Trans. R. Soc. Lond. A.* **357** 2495-2509.

[9] Candès, E. J. and Donoho, D. L. (2000). Curvelets: a surpisingly effective nonadaptive representation of objects with edges. in *Curve and Surface Fitting: Saint-Malo 1999* Albert Cohen, Christophe Rabut, and Larry L. Schumaker (eds.) Vanderbilt University Press, Nashville, TN. ISBN 0-8265-1357-3

[10] Cardoso, J.F. (1998) High-Order contrasts for independent component analysis. *Tutorial NIPS*98.*

[11] Cardoso, J.F. and Souloumiac, A. (1993) Blind Beamforming for non-Gaussian Signals. *IEE Proceedings-F.* **140** 352-370.

[12] Comon, P. (1994) Independent Component Analysis, a new concept? *Signal Processing* **36** 287-314.

[13] Donoho, D.L. (1981) On Minimum Entropy Deconvolution. In *Applied Time Series Analysis II*, D.F. Findlay, Ed. Academic Press.

[14] Donoho, D. L. (1999). Wedgelets: nearly-minimax estimation of edges. *Ann. Statist.* **27** 859-897.

[15] Donoho, D. L. (1998). *Orthonormal ridgelets and linear singularities* Technical Report, Department of Statistics, Stanford University. To appear, *SIAM J. Math. Anal.*

[16] Donoho, D. L. (1998). Sparse Components Analysis and Optimal Atomic Decomposition, Technical Report, Department of Statistics, Stanford University. To appear, *Constructive Approximation.*

[17] Donoho, D.L. (1999) Tight Frames of k-Plane Ridgelets and the Problem of Representing d-dimensional singularities in $\mathbf{R}^n$. *Proc. Nat. Acad. Sci. USA*, **96**, 1828-1833.

[18] Field, D.J. (1987) Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.* **4**, 2379-2394.

[19] Field, D.J. (1993) Scale-invariance and Self-similar 'Wavelet' transforms: an analysis of Natural Scenes and Mammalian Visual Systems. *Wavelets, Fractals and Fourier Transforms* M. Farge, J Hunt, and J.C. Vassilicos, eds. Oxford University Press.

[20] Fyfe, Colin, and Baddeley, R. (1995) Finding compact and sparse distributed representations of visual images. *Network* **6**, 333-344.

[21] Harpur, G.H and Prager, R.W. (1996) Development of low entropy coding in a recurrent network. *Network* **7** 277-284.

[22] van Hateren, J.H. and Ruderman, D.L. (1998) Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond.* B **265**

[23] van Hateren, J.H. and van der Schaaf, A. (1998) Independent component filters of natural images compared with simple cells in the primary visual cortex. *Proc. R. Soc. Lond.* B **265**:359-366.

[24] Hancock, P.J.B., Baddeley, R.J. and Smith, L. (1992) The principal components of natural images. *Network*, **2** 61-70.

[25] Hyvarinen, A. (1997) Independent Component Analysis by Minimization of Mutual Information. Report A46. Helsinki University of Technology.

[26] Karhunen, J., Hyvarinen, A, Vigario, R., Hurri, J. and Oja, E. (1997) Applications of Neural Blind Separation to Signal and Image Processing. *Proc. ICASSP '97* pp 131-134.

[27] Lewicki, M. and Olshausen, B. (1997) Inferring sparse, overcomplete image codes using an efficient coding framework. *Proc. NIPS*97.* Pages 815-821.

[28] Mallat, S. (1998) *A Wavelet Tour of Signal Processing.* Academic Press.

[29] Meyer, Y. (1992) Wavelets and Applications. Lecture at CIRM Luminy Meeting, Luminy, France, March 1992.

[30] Olshausen, B.A. and Field, D.J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381** 607-609.

[31] Ramsay, J.O. and Silverman, B.W. (1997) *Functional Data Analysis.* Springer: New York.

[32] Ruderman, D.L. (1993) The statistics of natural images. *Network*, **5**, 4, 517-548.

[33] Watson, A.B., Barlow, H.B., and Robson, J.G. (1983) What does the eye see best? *Nature*, **302** 419-422.

[34] Wickerhauser, M.V. (1993) *Adapted Wavelet Analysis: Theory and Algorithms.* A.K. Peters.

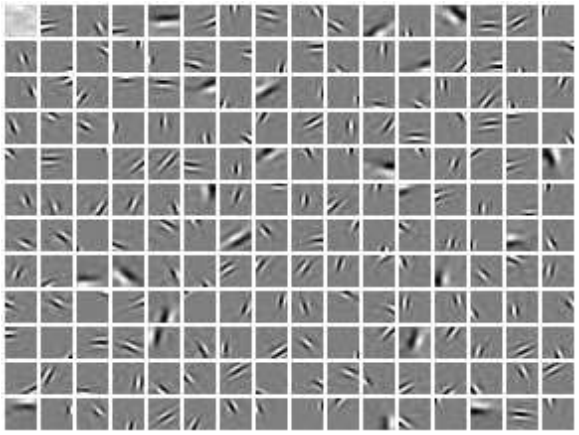

Figure 1: Basis for Image Patches. Result of Olshausen and Field.

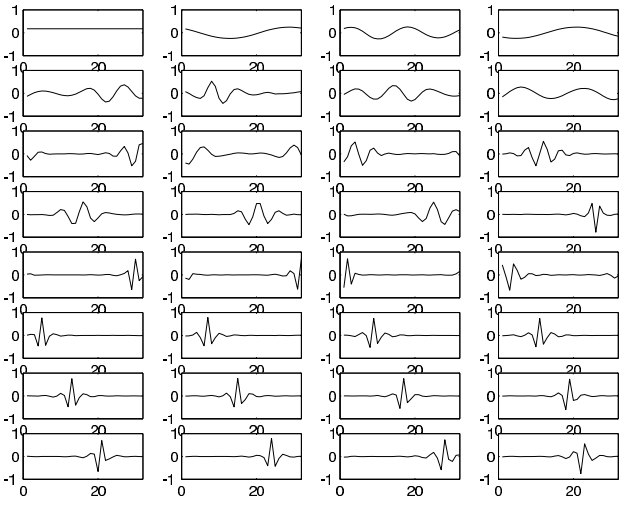

Figure 2: Orthonormal Basis Found by JADE

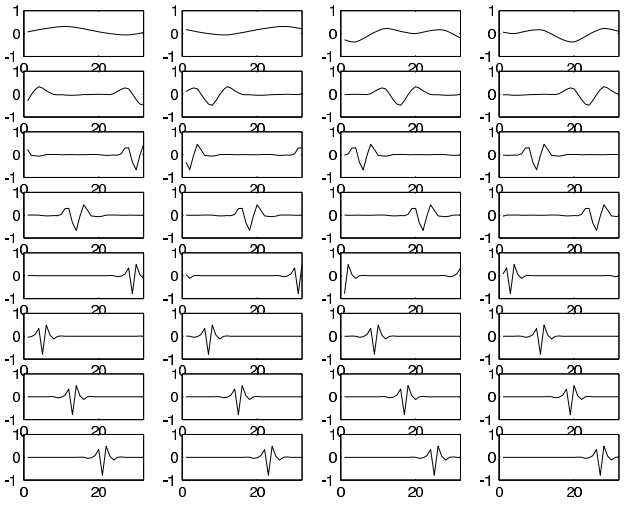

Figure 3: Daubechies Nearly Symmetric Wavelets (6)

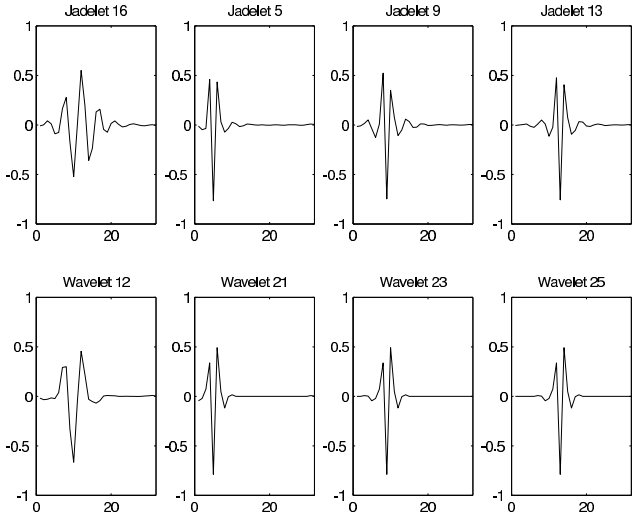

Figure 4: Comparison of a Few Basis Elements

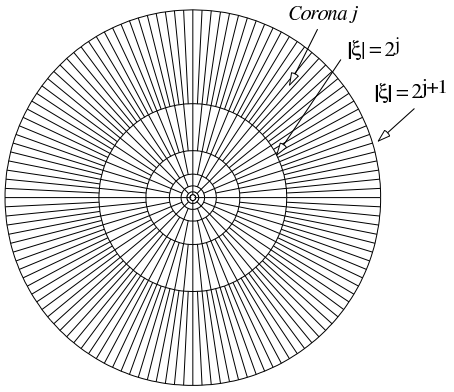

Figure 5: Sampling Scheme of the Ridgelet Transform

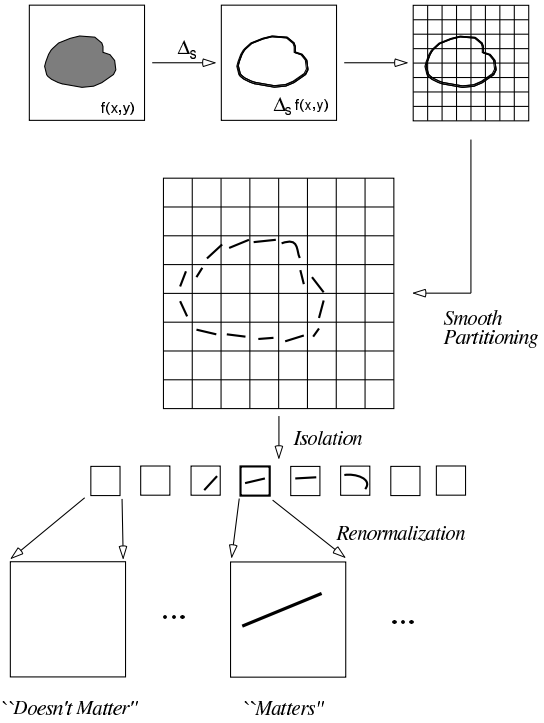

Figure 7: Curvelet Decomposition at a Single Scale

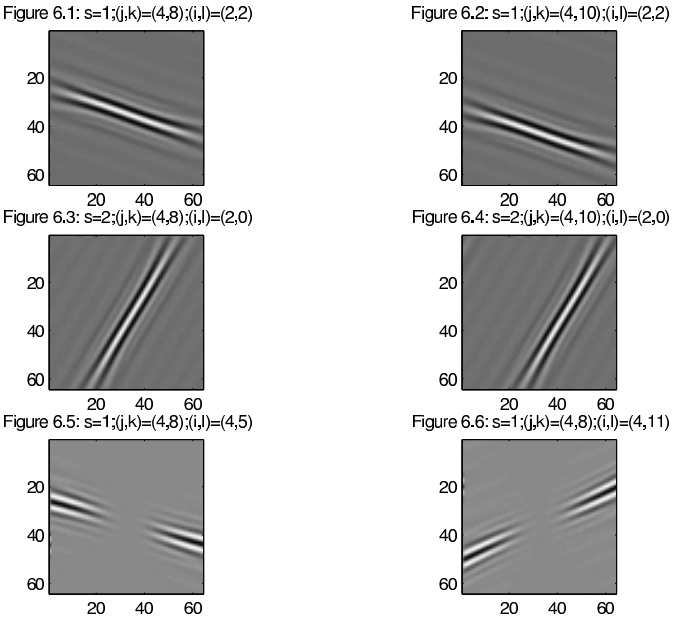

Figure 6: Several Ridgelets

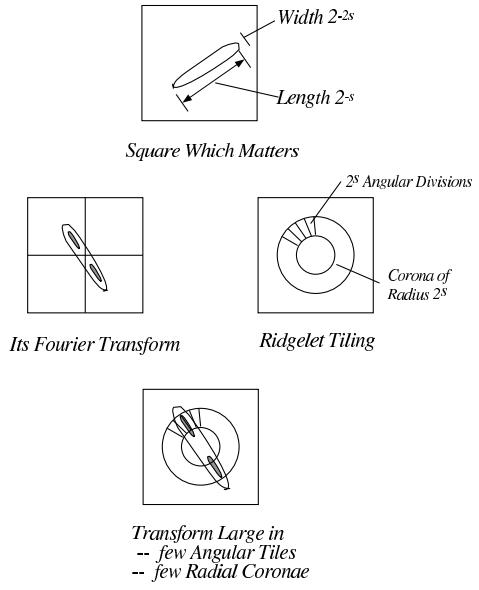

Figure 8: Ridgelet Analysis of a Ridge Fragment.