

Can recent innovations in harmonic analysis ‘explain’ key findings in natural image statistics?

David L Donoho and Ana Georgina Flesia

Statistics Department, Sequoia Hall, Stanford University, Stanford CA 94305-4065, USA

E-mail: donoho@stat.stanford.edu and flesia@stat.stanford.edu

Received 4 December 2000

Abstract

Recently, applied mathematicians have been pursuing the goal of sparse coding of certain mathematical models of images with edges. They have found by mathematical analysis that, instead of wavelets and Fourier methods, sparse coding leads towards new systems: *ridgelets* and *curvelets*. These new systems have elements distributed across a range of scales and locations, but also orientations. In fact they have highly direction-specific elements and exhibit increasing numbers of distinct directions as we go to successively finer scales.

Meanwhile, researchers in natural scene statistics (NSS) have been attempting to find sparse codes for natural images. The new systems they have found by computational optimization have elements distributed across a range of scales and locations, but also orientations. The new systems are certainly unlike wavelet and Gabor systems, on the one hand because of the multi-orientation and on the other hand because of the multi-scale nature.

There is a certain degree of visual resemblance between the findings in the two fields, which suggests the hypothesis that certain important findings in the NSS literature might possibly be explained by the slogan: edges are the dominant features in images, and curvelets are the right tool for representing edges.

We consider here certain empirical consequences of this hypothesis, looking at key findings of the NSS literature and conducting studies of curvelet and ridgelet transforms on synthetic and real images, to see if the results are consistent with predictions from this slogan.

Our first experiment measures the nonGaussianity of Fourier, wavelet, ridgelet and curvelet coefficients over a database of synthetic and photographic images. Empirically the curvelet coefficients exhibit noticeably higher kurtosis than wavelet, ridgelet, or Fourier coefficients. This is consistent with the hypothesis.

Our second experiment studies the inter-scale correlation of wavelet coefficient energies at the same location. We describe a simple experiment showing that presence of edges explains these correlations. We also develop a crude nonlinear ‘partial correlation’ by considering the correlation between wavelet parents and children after a few curvelet coefficients are removed.

When we kill the few biggest coefficients of the curvelet transform, much of the correlation between wavelet subbands disappears—consistent with the hypothesis.

We suggest implications for future discussions about NSS.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Mathematical concepts and mathematical language have, over the decades, helped shape discourse about research on the visual system and have even suggested research hypotheses.

- In the 1970s, there were claims that certain early stages of the human visual system acted by *Fourier analysis* [18];
- In the 1980s, there were claims that certain early stages of the human visual system acted by *Gabor analysis* [27, 28, 36], (and some claims of this kind are still made today); and
- More recently, there were claims that the human visual system acted, in early stages, by *wavelet analysis* [15, 16, 35] (we interpret the term *wavelets* in a broad sense, covering a range of multiscale decompositions from classic dyadic image pyramids to classic orthonormal wavelets [23, 25]).

However, it appears that the accuracy of such claims has been limited and the temporal influence of each specific mathematical idea has been short-lived. In our interpretation, much of the evidence presented at the recent Banbury Centre conference on Natural Stimulus Statistics [NSS 2000] points *away* from Fourier, Gabor, or wavelet models for simple cells in V1 (we will explain our interpretation further below).

We claim, moreover, that such short-livedness of mathematical influence ought to be expected. There is no *a priori* reason why *pre-existing* mathematical concepts, which were largely derived from solving problems in mathematical physics and related disciplines, should be accurate models or even helpful organizational principles in describing the human visual system.

If we are going to speak about productive interactions between the fields of mathematics and vision, it seems to us *a priori* far more plausible that *empirical data* about vision could shape *future mathematical research*, as mathematicians learn from empirical data about patterns which lead them to architect new organizational principles in mathematical analysis.

In short, we suggest that the most likely direction of *lasting* influence is *from* facts about the world *towards* development of new mathematical objects.

In fact one example of this direction of influence is well known. Work in the late 1970s and early 1980s in the field of vision on both scale space and on pyramid algorithms had a great deal of influence on mathematicians who developed wavelet methods in the mid 1980s; see [24, 25].

In further support of this direction of causality, we discuss in this paper two new systems of mathematical analysis—*ridgelets* and *curvelets*—which were developed *after* some of the key recent findings of the natural scene statistics (NSS) literature were known, and whose construction was definitely influenced by the NSS findings. It seems to us that these new systems may be better as models of early vision than classical models.

The essence of the new systems is that they were built to give optimally sparse codes for objects with edges; in the case of the ridgelets, straight edges, and in the case of the curvelets, curved edges. Within a mathematically precise setting, one can prove that they do a better job at sparse coding of edges than wavelets, Gabor, or Fourier methods [8].

Also, there is a visual resemblance between systems found in recent work in NSS, which seeks sparse codings of natural images, and the new mathematical systems. Both types of systems exhibit highly anisotropic elements, with increasing anisotropy at fine scales.

This coincidence is easily explained as follows:

- In both the mathematical setting and in the empirical setting, the effort is focused on developing sparse codes for images.
- In the mathematical setting, edges are the only features of interest in the synthetic imagery.
- In the empirical setting, edges are dominant features in real images.

Therefore, it would seem quite plausible that the systems, more than bearing a resemblance, might actually be in some sense the same. This brings to mind the possibility that the slogan

Natural images are dominated by edges ... and curvelets provide the sparsest code for images with *only* edges;

has considerable organizational power. If this slogan has value, it should be able to shed interesting light on the behaviour of other prominent phenomena in NSS. In this paper we focus on two of these.

- First, the ubiquity of nonGaussianity for distributions of coefficients of transforms which are local and have zero mean. We find that curvelet coefficients exhibit markedly greater nonGaussianity (as measured by kurtosis) on certain kinds of synthetic and real imagery.
- Second, the ubiquity of interscale correlations between coefficient amplitudes at adjacent scales at the same location. We find that the correlations are explainable as a result of a simple causal factor—the presence of edges. We also find that much of the correlations disappear when we perform a kind of nonlinear partial correlation transform, correlating residuals at adjacent scales after removing the fit of the few largest curvelets.

We believe these findings lend substantial credibility to our hypothesis.

Some of our statements are undoubtedly too bold, but they definitely force the discussion in the right direction. Certainly ridgelets and curvelets alone cannot provide the ‘true’ underpinnings of biological vision. Nevertheless, it would be very healthy for researchers in natural scene statistics to take these systems of analysis into account—both in the narrow sense of assimilating the findings that what we report here and in the broader sense of understanding that these findings suggest that there is an extensive territory lying beyond Fourier, Gabor, and wavelets, and that new systems of harmonic analysis may play a key role in interpreting natural image statistics.

2. Evidence from natural scene statistics

We begin by listing some findings of the natural image statistics community; these have been widely reported in the literature and were prominently in evidence at the NSS 2000 meeting.

The first set of findings is elementary in nature; each of the findings in the following list can be observed using numerous approaches.

- *Ubiquity of edges.* Edges are *the* dominant ‘phenomenon’ in images—not only in human perceptual experience (segmentation into objects), but also in raw statistical importance. Mumford and collaborators have shown that 1/2 or more of all ‘high-contrast’ 3-by-3 image patches in natural images can be interpreted as edge fragments.

- *Ubiquity of nonGaussianity.* A generic zero-mean short-range filter, applied to a generic natural image, gives an output filtered image that has a nonGaussian histogram. At NSS 2000, Mumford gave examples of this; at NSS 1997, Zetsche made similar claims. Observations of this phenomenon go back a long way, for example see Field (1987) [15] and Ruderman (1993) [30].
- *Ubiquity of interscale correlations.* Simoncelli and collaborators [3, 29], Baraniuk and collaborators [10], and others have shown that if one considers wavelet coefficients at two consecutive scales, then the presence of significant energy at a given location in the coarser of the two scales is predictive of the presence of significant energy at the same location at the next finer scale.

These phenomena might be considered as logically independent, in the sense that one might conceivably construct synthetic images which have some proper subset of these properties, rather than all three. However, it seems to us that on generic images they are all *manifestations of one single phenomenon*—namely the ubiquity of edges in images. That is, it seems to us simple and obvious to us that *nonGaussianity can be explained by edges alone*, and that *interscale correlations can be explained by edges alone*. Certainly, it is easy to derive mathematically that the presence of edges in a simple mathematical model of cartoon ‘images’ induces nonGaussianity and interscale correlations; moreover, despite extensive reading in the literature of natural scene statistics, we know of no other credible alternative explanation for either phenomenon.

The second set of findings is decidedly non-elementary; it concerns the existence of hidden components underlying natural image data. Such hidden components models seek a representation of image data according to an optimization principle: they seek a basis for the space of images which makes the coefficient representation of the image the most preferred in some sense. Such hidden components have now been widely reported and can be found by several known optimization principles:

- *Sparse coding.* Olshausen and Field [26] proposed that one can seek a basis in which the coefficients lead to the sparsest possible representation of the underlying image data.
- *Independent components.* Bell and Sejnowski [1] proposed that one can seek a basis in which the coefficients lead to the most independent possible representation of the underlying image data.
- *Exposing nonGaussianity.* One may seek a basis in which the coefficients lead to the most nonGaussian possible representation of the underlying image data [2]. The results of van Hateren and van der Schaaf [34] can be interpreted as of this kind.

The key finding, now replicated many times using many different databases of natural images and many different principles of decomposition [1, 17, 19, 20, 22], was obtained by Olshausen and Field [26]. They showed that sparse coding of natural image data led to a basis—in what follows, the *OF basis*—which was multiscale, multiorientation, with many highly anisotropic elements.

In view of our earlier remarks about edges, it is quite interesting that the OF basis seems also to be a collection of ‘edge detectors’, i.e. of elongated features which respond well to edges with specific orientations and locations. With our introduction in view, it is also quite interesting that these detectors are quite unlike the edge detectors commonly in use by the vision community, which lack this degree of direction specificity: for example Canny edge detectors and even wavelet-based edge detectors have receptive fields that are basically isotropic.

As far as comparisons with classical mathematical tools, the receptive fields of the OF basis:

- do not resemble Fourier basis functions, because they are localized and multiscale rather than global monoscale.
- do not resemble wavelet basis functions, because while they are localized and multiscale (like wavelets), they have many different orientations—increasingly many at finer scales—rather than a fixed number of orientations like wavelets or, for that matter steerable pyramids.
- do not resemble Gabor basis functions, because the Gabor system is monoscale, consisting of elements localized to various cells all at one single scale, whereas the OF basis is multiscale; and the Gabor basis is multi-oscillatory, containing elements that execute many cycles of oscillation within their support, whereas the receptive fields in the OF basis execute only a few oscillations within their support.

In short, *the OF basis, viewed as a whole, does not resemble any of the mathematically pre-existing bases listed in the first paragraph of the introduction as proposals for models of receptive fields.*

Our assertion will no doubt draw fire on several grounds; we are particularly thinking of the frequent oral assertions that ‘certain elements of the OF system look like wavelets’ or ‘certain elements of the OF system look like Gabor functions’. In our view such individual coincidences are not pertinent to the central question, which is the overall *architecture* of the analysing system.

Unfortunately, the empirical methods of determining hidden components are essentially limited to study of very small image patches—for example 16-by-16—because of the unfavourable scaling of storage demands and computational complexity with increasing image patch size. So it has not proved feasible as yet to investigate key architectural issues about the OF basis, for example to discern formulas giving the number of orientations available at each given scale, or the distribution of anisotropy. So current computation-constrained methods provide only tantalizing glimpses of what might emerge in the analysis of large-scale images.

(Finally, as a referee has pointed out, we should make clear that certain lesser-known pre-existing proposals for two-dimensional representation do exhibit increasing numbers of orientations with scale, although these are not in wide use. For example, in mathematical analysis, starting in the early 1970s with Fefferman’s work on the Riesz multiplier problem, and then intensively developed by Stein’s school [33, ch 8–9], subband decompositions have been used exhibit a scaling law for the number of orientations as a function of scale. In the late 1980s, working from the viewpoint of vision, Simoncelli and Adelson [31, 32] proposed a scheme of subband decomposition which allowed increasing numbers of oriented subbands at finer scales. Unfortunately, all these approaches are based on angular subdivision of the frequency domain, which, while intuitively appealing, fails to generate optimal sparsity. For the purpose of sparse decomposition in a mathematically precise model of images with edges, one can show that a pure frequency-domain subdivision scheme does not improve substantially on wavelets. It seems that a more subtle *joint space-frequency* subdivision scheme, described below, is necessary to get representations which are optimally sparse in a theoretical sense.)

3. Developments in computational harmonic analysis

3.1. Sparse coding of edges

In recent years, computational harmonic analysis has had some highly visible successes in ‘sparse coding of images’ in the broad sense. Transform coding of images has become widely used in the guise of JPEG, which performs a local Fourier analysis. This will soon yield to

improved image coding using wavelet analysis in JPEG-2000, soon to be deployed widely in web browsers and media players.

However, despite the worldly success of wavelets, it has been clear for some time that wavelets do not provide the be-all and end-all of image compression. One of the most important ways in which this is so has to do with their treatment of edges.

There is a simple theoretical model of two-dimensional ‘images’ which shows that wavelets do not provide the theoretically ‘sparsest representation’ of certain kinds of ‘images’. We put the key terms in quotes to emphasize that for aficionados of NSS, the model may seem rather remote.

Consider cartoon ‘images’ which consist of regions in which the image is smoothly varying, separated by edges which are themselves smooth, but across which the image makes a discontinuous jump. It is possible in such a model to make an information-theoretic analysis of the minimal number of bits required to represent such objects [12, 14] and to show that wavelet coding performs dramatically worse than the minimal the number of bits required to describe the image. Essentially the problem is that wavelets, being essentially supported in isotropic patches, are not able to become highly anisotropic and align with edges.

3.2. Sparse coding and curvelets

Recently, Candès and Donoho [7, 8] developed the curvelet transform, designed from the beginning to represent edges much more efficiently than traditional transforms such as wavelets. Here the definition of efficiency was based on sparse coding, i.e. on using very few coefficients for a given accuracy of reconstruction.

They in fact considered the model of cartoon ‘images’ which consist of regions in which the image is smoothly varying, separated by edges which are themselves smooth, but across which the image makes a discontinuous jump. They considered several methods of image representation (e.g. Fourier, local Fourier, wavelet and curvelet) for objects from this model, and studied the following criterion of sparse representation: the asymptotic behaviour of the m -term approximation error when the approximation is made by the m best-terms in the given basis. Basically, they found that for integrated squared error, wavelets obtained an m -term approximation error $O(1/m)$ as $m \rightarrow \infty$, Fourier was even worse at $O(1/\sqrt{m})$ as $m \rightarrow \infty$, and curvelets achieved $O(\log(m)^3/m^2)$. They also found information-theoretic bounds which no basis, frame, or even any reasonable dictionary could beat: $O(1/m^2)$. In short, curvelets were essentially the optimal sparse coding for cartoon ‘images’.

3.3. Properties of curvelets

In this short paper, it is not really possible to give a careful discussion of how the curvelet transform is constructed; several resources are given in the reference. Instead, we will give an impressionistic picture, which may perhaps pique the reader’s interest without causing a feeling of overwhelming infoglut.

Curvelets provide a multiscale collection of analysing elements with the following properties:

- *Frame property.* Curvelets provide an overcomplete system of analysing elements—a *frame*—and any image can be reconstructed from its curvelet coefficients. The reconstruction is stable under perturbations of the coefficients, in particular under thresholding. In fact, a Parseval-like relationship holds, giving an identity between the energy of a coefficient sequence and that of a reconstruction.

- *Multiscale*. The curvelet system, like the wavelet system, is multiscale, with basis elements occupying a wide range of dyadic lengths and locations.
- *Direction specificity*. The curvelet system consists of elements which are typically anisotropic, i.e. whose support has a very large length/width ratio.
- *Anisotropy scaling*. The curvelet system becomes increasingly anisotropic at fine scales; in fact the width scales as the square of the length, so that as curvelets become short they become to even a greater degree very narrow and very directionally specific.

If one were very sloppy about terminological matters one might say that curvelets look like ‘Gabor functions’ arranged in a multiscale pyramid, where the Gabors at a certain scale occupied a range of locations and orientations, and the envelopes of the Gabors were very anisotropic at fine scales. However, while this might convey some meaning to certain readers, we would warn against it, since in our view this is both stretching the notion of Gabor analysis to the point of meaningless generality and because certain specific features of traditional Gabor analysis—e.g. the existence of a whole range of frequencies associated with a given spatial cell, rather than just low frequencies, are very much not exhibited by curvelets.

In fact, the construction of curvelets is rather involved, much more involved than an analogy to ‘multiscale Gabor pyramids’ could ever suggest. At the heart is a kind of multiscale deployment of yet another system of analysis—ridgelets.

3.4. Ridgelets

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

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

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. The key point is the analysis by functions which are global in one direction and local in the other direction. For further developments, see [6]. The article [13] 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. Some of the ortho-ridgelets are localized near the centre of the image; others are fragments localized near the edge. The behaviour of these ridgelets is controlled by a five-parameter index set $i, j, k, \ell, \varepsilon$ which we do not have space to review here.

For a typical example, see figure 1. The reader who compares this figure with certain images presented in the NSS literature, for example in [26, 34], will notice a resemblance not just in the ridgelets localized near the centre of the image, but also in the edge fragments.

Underlying the ridgelet system is *ridgelet tiling principle*:

- Divide the frequency domain in dyadic coronae (rings); i.e. with ξ the two-dimensional frequency variable, the j th corona is the frequency band $\{|\xi| \in [2^j, 2^{j+1}]\}$.
- In the angular direction, sample the j th corona at least 2^j times.

This tiling is depicted in figure 2. For comparison, wavelet and steerable pyramid tilings have a constant number of angular subdivisions independent of radial corona.

It is interesting to compare the ridgelet tiling with frequency distribution of hidden components in the NSS literature. In [26] and [34], the reader will find graphs depicting the distribution of hidden components in frequency space, for example measuring the number of components in various radial frequency bands. Ridgelet tiling suggests a specific scaling law for this number; it would be interesting to test the distribution of hidden components against the ‘ridgelet hypothesis’.

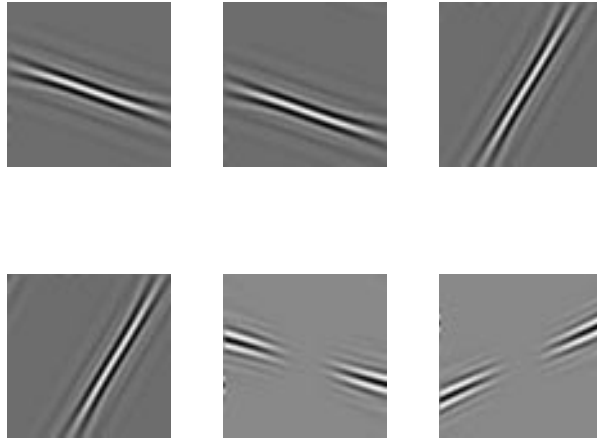


Figure 1. Several Ridgelets.

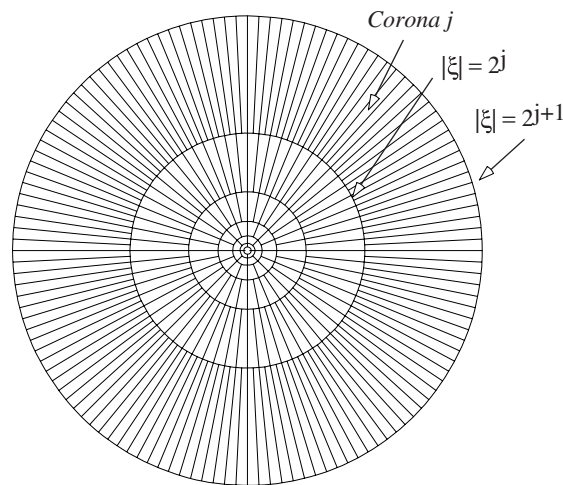


Figure 2. Tiling scheme of the Ridgelet transform—shown as subdivision of the frequency domain.

3.5. Curvelet analysis

Since the NSS community is very interested in whether a basis derived for sparse coding could be realized ‘neurally’, we briefly sketch the architecture we have used for the curvelet transform. It is based on combined space–frequency localization (compare figure 3).

The image is first separated into dyadic frequency subbands; in that figure, $s = 0, 1, 2$. These are like wavelet subbands, except that they are two, rather than one, octaves wide. See the second column of panels in the figure.

Each subband is then spatially partitioned into square subimages whose width is keyed to the subband scale, in such a way that when the subband filter has scale $n/2^{2s}$ pixels on an n by n image the size of the square is $n/2^s$ pixels. So at fine scales, the squares are comparatively broad (see the third column of panels in the figure).

To each square subimage we then apply the discrete ridgelet transform. The coefficients are displayed in the fourth column of panels.

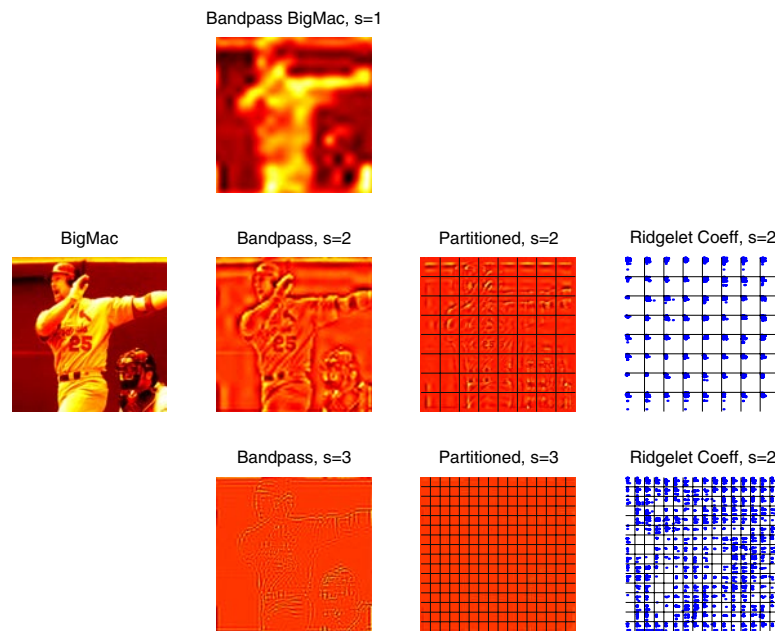


Figure 3. BigMac image and stages of curvelet transform.

The intuitive rationale for the organizing the curvelet transform in this way is indicated in figures 4 and 5. In the partitioning of subbands, many of the resulting subimages will be in regions where there is no edge present, and so the subimage will be essentially negligible. Others, where there is an edge, will have the appearance of ridges. Such subimages are non-negligible, but can be represented very efficiently by the ridgelet transform.

3.6. Implications of curvelets

We have now brought into play several pre-existing streams of evidence: we have pointed out that curvelets provide sparse codes for models of images with curved edges, ridgelets provide sparse codes for models of images with straight edges, and we have referred to evidence for ubiquity of edges in natural images.

We have also proposed that edges alone explain the nonGaussianity and interscale correlations observed in the NSS literature. We now turn to the question of whether curvelets and ridgelets can in some sense explain the nonGaussianity and interscale correlations.

4. Curvelets and nonGaussianity

We mentioned earlier the NSS finding that almost any zero-mean localized filter, applied to a natural image, gives a nonGaussian output. Thus, the wavelet coefficients of natural images have histograms which are highly nonGaussian, with sharp central peaks and heavy tails. It is likewise true that for a Gabor transform with sufficiently small spatial cells, the Gabor coefficients exhibit nonGaussianity, although less pronounced than in the wavelets case. Finally, a global Fourier analysis, with basis functions extending across the whole image, exhibits a relatively small degree of nonGaussianity (if at all), if we made histograms over reasonably defined subbands.

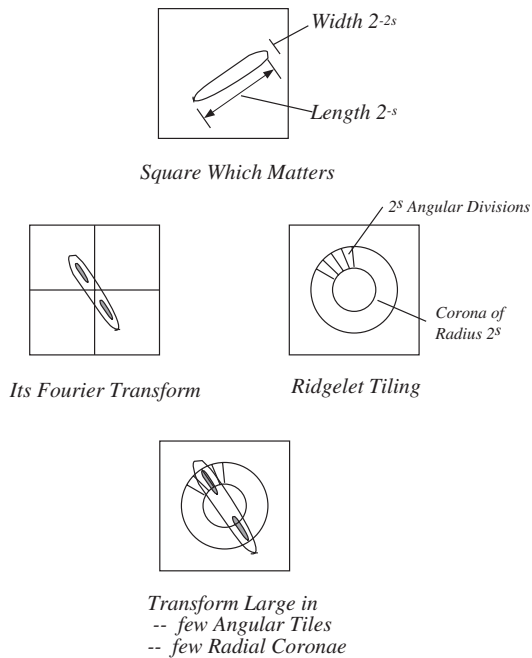


Figure 4. Ridgelet analysis of a ridge fragment.

4.1. The independent component analysis heuristic

At this point we introduce an important heuristic observation, which has been valuable in many fields of science and engineering [9, 11]. For certain nonGaussian probability distributions, there is a unique orthogonal basis whose coefficients are more nonGaussian than in any other orthogonal basis. These probability distributions are the *independent components models*, where the data are generated as a random superposition $X = \sum_j a_j \phi_j$ of waveforms ϕ_j with independent nonGaussian coefficients a_j . In this setting, the analysis of X into the independent components basis yields coefficients which are more nonGaussian than the analysis in any other basis. This basis, in addition to be the ‘correct’ basis for understanding X , is also the ‘most revealing’ of the existence of nonGaussianity. In every other basis, the nonGaussianity is less well exposed [2]. In this setting, searching for maximal nonGaussianity is equivalent to searching for the fundamentally correct representation.

The pattern we have described for natural scene statistics agrees with this picture: there is an ordering of bases by locality, and the nonGaussianity of the coefficients seems to be reflected in that. It seems superficially that the behaviour is consistent with an independent components model [17, 19], and that perhaps, from this limited information, one might speculate that we have ‘wavelet independent components’. David Field [16] has, more or less, proposed independent components models for images based on wavelets, and, as emphasized by Bell and Sejnowski [1], Olshausen and Field [26] have essentially derived their sparse coding objective function from an independent components model in which the waveforms are not assumed to be known wavelets, but instead, are objects to be found by optimization. Lewicki and Olshausen continued this line of work in [22]. (Of course, no one believes that the independent components model is literally true for natural scenes—images are formed by occlusion rather than superposition—only that it is a useful model at this stage of scientific study.)

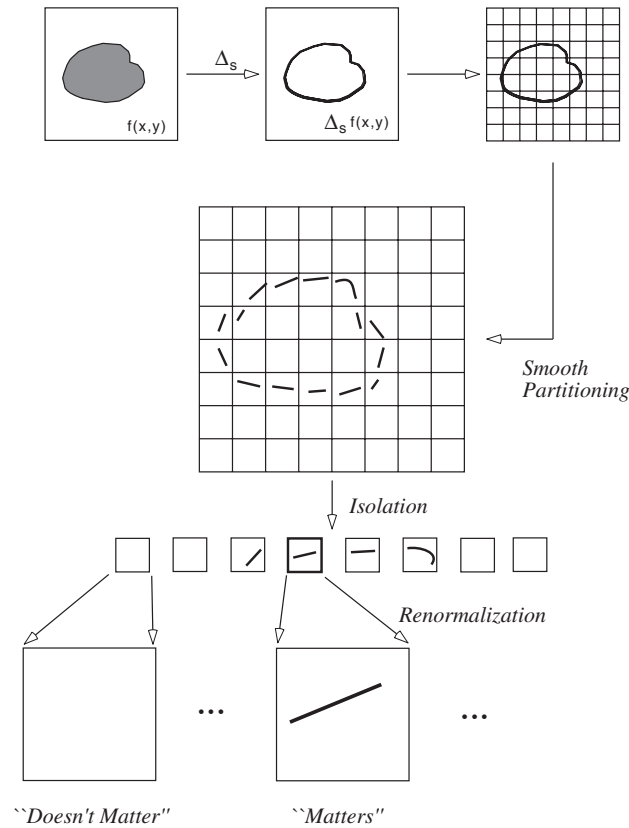


Figure 5. Decomposition at a single scale.

4.2. Measuring nonGaussianity

Motivated by the independent component analysis (ICA) viewpoint we now ask: what basis provides the most nonGaussian coefficient distributions? Under the ICA model, this ‘most revealing’ basis would be in some sense a candidate for the ‘correct’ basis for understanding the dataset.

More particularly, we could ask: how do curvelets and ridgelets fare in yielding nonGaussian coefficient distributions? Are they better than, or worse than wavelets?

We propose to approach such questions by using *kurtosis* (a sort of normalized fourth moment) as a measure of nonGaussianity. This can be justified from the independent components point of view. Suppose that we have a random process which is generated by independent components with a common distribution, and that the independent component vectors are orthogonal. Then the coefficients of that random process in a certain basis are themselves random variables, and have a theoretical kurtosis. The kurtosis would be zero if the process is Gaussian, and would be positive if the process is nonGaussian with heavy tails. We can measure the ‘amount of nonGaussianity exposed by a basis’ by the sum of those coefficient kurtosis. It turns out that this exposure of nonGaussianity is maximized, among all orthobases, by the underlying ICA basis. In short, under a model where ICAs exist, we can find them by looking for bases which maximize coefficient kurtosis.

Our interest in kurtosis can also be justified from a sparse coding viewpoint. 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 coefficients of a typical object are naturally grouped in subbands, and that, in each subband, the individual coefficient amplitudes either vanish or else approximately take a certain common nonzero value v_s . Both wavelets and curvelets have this property when applied to simple edge-dominated imagery.

Assume that the transform coefficients are equally likely to be positive and negative. Let ε_s be the fraction of nonzero items in the subband.

$$\varepsilon_s = \frac{\# \text{ signif. coeff.}}{\# \text{coeff.}}.$$

Define the normalized empirical kurtosis at subband s by

$$K_{4,s} = \text{Ave}[(X_i - \text{Ave}(X_i))^4] / \text{Ave}[(X_i - \text{Ave}(X_i))^2]^2 - 3,$$

where Ave denotes average across subscript i . 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.

Appealing to our sparse code hypothesis, we conclude that when two different bases provide different representations of essentially the same subband, the one achieving the sparser representation of the subband should exhibit coefficients with larger kurtosis at that subband.

4.3. Kurtosis in a theoretical model

We now apply the heuristic calculation of kurtosis to the model of cartoon ‘images’ with edges along smooth curves, for four bases: wavelets, sinusoids, ridgelets and curvelets. In all these cases, the above heuristic is reasonable. If we consider subband $2s$ of a separable two-dimensional wavelet basis, we have order $O(4^{2s})$ coefficients, and no more than $O(2^{2s})$ non-zero coefficients in each orientation. The Fourier corona $2s$ has also a total of $O(4^{2s})$ coefficients, and at least $O(2^{2s})$ non-zero coefficients, and possibly many more if edges span a wide range of directions.

For curvelets chosen from a matching frequency range and the specific ridgelet subband with ridge and angular scale $j = i = \log(n) - s$, there are roughly $4^s \times (2 \times 4^j)$ coefficients, since there are 4^s spatial boxes in the bandpass subband and (2×4^j) ridgelet coefficients in the ridgelet subband $i = j$, $\varepsilon = 0$. The number of non-zero coefficients depends on the number of boxes ‘hit’ by the edges. Crudely estimating that only $O(1)$ substantial ridgelet coefficients will be needed per box, and that L is the length of the curve, the number of non-zero coefficients could be estimated by $2 \times 2^s \times L$.

The matching ridgelet subband (j, i) is the one with ridge scale j and angular scale i equal to $2s$, $\varepsilon = 0$. The number of non-zero coefficients in that subband is 2×4^{2s} , and for objects with curved edges, the number of non-zero coefficients is as large as the wavelet and Fourier basis, 2×2^{2s} . (If we fortuitously analyse an image with only perfect straight edges, we can expect a few coefficients per subband.)

These yield the prediction that in analysing smooth objects with discontinuities along C^2 curves:

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

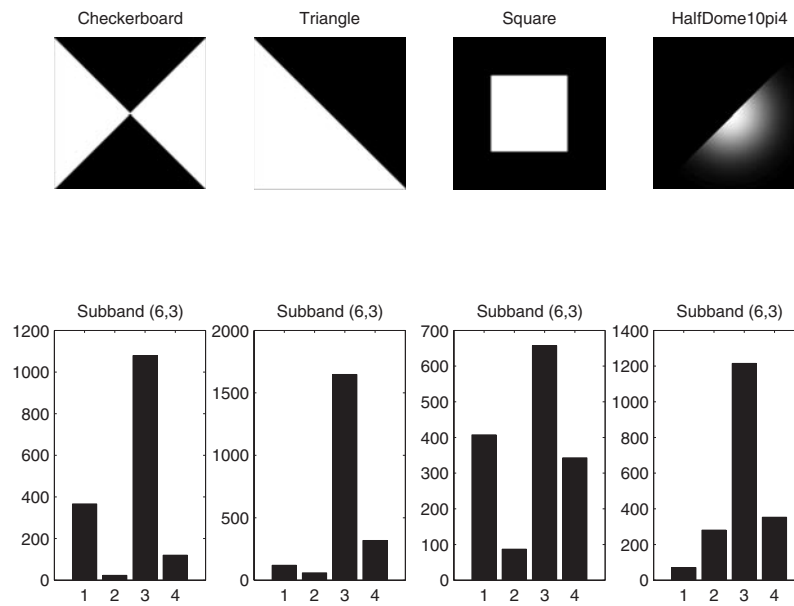


Figure 6. Kurtosis of Fourier (1), wavelet (2), ridgelet (3) and curvelet (4) coefficients; objects having only perfect straight edges, subband $2s = 6$, $j = i = 3$.

- Kurtosis in the wavelet basis scales with subband index $2s$ like 2^{2s} .
- Kurtosis in the curvelet basis scales with subband index s like $O(4^{3s/2})$.
- Kurtosis in the ridgelet basis scales with subband index $2s$ like 2^{2s} , (it could fortuitously scale as 4^{2s} on a model with perfect straight lines).

We conclude that on generic model images with smooth edges, kurtosis should grow substantially more rapidly at fine scales in the curvelet basis than it would grow at comparable scales in other bases.

4.4. Empirical kurtosis results

In our experience there is rough agreement between this heuristic calculation and what we see on synthetic and real images. We observe in figures 6–8 that this prediction is consistent with the direction of the empirical kurtosis differences, although the measured kurtoses do not vary as widely as our formulas suggest, the differences are marked and in the right direction.

On synthetic images with discontinuities along perfect straight lines (e.g. images like square, or halfdome, a discretization of a mutilated Gaussian), the kurtosis of the ridgelet coefficients is always the highest, indicating that ridgelets, not curvelets, provide the best basis for exposing nonGaussianity and sparsity in these cases. This makes sense, because it agrees with our sparse coding interpretation: we know that ridgelets are better than curvelets at representing discontinuities along perfectly straight lines.

On synthetic images containing only singularities along curved lines, the curvelet transform exposes twice the kurtosis of the other transforms, confirming in a general way our predictions.

Of course natural images do not consist merely of edges along straight lines, or even of edges along curves. Since curvature of edges has a wide range of variation, we can expect our predictions to be accurate for parts of many images, but not all of them.

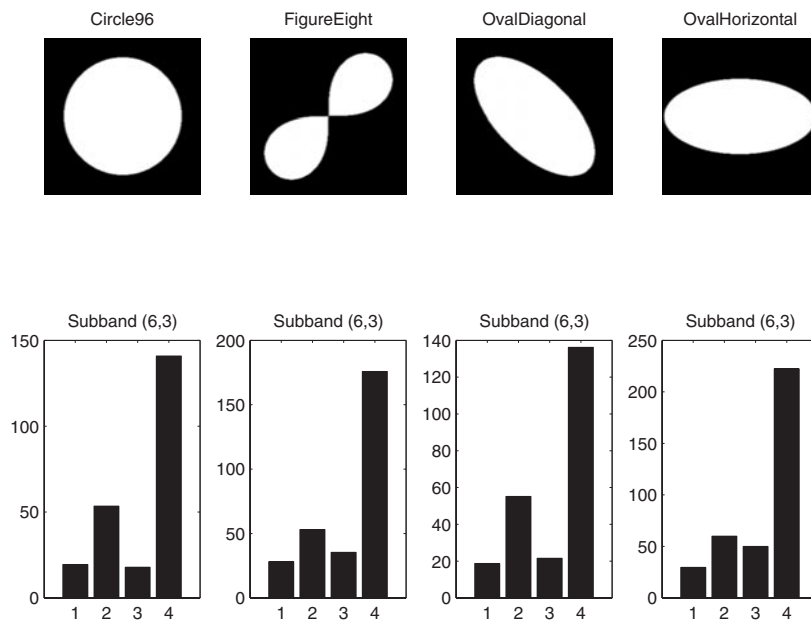


Figure 7. Kurtosis of Fourier (1), wavelet (2), ridgelet (3) and curvelet (4) coefficients; objects with curved edges, subband $2s = 6$, $j = i = 3$.

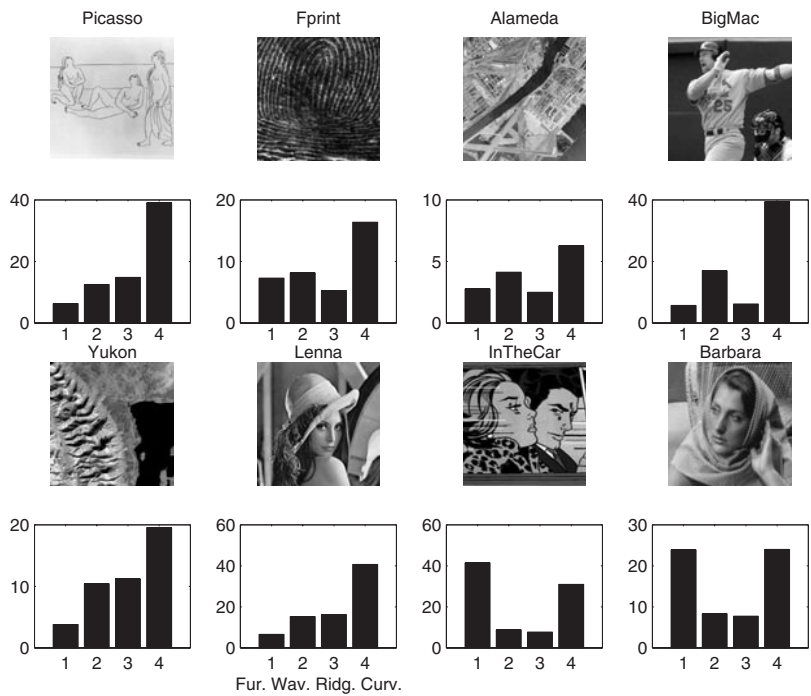


Figure 8. Kurtosis of Fourier, wavelet, ridgelet and curvelet coefficients of several non-synthetic images, subband $2s = 6$, $j = i = 3$.

An example is *InTheCar*, a cartoon, which has both straight lines, and curves. On this image the curvelet transform exposes more kurtosis than the ridgelet transform, but not more than the Fourier transform! This surprising performance of the Fourier transform could be explained by the presence of families of many parallel lines at roughly equal spacing. The Fourier transform is able to represent such features with only a few coefficients, while the curvelet and ridgelet transform need more coefficients for representing those details.

We are of course aware of the substantial objections which can be lodged against the use of kurtosis in analysis of empirical data. Primarily, these have to do with the fact that kurtosis of a distribution is highly sensitive to a small fraction of very extreme data. In our setting, we regard this as a virtue: we want highly kurtic output distributions. If, in an array of receptive fields, the overwhelming majority give no output and only one gives an output, then the output distribution will be highly kurtic, and we will be satisfied with that sort of distribution.

5. Curvelets and interscale correlation

Buccigrossi and Simoncelli [3] have developed a probability model for natural images, based on empirical observation of their statistics in the wavelet domain. We borrow the definitions from that paper, but we use only the simplest wavelets—Haar wavelets.

We use the rather obvious terminology that two wavelet coefficients at the same spatial location but adjacent scales are *parent* and *child*, with other family relationship terminology used in the obvious way.

The coefficients of wavelet subbands are approximately decorrelated, nevertheless it is clear from visual inspection that wavelet coefficients are not statistically independent. Large magnitude coefficients tend to occur at neighbouring spatial locations, and also at the same relative spatial location of subbands at adjacent scales and orientations.

Consider two coefficients representing information at adjacent scales but the same orientation (e.g. horizontal) and spatial location. Buccigrossi and Simoncelli displayed a conditional histogram of the child magnitudes conditioned on the parent magnitudes, and found that the presence of substantial amplitude at the parent is diagnostic for substantial amplitude at the child. Figure 9 shows that the pattern is surprisingly robust across a range of images.

In our study, we will define parent/child statistics as follows. To characterize parent energy, we sum the squares of each of the oriented parents at a given location and scale. To characterize the corresponding child energy, we sum the squares of the child coefficients across all orientations and all children at that location/scale. Thus, on an image of size $N = 2^j$, we will have a pair of $2^j \times 2^j$ matrices constructed on scale j . In figure 9 we display joint histograms of six different images of the database. Bright spots correspond to bins containing many pairs. Previous work of Simoncelli *et al* [3] and Huang and Mumford [21], using different definitions of parent/child statistics, published similar histograms, all displaying strong linear correlations.

5.1. Correlation versus causality

In statistical work it is common to worry about *causal explanation* of a correlation; one does not simply take an observed correlation at face value. Instead one asks if a certain correlation might be simply explained, not as expressing an intrinsic relation between the variables being correlated, but by a hidden or lurking variable that drives both of the measured variables and has not been taken into account. There are many classical examples of this theme; typically described in sections of statistics textbooks labelled ‘fallacies of correlation analysis’

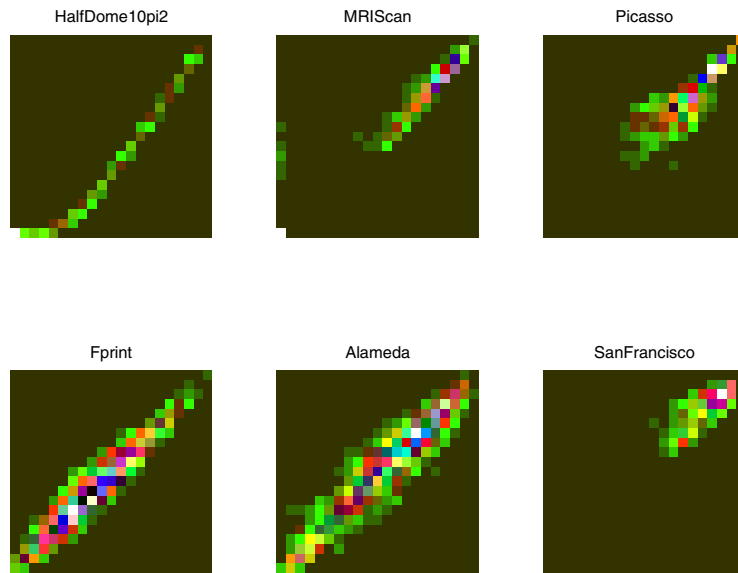


Figure 9. Joint frequency histograms of the log of parent and child energies in adjacent scales, on six different images, using 20^2 square bins.

or something similar. A typical example of meaningful correlation is the fact that country-by-country sales of cigarettes per capita in 1930 are correlated with country-by-country incidence of lung cancer per capita in 1960; the meaning is of course the underlying fact that smoking causes lung cancer. A typical example of spurious correlation is the fact that among US metropolitan urban areas, the per capita rate of PhDs is correlated with the per capita murder rate. There is of course no causal relationship: a third factor—urban density—is responsible for attracting both PhDs and criminals to urban areas. If we remain always at the superficial level of unexplained correlations, we can easily proclaim misleading relationships.

We believe there can be little doubt that the parent/child correlation phenomenon is caused by the presence of edges in images. To support this claim, we propose two types of evidence. First, we show that edges are indeed a causal factor for correlation. Second, we show that removing edges reduces parent/child correlation.

5.2. Adding edges causes correlation

We now consider an image where parent/child correlations are absent and then additively superpose edges, showing that parent/child correlations are created, and that the strength of the correlation varies in direct relation to the amplitude of the edges.

As our stereotyped ‘image’ where parent–child correlation is absent, we consider a Gaussian white noise. Indeed, since the orthonormal wavelet coefficients of a white noise are themselves a white noise, each distinct pair of wavelet coefficients of white noise will be stochastically independent—in particular parent/child pairs. Independence implies in particular uncorrelatedness of amplitudes or of squared amplitudes. Hence the parent/child correlation plot of a white noise must indeed show no correlation.

Now consider superposing on such a white noise a member of an edge database, i.e. consider a 16-by-16 image of the form $I = \alpha \cdot E + Z$ where Z is a white noise image, E is an image patch chosen at random from a database of 16-by-16 image patches containing

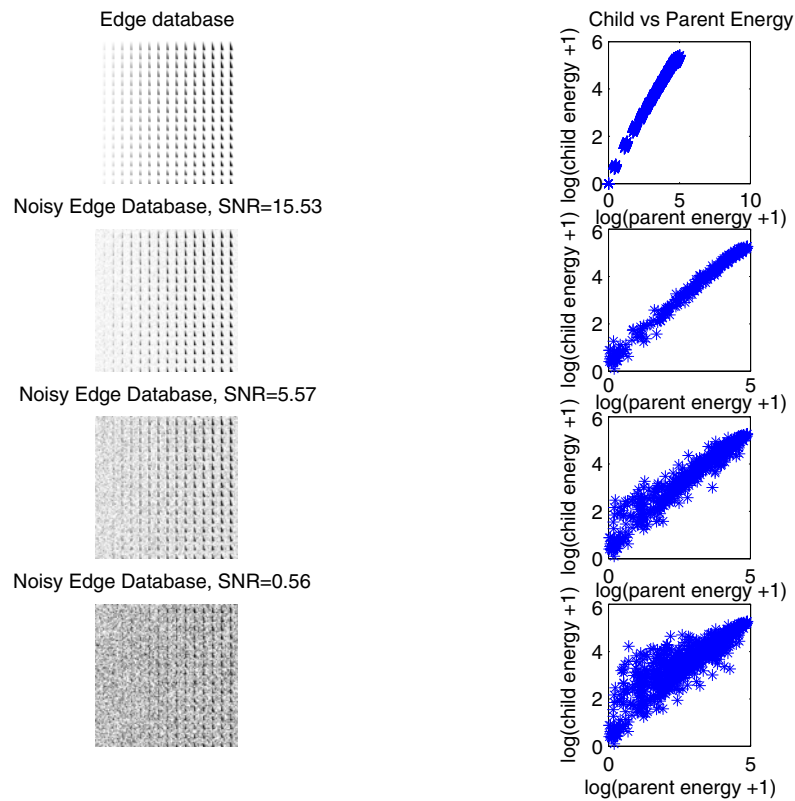


Figure 10. Image databases containing edges and noise and corresponding parent/child energies scatterplots at the finest two scales.

edges (examples below), and α is an amplitude (‘edge strength’) factor. Then depending on α we have something which appears to be a white noise (if α is small) or a slightly noisy edge (if α is larger). In figure 10, we see that indeed for small α there is a small correlation between parent and child, and that as the edge strength α increases, the correlation between parent energy and child energy increases.

5.3. Removing edges attenuates correlation

Our second method to illustrate that edges are indeed a causal factor for observed correlations is motivated by partial correlation. The notion of partial correlation was invented for the purpose of measuring the correlation between two variables *after* the joint influence of a third variable has been taken into account. The definition goes as follows:

- Let X and Y be the variables under study, and Z the third variable.
- Let \dot{X} be X adjusted for the value of Z , simply the residual $\dot{X} = X - E\{X|Z\}$ from predicting X using Z .
- Let \dot{Y} be Y adjusted for the value of Z , simply the residual $\dot{Y} = Y - E\{Y|Z\}$ of predicting Y using Z .
- The partial correlation between X and Y , given Z is then the ordinary correlation between \dot{X} and \dot{Y} .

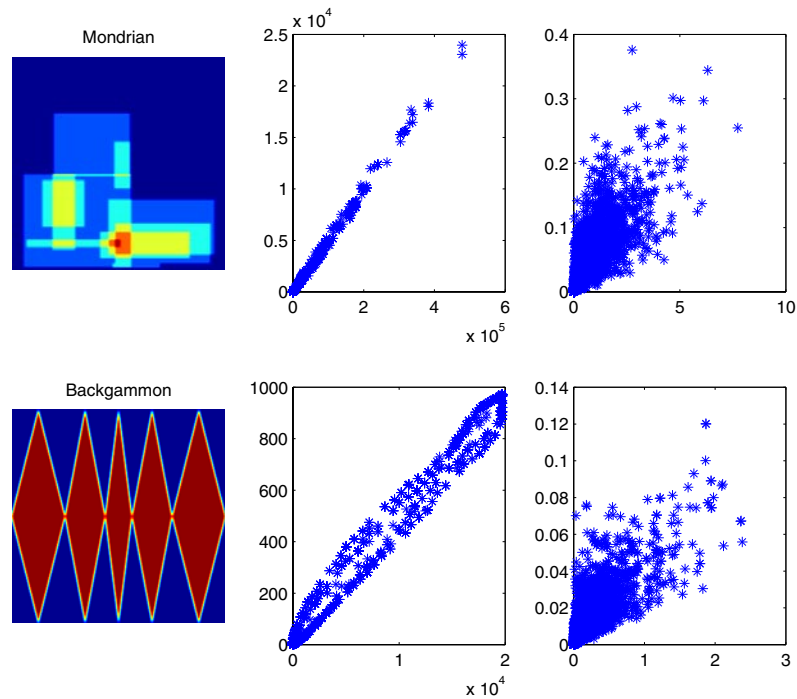


Figure 11. Scatterplots of wavelet coefficients of subbands 6 and 7 of two synthetic images. In each row, the left scatterplot shows raw child energy against parent energy for the original data, while the right scatterplot is for curvelet-adjusted subbands.

There are many easy examples where two correlated random variables turn out to have zero-partial correlation, conditional on a third variable. In fact, the subject of Markov processes in probability theory is essentially about defining special collections of random variables with such properties.

In essence, we would like to have a (*pseudo-*) *partial correlation* as a tool to answer: *what is the (pseudo-) partial correlation between X and Y , when the presence of edges has been taken into account?* If we could do this, we would be able to check the hypothesis that the appearance of correlations between parent/child amplitudes may be merely caused by the fact that edges are dominant features in images.

Unfortunately, there appears to be no appropriate pre-existing notion of partial correlation appropriate to our problem. We therefore propose our own method, based on the theoretical fact that, if there are edges in an image, they will be captured by the top few curvelet coefficients. Informally, we will define below, for an image I , the operator

$$\hat{E}\{I|\text{Curvelets}\}$$

as the reconstruction of the image using the top few curvelet coefficients. Based on an analogy with partial correlation, we proceed as follows:

- Let I_s and I_{s+1} be the subbands under study.
- Let $\hat{I}_s = I_s - \hat{E}\{I_s|\text{Curvelets}\}$.
- Let $\hat{I}_{s+1} = I_{s+1} - \hat{E}\{I_{s+1}|\text{Curvelets}\}$.
- Display an ordinary x - y plot of magnitudes of \hat{I}_s against \hat{I}_{s+1} .

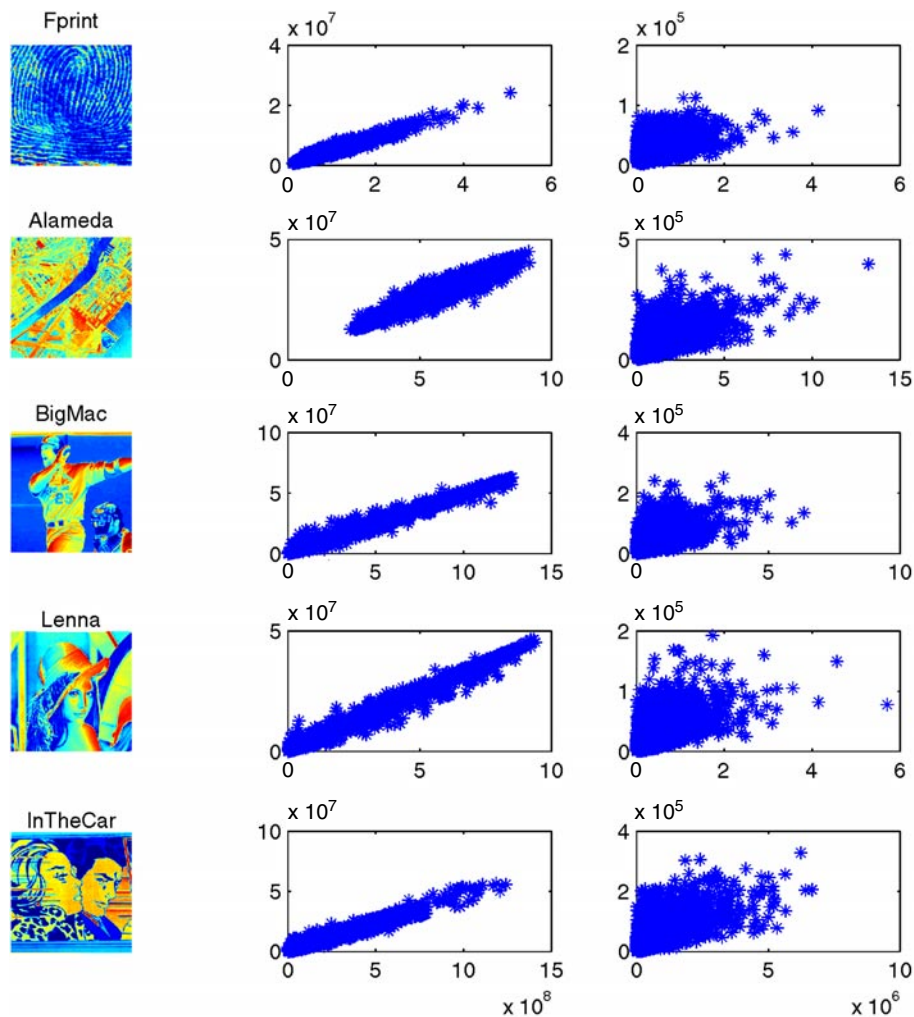


Figure 12. Scatterplots of wavelet coefficients of subbands 6 and 7 of five non-synthetic images. In each row, the left scatterplot shows raw child energy against parent energy for the original data, while the right histogram is for curvelet-adjusted subbands.

We note that this is a fair analogy to the case of partial correlation, in which we are predicting the image subband not from a fixed variable Z , but adaptively from the most important variables, in the sense of largest coefficients.

5.3.1. Heuristic theoretical calculation. We consider a theoretical model of images containing edges, and show heuristically that if, in defining the operator $\hat{E}\{I|\text{Curvelets}\}$, we use $O(2^{j/2})$ coefficients, the curvelet subband adjustment will essentially remove all the correlation between parent and child wavelet coefficients. In this model a random smooth image is augmented with an edge along a smooth curve. Write the image $I = S + D$, where I , S , and D are defined on the continuum, rather than on a discrete pixel grid; where S is a Gaussian random field which is twice differentiable, and D is the indicator function of a region R bounded by a twice-continuously differentiable edge. Hence I is smooth away from the boundary of

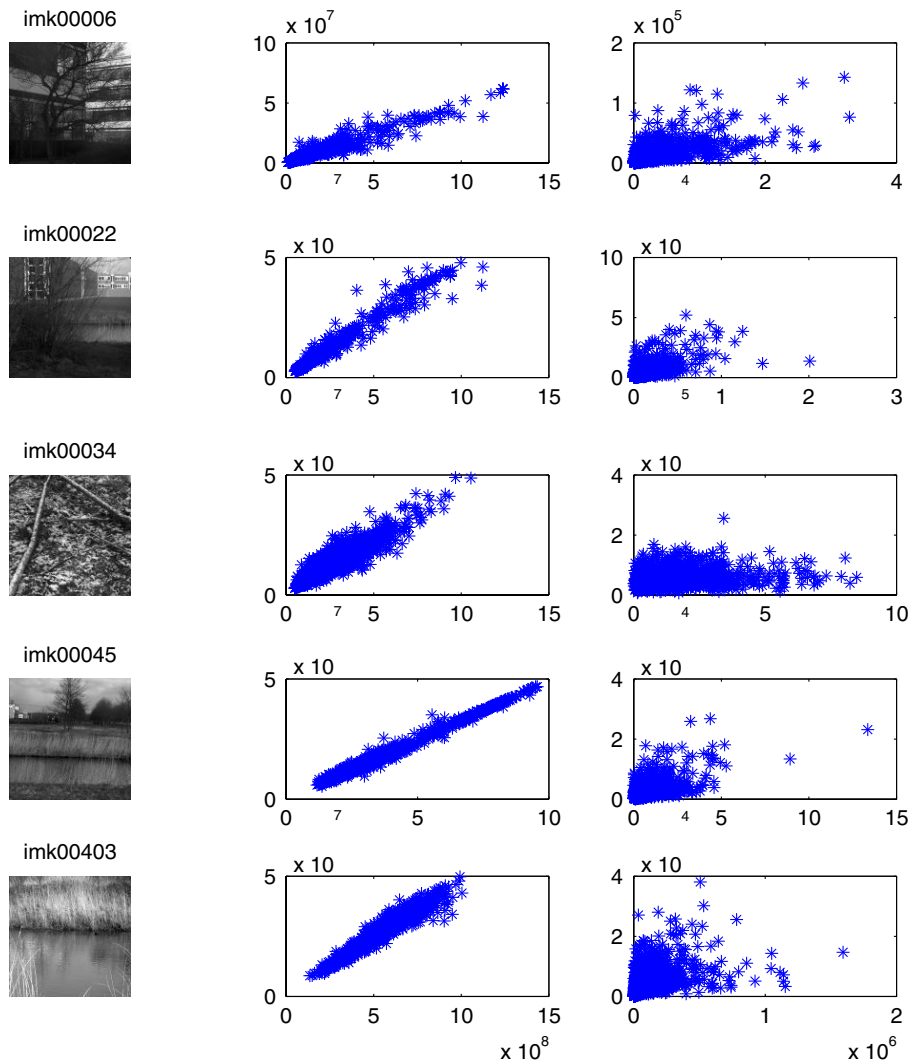


Figure 13. Scatterplots of wavelet coefficients of subbands 6 and 7 of five natural scenes. In each row, the left scatterplot shows raw child energy against parent energy for the original data, while the right histogram is for curvelet-adjusted subbands.

the region R . The smooth random field S is constructed by taking independent identically distributed Gaussian-wavelet coefficients at each wavelet subband and scaling the amplitudes according to a power law which generates the required twice differentiable behaviour.

Consider the wavelet analysis of such an object I . At the fine scale subbands with j large, the wavelet coefficients come in two groups: coefficients which do not ‘feel’ the edge, and coefficients which do ‘feel’ the edge (here ‘feel’ means that the support of the wavelet overlaps the edge). Coefficients that do not ‘feel’ the edge, instead measure properties of just the process S , and so parent/child pairs at such locations are independent Gaussians, each element of the pair is of size approx 2^{-2j} . We shall call such pairs ‘jointly small pairs’. Coefficients that do ‘feel’ the edge feel essentially measure properties only of the process D , and the coefficients

in such parent/child pairs are of size approx 2^{-j} . We call such pairs ‘jointly large pairs’. There are about $O(2^j)$ coefficients from ‘jointly large pairs’ out of 4^j parent–child energy pairs at at subband j . It follows that a display of the parent/child energies at subband j is made up overwhelmingly of jointly small pairs, with a small fraction of jointly large pairs.

Consider now the curvelet adjustment of such an object, thinking of this as acting wavelet subband by wavelet subband. The object I_j made up of wavelet components of I at scale j will display ridges which coincide with the edges in the image (compare the bottom two rows in the middle column of figure 3). When a wavelet subband is analysed by the curvelets frame, the image I_j is smoothly partitioned into squares of side approximately $2^{-j/2}$ by $2^{-j/2}$. The squares which overlap the edge will have the appearance of ridges. For curvelet analysis, the contents of each square are subjected to a ridgelet analysis. Operating heuristically, we assume that only $O(1)$ ridgelets are needed to capture the ridge associated with each square. Only $O(2^{j/2})$ squares intersect the edge curve. Consequently, if in the curvelet adjustment process, one removes $O(2^{j/2})$ curvelet terms in adjusting wavelet subband j , then one has effectively removed the presence of the ridge from that subband.

Suppose we now we look at the wavelet coefficients of the adjusted subbands. These will contain some residual of D , as well as the effects of the smooth random function S . Now the smooth random function S has wavelet coefficients which are pairwise independent, and so it generates parent/child pairs which are independent. Therefore, ignoring the residual of the D term, the parent/child wavelet pairs generated from the curvelet-adjusted subbands are independent.

Hence we have a heuristic calculation in a theoretical model which says that removing a small number $O(2^{j/2})$ of curvelet coefficients associated with scale subband j will result in essentially uncorrelated parent/child pairs at parent scale j . This heuristic calculation can be made rigorous by using ideas in Candès and Donoho [8]; complete rigour seems to require extracting logarithmically more terms than the $O(2^{j/2})$ terms suggested by the heuristic argument. Recall that there are $O(2^j)$ big wavelet coefficients at each scale subband, yet these are accounted for by many fewer— $O(2^{j/2})$ —curvelet coefficients.

5.3.2. Empirical results. We now ask: do the curvelet-adjusted subbands of real imagery display the same parent/child correlation structure as unadjusted subbands? In figures 11, 12 and 13 we have plotted scatterplots of parent/child energies, for both raw subbands and curvelet adjusted subbands. The asymptotic formula $O(2^{j/2})$ being too crude to provide detailed guidance, in this particular processing, we removed the 5% largest curvelet coefficients in adjusting each subband. It is visually quite evident that the degree of linear relationship is substantially reduced in the parent/child plots based on curvelet-adjusted subbands. This is true both for synthetic and real images.

6. Implications for discourse

Existing evidence shows that images are edge-dominated, and that curvelets and ridgelets offer sparse coding bases in a mathematically precise model of images with edges. Evidence presented in this paper shows that curvelets and ridgelets are more powerful for revealing nonGaussian structure than classical methods such as wavelets. Under the ICA hypothesis of section 4.1, this means they are more fundamentally connected to the structure of images. Evidence presented here shows that interscale correlations between parent and child wavelet coefficients are explainable by a simple underlying causal factor—namely the existence of edges; several types of evidence support this.

What are the implications for discourse in NSS? At least in the area of understanding empirical efforts to derive sparse coding bases, we see several implications.

- *Do not restrict attention just to the usual suspects.* Wavelets, Fourier, and Gabor are not the only schemes which might be useful in vision, or in natural scene analysis. In fact the evidence seems to point away from these schemes. Mathematicians are investigating completely new tiling schemes. Perhaps those are more appropriate.
- *Talk about tilings, not about individual elements.* A basis, or an organizational scheme for a visual cortex, is a whole system. Pointing out that individual basis elements ‘look like elongated Gabor functions’ is not really useful, because it draws attention away from learning about the overall organization, which is poorly modelled by the Gabor system. This sort of speech, describing a single basis element, perhaps accurately, runs the risk of suggesting that the Gabor organizational scheme is relevant, even when it is not.
- *Speak about tests for tilings.* Think of organizational principles and the way they may have empirically testable outcomes. Derive tests for those outcomes. Carry out and discuss the tests.
- *Talk about causal explanations for statistical phenomena.* Finally, when a certain statistical phenomenon is identified, speak about simple tangible features that could be causing it and about the extent to which those simple tangible features alone can explain the phenomenon in question.

Acknowledgments

This research was supported by National Science Foundation grants DMS 98–72890 (KDI); by AFOSR MURI-95-P49620-96-1-0028, and by DARPA BAA-98-004. DLD would like to thank the Mortimer and Raymond Sackler Institute of Advanced Studies at Tel Aviv University for hospitality during preparation of this paper, and AGF would like to thank the Statistics Department at UC Berkeley for its hospitality.

References

- [1] Bell A J and Sejnowski T J 1995 An information-maximization approach to blind separation and blind deconvolution *Neural Comput.* **7** 1129–59
- [2] Buckheit J and Donoho D 1996 Time frequency tilings which best expose the nonGaussianity of a stochastic process *IEEE Symp. Time-Freq. Time-Scale Anal.* (Piscataway, NJ: IEEE)
- [3] Buccigrossi R and Simoncelli E 1999 Image compression via joint statistical characterization in the wavelet domain *IEEE Trans. Image Process.* **8** 1688–701
- [4] Candès E 1999 Harmonic analysis of neural networks *Appl. Comput. Harmon. Anal.* **6** 197–218
- [5] Candès E 1998 Ridgelets: theory and applications *PhD Thesis* Department of Statistics, Stanford University
- [6] Candès E and Donoho D L 1999 Ridgelets: the key to high-dimensional intermittency? *Phil. Trans. R. Soc. A* **357** 2495–509
- [7] Candès E and Donoho D L 2000 Curvelets: a surprisingly effective nonadaptive representation of objects with edges *Curve and Surface Fitting: Saint-Malo 1999* ed A Cohen, C Rabut and L L Schumaker (Nashville, TN: Vanderbilt University Press)
- [8] Candès E J and Donoho D L 2000 Curvelets: optimally sparse representation of objects with edges *Technical Report* Department of Statistics, Stanford University
- [9] Comon P 1994 Independent Component Analysis, a new concept? *Signal Process.* **36** 287–314
- [10] Crouse M, Nowak R and Baraniuk R 1998 Wavelet-based statistical signal processing using hidden Markov models *IEEE Trans. Signal Process.* **46** 886–902
- [11] Donoho D L 1981 On minimum entropy deconvolution *Applied Time Series Analysis* vol 2, ed D F Findlay (New York: Academic) pp 565–608
- [12] Donoho D L 1999 Wedgelets: nearly minimax estimation of edges *Ann. Stat.* **27** 859–97
- [13] Donoho D L 2000 Orthonormal ridgelets and linear singularities *SIAM J. Math. Anal.* **31** 1062–99

- [14] Donoho D L 1998 Sparse components analysis and optimal atomic decomposition *Constructive Approximation* **17** 353–82
- [15] Field D J 1987 Relations between the statistics of natural images and the response properties of cortical cells *J. Opt. Soc. Am.* **4** 2379–94
- [16] 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* ed M Farge, J Hunt and J C Vassilicos (Oxford: Oxford University Press)
- [17] Fyfe C and Baddeley R 1995 Finding compact and sparse distributed representations of visual images *Network* **6** 333–44
- [18] Graham N 1981 The visual system does a crude Fourier analysis of patterns *Mathematical Psychology and Psychophysiology SIAM-AMS Proc. vol 13 (Philadelphia, 1980)* (Providence, RI: American Mathematical Society) pp 1–16
- [19] Hancock P, Baddeley R and Smith L 1992 The principal components of natural images *Network* **2** 61–70
- [20] Harpur G and Prager R W 1996 Development of low entropy coding in a recurrent network *Network* **7** 277–84
- [21] Huang J and Mumford D 1999 Statistics of natural images and models. Available at: <http://www.dam.brown.edu/people/mumford/Papers/cvpr99Huang.pdf>
- [22] Lewicki M and Olshausen B 1997 Inferring sparse, overcomplete image codes using an efficient coding *Framework Proc. NIPS*97* 815–21
- [23] Mallat S 1998 *A Wavelet Tour of Signal Processing* (New York: Academic)
- [24] Meyer Y 1993 Review of an introduction to wavelets and ten lectures on wavelets *Bull. Am. Math Soc* **28** 350–60
- [25] Meyer Y 1993 *Wavelets. Algorithms and Applications* (Philadelphia: Society for Industrial and Applied Mathematics)
- [26] Olshausen B and Field D 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images *Nature* **381** 607–9
- [27] Pollen D and Ronner S 1982 Phase relationship between adjacent simple cells in the visual cortex *Science* **212** 1409–11
- [28] Pollen D and Ronner S 1983 Visual cortical neurons as localized spatial frequency filters *IEEE Trans. Syst. Man Cybern.* **13** 907–16
- [29] Portilla J and Simoncelli E 1999 Texture modeling and synthesis using joint statistics of complex wavelet coefficients *Proc. IEEE Workshop Statist. Comput. Theories Vision (Fort Collins CO, June 1999)*
- [30] Ruderman D L 1993 The statistics of natural images *Network* **5** 517–48
- [31] Simoncelli E and Adelson E 1990 Non-separable extensions of quadrature mirror filters to multiple dimensions *Proc. IEEE* **78** 652–64
- [32] Simoncelli E and Adelson E 1991 Subband transforms *Subband Image Coding* ed John Woods (Norwell, MA: Kluwer) pp 143–92
- [33] Stein E 1993 *Harmonic Analysis: Real Variable Methods, Orthogonality, and Oscillatory Integrals* (Princeton: Princeton University Press)
- [34] 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. B* **265** 359–66
- [35] Watson A B The cortex transform: rapid computation of simulated neural images *Comput. Vis. Graph. Image Process.* **39** 311–27
- [36] Watson A B Barlow H B and Robson J G 1983 What does the eye see best? *Nature* **302** 419–22