Bayesian Comparison of Models for Images

Alex H. Barnett and David J.C. MacKay
Cavendish Laboratory
Cambridge, CB3 0HE. United Kingdom.

ABSTRACT. Probabilistic models for images are analysed quantitatively using Bayesian hypothesis comparison on a set of image data sets. One motivation for this study is to produce models which can be used as better *priors* in image reconstruction problems.

The types of model vary from the simplest, where spatial correlations in the image are irrelevant, to more complicated ones based on a radial power law for the standard deviations of the coefficients produced by Fourier or Wavelet Transforms. In our experiments the Fourier model is the most successful, as its *evidence* is conclusively the highest. This ties in with the statistical scaling self-similarity (fractal property) of many images. We discuss the invariances of the models, and make suggestions for further investigations.

1   Introduction

This paper's aim is to devise and search for 'good' statistical descriptions of *images*, which are greyscale pictures digitized from a camera, stored as an array of integers (representing the *intensities* of light falling on the camera's sensitive array). All of the images analysed in this paper have a fixed number of greyscale levels, $\lambda = 256$, corresponding to quantization to eight bits (one byte) of information per pixel. We assume that the image data is a linear function of physical intensity, and free of noise and blurring.

The statistical properties considered in the project are so general that what the images depict is largely unimportant, and we chose easily recognisable pictures (figure 1) such a face, natural objects, astronomical images, and the kind of images our eyes are subjected to frequently.

The development of the models is driven by intuitive ideas and by observations of real images, and is regulated by certain criteria for *invariance*, that is, operations on the image which should not affect its likelihood.

Bayesian analysis allows quantitative manipulation of data and prior beliefs to give a numerical result, the *evidence*, which reflects the probability of a hypothesis, and therefore how 'good' a model is.

Each model comprises a hypothesis H, with some free parameters denoted by the vector $\mathbf{w} = (\alpha, \beta, ...)$, which assigns a probability density $p(\mathbf{f}|\mathbf{w}, H)$, the likelihood, over the image space of $\mathbf{f}$, normalized so as to integrate to unity. The density's units are that of $[intensity]^{-n}$, since each pixel component f_i has units of $[intensity]$.

In most models the free parameters are initially unknown (*i.e.* they are assigned very wide prior distributions), and we search for their best fit value $\mathbf{w}_{BF}$, which has the largest likelihood given the image. Bayes' Theorem gives

$$p(\mathbf{w}|\mathbf{f}, H) = \frac{p(\mathbf{f}|\mathbf{w}, H)}{p(\mathbf{f}|H)} p(\mathbf{w}|H) \qquad (1)$$

1

'susie'   'mouse'   'redspot'   'trees'

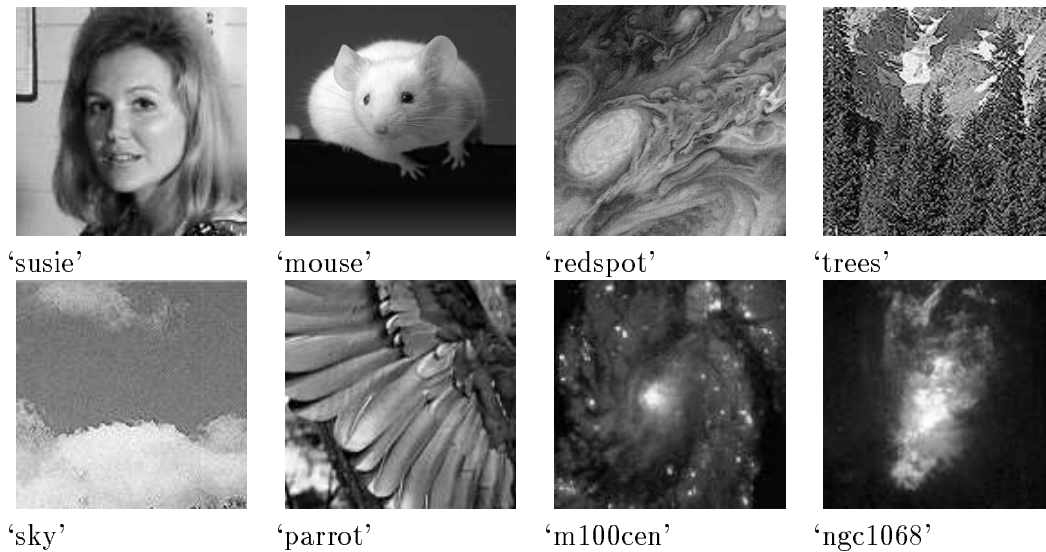'sky'   'parrot'   'm100cen'   'ngc1068'

Figure 1: The images analysed

This shows each image data set f_i as an intensity array of n_x by n_y pixels (n_x and n_y being either 128 or 256 in these images).

The denominator is independent of $\mathbf{w}$, the numerator is the *likelihood*, *i.e.* the probability of the observed image $\mathbf{f}$ as a function of $\mathbf{w}$, and the final term $p(\mathbf{w}|H)$ is the *prior* distribution on the the free parameters. This prior *has* to be assigned (based on our beliefs about images), even if seemingly arbitrarily, but has negligible effect on the $\mathbf{w}_{BF}$ found because the likelihood dominates. We know that $p(\mathbf{w}|\mathbf{f}, H)$ is normalized to 1, giving an expression for the denominator of (1), which we now call the *evidence for H*:

$$p(\mathbf{f}|H) = \int_{\mathbf{w}} p(\mathbf{f}|\mathbf{w}, H)p(\mathbf{w}|H)d\mathbf{w} \qquad (2)$$

This evidence is often dominated by the value of $p(\mathbf{f}|\mathbf{w}_{BF}, H)$ (the best fit likelihood). The evidence is equal to the best fit likelihood multiplied by a smaller factor known as the "Occam Factor". Applying Bayes' Theorem again gives us the probability of H (to within a constant factor) as $p(H|\mathbf{f}) \propto p(\mathbf{f}|H)p(H)$. The prior $p(H)$ can incorporate our beliefs about the validity of each hypothesis *before* the data arrived, but we chose all $p(H_i)$ as equal (in fact, usually any such prior would have to be *very* extreme to outweigh the evidence).

So, we now have the relative plausibilities of competing hypotheses, and in this paper we evaluate $p(\mathbf{f}|H_i)$ for given images and a variety of H_i.

2   Description of Models

We now detail the models (roughly in order of increasing complexity), the first four of which assume *independence* of the pixels f_i, meaning

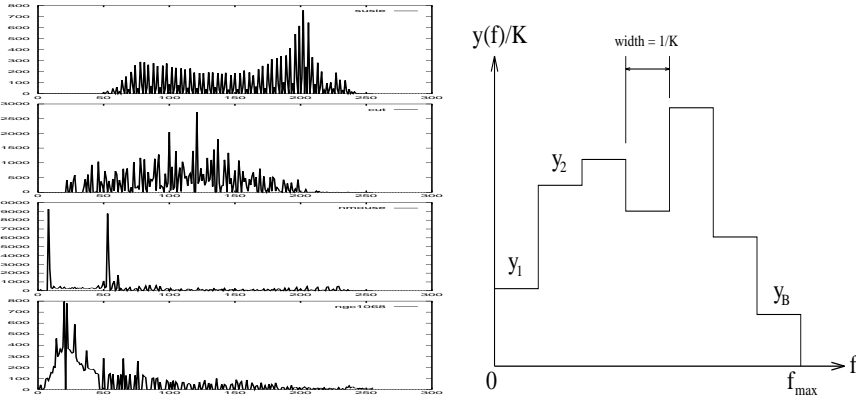$$p(f_0, f_1, ...f_{n-1}) = \prod_i p(f_i) . \qquad (3)$$

Figure 2: Some image intensity histograms, and a free-form distribution using B "bins"

Random Bits Model (RB)

The simplest distribution we can assign is uniform across all the allowed values of all n pixels, so $p(\mathbf{f}|H) = constant$ within the n-dimensional hypercube from 0 to f_{max} on each axis, and zero elsewhere. Our integer data f_i is assumed to have been truncated from a real continuous variable in the range 0 to f_{max} (f_{max} being λ units of intensity), so although our image vector $\mathbf{f}$ is always quantized onto an integer lattice, we will deal with continuous density functions. This model is called "random bits" because it corresponds to a prior of exactly $\frac{1}{2}$ for the probability of each binary bit being set in the stored image. Written in log form, the evidence is

$$\log p(\mathbf{f}|H_{RB}) = \sum_i \log (1/\lambda) = -n \log \lambda \ . \tag{4}$$

Free-Form Distribution Model (FF)

Figure 2 shows the frequency of occurrence of each intensity level in some images. It is clear that the distributions are far from flat, and are inconsistent from image to image, depending on properties of the camera and the digitization process. Therefore a model with a flexible, parametrized probability distribution function $y(f)$ over intensity f would be able to fit real images better than one with a uniform distribution. The figure also shows a simple $y(f)$ with a finite number (B) of variables, namely $\{y_b\} = \{y_1, y_2, ...y_B\}$, which give the probabilities of f falling into each "bin" of width 1/K. This probability is applied independently to each f_i of the image, so that

$$p(\mathbf{f}|\{y_b\}, H) = K^n \prod_b y_b^{N_b} \ , \tag{5}$$

where N_b is the number of pixels with intensity falling into bin b. This is substituted in (2) using $\{y_b\}$ as the parameter vector $\mathbf{w}$, and with a flat prior over all the normalized $\{y_b\}$ (but zero if not normalized, as used by Gregory and Loredo (1992)).

Approximating Gamma functions using logs eventually gives,

$$\log p(\mathbf{f}|H_{FF}) = n \log K + \sum_b (N_b + 1)\frac{\log(N_b + 1)}{n + B} - \left[\frac{1}{2}\sum_b \log(N_b + 1) - \frac{1}{2}\log\left(\frac{n+B}{B}\right)\right] \ .$$

The first two terms are the best fit likelihood, and the last term (in square brackets) is the log Occam factor. In order to disregard the statistical and digitization fluctuations in the histogram, but retain some flexibility, we have chosen $B = 16$ for this analysis (B must be between 1 and λ). A Bayesian choice of B might also be made.

Gaussian Distribution Model (GD)

This model applies the Gaussian probability distribution $N(\mu, \sigma^2)$ to each f_i, and is the first of a general class (that we'll call G) of Gaussian models which use the likelihood

$$p(\mathbf{f}|\mathbf{w}, H) = \frac{1}{Z} e^{-\frac{1}{2}(\mathbf{f}-\mathbf{a})^T \mathbf{C}(\mathbf{f}-\mathbf{a})} \ , \tag{6}$$

where $\mathbf{w}$ controls some properties of the square (order n) matrix $\mathbf{C}$ and the mean vector $\mathbf{a}$, and Z a normalizing constant. This gives for G models the evidence

$$\log p(\mathbf{f}|\mathbf{w}, H) = -\frac{1}{2}(\mathbf{f} - \mathbf{a})^T \mathbf{C}(\mathbf{f} - \mathbf{a}) + \frac{1}{2}\left[\log(\det \mathbf{C}) - n\log(2\pi)\right] \ . \tag{7}$$

In this model, GD, the values μ and σ are constant for all pixels (this is not only simplest, but desirable for invariance under spatial transformations), giving $\mathbf{C} = \mathbf{I}/\sigma^2$ and $\mathbf{a} = (\mu, \mu, ...\mu)$. The parameter $\mathbf{w}$ is (μ, σ), and solving $\nabla_{\mathbf{w}} \log p(\mathbf{f}|\mathbf{w}, H) = \mathbf{0}$ gives the best fit values:

$$\mu_{BF} = \frac{1}{n}\sum_i f_i \ , \qquad\qquad \sigma^2_{BF} = \frac{1}{n}\sum_i (f_i - \mu_{BF})^2 \ . \tag{8}$$

The approximation (usually a very good one, which we will use in all our G models — *e.g.* see Figure 4) that the peak about these best fit values is *Gaussian* makes equation 2 easy to evaluate. We assume

$$p(\mathbf{f}|\mathbf{w}, H) = p_{BF}\, e^{-\frac{1}{2}\mathbf{w}^T \mathbf{A}\mathbf{w}} \tag{9}$$

with $p_{BF} = p(\mathbf{f}|\mathbf{w}_{BF}, H)$, and $\mathbf{A}$ its Hessian matrix at $\mathbf{w}_{BF}$. Substituting (9) into (2) and assuming a constant prior $p(\mathbf{w}|H)_{BF}$ near the peak gives the general G model evidence

$$\log p(\mathbf{f}|H) = \log p_{BF} - \frac{1}{2}\log\left(\det\frac{\mathbf{A}}{2\pi}\right) + \log p(\mathbf{w}|H)_{BF} \tag{10}$$

We assigned a Gaussian prior on the *logarithm*[1] of each component, $\log w_i$, of standard deviation $\sigma_{\log w_i}$ about the best fit value $\log w_i^{BF}$. For this model, normalizing $p(\mathbf{w}|H)$ gave the prior at best fit

$$p(\mathbf{w}|H)_{BF} = (\ 2\pi\,\mu_{BF}\,\sigma_{\log\mu}\,\sigma_{BF}\,\sigma_{\log\sigma}\ )^{-1} \ . \tag{11}$$

Substituting the best fit values, the Hessian $\mathbf{A}$ and this prior into (10) enabled us to calculate this model's evidence $p(\mathbf{f}|H_{GD})$. Based on the largest and smallest conceivable μ and σ (given integer f_i from 0 to f_{max}), we set both the standard deviations $\sigma_{\log\mu}$ and $\sigma_{\log\sigma}$ to 4 (a value we used in all the G models). However, in our results the prior, and indeed the whole Occam Factor, is almost completely negligible compared to the relative likelihoods of different hypotheses, so we will not devote so much rigor to assigning priors in the coming models. (One would need to constrain $O(n)$ free parameters before this became significant.)

Although GD fits most images less well than the FF model, the above G class includes new, more powerful models (FP and WP), where pixels are no longer independent.

---

[1] This is appropriate since we initially have an uncertainty on $\mathbf{w}$ of orders of magnitude.
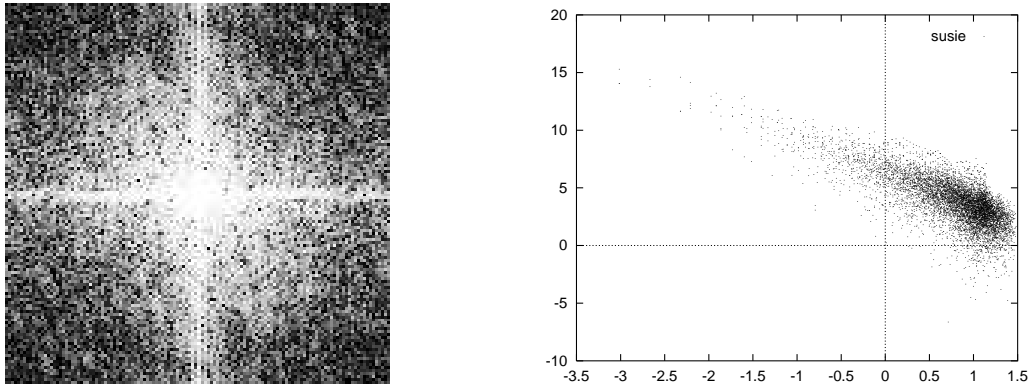
Figure 3: The 2D Fourier power spectrum $|F_i|^2$ of 'susie'

On the left it is displayed as a 2D histogram-equalized image (with $\mathbf{k} = \mathbf{0}$ central, and k_x and k_y in the range $[-\pi, +\pi]$), and on the right as a scatter plot of $\log |F(\mathbf{k})|^2$ against $\log |\mathbf{k}|$.

GAUSSIAN DISTRIBUTION OF $\log f_i$ MODEL (LGD)

In many classes of images, notably astronomical, there are a very large number of low-intensity pixels and fewer at higher intensities, and a Gaussian distribution on f_i is clearly inappropriate. However, if we define a new image $g_i = \log(f_i + s)$, where s is some constant offset (to keep g_i finite in the case of $f_i = 0$; we chose s as half an intensity unit), then $N(\mu_g, \sigma_g^2)$ in g_i-space corresponds to a suitably biased smooth distribution in f_i-space, which also has the desirable property of enforcing positivity of the intensities.

To transform probability densities we use $p(\mathbf{f}) = \det(\mathbf{J}) p(\mathbf{g})$, the determinant of J the Jacobean being $\det(\mathbf{J}) = \prod_i (f_i + s)^{-1} = e^{-n \mu_{gBF}}$, so we can use all the previous GD theory on g_i to assign $\log p(\mathbf{g}|H_{GD})$ then add on $\log(\det \mathbf{J})$, to get the log evidence $\log p(\mathbf{f}|H_{LGD})$.

FOURIER SPECTRUM RADIAL POWER LAW MODEL (FP)

So far none of the models have cared about spatial *correlations* in an image, which are after all usually what makes them recognisable. However, the 2D Discrete Fourier Transform (from f_i to the complex array $F(\mathbf{k})$, $\mathbf{k} = (k_x, k_y)$) allows us to construct a hypothesis with correlation (*i.e.* a non-diagonal $\mathbf{C}$), within the general scheme of the G class.

Visual examination of the 2D power spectrum of a typical image shows three main features (Figure 3):

1. seemingly uncorrelated random speckle on the scale of one pixel,

2. an approximately radially symmetric upward trend towards the point $\mathbf{k} = \mathbf{0}$, and

3. brighter lines on or near the vertical and horizontal axes (these are artifacts caused by the non-periodicity of the image, and were found to have little effect on the evidence when removed).

The observation of radial symmetry motivated a log-log plot of the spectrum as a function of radius in $\mathbf{k}$-space, which shows a clear linear downward trend of mean log power with log radius. This, together with the uncorrelated nature of the speckle, led to a hypothesis
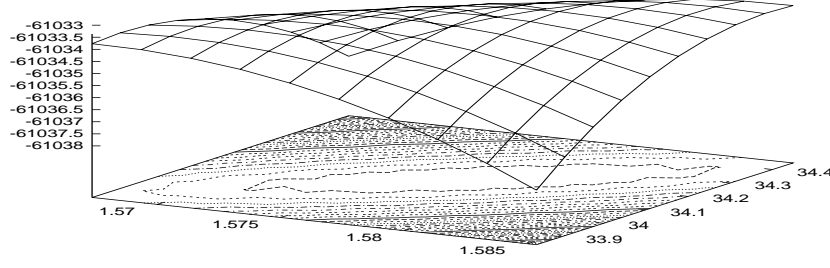
Figure 4: Peak of likelihood $\log p(\mathbf{F}|m, c, H)$ about best fit values for image 'susie'
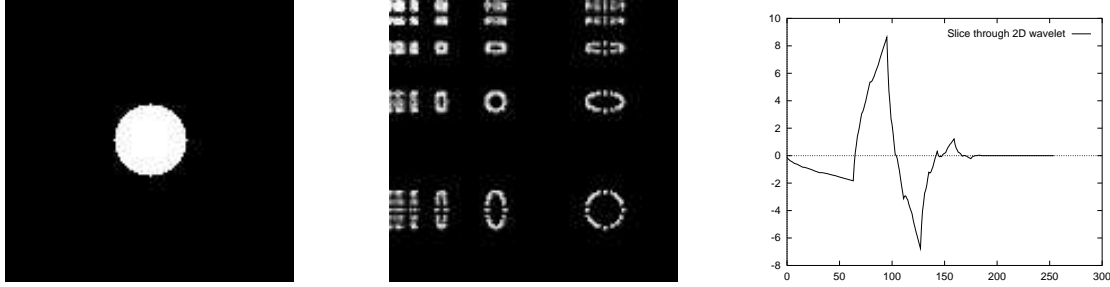

Figure 5: An artificial image, its 2D WT image, and a slice through a wavelet

Notice the structure of the WT: it is divided into rectangular regions of every possible binary (2^i by 2^j) size, each of which contains a map of the original image.

that the Fourier coefficients have real and imaginary parts which are both independently distributed like $N(0, \sqrt{2}\,\sigma(\mathbf{k}) = ck^{-m})$ where c and m are the *power law* constants and $k = |\mathbf{k}|$. We assigned $\sigma(\mathbf{0}) = \sqrt{n}f_{max}$ to avoid an infinity. This Gaussian distribution for the coefficients F was found to be very well justified when we histogrammed $Re[F]$ and $Im[F]$ for real images. Expressed as a density in $\mathbf{F}$-space, equation (7) becomes

$$\log p(\mathbf{F}|m, c, H) = -\sum_i \frac{|F_i|^2}{2\sigma_i^2} - \sum_i \log \sigma_i - \frac{n}{2}\log(2\pi) , \qquad (12)$$

and the orthogonality of the FT ($\det(\mathbf{J}) = 1$) means this is equal to $\log p(\mathbf{f}|m, c, H)$, and from this the evidence $p(\mathbf{f}|H_{FP})$ was found in a similar way to the GD model.

There was no simple analytic solution for m_{BF} and c_{BF}, so a Newton-Rapheson iterative approach was used to find $\mathbf{w}_{BF}$ and the Hessian $\mathbf{A}$ in the 2D space $\mathbf{w} = (m, c)$. Figure 4 confirms that $\log p(\mathbf{F}|m, c, H)$ has a Gaussian peak about $\mathbf{w}_{BF}$.

Wavelet Transform Power Law Model (WP)

The Wavelet Transform (WT) is linear, orthogonal, and operates on a real vector $\mathbf{f}$ (of n components, an integer power of two), converting it to another real vector $\mathbf{F}$ of n components. For a good, practical introduction see (Press *et al.* 1993:section 13.10), or (Strang

Figure 6: Typical sample images from models' **f**-space distributions

On the left is a sample from any of the RB, FF, GD or LGD models, which is spatially uncorrelated. On the right is a sample from the FP model, with $m = 1.5$ and $c = 10$, showing structure at all length scales (in fact it is a fractal).

1989) and (Heil and Walnut 1989) for more background. Wavelets have the property of being localized in both real and frequency space, so can efficiently represent both discontinuities and periodic features (see figure 5). They have many applications in lossy image-compression techniques, because they often reduce images to a few large coefficients and many small ones[2]. In this final model (also in the G class) we used the 4-coefficient Daubechies WT to replace the FT from the previous model, assigning Gaussian distributions to the WT coefficients F_i but with σ_i uniform within each of the 'binary regions' evident in figure 5. For each region an approximate k_i was used, based on the minimum wavelet dimension (in x or y), and the power law $\sigma_i = c k_i^{-m}$ was used as before. Skipping over details, this allowed equation (12) to be used to compute the evidence $p(\mathbf{f}|H_{WP})$ in an identical way to the FP model.

3   Results

Table 1 presents some of the results for different images, with e_i short for the log evidence of model i. For easy comparison, the uncorrelated models FF, GD and LGD are shown as ratios to the standard RB model (so that a number greater than 1 implies more evidence than for RB). Similarly, the ratios R_F and R_W are defined as $\frac{e_{GD}}{e_{FP}}$ and $\frac{e_{GD}}{e_{WP}}$ respectively, since the FP and WP models are closest in form to GD (of the uncorrelated models). So R gives us a guide to how much improvement[3] has been obtained by introducing correlation. Note that the evidences are extremely small numbers, and that small differences in R values correspond to *huge* factors of relative evidence, of the order of e^{10000} in our case, so that one hypothesis is *overwhelmingly* the most likely for a given image. The table also gives the best fit power law gradients m_F for FP and m_W for WP.

Three random computer-generated images were first analysed: 'A', with independent pixels with a flat distribution from 0 to f_{max} ($= \lambda = 256$); 'B', likewise but with a Gaussian distribution of $\sigma = 20$; and 'C', with a correlated power law distribution of $m = 1.5$ and

---

[2]Later we show that this is exactly the criterion required in a good model.

[3]Note that, because of the units of f_i chosen, e_i is equal to the optimum *message length* in *nats* needed for lossless communication of image **f** (to a precision of one intensity unit) using an encoding based on the hypothesis H_i. Thus R is the information compression ratio.

image	n_x	e_{RB}	e_{RB}/e_{FF}	e_{RB}/e_{GD}	e_{RB}/e_{LGD}	R_F	m_F	R_W	m_W
				uncorrelated models			correlated models		
A	256	-363409	1.003	0.969	0.931	1.000	-0.001	0.999	-0.001
B	256	-363409	1.248	1.255	1.249	1.001	-0.002	1.000	-0.004
C	256	-363409	1.202	1.204	1.189	1.796	1.488	1.635	1.453
susie	128	-90852	1.078	1.040	1.025	1.431	1.577	1.319	1.316
mouse	256	-363409	1.113	0.996	1.046	2.082	1.572	1.877	1.617
redspot	256	-363409	1.084	1.079	1.054	1.320	1.088	1.287	1.162
trees	128	-90852	1.044	0.985	0.949	1.026	0.350	1.027	0.420
sky	128	-90852	1.184	1.095	1.107	1.256	0.717	1.275	0.813
parrot	128	-90852	1.056	1.044	1.012	1.260	1.544	1.189	1.323
m100cen	128	-90852	1.163	1.113	1.094	1.261	0.858	1.240	0.900
ngc1068	128	-90852	1.158	1.033	1.137	1.549	1.215	1.490	1.269

Table 1: Log evidence results for simulated and real images

$c = 10$. So A,B and C are typical samples from the RB, GD and FP model distributions respectively (see Figure 6). These test images behaved as expected: for A and B we find $R_{F,W} \approx 1$ (since they are uncorrelated), whereas for C, $R_F \approx 1.8$ so the FP model shows a vastly higher evidence, and a best fit m close to the predicted value. For B, evidence gains in the uncorrelated models over RB are due to a better fitting of the narrower intensity range.

Analysis of the eight real images gave the general results:

- Correlated models are vastly more successful than uncorrelated, with FP consistently ahead of WP.

- $R_{F,W}$ tend to be larger the higher the best fit gradient m is.

- m_F and m_W loosely match for a given image.

- Of the uncorrelated models, FF invariably has the most evidence (although not always by a large margin), and RB usually the least.

- LGD has no convincing advantage over GD for the last two (astronomical) images.

4 Discussion

To understand the increase in R with m, we consider a general (G class) model where distributions $N(\mu_i, \sigma_i^2)$ are applied to the coefficients F_i produced by some orthogonal linear transform on the image f_i (FP and WP are special cases of this). Making the crude assumption that the F_i are distributed in this way implies that maximizing p_{BF} (and therefore the evidence) is equivalent to minimizing $\prod \sigma_i$ (under the orthogonality constraint $\sum \sigma_i^2 = constant$). This can best be achieved by having only a few large σ_i and many small σ_i, i.e. choosing a transform which concentrates the image 'power' $\sum f_i^2$ into as few coefficients as possible, and a higher m does this better than a low one in the FP or WP model.

It is interesting to realise how the power law found in many of our images relates to a fractal property. Based on Mandelbrot's (1982) statement (p. 254) in the 1D case, we derived that for an image sampled in N dimensions which obeys a statistical scaling law $f(\mathbf{x}) \sim h^{-\alpha} f(h\mathbf{x})$ then one would expect the power spectrum $<|F(k)|^2> \propto k^{-2m}$ (in the case of directional isotropy), with the relation $\alpha = m - N/2$. For this case, $N = 2$ and m is that of the FP model, m_F. This power law spectrum is surprisingly common in much of nature, for instance the rough fracture surfaces of metals (Barnett 1993), which initially led us to investigate the FP model.

Also worthy of discussion are the invariances that were considered in regulating the choice of models for this investigation. If a model had a likelihood function invariant under translation, rotation and scaling of the image, then it could not induce unnatural preferences for particular positions, angular directions or length scales when used as a prior in image reconstruction (or other such inverse problems). Apart from the axes-dependent behavior of the wavelets in the WP model, all the models in this paper share this invariance. However, models where correlation is introduced via a Gaussian ICF (intercorrelation function), for instance, are not scale-invariant and will be prone to favour length scales similar to the ICF radius. We believe that our FP model can be expressed in terms of an ICF, which will however have an asymptotic, power law form.

There are a huge number of directions for further investigation into models for images, but among the more fruitful we suggest:

1. Develop new models that incorporate positivity, since we are dealing with physical intensities which cannot be negative.

2. Search for new formulations of what 'correlation' is, and what makes images recognisable. Borrow ideas from good image compression techniques, as these rely on identifying correlations.

3. Investigate Gabor functions (Gabor 1946), which are forms of wavelets, and which, as Daugman (1985) discusses, seem to match the receptive fields of neurons in the primary visual cortex. We suggest that, since evolution has optimized so many biological design problems, the workings of our own perceptual system should be studied and mimicked to find good image processing and modelling techniques. It is, after all, our own perception that tells our consciousness that we are looking at a recognisable image.

5   Conclusion

A framework of simple models for images has been built up, and their Bayesian *evidence* has been evaluated for a set of image data. The results show a conclusively massive increase in evidence for correlated models (FP and WP) over uncorrelated (RB, FF, GD and LGD), with the FP model almost always the most successful, especially at higher m_F. This reflects a power law dependence of Fourier components apparent in images and implies a statistical scaling self-similarity, that is, a general fractal property.

References

BARNETT, A. H. (1993) Statistical modelling of rough crack surfaces in metals. *Internal Report for Non-Destructive Testing Applications Centre, Technology Division, Nuclear Electric plc.*

DAUGMAN, J. G. (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *J. Opt. Soc. Am. A* **2** (7): 1160–1169.

GABOR, D. (1946) Theory of communication. *J. Inst. Electr. Eng.* **93**: 429–457.

GREGORY, P. C., and LOREDO, T. J. (1992) A new method for the detection of a periodic signal of unknown shape and period. In *Maximum Entropy and Bayesian Methods,*, ed. by G. Erickson and C. Smith. Kluwer. also in The Astrophysical Journal, Oct 10, 1992.

HEIL, C. E., and WALNUT, D. F. (1989) Continuous and discrete wavelet transforms. *SIAM Review* **31** (4): 628–666.

MANDELBROT, B. (1982) *The Fractal Geometry of Nature*. W.H. Freeman and Co.

PRESS, W., B.P. FLANNERY, S. T., and VETTERLING, W. (1993) *Numerical Recipes in C, Second Edition*. Cambridge.

STRANG, G. (1989) Wavelets and dilation equations: A brief introduction. *SIAM Review* **31** (4): 614–627.