BAYESIAN MAXIMUM ENTROPY IMAGE RECONSTRUCTION

John Skilling
Dept. of Applied Mathematics and Theoretical Physics
Silver Street
Cambridge CB3 9EW, U.K.

Stephen F. Gull
Cavendish Laboratory
Madingley Road
Cambridge CB3 0HE, U.K.

ABSTRACT

This paper presents a Bayesian interpretation of maximum entropy image reconstruction and shows that $\exp(\alpha S(f,m))$, where $S(f,m)$ is the entropy of image f relative to model m, is the only consistent prior probability distribution for positive, additive images. It also leads to a natural choice for the regularizing parameter α, that supersedes the traditional practice of setting $\chi^2 = N$. The new condition is that the dimensionless measure of structure $-2\alpha S$ should be equal to the number of good singular values contained in the data. The performance of this new condition is discussed with reference to image deconvolution, but leads to a reconstruction that is visually disappointing. A deeper hypothesis space is proposed that overcomes these difficulties, by allowing for spatial correlations across the image.

1. Introduction

The Maximum Entropy method (MaxEnt) has proved to be an enormously powerful tool for reconstructing images from many types of data. It has been used most spectacularly in radio-astronomical interferometry, where it deals routinely with images of up to a million pixels, with high dynamic range. A review of the method, together with many examples taken from fields such as optical deblurring and NMR spectroscopy is given by Gull & Skilling (1984a).

The purpose of this paper is to present the underlying fundamental justification for the maximum entropy method in image processing and to give it a Bayesian interpretation. The advantage of this probabilistic formulation is that it now allows us to quantify the *reliability* of MaxEnt images. We also extend the power of MaxEnt in the field of imaging by introducing spatial statistics into the formalism.

In Section 2 we review the work of Cox (1946), who demonstrated that any consistent method for making inferences using real numbers must be equivalent to ordinary probability theory. This forces us to formulate our preferences for images $f(x)$ as Bayesian prior probabilities $\mathrm{pr}\,(f)$.

In Section 3, we observe that the only generally acceptable procedure for assigning prior probabilities is MaxEnt: it is the only method that gives acceptable results in simple cases. However, MaxEnt applies not just to probability distributions, but more generally to any positive, additive distribution such as an image, giving a direct justification for the use of entropy in imaging. However, it is too simplistic just to select that single image which has maximum entropy, because the Bayesian methodology forces us to consider quantified (probabilistic) distributions of images.

Once more taking a simple case, we proceed in Section 4 to quantify the entropy formula, finding that the prior probability $\mathrm{pr}\,(f)$ of any particular image $f(x)$ must be of the very specific form $\exp(\alpha S(f,m))$, where S is the entropy and $m(x)$ is the measure on x which must be assigned in order to define the entropy properly. m can be thought of as an initial model for f, away from which $S(f,m)$ measures (minus) the deviation, and it is often chosen to be constant. Finally, α is a dimensional constant which can certainly not be assigned a priori.

With noisy data, traditional practice has been to select a value of α that makes the χ^2 misfit statistic equal to the number of observations, but this is ad hoc and does not allow for the reduction in effective number of degrees of freedom caused by fitting accurate data. In Sections 5 and 6, we complete the derivation of "Classic" MaxEnt with the Bayesian determination of α, finding that the amount of structure in the image, quantified as $-2\alpha S$, must equal the number of "good" (accurate) singular vectors contained in the data. The value of χ^2 is not relevant to the choice of α, but instead allows an estimate of the overall noise level if it is unknown.

The application of this method is discussed (Section 7) by reference to a specific deconvolution example. Disconcertingly, the "Classic" reconstruction is visually disappointing, with an unfortunate level of "ringing". This can only be due to a poor choice of initial model m. Indeed, the initial, flat model is very far from the final reconstruction. In order to allow the "good" singular data vectors to be fitted, α must be small, so that there is little entropic smoothing, and the consequence is under-smoothing of the "bad" noisy data.

The next step must be a better model, incorporating some expectation of correlated

spatial statistics in a deeper hypothesis space. Section 8 rationalizes our approach to this, within the Bayesian MaxEnt framework, and Section 9 quantifies it. We introduce a set of "hidden variables" $\widetilde{m}(x)$ which are then blurred to make the model $m(x)$ used in "Classic". The prior for these hidden variables must also be of entropic form $\exp(\beta S(\widetilde{m}, \text{flat}))$. The new multiplier β and the width of the hidden blur are determined by Bayesian methods.

The results from this deeper hypothesis space are excellent, and provide a coherent rationale for some of the manipulations of the model m that have been found useful in current practice.

2. Bayesian probability theory: The Cox axioms

Whatever the content of our discussions, be it Raman spectroscopy or Roman history, we wish to be able to express our preferences for the various possibilities $i, j, k, \ldots$ before us. A minimal requirement is that we be able to rank our preferences consistently (i.e. transitively)

$$(\text{Prefer } i \text{ to } j) \text{ AND } (\text{Prefer } j \text{ to } k) \Longrightarrow (\text{Prefer } i \text{ to } k)\,. \tag{2.1}$$

Any transitive ranking can be mapped onto real numbers, by assigning numerical codes $P(i)$, $P(j), \ldots$ such that

$$P(i) > P(j) \Longleftrightarrow (\text{Prefer } i \text{ to } j). \tag{2.2}$$

Now, *if* there is a common general language, it must apply in simple cases. Cox (1946) formulated two such simple cases as axioms, which we restate briefly. It is difficult to argue against either.

Axiom A:

If we first specify our preference for i being true, and then specify our preference for j being true (given i), then we have implicitly defined our preference for i and j together.

This refers to a particularly simple set of hypotheses involving just two propositions i and j. In terms of the numerical codes,

$$P(i,j)|h) = F(P(i|h), P(j|i,h)), \tag{2.3}$$

where h is the given evidence and F is some unknown function. Using the Boolean rules obeyed by logical conjunction of propositions, Cox was able to manipulate this axiom into the associativity functional equation

$$F(p, F(q,r)) = F(F(p,q), r). \tag{2.4}$$

As a consequence of this (see also Aczel 1966), there exists some monotonically increasing non-negative function π of the original preferences p, in terms of which F is just scaled multiplication.

$$\pi(i,j|h) = C\pi(i|h)\pi(j|i,h), \tag{2.5}$$

where C is a constant. We may as well use this new numerical coding π in place of the more arbitrary original coding P.

However, π is not yet fully determined, because $C^{r-1}\pi^r$ (with $r > 0$) also obeys (2.5). Although this is the only freedom, it is still too much and another axiom is needed.

Axiom B:

If we specify our preference for i being true, then we have implicitly specified our preference for its negation $\sim i$.

In terms of the numerical codes,

$$\pi(\sim i|h) = f(\pi(i|h)), \tag{2.6}$$

where f is some unknown function. As a consequence, Cox showed that there is a particular choice of r and C which turns the codes π into other codes pr obeying

$$\mathrm{pr}\,(i|h) + \mathrm{pr}\,(\sim i|h) = 1. \tag{2.7}$$

Equation (2.5) then becomes

$$\mathrm{pr}\,(i,j|h) = \mathrm{pr}\,(i|h)\,\mathrm{pr}\,(j|i,h) \tag{2.8}$$

with its corollary, Bayes' Theorem

$$\mathrm{pr}\,(i|j,h) = \mathrm{pr}\,(i|h)\,\mathrm{pr}\,(j|i,h)/\,\mathrm{pr}\,(j|h). \tag{2.9}$$

We also have

$$0 \leq \mathrm{pr} \leq 1 \tag{2.10}$$

and may identify

$$\mathrm{pr}\,(\text{falsity}) = 0, \quad \mathrm{pr}\,(\text{certainty}) = 1. \tag{2.11}$$

There is no arbitrariness left. Thus there must be a mapping of the original codes P into other codes pr that obeys the usual rules of Bayesian probability theory. Therefore, *if* there is a common language, then it can only be this one, and in accordance with historical precedent set by Bernoulli and Laplace (Jaynes 1978) we call the codes pr thus defined "probabilities". Logically, of course, there may be no common language. There may be a lurking "Axiom C", just as convincing as Axioms A and B, which contradicts them. Although much effort has been expended on such arguments (Klir 1987), no such contradictory axiom has been demonstrated to our satisfaction, and accordingly we submit to the Bayesian rules.

Bayes' Theorem itself, which is simple corollary of these rules, then tells us how to *modulate* probabilities in accordance with extra evidence. It does not tell us how to *assign* probabilities in the first place. It turns out that such prior assignments should be accomplished by MaxEnt.

3. Maximum Entropy: The assignment of positive, additive distributions

The probability distribution $\mathrm{pr}\,(x)$ of a variable x is an example of a positive, additive distribution. It is positive by construction. It is additive in the sense that the overall probability in a domain D equals the sum of the probabilities in any decomposition into sub-domains, and we write it as $\int_D \mathrm{pr}\,(x)\,dx$. It also happens to be normalized, $\int_{\mathrm{all}\,x} \mathrm{pr}\,(x)\,dx = 1$.

Another example of a positive, additive distribution is the intensity or power $f(x,y)$ of incoherent light as a function of position (x,y) in an optical image. This is positive, and additive because the integral $\int\int_D f(x,y)\,dx\,dy$ represents the physically meaningful power in D. (By contrast, the amplitude of incoherent light, though positive, is not additive.) For brevity, we shall call a positive, additive distribution a "PAD".

It turns out to be simpler to investigate the general problem of assigning a PAD than the specific problem of assigning a probability distribution, which carries the technical complication of normalization. Accordingly, we investigate the assignment of a PAD $f(x)$, given some definitive but incomplete constraints on it: such constraints have been called "testable information" by Jaynes (1978). Now *if* there is a general rule for assigning a single PAD, *then* it must give sensible results in simple cases. The four "entropy axioms" –so called because they lead to entropic formulae–relate to such cases. Shore and Johnson (1980) and Tikochinsky, Tishby and Levine (1984) give related derivations pertaining to the special case of probability distributions. Proofs of the consequences of the axioms as formulated below appear in Skilling (1988a), though our phraseology improves upon that paper.

Axiom I: "Subset Independence"

Separate treatment of individual separate distributions should give the same assignment as joint treatment of their union.

More formally, if constraint C_1 applies to $f(x)$ in domain $x \in D_1$ and C_2 applies to a separate domain $x \in D_2$, then the assignment procedure should give

$$f[D_1|C_1] \cup f[D_2|C_2] = f[D_1 \cup D_2|C_1 \cup C_2], \tag{3.1}$$

where $f[D|C]$ means the PAD assigned in domain D on the basis of constraints C.

For example, if $f(x) = 4 (0 < x < 1)$ is assigned under the constraint $\int_0^1 f\,dx = 4$, and $f(x) = 2 (1 < x < 2)$ from $\int_1^2 f\,dx = 2$, then the joint assignment under the double constraint $(\int_0^1 f\,dx = 4, \int_1^2 f\,dx = 2)$ should be $f(x) = (4 \text{ for } 0 < x < 1,\ 2 \text{ for } 1 < x < 2)$.

Consequence: The PAD f should be assigned by maximizing over f some integral of the form

$$S(f,m) = \int dx\, m(x) \Theta(f(x),x). \tag{3.2}$$

Here Θ is a function, as yet unknown, and m is the Lebesgue measure associated with x which must be given before an integral can be defined. The effect of this basic axiom is to eliminate all cross-terms between different domains.

Axiom II: "Coordinate invariance"

The PAD should transform as a density under coordinate transformations.

For example, if $f(x) = 4\ (0 < x < 1)$ is assigned under the constraint $\int_0^1 f(x)\,dx = 4$, and x is transformed to $y = 2x + 1$, then the corresponding constraint $\int_1^3 F(y)\,dy = 4$ should yield the reconstruction $F(y) = f(x)dx/dy = 2 (1 < y < 3)$.

Consequence: The PAD f should be assigned by maximizing over f some integral of invariants

$$S(f,m) = \int dx\, m(x) \phi(f(x)/m(x)), \tag{3.3}$$

where ϕ is a function, as yet unknown. The crucial axiom is the next.

Axiom III: "System independence"

If a proportion q of a population has a certain property, then the proportion of any sub-population having that property should properly be assigned as q.

For example, if 1/3 of kangaroos have blue eyes (Gull and Skilling 1984b), then the proportion of left-handed kangaroos having blue eyes should also be assigned the value 1/3.

Applying this to a two-dimensional PAD on the unit square with constant (unit) measure $m(x, y) = 1$, we see that if the marginal distribution along x is known to be

$$\int_0^1 f(x, y)\, dy = a(x), \tag{3.4}$$

then the x-variation of f at any particular y must be assigned as $a(x)$. In other words, $f(x, y) = a(x)\psi(y)$ for some ψ. If, additionally, the marginal y-distribution is known to be

$$\int_0^1 f(x, y)\, dx = b(y), \tag{3.5}$$

then the overall assignment must be

$$f(x, y) = Ka(x)b(y), \tag{3.6}$$

where the constant K takes account of overall normalization.

Consequence: The only integral of invariants whose maximum always selects this assignment is

$$S(f, m) = -\int dx\, f(x) \log(f(x)/\operatorname{cem}(x)), \tag{3.7}$$

where c is a constant, scaled by e for convenience.

Axiom IV: "Scaling"

In the absence of additional information, the PAD should be assigned equal to the given measure.

Without this axiom, the PAD is assigned as $f(x) = cm(x)$, so the axiom fixes $c = 1$, and states in effect that f and m should be measured in the same units. This is a practical convenience rather than a deep requirement.

Consequence: The PAD f should be assigned by maximizing over f

$$S(f, m) = \int dx(f(x) - m(x) - f(x) \log(f(x)/m(x))). \tag{3.8}$$

The additive constant $\int m\, dx$ in this expression ensures that the global maximum of S, at $f(x) = m(x)$, is zero, which is both convenient and required for other purposes (Skilling 1988a).

Because of its entropic form, we call S as defined in (3.8) the *entropy* of the positive, additive distribution f. It reduces to the usual cross-entropy formula $-\int dx\, f \log(f/m)$

if f and m happen to be normalized, but is actually more general. (Holding that the general concept should carry the generic name, we deliberately eschew giving (3.8) a qualified name.)

We see that MaxEnt is the only method which gives sensible results in simple cases, so *if* there is a general assignment method, it *must* be MaxEnt. (Logically, there may be a lurking, contradictory "Axiom V", but we have not found one, and accordingly we submit to this "principle of maximum entropy".) Two major applications follow from this analysis. Firstly, MaxEnt is seen to be the proper method for assigning probability distributions $\mathrm{pr}\,(x)$, given testable information. Secondly, in practical data analysis, if it is agreed that prior knowledge of a PAD satisfies axioms I-IV, and if testable information is given on it, then any single PAD to be assigned on this basis must be that given by MaxEnt.

However, the arguments above do not address the *reliability* of the MaxEnt assignment: would a slightly different PAD be very much inferior?. Furthermore, experimental data are usually noisy, so that they do not constitute testable information about a PAD f. Instead, they define the likelihood or conditional probability $\mathrm{pr}\,(\mathrm{data}|f)$ as a function of f. In order to use this in a proper Bayesian analysis, we need the quantified prior probability $\mathrm{pr}\,(f)$–or strictly $\mathrm{pr}\,(f|m)$ because we have needed to set a measure m.

4. Quantification

The reliability of an estimate is usually described in terms of ranges and domains, leading us to investigate probability integrals over domains V of possible PADs $f(x)$, digitized for convenience into r cells as $(f_1, f_2, \ldots, f_r)$.

$$\mathrm{pr}\,(f \in V|m) = \int_V d^r f M(f)\,\mathrm{pr}\,(f|m), \tag{4.1}$$

where $M(f)$ is the measure on the space of PADs. By definition, the single PAD we most prefer is the most probable, and we identify this with the PAD assigned by MaxEnt. Hence $\mathrm{pr}\,(f|m)$ must be of the form

$$\mathrm{pr}\,(f|m) = \text{ monotonic function } (S(f,m)), \tag{4.2}$$

but we do not yet know which function. Now S has the units (dimensions) of f, so this monotonic function must incorporate a dimensional constant, α say, *not* an absolute constant, so that

$$\mathrm{pr}\,(f \in V|m) = \int_V d^r f M(f)\Phi(\alpha S(f,m))/Z_S(\alpha,m), \tag{4.3}$$

where Φ is a monotonic function of dimensionless argument and

$$Z_S(\alpha,m) = \int_\infty d^r f M(f)\Phi(\alpha S(f,m)) \tag{4.4}$$

is the partition function which ensures that $\mathrm{pr}\,(f|m)$ is properly normalized.

In order to find Φ, we consider a simple case, satisfying axioms I–IV, for which the probability is known. Let the traditional team of monkeys throw balls (each of

quantum size q) at r cells ($i = 1, 2, \ldots, r$), at random with Poisson expectations μ_i. The probability of occupation numbers n_i is

$$\mathrm{pr}\,(n|\mu) = \Pi_i \mu_i^{n_i} e^{-\mu_i}/n_i!\,. \tag{4.5}$$

Define $f_i = n_i q$ and $m_i = \mu_i q$ to remain finite as the quantum size q is allowed to approach zero. Then the image-space of f becomes constructed from microcells of volume q^r, each associated with one lattice-point of integers $(n_1, n_2, \ldots, n_r)$. Hence we have, as q tends to 0,

$$\begin{aligned}\mathrm{pr}\,(f \in V|m) &= \sum_{\text{lattice points in } V} \mathrm{pr}\,(n|\mu) \\ &= \int_V (d^r f/q^r)\Pi_i \mu_i^{n_i} e^{-\mu_i}/n_i!\,. \end{aligned} \tag{4.6}$$

Because we are taking n large, we may use Stirling's formula

$$n_i! = (2\pi n_i)^{1/2} n_i^{n_i} e^{-n_i} \tag{4.7}$$

to obtain (accurately to within $0(1/n)$)

$$\mathrm{pr}\,(f \in V|m) = \int_V \frac{d^r f}{\Pi(2\pi q f_i)^{1/2}} \exp \frac{\Sigma(f_i - m_i - f_i \log(f_i/m_i))}{q} \tag{4.8}$$

Here we recognize the entropy on r cells,

$$\Sigma(f_i - m_i - f_i \log(f_i/m_i)) = S(f, m), \tag{4.9}$$

so that

$$\mathrm{pr}\,(f \in V|m) = \int_V \frac{d^r f}{\Pi f_i^{1/2}} \frac{\exp(S(f,m)/q)}{(2\pi q)^{r/2}}\,. \tag{4.10}$$

Comparing this with the previous formula (4.3), we must identify

$$q = 1/\alpha, \quad \Phi(\alpha S(f,m)) = \exp(\alpha S(f,m)) \tag{4.11}$$

and

$$Z_S(\alpha, m) = (2\pi/\alpha)^{r/2}, \quad M(f) = \Pi f_i^{-1/2}\,. \tag{4.12}$$

save possibly for multiplicative constants in Φ, Z_S, M which can be defined to be unity. Note how the often-ignored "square-root" factors in Stirling's formula have enabled us to derive the measure M, which allows us to make the passage between pointwise probability comparisons and full probability integrals over domains.

A natural interpretation of the measure is as the invariant volume $(\det g)^{1/2}$ of a metric g defined on the space. Thus the natural metric for the space of PADs is

$$g_{ij} = \begin{cases} 1/f_i & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \tag{4.13}$$

which happens to equal (minus) the entropy curvature $\nabla\nabla S \equiv \partial^2 S/\partial f\,\partial f$. This result was previously obtained from an alternative viewpoint by Levine (1985).

Although this analysis has used large numbers of small quanta q, so that α is large, this limit also ensures that each n_i will almost certainly be close to its expectation μ_i. Indeed, the expected values of αS remain 0(1), so that the identification

$$\Phi(u) = \exp(u) \tag{4.14}$$

holds for finite arguments u. Finally, *if* there is a general form of Φ, it must be valid for the small quantum case, so Φ must be exponential.

To summarize, *if* there is a general prior for positive, additive distributions f, it must be

$$\mathrm{pr}\,(f|m) = \exp(\alpha S(f,m))/Z_S(\alpha) \tag{4.15}$$

and furthermore

$$\mathrm{pr}\,(f \in V|m) = \int_V \frac{d^r f}{\Pi f_i^{1/2}} \frac{\exp(\alpha S(f,m))}{Z_S(\alpha)}, \tag{4.16}$$

where

$$Z_S(\alpha) = \int_\infty \frac{d^r f}{\Pi f_i^{1/2}} \exp(\alpha S(f,m)). \tag{4.17}$$

This quantified prior contains just one undetermined, dimensional parameter α.

5. Classic MaxEnt–the choice of α

The only remaining parameter in our "Classic" hypothesis space is the constant α. We do not believe that we can determine α a priori by general arguments. Not only is α dimensional, so that it depends on the scaling of the problem, but its best-fitting value varies quite strongly with the type and quality of the data available. It can only be determined a posteriori.

We therefore turn for a moment to the other side of the problem, the likelihood, which we write as:

$$\mathrm{pr}\,(D|f) = \exp(-L(f))/Z_L, \tag{5.1}$$

where

$$Z_L = \int d^N D \exp(-L), \tag{5.2}$$

N being the number of data. The log-likelihood $L(f)$ defined by this expression contains all the details of the experimental setup and accuracies of measurement. For the common case of independent, Gaussian errors, this reduces to $L = \chi^2/2$, but other types of error such as Poisson noise are also important. Quite frequently, the overall level of noise is not well-known, so we will eventually generalize to

$$\mathrm{pr}\,(D|f,\sigma) = \exp(-L(f)/\sigma^2)/Z_L(\sigma), \tag{5.3}$$

but for now we assume that the errors are known in advance, so that $\sigma = 1$.

We now write down the joint p.d.f. of data and image:

$$\mathrm{pr}\,(f,D|\alpha,m) = Z_L^{-1} Z_S^{-1} \exp(\alpha S - L). \tag{5.4}$$

Byes' Theorem tells us that this is also proportional to the posterior probability distribution for f : $\mathrm{pr}\,(f|D,\alpha,m)$. The maximum of this distribution as a function of f is then our "best" reconstruction, and occurs at the maximum of

$$Q = \alpha S - L. \tag{5.5}$$

This brings us back once again to the choice of α, which can now be viewed as a regularizing parameter. When seen this way, α controls the competition between S and L: if α is large, the data cannot move the reconstruction far from the model–the entropy term dominates. If α is low there is little smoothing and the reconstruction will show wild oscillations as the noise in the data is interpreted as true signal. We have to control α carefully, but there is usually a large range of sensible values.

Our practice hitherto (Gull & Daniell 1978, Gull & Skilling 1984a) has been to set α so that the misfit statistic χ^2 is equal to the number of data points N. Although this has a respectable pedigree in the statistical literature (the discrepancy method (Tikhonov & Arsenin 1977)), it is ad hoc, and can be criticized on several grounds.

(1) the only "derivation" of the $\chi^2 = N$ condition that has been produced is a frequentist argument. If the image was known in advance and the data were then repeatedly measured, $\chi^2 = N$ would result on average. However, the data are only measured once and the image is not known a priori, but is instead estimated from the one dataset we have.

(2) There is no allowance for the fact that good data cause real structure in the reconstruction f. These "good" degrees of freedom are essentially parameters that are being fitted from the data, so that they no longer contribute to the variance. This leads, in general terms, to "under-fitting" of data (Titterington 1985). This is particularly apparent for imaging problems where there is little or no blurring. The $\chi^2 = N$ criterion leads to a uniform, one standard deviation bias towards the model. This bias is very unfortunate: it is the job of a regularizer such as entropy to cope with noise and missing information, not to bias the data that we do have.

(3) For many problems (such as radioastronomical imaging, where we started) the data are nearly all noise, so that $\chi^2 \approx N$ for any reasonable α. The statistic χ^2 is in any case expected to vary by $\pm\sqrt{N}$ from one data realization to another, and this can easily swamp the difference between χ^2 at $\alpha = \infty$ and the χ^2 appropriate to a sensible reconstruction.

For these reasons we now believe that there is no acceptable criterion for selecting α that looks only at the value of a misfit statistic such as χ^2.

6. Bayesian choice of α

Within our Bayesian framework there is a natural way of choosing α. We simply treat it as another parameter in our hypothesis space, with its own prior distribution. The joint p.d.f. is now

$$\mathrm{pr}\,(f, D, \alpha|m) = \mathrm{pr}\,(\alpha)\,\mathrm{pr}\,(f, D|\alpha, m). \tag{6.1}$$

To complete the assignment of the joint p.d.f. we select an uninformative prior, uniform in $\log(\alpha)$: $\mathrm{pr}\,(\log\alpha)$ = constant over some "sensible" range $[\alpha_{\min}, \alpha_{\max}]$. We shall return to the definition of "sensible" later.

Using Bayes' Theorem, this joint distribution is also proportional to the posterior distribution $\mathrm{pr}\,(f, \alpha|D, m)$ and we proceed to estimate the best value of α by marginalization over the reconstruction f:

$$\mathrm{pr}\,(\alpha|D, m) = \int d^r f\, \Pi f^{-1/2}\, \mathrm{pr}\,(f, \alpha|D, m).$$

$$\propto Z_Q Z_S^{-1} Z_L^{-1}, \tag{6.2}$$

where $Z_Q = \int d^r f \Pi f^{-1/2} \exp(\alpha S - L)$.

It is essential to perform this integral carefully, rather than estimating α by maximizing the integrand with respect to f and α simultaneously, because the distribution in $f - \alpha$ space is significantly skew. In fact, the maximum of $\mathrm{pr}\,(f, \alpha | D, m)$ is usually at $\alpha = \alpha_{\max} =$ large, $f \approx m$, which is certainly not what we want.

We now evaluate the integrals involved. The integrand for Z_S has a maximum at $f = m$ and, using Gaussian approximations, we find that for all α a reasonable approximation to $\log Z_S$ is:

$$\log Z_S = r/2 \ \log(\alpha/2\pi). \tag{6.3}$$

In performing this integral, the terms from the volume element cancel with those from the curvature $\nabla\nabla S$. This is a happy consequence of the fact that the entropy curvature is also the natural metric tensor of the f space.

The Z_Q integral is done similarly, expanding about the maximum of $Q(f, m, \alpha)$ at $\hat{f}$. We can aid our understanding by introducing at this point the eigenvalues $\{\lambda_i\}$ of the symmetric matrix

$$A = \mathrm{diag}\,(f^{1/2}).\nabla\nabla L.\ \mathrm{diag}\,(f^{1/2}), \tag{6.4}$$

which is the curvature of L viewed in a the entropy metric. The eigenvalues λ and eigenvectors in f space define the natural coordinates for our problem, and the $\lambda^{1/2}$ are the appropriate "singular values". A large value of λ implies a "good" or measured direction, whereas a low or zero λ corresponds to a poorly measured quantity.

Evaluating the integrals in the Gaussian approximation, we find

$$\begin{aligned} \log \mathrm{pr}\,(\alpha | D, m) &= \text{constant} + r/2 \ \log(\alpha) - 1/2 \ \log\det(\alpha I + B) + Q(\hat{f}, m, \alpha) \\ &= \text{constant} + 1/2 \sum_j \log(\alpha/(\alpha + \lambda_j)) + \alpha S(\hat{f}, m) - L(\hat{f}). \end{aligned} \tag{6.5}$$

For large datasets this has a sharp maximum at a particular value of α. Differentiating with respect to $\log \alpha$, and noting that the $\hat{f}$ derivatives cancel, we find the condition:

$$-2\alpha S(\hat{f}, m) = \sum_j \lambda_j/(\alpha + \lambda_j). \tag{6.6}$$

This fixes our estimate of $\alpha = \hat{\alpha}$ quite closely, provided we have many data, so that we can return to the determination of the reconstruction $\hat{f}$. Strictly, having already integrated out f to determine $\mathrm{pr}\,(\alpha)$, the formalism does not allow us to return with a single value $\hat{\alpha}$. However, we are allowed to find the distribution of any integral $\int d\alpha f(x) r(x)$ by integrating the joint p.d.f. successively over f and then α. Because $\mathrm{pr}\,(\alpha)$ is so sharply peaked, the effect on R is just as if α were set equal to $\hat{\alpha}$. We may as well simplify the notation by setting $\alpha = \hat{\alpha}$ in the derivation of f itself:

$$\begin{aligned} \mathrm{pr}\,(f | D, m) &= \int d\alpha \ \mathrm{pr}\,(f, \alpha | D, m) \\ &= \int d\alpha \ \mathrm{pr}\,(\alpha | D, m) \ \mathrm{pr}\,(f | \alpha, D, m) \\ &\cong \mathrm{pr}\,(f | \hat{\alpha}, D, m) \\ &= Z_Q^{-1} \exp(\hat{\alpha} S(\hat{f}, m) - L(\hat{f})). \end{aligned} \tag{6.7}$$

The fluctuations (uncertainty) of f about $\hat{f}$ can also be investigated, at least in principle, by using the known curvature:

$$\langle \delta f \delta f^t \rangle = [\nabla\nabla Q]^{-1}. \tag{6.8}$$

We can understand our Bayesian formula for the best value $\hat{\alpha}$ as follows.

(1) The statistic $\lambda/(\alpha+\lambda)$ is a measure of the quality of the data along any given singular vector. If $\lambda \gg \alpha$ the data are good and $\lambda/(\alpha+\lambda)$ adds one to the statistic. If, on the other hand, $\lambda \ll \alpha$, then the regularizing entropy dominates the observations and the contribution is approximately zero. We can therefore say that $\Sigma\lambda/(\alpha+\lambda)$ specifies the number of good, independent data measurements, or the number of degrees of freedom with the entropy rather than the likelihood because these are the directions (dimensions) that contribute to the entropy. We shall see later (equ. 9.3–4) the reason for this apparently perverse choice of notation.

(2) The quantity $-2\alpha S$ is a dimensionless measure of the amount of structure in the image relative to the model, or the distance that the likelihood has been able to pull the reconstruction away from the starting model.

The formula thus has a very plausible interpretation: the dimensionless measure of the amount of structure demanded by the data is equal to the number of good, independent measurements. We also note that, as we indicated earlier, the value of the misfit statistic L is irrelevant to the choice of α. However, it too has a role to play. To see this we now generalize to the case of unknown overall noise level

$$\mathrm{pr}\,(D|f,\sigma) = \exp -L(f)/\sigma^2/Z_L(\sigma), \tag{6.9}$$

and this time keeping all terms involving σ find:

$$\mathrm{pr}\,(\alpha,\sigma) = \text{ constant } -N\log(\sigma) + 1/2\ \Sigma\alpha'/(\alpha'+\lambda_j) + \alpha S = L/\sigma^2, \tag{6.10}$$

where $\alpha' = \alpha\sigma^2$. There is now an additional Bayesian choice for σ and its estimate $\hat{\sigma}$,

$$2L(\hat{f})/\sigma^2 = N - \Sigma\lambda/(\alpha+\lambda). \tag{6.11}$$

The interpretation of this condition is also very plausible: the expected $\chi^2(=2L)$ is equal to the number of degrees of freedom controlled by the entropy, that is, the poorly measured "bad" directions of f space. This is less than the number of data, thereby answering our first objection to $\chi^2 = N$, and showing that the χ^2 (or L) is really suited to estimation of the noise level, not α. Notice also how there is a clean division of degrees of freedom between S and L, so that

$$N = ndf(S) + ndf(L). \tag{6.12}$$

The choice of regularizing parameters has been much debated in the statistical literature (Titterington 1985 gives a review). Our arguments in this section have reproduced (albeit for an entropic variation) one of these prescriptions, known elsewhere as Generalized Maximum Likelihood (Davies & Anderssen 1986).

7. Performance of the Bayesian α

To illustrate both the power and the shortcomings of the Bayesian choice for α, we turn now to a practical example, a picture of "Susie". Figure 1 shows Susie, digitized on a 128 × 128 pixel grid, with grey-level values between 40 and 255. This picture was blurred with a 6-pixel radius Gaussian point-spread function (PSF) and noise of unit variance added. This is a traditional example for MaxEnt processing (e.g. Daniell & Gull 1980, Gull & Skilling 1984a), and we show a $\chi^2 = N$ reconstruction. Our previously-published "Susie" have used a disc PSF, appropriate to an out-of-focus camera, and for which the MaxEnt results at this signal-to-noise are more impressive visually. A Gaussian PSF gives less improvement in resolution because the eigenvalues of $\nabla\nabla L$ fall off very fast.

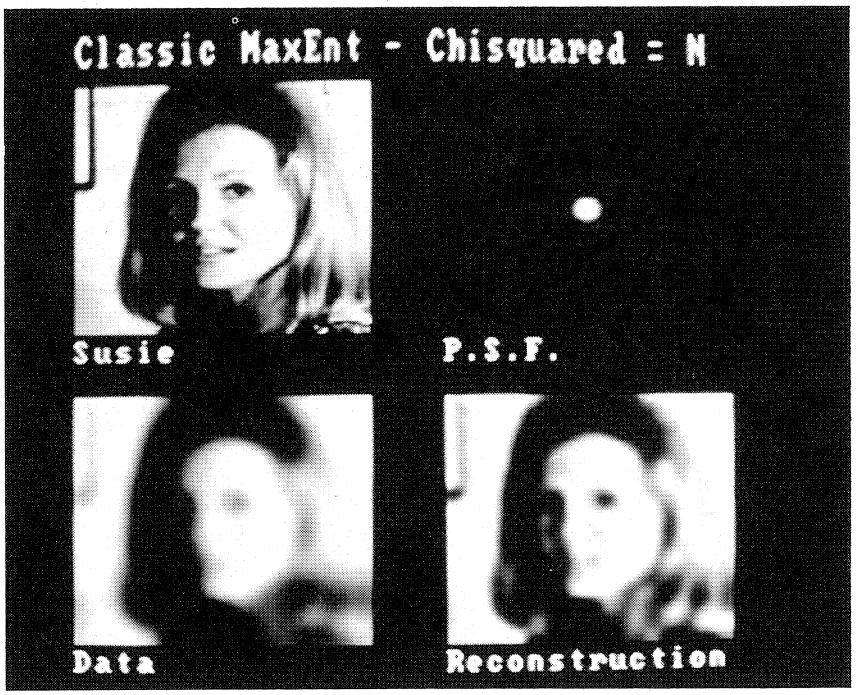

Figure 1. 128 × 128 image of Susie, blurred with a 6-pixel Gaussian PSF. MaxEnt reconstruction using $\chi^2 = N$.

We now reach the first practical difficulty associated with our Bayesian answer. The log-determinant and the $ndf(S)$ statistic require a knowledge of the eigenvalue spectrum of $f^{1/2}\nabla\nabla L f^{1/2}$. For the present case, this is a 16384×16384 matrix, a size which is well in excess of the limits for conventional computational methods of calculating eigenvalues. However, Skilling (1988b) has recently developed a method based on the application of the matrix to random vectors, together with the use of Maxent, that allows an estimate of the eigenvalue spectrum to be obtained. In particular, the accuracy of estimation

of scalars such as $ndf(S)$ is excellent using this technique. It seems, therefore, that practical computation of the Bayesian solution is in general possible.

For the moment, the problem of the eigenvalues is avoided in a different way: we change the definition of S. All of the Bayesian analysis of the last section applies equally to any regularizing function, so we select a simple one that allows us to diagonalize $\nabla\nabla L$ and $\nabla\nabla S$ simultaneously. This is the case for a spatially-invariant, circulant PSF and for the quadratic

$$S = -1/2 \sum_j (f_j - m_j)^2, \tag{7.1}$$

which is a linearized version of the correct form, and a reasonable approximation for a low-contrast image such as Susie. The computations can now be performed easily in eigenvector (Fourier Transform) coordinates. The change in the definition of S makes no difference to the formulae, except that the metric is now flat, the $f^{1/2}$ terms disappear and f might possibly go negative. The change makes no difference whatever to our conclusions about the performance of the Bayesian solution.

Figure 2 shows the reconstruction from blurred Susie for a selection of α values. When α is high the reconstruction looks like the original blurred data, and when α is too low unsightly ripples appear due to the amplification of noise. Note, however that this behavior covers a wide range of $\alpha(\sim 10^4)$ and that there is a large region where the reconstruction is generally satisfactory.

For our example the Bayesian solution suggests that there are ~ 790 good degrees of freedom out of the total 16384. As might be expected, this is somewhat greater than $16384/36\pi = 145$ independent PSFs contained in the image, the excess being a rough measure of the degree of deconvolution obtained. Its estimate of the noise level was correct to within the expected error and, indeed, we have always found that the noise level prediction performance of the Bayesian solution is excellent. Figure 3 shows a plot of the posterior probability of α, both as its logarithm, and also linearly, to emphasize the discrimination in the determination of α, which is better than $1\,db$ for this dataset. The posterior p.d.f. is normalizable as α approaches zero (towards the left of Figure 3a, b,) if the noise level is known, but a global view (towards the right of Figure 3c) shows that it levels off once α exceeds the highest eigenvalue, resulting in a technically improper distribution. We therefore return to the definition of a "sensible" cutoff for $\alpha_{\max}$ referred to earlier. The scale of Figure 3c is rather large: in order to make a 50 per cent contribution to the probability integral, the $\alpha_{\max}$ cutoff has to exceed $\exp(\exp(1.4 \times 10^7))$. Such numbers are typical of the "singularities" encountered in this type of Bayesian analysis. We are content to take $\alpha_{\max}$ less than this bizarre value, so that we are unconcerned by this technical impropriety.

The reconstruction $\hat{f}(\hat{\alpha})$ is shown as Figure 4. It is visually disappointing, and is clearly in the range of the "over-fitted" solutions for which α is too low. It is very easy to understand why this is so. The initial model used for these reconstructions was everywhere uniform, at approximately the mean of the data. This model is very far from the final reconstruction, because there is plenty of real structure in the picture produced by the 790 good measurements in the data. α must be reduced sufficiently to accommodate this structure, or a large penalty in L results. An unfortunate consequence is that α now becomes too low to reject noise properly along the "bad" directions. In general terms, the Bayesian solution will tend to allow fluctuations of the same order of magnitude as the deviation of the reconstruction from the initial model.

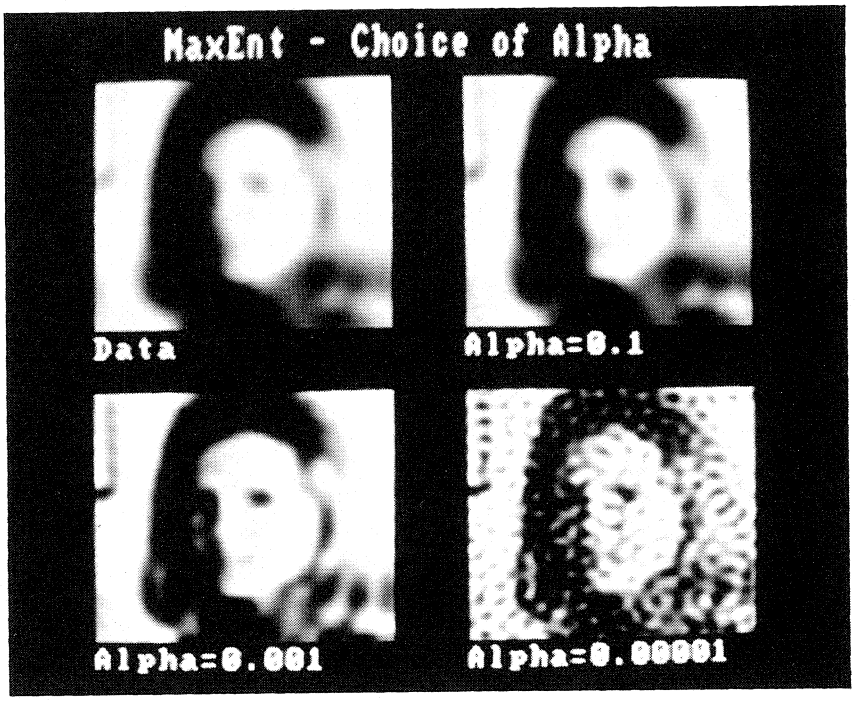

Figure 2. Susie images showing the behavior of reconstruction quality as α is varied.

8. Towards a better model

We have seen that the Bayesian choice of α will often lead to a reconstruction that is over-fitted. Despite this, we feel that this "Classic" choice is the correct answer to the problem that we have so far formulated. In fact, it was the purity of this derivation, combined with problems of its performance that led us to propose the name "Classic" for it. We have derived a joint p.d.f. $\mathrm{pr}\,(D, f, \alpha, \sigma|m)$ which is still conditional on the knowledge of an initial model m. This m was first introduced as a "measure" on the x-space of pixels, but it is a point in f-space and acts as a "model" there. The only freedom that we have left in our hypothesis space is to consider variations in this model, which we recall was a flat, uniform picture set to the average of the data m_0. The fact that the model was flat expresses our lack of prior information about the structure of the picture, but where did the brightness level m_0 come from?

The answer is again: Bayes' Theorem. We expand the hypothesis space to $\mathrm{pr}\,(D, f, \alpha, m_0|\text{ flat})$ and select an uninformative prior for $\mathrm{pr}\,(m_0|\text{ flat})$. The posterior distribution for m_0 (Figure 5) is again sharply peaked and in the Gaussian approximation has a maximum at exactly the mean of the data. Reconstructions using values of m_0 different from this Bayesian optimum exacerbate the over-fitting problem, as one would expect. However, this exercise of varying the model is very instructive, because it emphasizes the cause of the problem; the picture is very non-uniform. There are large areas of the

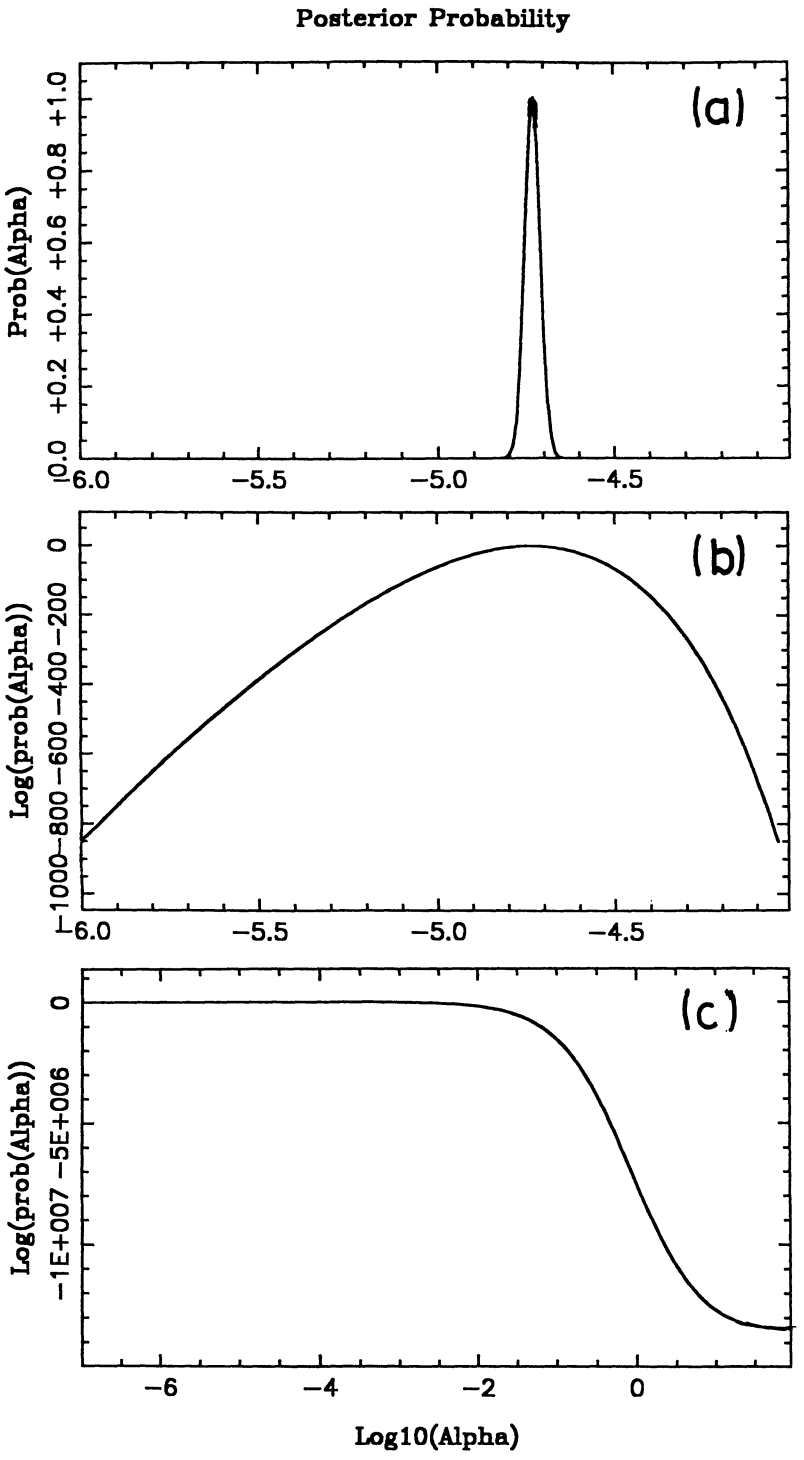

Figure 3. Posterior distribution of the smoothing parameter α for the Susie image, plotted *(a)* logarithmically, *(b)* linearly, *(c)* logarithmically over a large range of α.

picture where the lighting is generally light or dark, with interesting details superimposed. There are correlations from pixel to pixel present in the image that we have so far ignored. Indeed, our earlier MaxEnt Axiom I forbids us to put pixel-to-pixel correlations directly into our prior $\mathrm{pr}(f|m,\alpha)$. We wish to circumvent this axiom, but we must be subtle.

Figure 4. "Classic MaxEnt" reconstruction of Susie.

Suppose we imagine a silly case where the left half of our picture is Susie, but the right half is a distant galaxy. Axiom I is designed to protect us from letting the reconstruction of Susie influence our astrophysics, or vice-versa. But there is nothing stopping us from having a different m_0 level for each half. In fact, in view of the grossly different luminance levels involved, it would be extremely desirable to have different levels of m_R and m_L. When seen this way, there is nothing to prevent us considering the right and left halves of the original Susie picture separately, because the average luminance levels are different. A new hypothesis space involving $\mathrm{pr}(m_R, m_L|\ \mathrm{flat}, L/R)$ will again fix suitable levels for m_R and m_L a posteriori. If there is a strong right/left brightness variation across the picture, then this two-value model will be closer to the reconstruction and $\hat{\alpha}$ will increase, reducing the ripples. But in that case why not use 4 subdivisions (top/bottom, left/right), or 8, or more?

If we continue to subdivide, we can get a better model, closer to the reconstruction, so we expect that $\hat{\alpha}$ will increase. However, we are introducing extra parameters, so

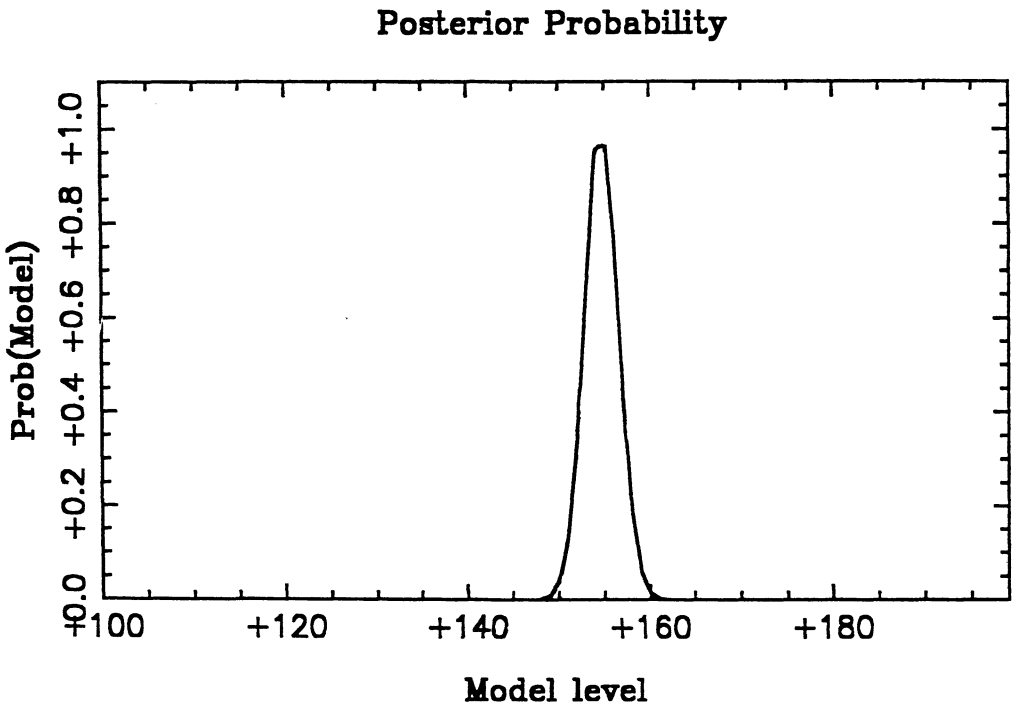

Figure 5. Posterior probability distribution of the initial model level m_0 for the Susie image. The maximum occurs at the mean of the data.

that we would expect there to be a penalty for this, and that it would be likely to have some effect on the choice of $\hat{\alpha}$. A further consideration is that, if at all possible, we should like to avoid the sharp boundaries that such a crude division of the model would involve.

9. New MaxEnt

We are now in a position to formulate a new, flexible hypothesis space that is suitable for pictures such as Susie. We suppose that the model m for use in "Classic" MaxEnt is itself generated from a blurred image of hidden variables $\tilde{m}$:

$$m = \tilde{m} * b = b\tilde{m}, \tag{9.1}$$

where b is our "model-blur" PSF, which can also be written as a circulant matrix B. For the case of Susie we might like to think of $\tilde{m}$ as the source of background lighting. If this model-blur is broad, then our model in "Classic" is smooth, and there are effectively very few parameters in it. If b is narrow, there are many parameters. The shape and width of the model-blur are to be determined by Bayesian methods as well. We do not expect the shape of this blur to matter greatly and we arbitrarily restrict it to be a Gaussian. The crucial parameter is the width and we expect that the most useful width will be about equal to the size of the correlation-length that is actually present in the picture. Our Bayesian analysis of the larger, richer hypothesis space will then tell us how useful is the freedom provided by the hidden variables. The final probability levels

will quantify for us the level of improvement relative to "Classic", which is contained in our new space as limiting cases.

To complete the analysis we must assign a prior for the "pre-model" $\widetilde{m}$. We treat $\widetilde{m}$ as an image and again use the entropic prior:

$$\text{pr}\,(\widetilde{m}|\beta,\ \text{flat}) = Z_T^{-1}\exp(\beta T), \tag{9.2}$$

where $T = S(\widetilde{m},\ \text{flat})$ and we have introduced β as a new Lagrange multiplier for the $\widetilde{m}$-space entropy T. We again restrict ourselves to the mathematically tractable (but still interesting) case of quadratic S and T, circulant blurs and spatially uniform noise level, for which the $\nabla\nabla L$, $\nabla\nabla S$ and $\nabla\nabla T$ matrices are all simultaneously diagonal in Fourier transform space. The Bayesian calculation of $\hat{\alpha}$ and $\hat{\beta}$ now yields:

$$-2\alpha S(\hat{f},\hat{m}) = ndf(s) = \sum_i \beta\lambda_i/(\alpha\beta + \beta\lambda_i + \alpha\lambda_i b_i^2), \tag{9.3}$$

$$-2\beta T(\hat{\widetilde{m}},\ \text{flat}) = ndf(T) = \sum_i \alpha\lambda_i b_i^2/(\alpha\beta + \beta\lambda_i + \alpha\lambda_i b_i^2) \tag{9.4}$$

where b_i^2 are the eigenvalues of B^tB and

$$\log\text{pr}\,(\alpha,\beta,b|D) = \text{constant} + 1/2\sum_i \alpha\beta/(\alpha\beta + \beta\lambda_i + \alpha\lambda_i b_i^2) + \beta T + \alpha S - L. \tag{9.5}$$

The noise level σ can also be estimated as before:

$$\chi^2 = 2L(\hat{f})/\hat{\sigma}^2 = N - ndf(S) - ndf(T). \tag{9.6}$$

Notice how there is once again a neat division of the degrees of freedom between S, T and L.

We have tested the performance of New MaxEnt on the Susie picture. Classic MaxEnt is contained in new Maxent in several ways:

(1) As $\beta \to \infty$, because $\widetilde{m}$ cannot move from the initial m_0.

(2) As $b \to \infty$, because the model becomes flat.

(3) (rather surprisingly) As $b \to 0$. This last case illustrates a general peculiarity of

$$\log\text{pr}\,(\alpha,\beta,b|D) = \text{constant} + \log(\det) + \alpha S + \beta T - L, \tag{9.7}$$

an object which would be known elsewhere in physics as a Gibb's surface. Our new hypothesis space has sufficient structure to contain phase transitions and one such occurs for the Susie image as the width of b is reduced below 4.27 pixels. Below this value of the model-blur, the model is sufficiently detailed to cope with all the structure in the image demanded by the data, and $S(f,m)$ no longer adds anything that is useful. The New MaxEnt $\hat{\alpha}$ increases to infinity at this pint; S switches off and the reconstruction is the model $m = \widetilde{m} * b$. This is illustrated in Figure 6, which shows the posterior distribution of α and β for $b = 3$ and $b = 7$ pixels.

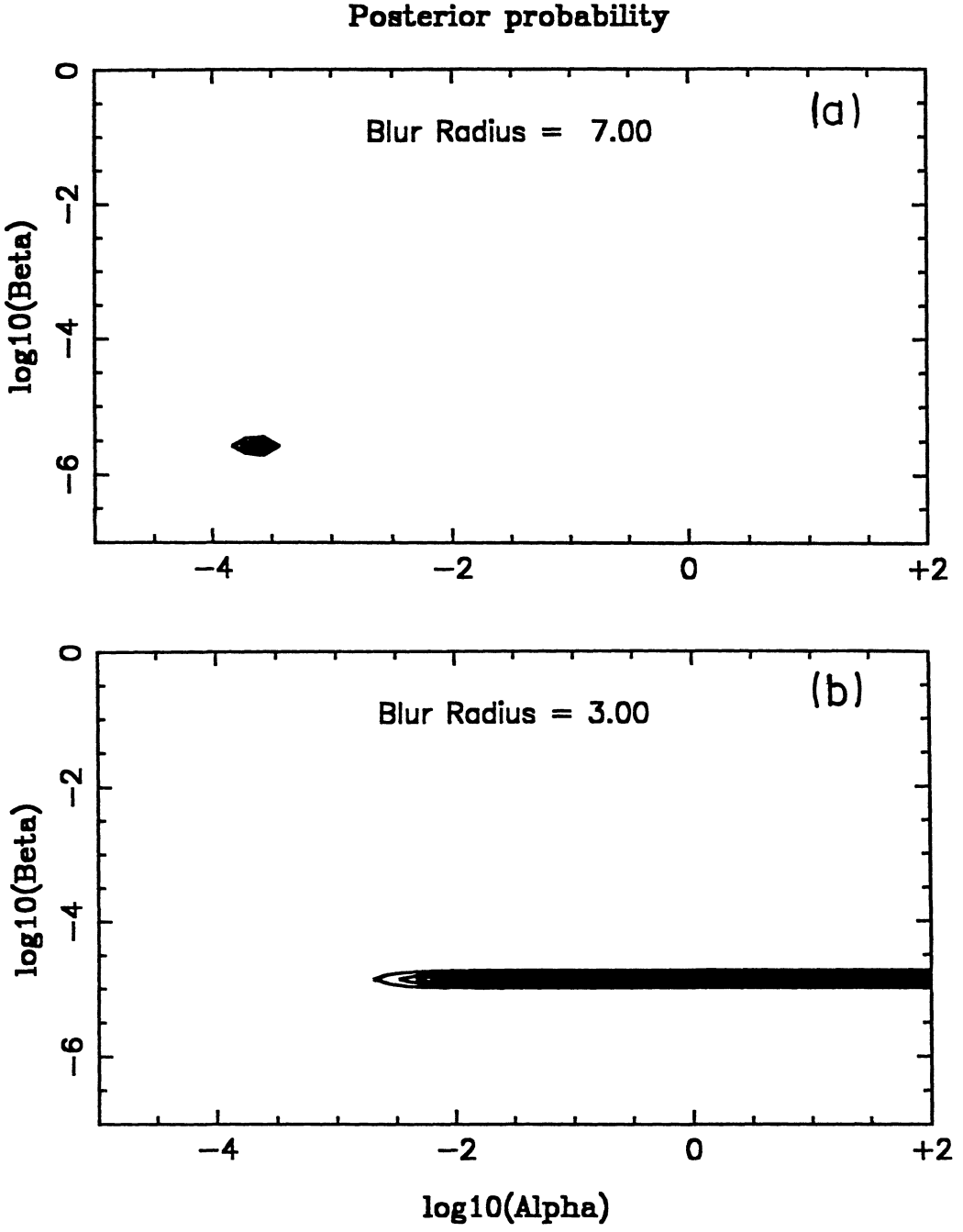

Figure 6. Posterior p.d.f. of the Lagrange multipliers α and β for the New MaxEnt reconstruction of Susie, having $b = 3$ and $b = 7$ pixels. The contours' intervals are linear.

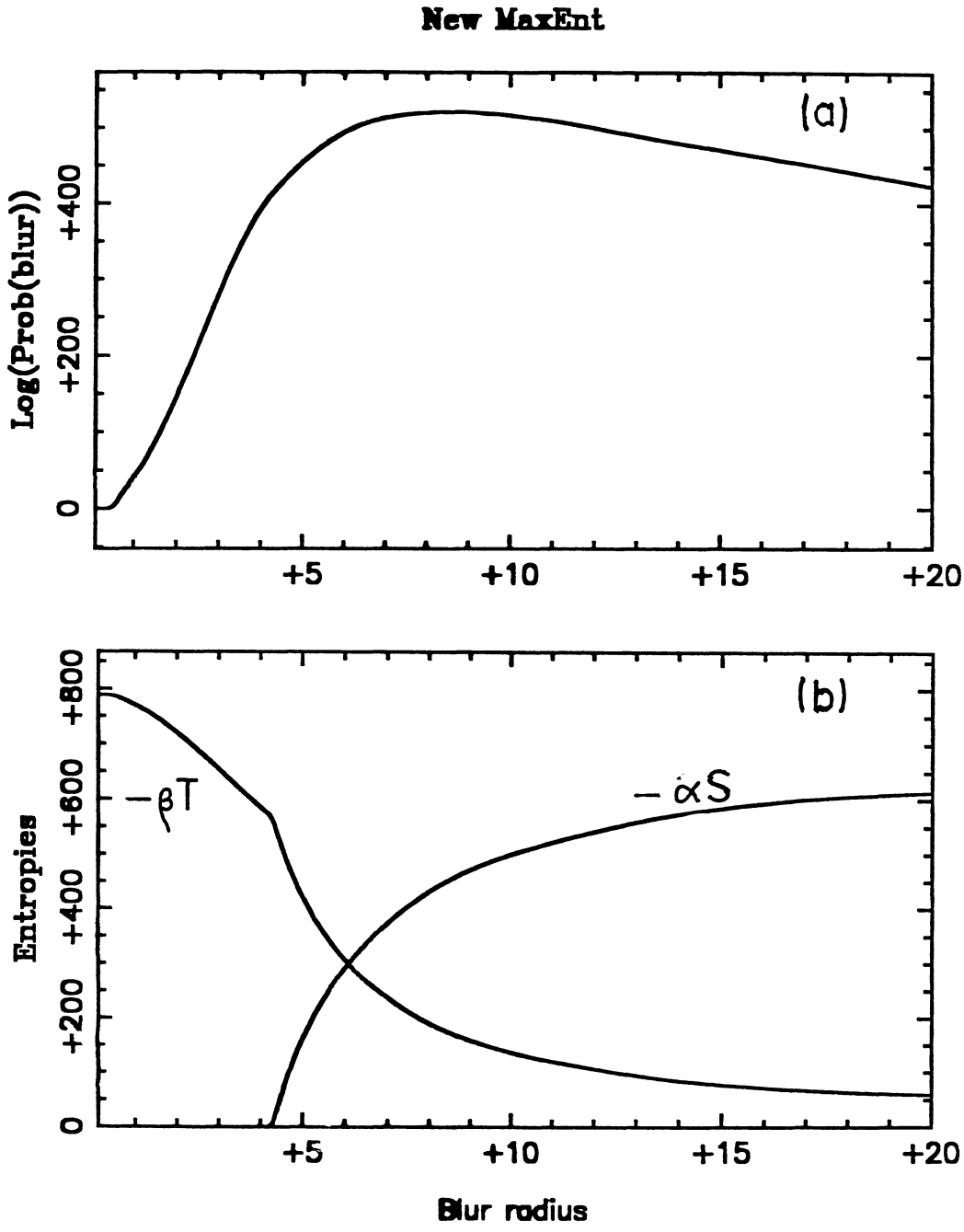

Figure 7. *(a)* Posterior distribution of the model-blur width for New MaxEnt Susie images. *(b)* Image-space entropy S and model-space entropy T. Note that S is zero for model-blurs narrower than 4.27 pixels.

Figure 8. *(a)* Posterior distribution of the model-blur width for New MaxEnt Susie Images. *(b)* Image-space entropy S and model-space entropy T. Note that S is zero for model-blurs narrower than 4.27 pixels.

Figure 7 shows the posterior distribution of the width of b, which rises to a maximum at $\sim$ 8.5 pixels. This diagram also answers the question of how useful our new hypothesis space is. It is useful to the extent of being more probable than Classic MaxEnt by exp(520). The extrinsic variables S, T and χ^2 are also plotted, showing a change of slope at the phase transition. There is no specific heat associated with this phase change! Inspection of the reconstruction and effective model $m = \widetilde{m} * b$ for the optimum width of b (Figure 8) confirms that the New MaxEnt has indeed achieved its promise.

Of course, our New MaxEnt can be used to encourage smoothness in any image, whether or not it is actually blurred. Indeed, our failure to offer a solution the problem to analyzing noisy, but unblurred pictures has been a continual source of frustration over the years. We test the noise-smoothing properties of the method with a picture of Susie which is in focus, but which has had 25 units of noise added. For this type of problem, the Classic MaxEnt reconstruction is almost identical to the data. The best value of the model blur is now $\sim$ 3 pixels, and it can be seen from Figure 8 that there is an increase in probability of exp(10000) over Classic for this case. The picture produced (Figure 9) is also very good, and shows all the structure that can be reliably produced from this noisy dataset. A detail from this (Figure 10) confirms that the pixel-to-pixel noise has been greatly reduced, without degrading the information content of the picture in any

way.

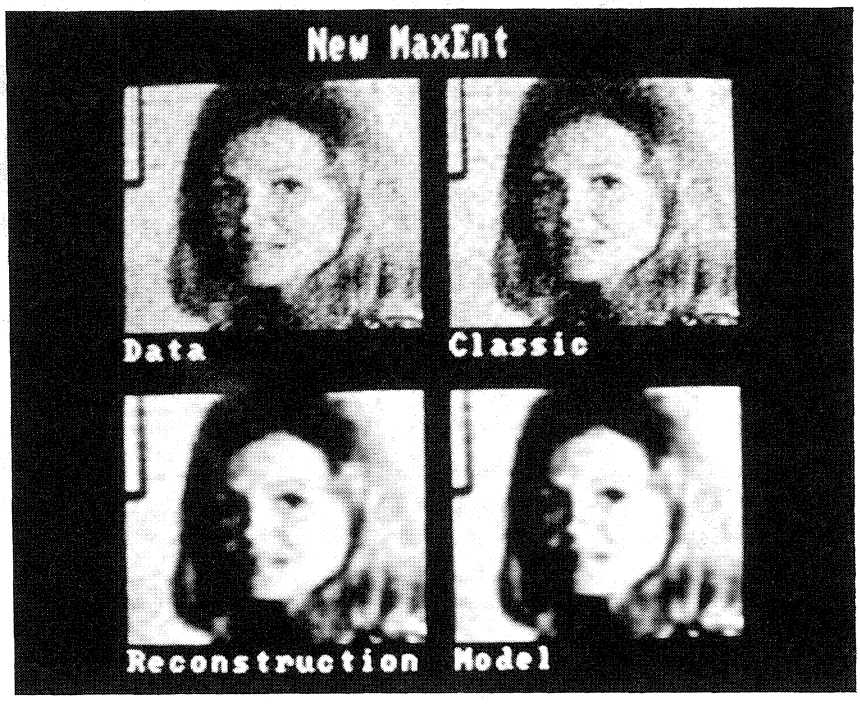

Figure 9. Comparison of Classic and New MaxEnt reconstruction of a noisy Susie picture.

10. Discussion

Our New MaxEnt approach is related to other methods of introducing spatial smoothing that have been found useful in practice. Within the context of maximum entropy image processing, there are now many examples of "reconstruction-dependent" models $m(\hat{f})$. A particularly successful application to tomographic mapping of stellar accretion discs is presented by Marsh and Horne (1988), following Horne (1985). To improve the quality of the images, they used a model that was a blurred form of their current reconstruction. We have also found such techniques useful: Charter & Gull (1988) give an example of studies of drug absorption rate into the bloodstream, in which a blurred version of the reconstruction is again used as the model.

Such tricks have previously lacked any rigorous justification, because the development of the MaxEnt story treats m as a point in f-space that is given a priori. It was thus difficult to see how we could legally let it depend on $\hat{f}$. However, in New MaxEnt, the effective model m looks very much like a blurred version of $\hat{f}$, although it is actually a blurred version of the hidden variables $\widetilde{m}$. We can now justify the above tricks in

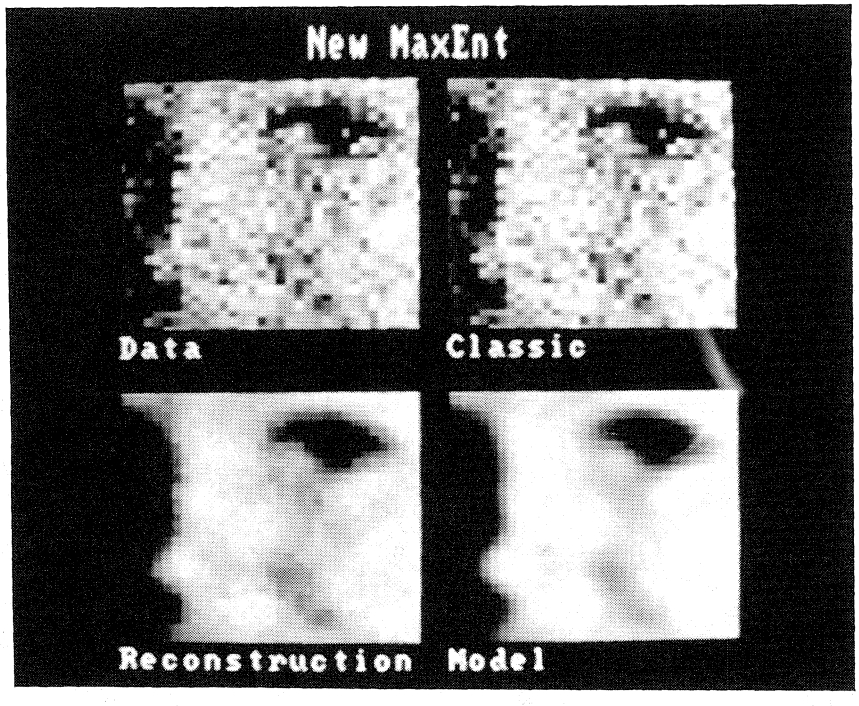

Figure 10. Detail of Figure 9, showing the improvement due to noise suppression.

terms of New MaxEnt. Thus in the drug absorption problem, f represents the rate of absorption into the bloodstream, $\widetilde{m}$ is the rate at which the tablets break down in the stomach, and b represents the time delay as the drug passes through the liver. Charter (private communication) also gives another, intriguing example, in which he simply pretends that the data are more blurred than is actually true, adding an additional "pre-blur" to the real PSF. Often the results are improved by this device, encouraging smoothness and eliminating noise. We can now see that this trick too is covered in New MaxEnt as the degenerate case $\alpha \to \infty$ that occurs in the case of Susie for small model-blurs. The New MaxEnt hypothesis space provides a natural justification for these variants, and automatically includes any consequential effect upon the stopping criterion due to the additional parameters in the model.

It is also useful to examine our new procedure in the context of spatial statistics. Indeed, much of our motivation for the New MaxEnt was provided at this very meeting where, although our practical results were well received, the MaxEnt Axiom I was considered unhelpful, to say the least. In this field of spatial statistics, the currently favoured techniques are things such as Markov random fields (Kinderman and Snell 1980, Geman & Geman 1984) and smoothness-enforcing regularizers (Titterington 1985). We can compare New MaxEnt with these techniques by marginalizing out $\widetilde{m}$ to get an effective prior for $\mathrm{pr}\,(f|\alpha,\beta,b,\ \mathrm{flat})$. We have not so far done this, because it would obscure the real structure of our hypothesis space, which is still faithful to the spirit of

Axiom I. When we do it, we find

$$\mathrm{pr}\,(f|\alpha,\beta,b,m_0) \propto \exp -1/2\delta f^t R^{-1}\delta f, \tag{10.1}$$

where $\delta f_i = f_i - m_0$ and R is a circulant matrix that has eigenvalues $1/\alpha + b^2/\beta$.

By varying the shape of the model-blur b we can clearly mimic any given spectral behavior of spatial smoothing. Markov random fields correspond to particular functional forms of b. New MaxEnt contains these techniques as special cases. However, we prefer the rationale of our new hypothesis space, because we feel it is more closely related to our prior knowledge of the imaging problem.

11. Conclusions

The Bayesian choice of the regularizing parameter α completes the derivation Classic MaxEnt and represents a major advance over our previous practice of setting $\chi^2 = N$. The resulting formula $-2\alpha S = ndf(S)$ is theoretically appealing, and expresses the fact that the amount of structure produced in the reconstruction is equal to the number of good, independent measurements present in the dataset.

For some problems we have found the Classic stopping criterion to be satisfactory, but there are general grounds for supposing that it leads to overfitting, because α has to be reduced to allow for the structure produced by good data. This leads to undersmoothing of bad data, as we have illustrated with our picture of Susie.

The New MaxEnt hypothesis space which incorporates spatial correlations is sufficiently powerful to correct these problems and is considerably more probable than Classic, showing that the inclusion of spatial information is useful.

New MaxEnt also provides a consistent rationale for a wide class of model manipulations that are found to be useful in practical applications. Although we have, for reasons of computational expediency, illustrated the New MaxEnt only in the quadratic (Wiener filter) approximation, the results are already excellent. We do not expect our conclusions to change when the correct entropic forms are used, indeed the results can only improve.

Finally, we ask the question: "Is our hypothesis space good enough?" Of course, the answer depends on what we are trying to achieve. Certainly our new procedure is good enough to overcome the over-fitting problems of Classic MaxEnt and produce a good reconstruction of Susie. However, looking at the images produced for different values of the model-blur width, our eyes tell us that the reconstruction for $b = 5$ pixels is visually slightly better than that for the Bayesian optimum $b = 8.5$ pixels, although the probability of $b = 5$ is lower by exp(50). This is a warning that we may eventually find another, deeper hypothesis space even more useful for the imaging problem (as envisaged by Jaynes 1986). We speculate that the improvement we get by going to $b = 5$ tells us something about human vision. We pay attention to the fine details present in Susie's face and relatively ignore the background. The computer, with its spatially-invariant model PSF sees the smooth surfaces in the background and weights them equally, thereby arriving at a slightly large correlation length than our eyes would like.

Acknowledgements

We are grateful to all the past and present members of the Cambridge MaxEnt Group for discussions about the MaxEnt stopping criterion during the last twelve years. This work was partly supported by Maximum Entropy Data Consultants Ltd.

References

Aczel, J. (1966). Lectures on functional equations and their applications, Section 6.2, Academic Press, New York.

Charter, M. K. & Gull, S. F. (1988) Maximum entropy and its application to the calculation of drug absorption rates. J. Pharmacokinetics and Biopharmaceutics, *15*, 645–655.

Cox, R. P. (1946). Probability, frequency and reasonable expectation. Am. Jour. Phys. *17*, 1–13.

Daniell, G. J. & Gull, S. F. (1980). Maximum entropy algorithm applied to image enhancement, IEEE Proc., *127E*, 170–172.

Davies, A. R. & Anderssen, R. S. (1986). Optimization in the regularization of ill-posed problems. J. Austral. Math. Soc. Ser. B., *28*, 114–133.

Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Trans *PAMI-6*, 721–741.

Gull, S. F. & Daniell, G. J. (1978). Image reconstruction from incomplete and noisy data. Nature, *272*, 686–690.

Gull, S. F. & Skilling, J. (1984a). Maximum entropy method in image processing. IEE Proc. *131 (F)*, 646–659.

Gull, S. F. & Skilling, J. (1984b). The maximum entropy method. In Indirect Imaging, ed. J. A. Roberts, pp 267–279. Cambridge University Press.

Horne, K. (1985) Images of accretion discs I: The eclipse mapping method, Mon. Not. R. astr. Soc., *213*, 129–141.

Jaynes, E. T. (1978). Where do we stand on maximum entropy? *Reprinted in* E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics, ed. R. Rosenkrantz, pp 211–314. Dordrecht 1983: Reidel.

Jaynes, E. T. (1986). Bayesian methods: general background. In Maximum Entropy and Bayesian Methods in Applied Statistics. ed. J. H. Justice., pp 1–25. Cambridge University Press.

Kinderman, R. & Snell, J. L. (1980) Markov random fields and their applications. Amer. Math. Soc. Providence, RI.

Klir, G. J. (1987). Where do we stand on measures of uncertainty, ambiguity, fuzziness and the like? Fuzzy sets and systems, *24*, 141–160.

Levine, R. D. (1986). Geometry in classical statistical thermodynamics, J. Chem. Phys., *84*, 910–916.

Marsh, T. R. & Horne, K. (1988) Maximum entropy tomography of accretion discs from their emission lines. In Maximum Entropy and Bayesian Methods: Cambridge 1988, ed. J. Skilling, Kluwer (in press)

Shore, J. E. & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans. Info. Theory, *IT-26*, 26–39 and *IT-29*, 942–943.

Skilling, J. (1988a). The axioms of maximum entropy. In Maximum Entropy and Bayesian Methods in Science and Engineering, Vol. 1., ed. G. J. Erickson & C. R. Smith, pp. 173–188. Kluwer.

Skilling, J. (1988b). The eigenvalues of mega-dimensional matrices. In Maximum Entropy and Bayesian Methods: Cambridge 1988, ed. J. Skilling, Kluwer (in press).

Tikhonov, A. N. & Arsenin, V. Y. (1977). Solutions of ill-posed problems. Wiley, New York.

Tikochinsky, Y., Tishby, N.Z. & Levine, R. D. (1984). Consistent inference of probabilities for reproducible experiments. Phys. Rev. Lett., *52*, 1357–1360.

Titterington, D. M. (1985) General structure of regularization procedures in image reconstruction. Astron. Astrophys. *144*, 381–387.