

Massive Inference and Maximum Entropy

John Skilling

Department of Applied Mathematics and Theoretical Physics
University of Cambridge
England CB3 9EW

January 1998

Abstract

In data analysis, maximum entropy (MaxEnt) has been used to reconstruct measures (*i.e.* positive, additive distributions) from limited data. The MaxEnt prior was originally derived from the “monkey model” in which quanta of uniform intensity could appear randomly in the field of view. To avoid undue digitisation, the quanta had to be small, and this led to difficulties with the Law of Large Numbers, and to unavoidable approximations in computing the posterior. A better way of avoiding digitisation is to give the quanta variable intensity with an exponential prior, that being the natural MaxEnt assignment. We call this technique “Massive Inference” (MassInf). Although the entropy formula no longer appears in the prior, MassInf results show improved quality. MassInf is also capable of assigning a simple prior for polarized images.

Key words: Maximum entropy, infinitely divisible, polarization, regularization.

1 History of maximum entropy

As presented by Jaynes (1957), the Principle of Maximum Entropy (PME) is a rule for assigning probability distributions: in making inferences on the basis of partial information we are to use that probability distribution which has maximum entropy subject to whatever ensemble-average constraints are known. Given some mean values

$$\sum_i R_{ki} p_i = D_k, \quad k = 1, 2, \dots \quad (1)$$

that include normalisation $\sum p_i = 1$, the probability distribution p is to be assigned by maximising its entropy

$$S = - \sum p_i \log p_i \quad (2)$$

(Shannon, 1948) subject to the given constraints. In statistical mechanics, the entropy S is derivable from the combinatoric number of ways

$$\Omega = N! / \prod n_i! \approx \exp(S), \quad S(n) = - \sum n_i \log(n_i/N) \quad (3)$$

of dividing an ensemble of N systems into cells with occupation n_i having mean Np_i . Alternatively, if individual mean values m_i are assigned, each occupation number can be given a Poisson distribution with correct mean, leading to

$$\Pr(n) = \prod e^{-m_i} m_i^{n_i} / n_i! \approx \exp(S), \quad S(n) = \sum (n_i - m_i - n_i \log(n_i/m_i)) \quad (4)$$

The PME amounts to the entirely reasonable prescription of giving equal intrinsic weight to each individual state. When the constraints D include a defining set of physical variables such as energy and volume, the maximum entropy distribution gives accurate predictions of other average quantities with different R . Fluctuations are generally $\mathcal{O}(N^{-1/2})$ small, and deviations larger than this indicate an unacknowledged extra constraint.

It is tempting and productive to apply this successful formalism to data analysis, and this has been done both directly and indirectly. Indirect applications use PME to assign the posterior probability distribution $\Pr(f|D)$ of the quantity f being sought. It is supposed that the data are ensemble-average constraints $D_k = \int R_k(f) \Pr(f) df$, so that the PME becomes applicable. The commonest such application is the derivation of a power spectrum from autocorrelation coefficients D of a time-series f . Actually, data tend to be observations of the particular object being investigated at the time, and the proper analysis is Bayesian. A prior probability $\Pr(f)$ needs to be assigned (by PME or other insight), but it is then not determined by the data, whose rôle is to modulate the prior through the usual likelihood function $\Pr(D|f)$.

For a direct application of MaxEnt, we suppose that we seek the distribution f of some positive, additive quantity such as the intensity of light in an image, or the flux of energy along a spectrum. Mathematicians call such an object a “measure”. We then proceed with the “monkey model” (Gull and Daniell, 1978) in which f is identified with a number n of quanta of some unknown but presumed small strength q . With m similarly rescaled, Stirling’s approximation for large n yields the Quantified Maximum Entropy (QME) prior

$$\Pr(f) \propto \exp(\alpha S) / \prod f_i^{1/2}, \quad S(f) = \sum (f_i - m_i - f_i \log(f_i/m_i)) \quad (5)$$

where m is a set of weights that models any *a priori* non-equivalence of the cells i , and $\alpha = 1/q$ is an unknown but apparently large hyper-parameter. When applying this, α cannot in fact be particularly large, lest the prior dominate the likelihood. Fortunately, the same entropy form can be derived from symmetry arguments which avoid combinatoric modelling (Shore

and Johnson, 1980). Related arguments (Skilling, 1989) suggest the QME prior, implicitly assuming that the desirable entropy maximum should yield a useful selection from the posterior distribution.

Were the data observations $D_k = \sum R_{ki} f_i$ to be exact, the PME could be used to assign a corresponding f , just as in statistical mechanics. Even if the data are noisy and thus subsumed into a likelihood function $\Pr(D|f)$, the PME can still be used to assign a particular f from among those with some acceptable fit $\Pr(f) \geq P_0$ to the data. Visually, distributions assigned by maximum entropy are often of high quality and utility. Entropy is a good regulariser.

Yet the reliability of the inferred f cannot be deduced from any single assignment, however distinguished its provenance. To deal with uncertainty, we must use the probability distribution in 5, and not just regularise. In each particular application, α is to be estimated probabilistically (Gull, 1989). The “best” individual result \hat{f} is then taken to be that which maximises the product $\exp(\alpha S) \Pr(D|f)$, so it remains a PME selection with the corresponding visual appeal. Around it lies a probability distribution from which the uncertainty of any inference can be calculated. This methodology has been used with considerable success, but limitations have become apparent.

2 MaxEnt Polarization

Suppose the object f being inferred is the intensity pattern $I(x, y)$ of light across an image. Light can be polarized, most simply into potentially independent linear channels $X(x, y)$ and $Y(x, y)$ that sum to $I = X + Y$. Each individual polarization represents an observable spatial pattern to which the symmetries of QME plausibly apply, so that it is natural to assign QME priors to each of X and Y :

$$\Pr(X) \propto e^{\alpha(S(X))}, \quad \Pr(Y) \propto e^{\alpha(S(Y))} \quad (6)$$

The prior on the total I is then determined through

$$\Pr(I) = \iint \delta(I - X - Y) \Pr(X) \Pr(Y) dXdY \quad (7)$$

This does **not** evaluate to a QME form for I (even if the value of α differs), even though it would have been just as natural to insist on the latter assignment. Worse, if the prior on I is assigned by QME, it becomes impossible to find **any** subsidiary probability distribution that can be assigned to X and Y to yield the combination I . A subsidiary distribution can be found numerically, but parts of it are negative, which is impossible.

3 MaxEnt Pixellation

The QME prior is defined over some mesh of M cells that must be fixed in advance. It can happen that no choice of M allows a sensible reconstruction. For example, suppose we wish to infer a distribution $f(x)$ from inexact data on the lowest-order moments

$$\mu_0 = \int_0^1 f(x) dx, \quad \mu_1 = \int_0^1 xf(x) dx \quad (8)$$

The range $x \in [0, 1]$ is to be divided into M equal cells, within each of which f will be approximated by a single central value. Specifically, let the likelihood function be

$$\Pr(\text{Data} | f) = e^{-64\mu_1} (e^{-\mu_0} - e^{-2\mu_0})^8 \quad (9)$$

In detail, this form is chosen to enable a M -dimensional probability integral to be computed without approximation as the binomially-expanded sum of 9 integrals each of which is M -fold separable. In broad outline, the likelihood factors indicate a measurement of mean location

$$\langle x \rangle = \mu_1 / \mu_0 = 0.022 \pm 0.025 \quad (10)$$

along with normalisation $\mu_0 \approx 1$. Clearly, at least 40 or so cells are needed in order to make such a small mean x visible.

Unfortunately, the law of large numbers comes into play as the number of cells increases, and the prior forces $\langle x \rangle$ ever closer to $\frac{1}{2}$. Specifically, the prior variance is

$$\text{var}(x) \leq \frac{1}{6M} \quad \forall \alpha \quad (11)$$

Consequently, when M is 40, the “measurement” of $\langle x \rangle$ has already become an improbable 7σ outlier, according to the prior. The posterior fails to cover it well and the prior predictive “evidence” value becomes small. In fact, the favoured value of M is 4, which is much too coarse to fit the data well (see Fig.1). In our analysis, we should seek a method that behaves sensibly as the division into cells becomes finer. QME does not have this property.

This was not noticed at first because practical computations approximated the probability summit with a Gaussian, and Gaussian distributions do give error bars with sensible limits as the number of cells increases. Unfortunately, the approximation becomes indefinitely bad in this limit because it is being used out to $\mathcal{O}(f^{1/2})$ whereas f cannot decrease by more than itself without going negative, and typical f values are becoming small.

4 Massive Inference (MassInf)

When we seek a measure F , distributed with density f over a domain α of x , both the observed data and any subsequent inferences take the form of

integrals

$$F(\alpha) = \int_{x \in \alpha} R(x) f(x) dx \quad (12)$$

or, at worst, functions of such integrals. Indeed, the main aim of this branch of data analysis is to use imperfectly observed values of some integrals in order to predict others. An integral over a finite x -domain can be defined as the limit of the sum over its division into cells, as all these M cells become arbitrarily small — it being assumed that the limit exists. Hence the probability distribution for such an integral must be the limit of the convolution of the individual probability distributions $\text{pr}(F_i)$ over the cells: a consistent family of probability distributions over sub-domains is called a “process”, and a process that allows an arbitrary degree of sub-division is called “infinitely divisible”.

$$F(\alpha) = \lim \sum_{i=1}^M F_i, \quad (13)$$

$$\text{Pr}(F|\alpha) = \lim \int \cdots \int dF_1 \cdots dF_M \text{pr}(F_1) \cdots \text{pr}(F_M) \delta(F_1 + \cdots + F_M - F)$$

In writing this without extra conditionalities, we have already assumed spatial independence. It is convenient to assume spatial invariance also, so that the prior over any finite domain can be built from a single common prior assigned to an arbitrary microscopic cell. Analytically, the required convolution is conveniently performed by multiplying the Laplace transforms:

$$\widetilde{\text{Pr}}(s|\alpha) = \widetilde{\text{pr}}_1(s) \cdots \widetilde{\text{pr}}_M(s) = (\widetilde{\text{pr}}(s))^M \quad (14)$$

In macroscopic units, the mean flux associated with a small cell must be correspondingly small, $\mathcal{O}(M^{-1})$. So must its variance, which also divides equally among the cells so that it becomes $\mathcal{O}(M^{-1})$ small, and the standard deviation considerably exceeds the mean. This indicates that there will be a small probability of having a finite $\mathcal{O}(1)$ flux in a small cell, rather than a substantial $\mathcal{O}(1)$ probability of a small flux: this plausibility argument can be extended into a proof that most of the flux will be tightly localised. Accordingly, the flux in a small cell dx might be defined from a density $\lambda(F)$ through

$$\text{pr}(F|dx) = (1 - \Lambda dx) \delta(F) + \lambda(F) dx + \mathcal{O}(dx)^2 \quad (15)$$

where $\Lambda = \int_0^\infty \lambda(F) dF$ ensures normalisation. Laplace convolution up to a range α induces the macroscopic transform

$$\widetilde{\text{Pr}}(s|\alpha) = \exp\left(-\int_0^\infty F^{-1} L(F) (1 - e^{-sF}) dF\right) \quad (16)$$

known as the Lévy–Khinchin representation, $L(F) = \alpha F \lambda(F)$ being the Lévy measure. Actually, the latter integral can converge meaningfully to

Table 1: Analytic Lévy–Khinchin Representations.

	$L(F)$ $L(\beta F)e^{-\gamma F}$	$\Pr(F \alpha) = P(F)$ $P(\beta F)e^{-\gamma F} / \int_0^\infty P(\beta F)e^{-\gamma F} dF$
1	$\alpha\delta(F)$	$\delta(F - \alpha)$
2	$\alpha\delta(F - 1)$	$e^{-\alpha} \sum_{n=0}^\infty \frac{\alpha^n}{n!} \delta(F - n)$
3	$\alpha F e^{-F}$	$e^{-\alpha} \left(\delta(F) + e^{-F} \sqrt{\frac{\alpha}{F}} I_1(2\sqrt{\alpha F}) \right)$
4	$\frac{\alpha F^m e^{-F}}{(m-1)!}$	$e^{-\alpha} \left(\delta(F) + \frac{\alpha F^{m-1} e^{-F}}{(m-1)!} {}_0F_m \left(\frac{m+1}{m}, \frac{m+2}{m}, \dots, 2; \frac{\alpha F^m}{m^m} \right) \right)$
5	αe^{-F}	$\frac{F^{-1+\alpha} e^{-F}}{\Gamma(\alpha)}$
6	$\alpha I_0(F) e^{-F}$	$\frac{\alpha}{F} I_\alpha(F) e^{-F}$
7	$\frac{\alpha}{2\sqrt{\pi}} F^{-1/2}$	$\frac{\alpha}{2\sqrt{\pi F^3}} e^{-\alpha^2/4F}$
8	$\frac{\alpha \Gamma(1/3)}{2\sqrt{3}\pi} F^{-1/3}$	$\frac{\alpha^{3/2}}{3\pi F^{3/2}} K_{\frac{1}{3}}(2\sqrt{\alpha^3/27F})$
9	$\alpha e^F \operatorname{erfc}(\sqrt{F})$	$\sqrt{\frac{8}{\pi}} \alpha (2F)^{\alpha-1} e^{F/2} \mathcal{D}_{-2\alpha-1}(\sqrt{2F})$
10	$\alpha e^{-aF} + \beta e^{-bF}$	$\frac{\alpha^\alpha \beta^\beta}{\Gamma(\alpha+\beta)} F^{-1+\alpha+\beta} e^{-bF} M(\alpha; \alpha + \beta; (b-a)F)$
11	$\alpha e^{-aF} + \alpha e^{-bF}$	$\frac{\sqrt{\pi}(ab)^\alpha}{\Gamma(\alpha)} \left(\frac{F}{b-a} \right)^{\alpha-1/2} e^{-(a+b)F/2} I_{\alpha-\frac{1}{2}} \left(\frac{b-a}{2} F \right)$
12	$\alpha e^{-F} + \beta F e^{-F}$	$\left(\frac{F}{\beta} \right)^{(\alpha-1)/2} e^{-\beta-F} I_{\alpha-1}(2\sqrt{\beta F})$
13	$\alpha e^{-F} + \beta e^{-F}/\sqrt{\pi F}$	$\frac{2^\alpha}{\sqrt{2\pi}} F^{-1+\alpha} e^{2\beta-F-\beta^2/2F} \mathcal{D}_{1-2\alpha}(\beta\sqrt{2/F})$

define an infinitely divisible process even if λ is non-integrable: a process can exist at all finite resolutions without the infinitesimal limit necessarily having the form 15. Even so, 16 is restrictive. Many prior distributions, including QME, cannot be written in Lévy–Khinchin form, and consequently cannot be resolved onto arbitrarily small cells. A standard professional reference for this material is Feller (1971).

Table 1 lists the forms that have been found with a reasonably convenient algebraic form, with the available rescaling options in the top line. Form 1 is the trivial special case of a distribution F of rigidly constant density. Form 2 is a basic Poisson distribution in which “atoms” of fixed (unit) flux are placed randomly in x . Form 3 is a Poisson variant in which the atoms have an exponential distribution of flux. Form 4 generalises the Poisson variant 3 to a model in which atoms have $m \geq 1$ exponentially-distributed components (${}_0F_m$ is a hypergeometric function). Form 5 is a Gamma distribution, particularly popular among statisticians. Form 6 arises in the study of random walks on an integer lattice. Form 7 is known as the Lévy distribution. L being a simple power law, form 7 has the “stable” property (Lukacs, 1970) that self-convolution recovers the same shape. Form 8 is another stable distribution, apparently overlooked previously. Perhaps understandably, form 9 also appears to be absent from the literature (\mathcal{D} is the parabolic cylinder function). Forms 10 to 13 are combinations of the above that happen to

have closed form, 11 being a special case of 10. Of these, only the Poisson forms 2, 3 and 4 have the infinitesimal interpretation 15 with integrable λ .

The basic Poisson form 2 is, in fact, the original “monkey model” without any large- n approximation. However, it restricts all fluxes to integer multiples of some unit quantum, whereas we usually require flux to be on a continuous scale. To allow individual fluxes to be continuous, the natural prior distribution for a single atom is exponential

$$\lambda(F) = q^{-1} e^{-F/q} \quad (17)$$

because that is the PME assignment appropriate to a constraint q on the expected flux. This yields the Poisson variant form 3, now seen to be the natural prior distribution for a measure F . It has two hyper-parameters, q for the quantum size, and the Poisson mean number of atoms per unit range x (which could be a function of x). The associated macroscopic distribution over a range on which α atoms are expected is

$$\Pr(F|\alpha, q) = e^{-\alpha} \left[\delta(F) + e^{-F/q} \sqrt{\frac{\alpha}{qF}} I_1 \left(2\sqrt{\alpha F/q} \right) \right] \quad (18)$$

Interestingly, the delta function ensures that the posterior mode follows the prior mode at $F = 0$: unless the data utterly prohibit it, the most probable individual measure F will be null. This destroys the supposition of QME that the maximum might be a useful selection from the posterior distribution. Instead, the mean of the posterior is used for display purposes. Of itself, the mean lacks the symmetry properties that underlie a MaxEnt selection, but it is a sufficient statistic for reading off the mean value of any integral property of f . To determine deeper information such as uncertainty, the posterior distribution must be recorded more fully, usually as a set of random samples.

Because the flux in any sample is located as a set of “point mass” delta functions, we call 17/18 the “Massive Inference” (MassInf) prior. Computationally, the MassInf prior is convenient because each sample of F is fully defined by a finite number of atoms, whilst other forms have flux (mostly very small) everywhere. Also, the exponential prior on each atomic flux is conjugate to the Gaussian likelihood function that is commonly appropriate, so it can be folded in without un-necessary algorithmic complexity.

5 MassInf Pixellation

By construction, MassInf has no difficulty with pixellation. Any finite resolution with a restricted number of cells may damage the results, but inferences will have a well-defined continuum limit as the resolution increases. This is confirmed with the example 9 above, for which the upper curve in Fig.1 shows the prior predictive “evidence” values for various numbers of cells. As

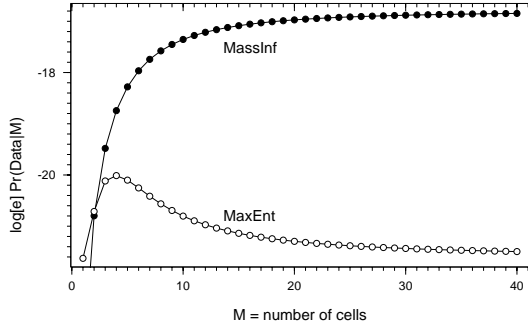


Figure 1: Test comparison of MaxEnt and MassInf “evidence” values (hyper-parameters being given their optimal values).

it ought, this increases gently towards the continuum limit, as opposed to the awkward behaviour of QME. Towards this limit, the posterior behaves sensibly.

6 MassInf Polarization

Even if an image has the extra complication of polarization, we can still model it with a Poisson distribution of point atoms, in the spirit of the original monkey model. Instead of simply assigning an exponential prior on a single intensity, though, an atom now has polarized structure on which a prior is needed.

Among the oscillating electric fields E_x and E_y comprising a beam of radiation along z , there can be up to four different correlation coefficients among the in-phase and out-of-phase components. These define four Stokes’ parameters I, Q, U, V (Stone, 1963). Of these, I is the (non-negative) total intensity. The other three are smaller and obey

$$P \equiv \sqrt{Q^2 + U^2 + V^2} \leq I \quad (19)$$

Geometrically, they lie within the “Poincaré sphere” of radius I . They can be rotated into each other by harmless coordinate rotation and time offsets, so we should assign the prior uniformly over the sphere.

$$\Pr(I, Q, U, V \mid 1 \text{ atom}) = \theta(I, P) \quad (20)$$

Let us restrict our attention to the special case $U = V = 0$. The polarization state now has two channels, linear x with intensity $X = \frac{1}{2}(I+Q)$, and linear

y with intensity $Y = \frac{1}{2}(I - Q)$. These states being orthogonal, it is natural to assign exponential priors $\Pr(X) = e^{-X}$ and $\Pr(Y) = e^{-Y}$ to each as in 17, scaling to $q = 1$ for convenience. In terms of I and Q , this sets

$$\Pr(I, Q | U = V = 0, 1 \text{ atom}) = e^{-I}/2, \quad |Q| \leq I \quad (21)$$

This implies that $\theta(I, P) \propto e^{-I}$. After normalising to unit total, we reach

$$\Pr(I, Q, U, V | 1 \text{ atom}) = \begin{cases} e^{-I}/8\pi & \text{if } P \leq I; \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

With the general prior in hand, we can now investigate special cases.

1. Fixed polarization state: $Q/I, U/I, V/I$ all known.

$$\Pr(I | \text{fixed polarization, 1 atom}) = e^{-I} \quad (23)$$

Lacking any sub-structure, the atom has a simple exponential prior.

2. Observe linear polarizations from arbitrarily polarized radiation: seek X and Y after marginalizing over U and V .

$$\Pr(X | 1 \text{ atom}) = X e^{-X} \text{ and independently } \Pr(Y | 1 \text{ atom}) = Y e^{-Y} \quad (24)$$

Each of X and Y show the effect of internal structure, as if each was the convolution of two independent exponentially-distributed fluxes. The corresponding macroscopic prior can be found by setting $m = 2$ in form 4 of Table 1.

3. Observe the total intensity from arbitrarily polarized radiation: seek I after marginalizing over Q, U and V .

$$\Pr(I | 1 \text{ atom}) = I^3 e^{-I}/6 \quad (25)$$

The intensity appears as if it were the convolution of four independent exponentially-distributed fluxes, though these cannot be identified individually among the four polarization parameters. The corresponding macroscopic prior can be found by setting $m = 4$ in form 4 of Table 1.

7 Examples

We present a simulation and an application to practical data. The simulation is due to Bretthorst (1990) and represents data

$$D_t = 100e^{-0.03t} + 50e^{-0.05t} + e_t, \quad t = 1, 2, \dots, 100 \quad (26)$$

where e is unit normally distributed noise. These data (Fig.2) are to be analyzed to recover the spectrum $f(x)$ of decay rates.

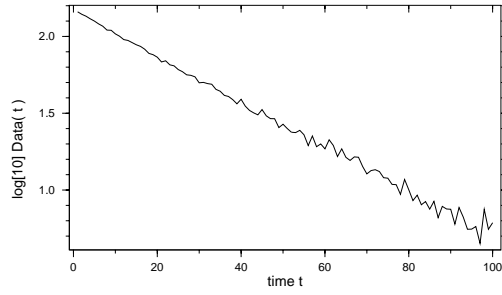


Figure 2: Bretthorst's decaying exponential data.

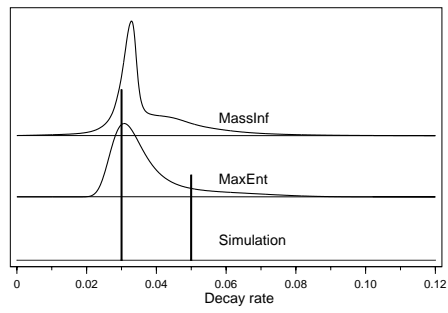


Figure 3: Comparison of MaxEnt with mean MassInf reconstructions from Bretthorst's data.

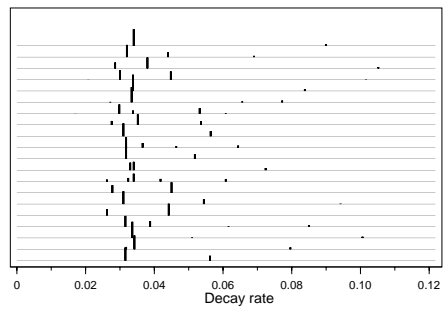


Figure 4: 20 typical MassInf samples with Bretthorst's data.

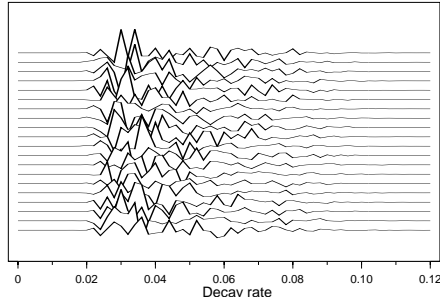


Figure 5: 20 typical QME samples (in Gaussian approximation) with Bretthorst's data.

Bretthorst's original analysis used a Gaussian prior, which loses power because the context supposes that f is a measure that cannot be negative. The MaxEnt selection (Skilling, 1991) is shown in Fig.3, where it may be compared with the simulation. Only a single hump is recovered, though the rightward tail might optimistically suggest rightward structure. The mean MassInf result, though, is sharper: the rightward extension almost gives a second local maximum. Fig.4 shows a selection of the random samples from which the mean MassInf result was accumulated. Although each of these is necessarily sharp and positive, the mean is smooth, and localised only to the degree imposed by the data. In as many samples as not, there are two important components, though sometimes only one dominates. One can easily accumulate the probabilities of having 0, 1, 2, 3, ... components.

A reason why the MassInf result is sharper than MaxEnt can be seen in Fig.5, which shows a selection of random samples from the MaxEnt probability distribution. Technically, these samples are drawn from the Gaussian approximation to the posterior, otherwise the pixellation into finite cells would have induced its bias towards uniformity. These approximated samples are **not** properly required to be positive, so that an important constraint is slackened. In effect, part of the prior information in MaxEnt has been relaxed in order to disguise the problem with pixellation, and this leads to a looser reconstruction.

A practical dataset from NMR spectroscopy is shown in Fig.6. The underlying signals have been blurred by a point-spread-function about 40 units wide (full-width-half-maximum), and there is also noise. Fig.7 compares the mean MassInf deconvolution with that from MaxEnt. The MassInf result is again sharper and thus higher, and has a cleaner, flatter background. A reason for this extra clarity can be seen by comparing the samples (Fig.8 for MassInf and Fig.9 for MaxEnt). MaxEnt samples can be locally negative,

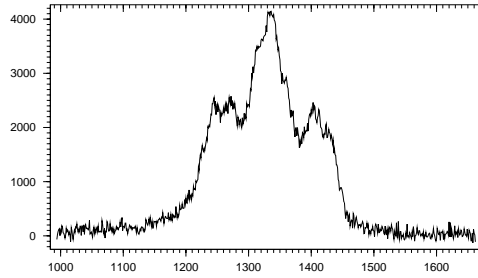


Figure 6: NMR spectroscopy data (courtesy Wellcome Research Laboratories).

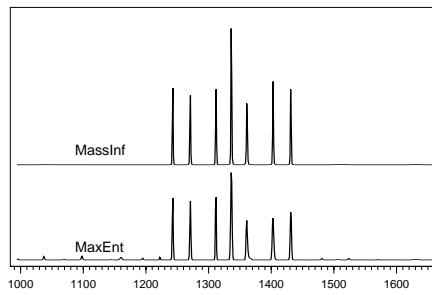


Figure 7: Comparison of MaxEnt with mean MassInf reconstructions from NMR data.

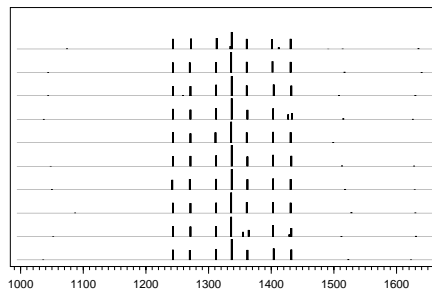


Figure 8: 10 typical MassInf samples with NMR data.

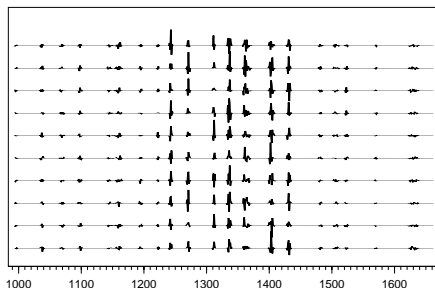


Figure 9: 10 typical QME samples (in Gaussian approximation) with NMR data.

Table 2: Quantified MassInf results from NMR data.

Peak	Prob(> 0)	Mean Position	Cumulant (%)
1	10%	1067.4 ± 41.6	0.0 ± 0.1
2	100%	1243.6 ± 0.7	12.1 ± 0.5
3	100%	1269.8 ± 0.8	11.6 ± 0.4
4	100%	1311.9 ± 1.0	11.3 ± 0.7
5	100%	1335.7 ± 0.9	26.3 ± 0.9
6	100%	1359.9 ± 0.9	11.7 ± 0.6
7	100%	1403.0 ± 0.9	13.5 ± 0.4
8	100%	1430.5 ± 0.7	13.0 ± 0.5
9	70%	1517.3 ± 8.2	0.4 ± 0.3
10	60%	1629.4 ± 11.4	0.2 ± 0.2

and this has allowed some low-level breakthrough at several locations across the MaxEnt reconstruction. MassInf prohibits such negatives, giving tighter control.

The MassInf analysis can be taken further by separating the mean reconstruction into constituent peaks (lying between adjacent minima). Accumulating the samples in each domain gives the usual mean and standard deviation uncertainty, but augmented with the probability that there is a signal present at all (see Table 2). This is the proportion of samples in which at least one atom is present. The seven obvious components are all definitely present, with 100% reliability. Their intensities, moreover, accord with the $1 : 1 : 1 : 2 : 1 : 1 : 1$ ratio expected from the chemistry, to within 1 or 2 standard deviations. However, there are also a couple of weaker sugges-

tions to the right, arising from a barely visible excess in the data. MassInf gives probabilities of 70% and 60% of these lines being present. Such analysis directly answers the basic question “Is structure there?” — a helpful preliminary to the usual question “How much is there?”.

Acknowledgements

This work was supported by MaxEnt Solutions Ltd. Theoretical work on Massive Inference was carried out with S. Sibisi, and the ideas on polarization were developed with S.F. Gull. My thanks to both for long-standing collaboration.

References

- Bretthorst, G.L.: 1990, ‘Bayesian Spectrum Analysis and Parameter Estimation’ in *Lecture Notes in Statistics* **48**, Springer-Verlag, New York.
- Feller, W.: 1971, *An Introduction to Probability Theory and its Applications, Vol. II*, Wiley, New York.
- Gull, S.F. and Daniell, G.J.: 1978, ‘Image reconstruction from incomplete and noisy data’, *Nature* **272**, 686–690.
- Gull, S.F.: 1989, ‘Developments in Maximum Entropy data analysis’ in *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.) Kluwer Academic Publishers, Dordrecht, 53–71.
- Jaynes, E.T.: 1957, ‘Information Theory and Statistical Mechanics, I’, *Phys. Rev.* **106**, 620–630.
- Lukacs, E.: 1970, *Characteristic Functions*, 2nd ed., Griffin, London.
- Shannon, C.E.: 1948, ‘A Mathematical Theory of Communication’, *Bell Systems Tech. J.* **27**, 379, 623.
- Shore, J.E. and 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.: 1989, ‘Classic Maximum Entropy’ in *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.) Kluwer Academic Publishers, Dordrecht, 45–52.
- Skilling, J.: 1991, ‘On Parameter Estimation and Quantified MaxEnt’ in *Maximum Entropy and Bayesian Methods*, W.T. Grandy, Jr. and L.H. Schick (ed.) Kluwer Academic Publishers, Dordrecht, 267–273.
- Stone, J.M.: 1963, *Radiation and Optics*, McGraw-Hill, New York, 313.