THE MAXIMUM ENTROPY METHOD

(Invited paper)

S.F. Gull
Mullard Radio Astronomy Observatory
Cavendish Laboratory, Cambridge, United Kingdom

G.J. Daniell
Department of Physics
University of Southampton, United Kingdom

1. INTRODUCTION

We have heard so far at this meeting from the practical-minded men of radio astronomy. Faced with the problem: 'my map has sidelobes and noise on it', they promptly reach for the nearest available weapons - 'cleaners, polishers and kneaders'. All these use varying amounts of subjective human judgement, as for example in 'window cleaning'.

The ideas of the maximum entropy method are intriguing, since they suggest that one can in principle remove the effects of sidelobes from the map, reduce the apparent noise, and give extra resolution. But maximum entropy is not just another image reconstruction technique, it is a technique for the idealist; the man who cares what should be done with his data. It is more fundamental than other techniques and deals with problems in a very general way. What the maximum entropy method sets out to do is to seek the most likely answer to a particular question that can be obtained from a given body of scientific data.

2. BAYES' THEOREM

We declare immediately our scientific philosophy. We are followers of Jeffreys (1) and are unashamed Bayesians. We further believe, following Jaynes (2), that the various techniques and ideas of maximum entropy should be regarded as simplified ways of applying Bayes' Theorem in probability theory. Although the fundamental ideas are very general, we refer here to a radioastronomical context; the problem is to make a map of the radio sky.

All our knowledge of the sky can be expressed, we believe, as the relative probabilities of various possible maps of the sky. Bayes' Theorem is then the basic theorem of probability theory that tells us how to modify our knowledge or prejudice about the sky when some

C. van Schooneveld (ed.), Image Formation from Coherence Functions in Astronomy, 219-225.

https://doi.org/10.1017/S025292110007490X Published online by Cambridge University Press

additional data are acquired. It says:

$$\text{Posterior Prob.(sky} \mid \text{data)} \propto \text{Prior Prob.(sky)} \times \text{Prob.(data} \mid \text{sky).}$$

The constant of proportionality can be found by normalizing the posterior probability.

The second term on the right is easy to obtain: we can calculate what we ought to observe from any given sky, and hence the probability of this giving rise to our measured data. For the special case when our measurements have independent Gaussian errors this second term is $\exp(-\chi^2/2)$ where:

$$\chi^2 = \sum_{\text{data}} \frac{(\text{misfit to data})^2}{(\text{error in data})^2} ,$$

so that:

$$\text{Posterior Prob. of sky} \propto \text{Prior Prob. of sky} \times \exp.(-\chi^2/2).$$

When the number of sample points in the sky (pixels) is large, and certainly in the continuous limit, this probability distribution of maps of the sky is far too complicated to comprehend. One, albeit very inadequate, way of summarizing it and displaying a single, useful map is to pick the map for which the posterior probability is a maximum.

This maximum in the distribution may, or may not, be sharp; there may even be several local maxima. If the maximum is broad a single map is not very useful, and if there are several maxima it is downright misleading. The occurrence of these situations simply means that the data collected do not determine the sky properly; you must do a better experiment!

We emphasize that in this formulation the data enters into a probability distribution, prior prejudice or prior knowledge of the sky is modulated by a combination of the data values and their errors. Highly uncertain data simply changes one's opinion less than more accurate data.

The vital importance of the errors in the data is not always realized; for example in the Burg (3) algorithm for maximum entropy spectral analysis the data are estimates of the autocorrelation function of a time series. These estimates will have errors, nevertheless the spectrum is made to agree exactly with the estimates. We suggest that the deficiencies of maximum entropy spectral analysis, the line splitting and line shifts, are basically due to ignoring the errors of estimation.

https://doi.org/10.1017/S025292110007490X Published online by Cambridge University Press

There is danger in interpreting any map of the sky that fits the data exactly, it will almost certainly contain features arising from the noise in the data. The above analysis shows the importance of the statistic χ^2 in any problem where the data values have independent Gaussian errors.

3. THE PRIOR AND MAXIMUM ENTROPY

The fundamental limitation of Bayesian statistics lies in the selection of a suitable expression for the prior probability; in our example the prior probability of various maps of the radio sky before we are given any data. If we write this exp (S/λ), where S is a function of the map and λ a fixed constant, we see that to display the single map that maximizes the posterior probability we have to maximize $S - \lambda\chi^2$. This is the recipe used by Gull and Daniell (4) and by Wernecke and D'Addario (5) for constructing maximum entropy maps.

We can therefore identify S as the 'entropy of a map of the sky' and the result of Bayes' theorem is identical with the maximum entropy algorithm in which we maximize an entropy expression, S, subject to the constraint that the final map fits the data. The constant, λ, has become a Lagrange multiplier. It follows, in accordance with the philosophy of Jaynes (2), that the maximum entropy method is a simplified Bayesian approach and the various entropy expressions in use are really statements of a prior prejudice about the sky, given no data.

The problem of determining appropriate prior distributions corresponding to such ignorance has led to considerable controversy since the time of Laplace and there is, as yet, no entirely satisfactory way of assigning priors, even in the simplest situations. Therefore, at this stage, we suggest that it is most useful to understand the properties of the prior distributions that arise from the various entropy expressions and to collect arguments that lead to different priors, without attempting to decide which, if any, are correct.

4. THE ENTROPY EXPRESSION

There are two schools of thought concerning the entropy of a map of the sky. If we label a set of areas on a map by an index i and the flux of radio waves from the ith area is f_i, then we can express the results thus: The H_1 school says that $S \propto \sum_i \log f_i$, and the H_2 school says $S \propto -\sum_i f_i \log f_i$. We are members of this second school and we can give an outline derivation of the entropy expression using a model incorporating prior assumptions about the sky.

For ease of counting maps we assume the flux f_i from each cell of the sky is quantized. We then employ a secret weapon - the canonical team of monkeys. Each monkey is assumed to be given a large number N of flux quanta which he scatters randomly across the cells of the sky to

https://doi.org/10.1017/S025292110007490X Published online by Cambridge University Press

make a trial map. Suppose n_i quanta land in the ith cell. Not all sets of n_i generated in this way by the monkeys are equally likely. The probability that a particular set occurs is proportional to $N!/\prod_i n_i!$. If N and all the n_i are large, we can use Stirling's approximation and obtain $S \propto \log(\text{probability}) \propto -\sum n_i \log n_i/N$. When the question of relating our artificial flux quanta to the continuous quantities f_i is considered we are led to the entropy expression:

$$S \propto -\sum_i \left[\frac{f_i}{\Sigma f}\right] \log \left[\frac{f_i}{\Sigma f}\right]$$

Suppose now we are given a body of data, for example samples of a spatial coherence function, having independent errors. The fit of a trial map to these data may be defined using χ^2 and the probability of the data given the map $\{f_i\}$ is then proportional to $\exp(-\chi^2/2)$. Maximizing $S - \lambda\chi^2$ yields:

$$f_i = \left[\exp\left\{\frac{\sum_j f_j \log f_j}{\sum_j f_j}\right\}\right] \exp(-\lambda\frac{\partial\chi^2}{\partial f_i}) .$$

The first term in brackets is just a constant and it ensures that the final map does not depend on the units in which the data are given. The value of λ is not determined by this argument and we have suggested (Gull and Daniell (4)) that λ is chosen so that χ^2 for the final map equals its statistically expected value, i.e. the number of data samples. In this way we get a safe map that is unlikely to contain artifacts arising from the noise in the data. Of course, if there are several maxima in the probability this map will still be misleading. However, for the problem of aperture synthesis, with phase information, or for deconvolution, this cannot happen. The prior based on the monkey mechanisms is strongly peaked about uniform maps so we can loosely say that we have constructed the most uniform map consistent with the data and their errors.

The above discussion shows in outline how a prior probability distribution or entropy expression follows from assumptions or prejudice about the sky. Other entropy expressions follow from other prejudices corresponding apparently to different types of 'complete ignorance' about the sky. The expression H_2 is a configurational entropy of the radio flux, although the above calculation is similar to those used in statistical mechanics. One cannot run a heat engine on an image!

5. OTHER ENTROPY EXPRESSIONS

We will not deal in such detail with the alternative expression H_1, but will give some idea of how it arose and the assumptions it makes. It has been used for many years in Burg's (3) algorithm and is based on

Shannon's information theory.

Burg is concerned with the estimation of the power spectrum of a Gaussian random process or time series. The relation of this to the autocorrelation function certainly involves a Fourier transform, and so indeed does aperture synthesis, but here the analogy ceases; there is no 'space series' in radio astronomy. The expression $H_1 = \int \log P(\nu)d\nu$ (Bartlett (6)) is the entropy rate of a noise source with power $P(\nu)$ at frequency ν. It is distinct from the configurational entropy of the spectrum itself. In more recent work on the Burg method the information theoretic aspects have been relatively neglected and interest has centred on the fact that the method is equivalent to fitting an autoregressive process to the time series. This has a natural interpretation for the estimation of power spectra of time series but it does not make sense to regard the signals from successive aerials in an interferometer as being derived by spatially filtering random noise. We conclude that Burg's and Shannon's arguments, whilst sound within their terms of reference for time series, are not appropriate to radio astronomy.

More recently Kikuchi and Soffer (7) have claimed that both H_1 and H_2 expressions are limiting cases of a single one, and the choice between them depends on a parameter n/z, n being the number of photons received and z their number of degrees of freedom. Their expression is based on a discussion of the entropy of the radiation field coming from the sky and the fact that photons obey Bose-Einstein statistics is central to their argument. They claim that if $n/z \gg 1$, which is true in radio astronomy, the H_1 expression is correct, whereas in optical astronomy where $n/z \ll 1$ the H_2 expression should be used. The trouble with this is that, in our language, they are saying that we should have a different prior prejudice about the shape of radio sources from the shape of optical sources, which is ludicrous. Surely we are interested in extracting information about the sky itself, not the radiation field. We cannot accept that the analysis of the shape of radio sources depends on the real quantized nature of light. We have used quanta to assist in the counting of states, as in elementary treatments of classical statistical mechanics, and have used this to derive H_2, but we then take the limit as the quanta become indefinitely small. The results of Kikuchi and Soffer seem to imply that, by studying the relative efficiency of the H_1 and H_2 algorithms on the restoration of a photograph, one could determine Planck's constant, or at least its order of magnitude.

6. PRACTICALITIES

So much for theory, can it be done in practice? Yes!

The H_1 algorithm has been used by Wernecke and D'Addario (5) and the H_2 by Gull and Daniell (4), on real astronomical data. In aperture synthesis the latter implementation takes 5 minutes on an IBM 370/165 for a 128 x 128 point image and a program for 256 x 256 points is

https://doi.org/10.1017/S025292110007490X Published online by Cambridge University Press

working. For deconvolution there exists a working program for 128 x 128 pixels and the results are significantly better than the ART algorithm. Gull and Daniell have also successfully applied their method to aperture synthesis without phase information, and to VLBI data.

7. CONCLUSIONS

1. The maximum entropy method can do what its inventors hoped, it is a fundamentally sound, but different, way of looking at data, derived from Bayes' theorem and rooted in the foundations of scientific methodology.

2. It can be used for any problem for which one can predict the data that would be observed from a trial map.

3. It provides an objective, uniquely defined, procedure for analysing data and therefore runs counter to the trend for interactive data processing.

REFERENCES

1. Jeffreys, H. Theory of Probability, 3rd edition, Oxford 1961.
2. Jaynes, E.T. Probability Theory in Science and Engineering, Field Research Laboratory, Socony Mobil Oil Co., Inc. 1959.
3. Burg, J.P. Maximum Entropy Spectral Analysis, presented at the 37th meeting of the Society of Exploration Geophysicists, Oklahoma City 1967.
4. Gull, S.F., and Daniell, G.J. Nature, 272, 686 1978.
5. Wernecke, S.J., and D'Addario, L.R. IEEE Trans. on Computers, C26, 351 1977.
6. Bartlett, M.S. An Introduction to Stochastic Processes, 2nd Edition, Cambridge University Press 1966.
7. Kikuchi, R., and Soffer, B.H. J. Opt. Soc. Am., 67, 1656 1977.

DISCUSSION

Comment C. VAN SCHOONEVELD.

1) What is the difference between the results when the H_1- and H_2-definition are applied to the same data? 2) How can we expect a resolution improvement if we start from the a-priori assumption of a most uniform sky?

Reply S.F. GULL.

1) We have not made calculations using H_1, but with nearly complete and accurate data the choice of prior distributions has comparatively little effect and both H_1 and H_2 will give similar results. (See also comment to question by CRANE.) 2) For accurate data a modest improvement results (never more than a factor of 2). This arises from the positivity and

https://doi.org/10.1017/S025292110007490X Published online by Cambridge University Press

from the fact that ungraded Fourier data can be used without creating sidelobes.

Comment R. GORDON.
In what sense is Max.Entropy the spatially smoothest solution?
Reply S.F. GULL.
In no sense whatsoever. The Max.Entropy solution gives the statistically most uniform map, in the sense that it has the most uniform histogram of flux values.

Comment L.R. D'ADDARIO
The sense in which Max.Entropy maps are "smooth" is made precise by a theorem due to Wernecke (1977, Radio Science, 12, 831-844), which says that given the ME-map with measurement discrepancy D, any linear smoothing of the map results in a new map with a discrepancy $\geq$ D.

Comment Y.G. BIRAUD.
1) Analogous work was done by R. Herschel and B.R. Frieden. 2) What is the importance of the pixel size? 3) What is the gain in resolution versus S/N?
Reply S.F. GULL.
2) Decreasing the pixelsize and increasing their number improves the detail visible on the map until the limit set by the signal-to-noise ratio. After that, smaller pixels simply interpolate smoothly between their neighbours. 3) We made quantitative tests for the deconvolution case. Until a signal-to-noise ratio of 3:1 is reached (per pixel), the MEM gives a smoother result than conventional analysis. Beyond that ratio the method gives increasing resolution until a super-resolution of a factor 2 is reached at S/N $\approx$ 100. After that, virtually no improvement is possible.

Comment P.C. CRANE.
Two versions of Wernecke and D'Addario's algorithm operating at NRAO, one implementing H_1 and the other H_2, give, in one case, results identical to $\lesssim$ 10%.
Reply S.F. GULL.
I am not at all surprised. For the radio astronomical case the data are often extremely good. The posterior probability distribution is then largely determined by the data, not by the prior. It is when only a small amount of noisy data is available that the influence of the prior distribution will be seen.

Comment U.J. SCHWARZ.
How stable is the MEM solution against the choice of λ?
Reply S.F. GULL.
Very stable. The quantity χ^2decreases as λ is increased but the map does not change much unless χ^2 is very different from its expected value.

https://doi.org/10.1017/S025292110007490X Published online by Cambridge University Press