IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Maximum entropy in data analysis with error-carrying constraints

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1987 J. Phys. A: Math. Gen. 20 2379

(http://iopscience.iop.org/0305-4470/20/9/025)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 129.252.86.83
The article was downloaded on 01/06/2010 at 05:32

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. 20 (1987) 2379-2388. Printed in the UK

Maximum entropy in data analysis with error-carrying constraints

R Lieu, R B Hicks and C J Bland

Department of Physics, The University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4

Received 9 August 1985, in final form 6 October 1986

Abstract. The problem of finding the maximum entropy solution which satisfies some poorly defined constraints is central to data processing. Hitherto all attempts to solve it have been conceptually somewhat incomplete. The present work follows the traditional Boltzmann-Gibbs approach to entropy. When data are available the various outcomes can be assigned *a priori* probabilities (or weights) according to how much each of them agrees with the measured distribution when errors are taken into account. The reasoning yields a statistic $S-\chi^2$, which is a precise expression for the entropy. The maximum $S-\chi^2$ solution defines a distribution uniquely. Even in circumstances when there is no information apart from those provided by the data, the inferred distribution differs from that given by the data mean values and bears a closer resemblance to the underlying population than do the raw data. Where additional constraints are present, this maximum entropy procedure is more effective in incorporating them than are traditional maximum likelihood techniques. $S-\chi^2$ maximisation may also be carried out iteratively up to a certain limit. It offers the most advantage in situations when one has a thorough understanding of the inherent noise behaviour.

1. Introduction

Maximum entropy methods have recently been applied to time series (Burg 1975), spectral analysis (Fougere 1977, Johnson *et al* 1984), speech processing (Johnson and Shore 1984), particle dynamics (Engel and Levine 1983), image reconstruction (Skilling *et al* 1979, Burch *et al* 1983) and geography (Wilson 1970). The general rationale for this technique has been studied in depth (Shore and Johnson 1980, Jaynes 1982). The present work seeks a more fundamental approach to problems involving uncertain constraints, notably those introduced by experimental data which are inevitably subject to errors.

2. Some methods of distribution determination in cases of insufficient knowledge

We discuss below three commonly used procedures.

(i) Maximum likelihood (least squares method). This minimises the statistic

$$\phi^2=\sum_{i=1}^{N}\frac{(y_i-y_i^{\text{data}})^2}{2\sigma_i^2}$$

0305-4470/87/092379+10$02.50 © 1987 IOP Publishing Ltd

where y_i^{data} and σ_i are respectively the mean values and standard deviations of the experimental data and y_i is the inferred distribution. Note that for a Gaussian error distribution $\phi^2=\chi^2$, and the minimisation procedure is based on the principle of maximum likelihood. In the absence of any constraints, the global minimum of ϕ^2 simply gives the mean values of the data as the inferred distribution, i.e.

$$y_i=y_i^{\text{data}}.$$

In more typical situations, such as hypothesis testing, one generally assumes a parametric representation $y_i=y_i(\alpha,\beta,\ldots)$ and minimises ϕ^2 with respect to α, β, etc. Normally the number of parameters involved is much less than the number of data values N.

There are also situations in which ϕ^2 is minimised subject to known constraints. For instance if the total value of the population is precisely known to be Y, the distribution would be obtained by solving a set of simultaneous equations

$$\frac{\partial}{\partial y_i}\left[\phi^2+\lambda\left(\sum_{i=1}^{N} y_i - Y\right)\right]=0$$

and

$$\sum_{i=1}^{N} y_i = Y.$$

This gives rise to the 'most likely' distribution derived from the data subject to the condition of a well defined mean (or sum total).

(ii) The Gull-Daniell-Skilling method. This had been used and discussed in detail with reference to image restoration (see Burch *et al* 1983, Gull and Daniell 1978). The suggestion is to search for the distribution of maximum entropy which is consistent with the available data within experimental errors. A convenient way of measuring this 'consistency' is the ϕ^2 statistic. For a set of N data values, normalised to a total $\Sigma_i y_i = Y$, there are $N-1$ degrees of freedom. The condition $\phi^2/(N-1)<1$ or

$$\phi^2_{\text{red}}<1$$

(where the subscript 'red' denotes 'reduced') ensures that each data value in the inferred distribution cannot, on average, depart from the experimental mean by more than unit standard deviation. In practice it is found that the maximum entropy solution always occurs on the contour of maximum ϕ^2. Thus, if $\phi^2_{\text{red}}=1$ is used as a constraint, the inferred distribution would be obtained by maximising the entropy function $S=-\Sigma_i y_i \log y_i$. In other words the following system of $N+2$ simultaneous linear equations needs to be solved:

$$\frac{\partial}{\partial y_i}\left[S+\mu(\phi^2-N)+\lambda\left(\sum_i y_i - Y\right)\right]=0$$

and

$$\phi^2=N$$

$$\sum_i y_i = Y$$

where λ and μ are Lagrange undetermined multipliers and, for simplicity, we dropped the details in the summation sign. It is also assumed that N is large, i.e. $N\approx N-1$. Clearly such a technique is a reasonable compromise between maximum likelihood and maximum entropy.

(iii) The minimum cross-entropy. This iterative process is largely due to Kullback (1959), also Johnson and Shore (see references cited earlier), and minimises the cross-entropy

$$S_c = \sum_i y_i \log(y_i/m_i)$$

where m_i is a 'prior' distribution which one introduces as a first-order approximation. If $m_i = 1$ for all channels $i = 1$ to N, $S_c = -S$ where S is the conventional definition of entropy. In such a case, minimising S_c would lead to a global entropy maximum for S, corresponding to a uniform distribution.

It is perfectly legitimate to use the available experimental data as the prior input (or ansatz), i.e. $m_i = y_i^{\text{data}}$. We will now show that, for such a starting point, and in the case of Poisson errors, minimum cross-entropy is not a procedure very different from least squares.

The proof is as follows. Setting $m_i = y_i^{\text{data}}$ and $\Delta y_i = y_i - y_i^{\text{data}}$ it is easy to show that

$$S_c = \sum_i (y_i^{\text{d}} + \Delta y_i) \log(1 + \Delta y_i / y_i^{\text{d}})$$

where $y_i^{\text{d}} = y_i^{\text{data}}$.

If we now assume that $\Delta y_i \ll y_i^{\text{data}}$ it is possible to take only the first term in the expansion of the logarithmic function to obtain

$$\begin{aligned} S_c &= \sum_i (y_i^{\text{d}} + \Delta y_i)\Delta y_i / y_i^{\text{d}} \\ &= \sum_i \Delta y_i + \sum_i (\Delta y_i)^2 / y_i^{\text{d}}. \end{aligned}$$

In this last equation the second term is equal to $2\phi^2$ for Poisson errors (i.e. $\sigma_i^2 = y_i^{\text{d}}$). The first term is of little significance in the majority of circumstances, being ≈ 0 close to the ϕ^2 minimum. The assertion that $S_c \approx \phi^2$ for Gaussian statistics is therefore justified.

It is very important to note that none of the three current methods described above provide a theoretically rigorous framework for utilising the maximum entropy principle in the analysis of data from everyday experiments. It was shown that minimum cross-entropy is very similar to least squares, which is founded on the principle of maximum likelihood. The Gull–Daniell–Skilling routine does bring in entropy. However, in order to avoid arriving at the global entropy maximum, these authors introduced some sort of 'maximum likelihood' constraints as a measure of misfit with data. The use of a rather *ad hoc* restriction $\phi^2 < N - 1$ assigns equal weighting to all distributions inside this region and zero weighting to those outside it. Data processed by the Gull–Daniell–Skilling routine appear smoother than the raw data and they retain all the genuine features of the population. This provides a key advantage in image restoration. However, it can also be shown that such data generally have less statistical agreement (in terms of χ^2) with the underlying population. It follows that considerable caution should be exercised in applying the method.

3. Boltzmann entropy with uncertain constraints

Let us take one step backward to review the original series of arguments of Boltzmann and Gibbs, arguments which ultimately lead to the present day definition of entropy as $S = -\Sigma\, p \log p$.

Consider a single experiment involving n trials, each having r possible *results* with unknown probabilities $p_1, p_2, \ldots, p_r$. Convenient examples include n tosses of a r-sided die, and distribution of n photons in r energy channels. For large n, ideally, $n_i = np_i$ $(i = 1, 2, \ldots, r)$ of the n trials will have result i. The number of possible *outcomes* (i.e. distinguishable experiments each involving n trials) consistent with a *distribution* $(n_1, n_2, \ldots, n_r)$ is given by

$$\Omega(n_1, n_2, \ldots, n_r) = \frac{n!}{n_1!\,n_2!\ldots n_r!}. \tag{1}$$

At this point the reader is urged to distinguish carefully between the terms 'results', 'outcomes' and 'distribution' as used in our present context. The total number of outcomes is

$$\Omega_{\text{total}} = \sum_{n_1, n_2, \ldots, n_r} \frac{n!}{n_1!\,n_2!\ldots n_r!} = r^n.$$

The chances of a distribution $(n_1, n_2, \ldots, n_r)$ turning up in a single experiment, assuming that all outcomes have equal *a priori* probabilities, is

$$P(n_1, n_2, \ldots, n_r) = \frac{\Omega(n_1, n_2, \ldots, n_r)}{\Omega_{\text{total}}}. \tag{2}$$

It is well known that maximising the entropy

$$S = -\sum_{i=1}^{r} p_i \log p_i \qquad \text{or} \qquad -\sum_{i=1}^{r} n_i \log n_i$$

is equivalent to maximising P or Ω, i.e. selecting the distribution which contains the largest number of possible outcomes.

In the presence of available data concerning the distribution, the assignment of equal probabilities is incorrect. For example, an astronomical detection device might gather a small number of photons n_i in each of the r energy channels, giving rise to *some* idea about $n_1, n_2, \ldots, n_r$, and hence some idea about the basic probabilities $p_1, p_2, \ldots, p_r$ associated with the r results. However, due to the finiteness of n, the n_i are known with errors governed by binomial statistics, which in most circumstances can be approximated by Gaussian functions. Thus, the statistical weighting for different possible values n_i may be determined experimentally as

$$g_i(n_i)\,\mathrm{d}n_i = \frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left(-\frac{(n_i - \bar{n}_i)^2}{2\sigma_i^2}\right) \mathrm{d}n_i \tag{3}$$

where $\bar{n}_i$ and σ_i^2 are respectively the mean and variance of n_i as inferred from the data and from knowledge of the experimental conditions. The corresponding statistical weighting for preconceived values of the basic probability p_i is likewise

$$f_i(p_i)\,\mathrm{d}p_i = \frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left(-\frac{(p_i - \bar{p}_i)^2}{2\sigma_i^2}\right) \mathrm{d}p_i. \tag{4}$$

Here σ_i is reduced by a factor n when compared with the σ_i in (3) (i.e. $\sigma_i^2 = p_i/n$).

This provides a means of assigning *a priori* probabilities to the outcomes. Consider one single experiment leading to an outcome consistent with the distribution $(n_1, n_2, \ldots, n_r)$. For a sufficiently large number of trials n, one can infer from this distribution a single underlying population of probabilities $(p_1, p_2, \ldots, p_r)$, where

$p_i = n_i/n$ $(i = 1, 2, \ldots, r)$. What is the likelihood that such a population is derived from our observed data as a result of statistical fluctuations? The answer for one trial is given by

$$f_1(p_1)f_2(p_2)\ldots f_r(p_r)\Delta p_1\Delta p_2\ldots\Delta p_r \equiv g_1(n_1)g_2(n_2)\ldots g_r(n_r)\Delta n_1\Delta n_2\ldots\Delta n_r.$$

The likelihood of getting this stable population from the data for all n trials is given by

$$(g_1(n_1)g_2(n_2)\ldots g_r(n_r)\Delta n_1\Delta n_2\ldots\Delta n_r)^n \equiv (g_1(n_1)g_2(n_2)\ldots g_r(n_r))^n \quad (5)$$

where in (5) we have set $\Delta n_i = 1$ $(i = 1, 2, \ldots, r)$. This introduces no harm as long as we are in the 'continuous' limit $\Delta n_i \ll n_i$. Equation (5) provides the necessary statistical weighting of individual outcomes. Using (5) together with (2), we now obtain the probability for the occurrence of a distribution $(n_1, n_2, \ldots, n_r)$ in a single experiment:

$$P(n_1, n_2, \ldots, n_r) \propto \Omega(n_1, n_2, \ldots, n_r)[g_1(n_1)g_2(n_2)\ldots g_r(n_r)]^n. \quad (6)$$

In (6) we have ignored an unimportant constant which enforces the normalisation

$$\sum_{n_1, n_2, \ldots, n_r} P(n_1, n_2, \ldots, n_r) = 1.$$

The logarithm of (6) is the required *data processing entropy* S_d which we must maximise:

$$S_d = \log P = \log \Omega + n \sum_{i=1}^{r} \log(g_i(n_i)).$$

We now substitute for Ω in (1) and $g_i(n_i)$ in (3). Using Stirling's approximation $\log n! = n \log n - n$, we deduce

$$S_d = \sum_{i=1}^{r} -n_i \log n_i - n \sum_{i=1}^{r} \frac{(n_i - \bar{n}_i)^2}{2\sigma_i^2}$$

where we have discarded a constant equal to

$$n \log n - n + n \sum_i \log\left(\frac{1}{(2\pi)^{1/2}\sigma_i}\right).$$

Note that S_d can be rewritten as

$$S_d = \sum_{i=1}^{r} -n_i \log n_i - n\chi^2.$$

If we write $n_i = np_i$ and are prepared to remove some more constants, it will be sufficiently trivial for the reader to verify that

$$S_d = \sum_{i=1}^{r} -p_i \log p_i - \chi^2 = S - \chi^2$$

is the correct statistic to maximise, if one wishes to adopt a rigorous approach to entropy.

We note that the above derivation is based upon the assumption of Gaussian noise. If errors do not obey the Gaussian distribution, the method is no longer accurate. If it is nevertheless attempted, the correct symbol for the statistic should be $S - \phi^2$. The procedure of maximum $S - \phi^2$ generates a unique output distribution. Henceforth we will denote the operation by a symbol $\mathcal{S}$. Note that the effect of $\mathcal{S}$ is to transform one distribution to another. $\mathcal{S}$ may also be applied subject to additional constraints which are independent of the data (see later for more discussion on this point).

The treatment can be extended beyond those constraints imposed by the availability of data. If there exists any condition

$$\zeta(n_1, n_2, \ldots, n_r) = 0$$

subject to a probability distribution functional

$$\xi[\zeta] = \xi(\zeta(n_i))$$

it is straightforward to repeat the foregoing arguments to deduce that

$$S_{\mathrm{d}} = \sum_i n_i \log n_i + n \log \xi[\zeta].$$

4. Applications to simulated data

To assess the performance of maximum entropy we have selected a variety of common patterns. For each of these we added noise to them and applied $\mathcal{S}$. The resulting processed distribution, together with the raw data, are compared with the original population. To this end, we assume that the variances due to noise are known. It is then possible to calculate the likelihood of the raw and processed data being derived from the population. In other words, we compute the coefficient of improvement

$$\alpha = \frac{\Delta\chi^2}{\chi^2} = \frac{\chi_1^2 - \chi_2^2}{\chi_1^2} \qquad (7)$$

where χ_1^2 and χ_2^2 are respectively the χ^2 difference between the population and raw data, and that between the population and processed data. If $\alpha > 0$ the processed data resemble the population more than the raw data do.

Table 1 describes the different kinds of population (signal) and noise being employed in our analysis. The procedure adopted for generating Poisson noise is as follows. We begin with a fixed total number of counts, Y. For the first of these counts a random number is generated. This number, together with the particular population (signal or pattern) under consideration, determine which channel the count will go. The process is then repeated for all remaining events. This creates 'binomial noise' which approximately obeys Poisson statistics when the probability of entry for any of the channels is small. The noise thus introduced has the following characteristics: (i) the error in each channel having a mean of five or more counts is very closely Gaussian; (ii) the total number of counts in any data sample is conserved, and equals that of the population (i.e. Y).

Concerning the second type of noise mentioned in table 1, namely white noise, it is strictly Gaussian in nature. Unlike Poisson noise, the standard errors are input parameters (see table 1) independent of the signal strength. Note also that, in the presence of such noise, the total number of counts varies from sample to sample.

The method of data processing is always to apply $\mathcal{S}$ subject to conservation of total counts, i.e. to solve the system of equations

$$\frac{\partial}{\partial y_i}\left[\sum_{i=1}^{r} y_i \log y_i + Y\left(\sum_{i=1}^{r} \frac{(y_i - y_i^{\mathrm{d}})^2}{2\sigma_i^2}\right) + \lambda\left(\sum_{i=1}^{r} y_i - Y\right)\right] = 0 \qquad (8)$$

where r is the total number of channels and $y_i^{\mathrm{d}} = y_i^{\mathrm{data}}$ determines uniquely the output distribution $\{y_i\}$. In order to be cautious, the σ_i used in (8) are inferred directly from the data. So in virtually all applications of the maximum entropy principle, the

Table 1.

I. Classification of signal populations	
(a) Straight line with positive slope	
$y=2x+3$	$x=1,2,\ldots,10$
(b) Straight line with negative slope	
$y=-x+15$	$x=1,2,\ldots,10$
(c) Sawtooth pattern	
$y=5$	$0<x<6$ and $12<x<18$
$y=5+10(x-5)$	$5<x<10$
$y=45-10(x-9)$	$9<x<13$
x is always an integer	
(d) Sharp sawtooth	
$y=5$	$0<x<6$ and $12<x<18$
$y=5+100(x-5)$	$5<x<10$
$y=405-100(x-9)$	$9<x<13$
(e) Sine function	
$y=7.5\sin(x)+10$	$x=20,40,\ldots,360$
x in degrees.	
II. Classification of noise	
(a) Poisson noise. Variance $=(1-p)n$. Total count is conserved. For more elaboration, see text.	
(b) White Gaussian noise. Variance $=\sigma^2$ where σ is an input parameter. In some cases there may be more than one sigma as input.	

normalisation to a total probability is included for completeness. During processing the total number of counts Y is inferred from the population. For samples containing Poisson noise this poses no additional constraint. However, when white noise is present the total number of counts is not conserved, and the condition

$$\sum_{i=1}^{r} y_i = Y \tag{9}$$

where Y is the *population* total, does serve as additional knowledge which will drive the processed data closer to the population. Thus, application of (9) will, in some cases, provide a means of assessing the potential applicability of $\mathscr{S}$ as an estimator which can incorporate constraints other than those imposed by the data.

We discuss the results. For each basic population (signal or pattern) we generated many noisy data samples in order to obtain a frequency distribution for the coefficient of improvement α in (7). Cases 1-4 of table 2 show these distributions for samples containing Poisson noise.

The peaks of the frequency curves lie in the region $\alpha>0$, meaning that there is a definite improvement. The amount in each case is too small (a few per cent) to warrant practical applications of the technique. However, the fact that $\alpha>0$ does provide incentive for further studies of $\mathscr{S}$ in order to identify domains where more substantial enhancements of the underlying structures are present (see later in this section).

Table 2. Signals contaminated with noises are processed and parameters for the statistical distributions of the coefficient of improvement (for definition, see text) are tabulated. Unless otherwise stated in the 'description' column, the method employed in data processing is $S-\chi^2$.

Case number	Description (reference to table 1)	Number of data samples	α (%) mean	α (%) mode	α (%) max	α (%) min
1	Class (a) signal and class (a) noise	343	1.7	1.3	3.1	−0.15
2	Class (b) signal and class (a) noise	340	1.0	1.2	2.65	−0.8
3	Class (c) signal and class (a) noise	327	0.4	−0.3	1.95	−1.32
4	Class (e) signal and class (a) noise	268	0.70	1.0	1.72	−0.43
5	Class (d) signal and class (a) plus class (b) noise	300	8.10	2.3	40	−0.4
6	Class (e) signal and class (a) noise plus 15 iterations of $S-\chi^2$	194	9.1	12.6	23.1	−8.0
7	Class (d) signal and class (a) plus class (b) noise. Here $\alpha=\Delta\chi^2/\chi^2$ where $\Delta\chi^2$ is the χ^2 difference between the least squares and $S-\chi^2$ processed data	368	1.26	0.3	5.1	−1.10
8	Class (c) signal and class (a) plus class (b) noise. Three-point smoothing technique	356	0.7	1.32	−0.37	1.6
9	Class (d) signal and class (a) plus class (b) noise. Three-point smoothing technique	287	−12.5	−7.15	0.0	−38.0

Turning to samples containing white noise as well as Poisson noise, case 5 of table 2 displays the frequency distribution for α. Evidently there is more improvement than previous cases and the reason is because the normalisation of total counts to a value which equals that of the population serves as additional constraint.

Case 6 describes situations where we applied $\mathscr{S}$ more than once. During each iteration the errors σ_i are inferred from the previous distribution. In other words the processed data from one iteration are treated as raw data for the next. Case 6 of table 2 shows the frequency curves for α in some typical situations. The fact that peak values of α now reach approximately 10% is a significant development and suggests that the method has potential for practical applications.

However, a word of caution is needed here. If the algorithm is repeated indefinitely, there will always exist an 'equilibrium point' beyond which α turns negative and remains so; in other words the processed data become progressively worse. The optimal or critical 'index' may be defined as the value n such that $\mathscr{S}^n$ provides the greatest improvement. This index depends on three main factors: (i) form of the underlying population, (ii) type of noise and (iii) how well the noise is estimated. For instance, the sawtooth pattern in case 3, when it carries Poisson noise, gives rise to little improvement even at the first trial $n=1$ (as is evident from table 2). It was found that repeated applications of $\mathscr{S}$ did more harm than good, so the critical index is $n=1$. We found that the same situation becomes substantially more favourable if white noise is introduced in addition to Poisson noise, and if we assume exact knowledge of the noise. The difficulty with Poisson noise alone is that such an assumption cannot be made, since noise and signal are not independent. It is therefore clear that certain patterns contaminated only with Poisson noise may yield a low critical index n when processed with $\mathscr{S}$. Once the signal is estimated wrongly, error for the next iteration will also be estimated wrongly, so the sequence diverges away from the truth.

5. Comparison of maximum entropy with other existing methods

(*a*) Least squares. For Poisson noise, which conserves the total number of photons in the data, the procedure of minimum χ^2 subject to the constraint $\Sigma\, y_i = Y$ gives nothing but the global χ^2 minimum ($\chi^2 = 0$) during which $y_i = y_i^{\text{data}}$. If white noise is introduced, the constraint $\Sigma_i\, y_i = Y$, where Y equals the population total, does serve as useful additional knowledge which should improve the resemblance between data and population. Case 7 of table 2 gives the frequency distribution of $\alpha = \Delta\chi^2/\chi^2$ where now $\Delta\chi^2$ is the difference between the χ^2 of the least squares processed data, and that of the $\mathscr{S}$ operator processed data. It is apparent from the figure that α is once again positive, meaning that $\mathscr{S}$ is a better technique for incorporating constraints other than those already imposed by the data.

(*b*) Gull-Daniell-Skilling method. Comparison of $\mathscr{S}$ with this method is inappropriate. As noted earlier, this method consistently drives the processed data away from the population. Clearly its merits have to be assessed in a very different way.

(*c*) Three-point smoothing. Simple smoothing techniques such as

$$y_i \rightarrow y_i' = \tfrac{1}{4}y_{i-1} + \tfrac{1}{2}y_i + \tfrac{1}{4}y_{i+1}$$

give substantial improvements for mildly varying patterns (case 8). For patterns containing high-frequency components, the method fails completely (case 9). However, the $\mathscr{S}$ operator is capable of providing improvements even under such circumstances (case 5).

6. Conclusion

The operator has potential merits in data processing. Even in situations when there is completely no information apart from that provided by the measurements (i.e. signals and errors), the method yields a distribution which differs from that given by the data mean values. It was shown that, for a variety of circumstances, the processed distribution resembles the underlying population more than the raw data do. Repeated use of $\mathscr{S}$ yields progressively better results, up to a certain limit. Precautions must be taken not to exceed a critical number of applications. It was also shown that the method offers an effective way of incorporating additional constraints (e.g. constraints other than those provided by the data and its errors). The method is most useful when a good knowledge of the data noise level is available.

Acknowledgment

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. We are grateful to the Nuclear Medicine Division of Foothills Hospital for the computing facilities made available to us.

References

Burch S F, Gull S F and Skilling J 1983 *Computer Vision, Graphics and Image Processing* **23** 113
Burg J P 1975 *PhD thesis* Stanford University

Engel Y M and Levine R D 1983 *Phys. Rev.* C **28** 2321
Fougere S F 1977 *J. Geophys. Res.* **82** 1051
Gull S F and Daniell G J 1978 *Nature* **272** 686
Jaynes E T 1982 *Proc. IEEE* **70** 939
Johnson R W and Shore J E 1984 *IEEE Trans. Acoust., Speech Signal Process.* **ASSP-32** 129
Johnson R W, Shore R W and Burg J P 1984 *IEEE Trans. Acoust., Speech Signal Process.* **ASSP-32** 531
Kullback S 1959 *Information Theory and Statistics* (New York: Wiley)
Shore J E and Johnson R W 1980 *IEEE Trans. Information Theor.* **IT-26** 26
Skilling J, Strong A W and Bennett K 1979 *Mon. Not. R. Astron. Soc.* **187** 145
Wilson A G 1970 *Entropy in Urban and Regional Modelling* (London: Pion)