

Introduction to maximum entropy

D. S. Sivia

Theoretical Division and Los Alamos Neutron Scattering Center
 Los Alamos National Laboratory
 Los Alamos, New Mexico 87545
 USA

ABSTRACT: The maximum entropy (MaxEnt) principle has been successfully used in image reconstruction in a wide variety of fields. We review the need for such methods in data analysis and show, by use of a very simple example, why MaxEnt is to be preferred over other regularising functions. This leads to a more general interpretation of the MaxEnt method, and its use is illustrated with several different examples. Practical difficulties with non-linear problems still remain, this being highlighted by the notorious phase problem in crystallography. We conclude with an example from neutron scattering, using data from a filter difference spectrometer to contrast MaxEnt with a conventional deconvolution.

1. Introduction

In many scientific experiments, the quantity of interest f is related to the data d through some transformation O and noise σ :

$$d = O.f + \sigma.$$

For example, f might be the radio-flux distribution of an astronomical source, the momentum distribution of atoms in liquid helium, or the scattering law in a neutron scattering experiment, and so on. The transformation operator O might represent a Fourier transform or a convolution with an instrumental resolution function. The job of data analysis is to infer the desired quantity f from the data d .

The simplest way of deriving an estimate of f , \hat{f} , from the data is to apply the inverse transform O^{-1} to the data: $\hat{f} = O^{-1} .d$. In many cases, however, we cannot do this because the inverse operator does not exist, often because we have missing data. We cannot Fourier transform a data set, for example, if we have unmeasured data. Even if we can compute the inverse transform, our reconstruction will have many artifacts because we have not taken into account the fact that the data were noisy:

$$\hat{f} = O^{-1} .d = f + O^{-1} .\sigma .$$

We will illustrate the effects of noise on the direct inverse graphically in Section 6.

The fact that the data are both noisy and incomplete means that our problem is fundamentally ill-posed—there are many reconstructions of f permitted by the data. We can consider all the reconstructions that would give data consistent with those

actually measured by setting up a misfit statistic— χ^2 is often appropriate:

$$\chi^2 = \sum_{k=1}^N \frac{|\hat{d}_k - d_k|^2}{\sigma_k^2},$$

where d_k is the k^{th} measured datum, with error-bar σ_k , and \hat{d}_k is the corresponding datum that a trial reconstruction \hat{f} would produce in the absence of noise: $\hat{d}_k = [\mathbf{O} \cdot \hat{f}]_k$. Those reconstructions that give $\chi^2 \leq N$ are deemed to have "fit the data" and constitute the *feasible set* of \hat{f} . This *feasible set*, however, is incomprehensibly large: suppose we wish to reconstruct a 2-d image on an 8x8 pixel grid with just 16 grey levels; this gives a total number of 10^{77} possible reconstructions. Even if the data restricted the intensity of each pixel to vary only by (\pm) one level on average, the *feasible set* would still consist of 10^{30} possible reconstructions. This is enormous if you compare it with age of the universe, which is only 10^{17} seconds. Real problems are typically 128x128 pixel grids with 256 grey levels!

As we cannot even comprehend the total number of solutions, let alone compute and display them, we are forced to make a selection. We would like to say this is our ("best") estimate of the true f . Which solution should we select?

2. The principle of maximum entropy

If f is a *positive* and *additive* quantity—for example, a probability density function, or the intensity distribution of an optical picture, or the radio-flux distribution of an astronomical source—then the MaxEnt principle states we should choose that solution which maximises the Shannon-Jaynes entropy S (Jaynes 1983, Skilling 1988):

$$S = \sum_j f_j - m_j - f_j \log(f_j/m_j) ,$$

where f_j is the flux in the j^{th} pixel of the digitized reconstruction of f , and $\{m_j\}$ is a starting model which incorporates any prior knowledge we have about f ; in the absence of any such knowledge, all the m_j are set equal. If f is a normalised quantity such that $\sum f_j = 1$ and $\{m_j\}$ is constant, then entropy reduces to the more familiar form— $\sum f_j \cdot \log(f_j)$.

But why should we choose the MaxEnt solution? We shall try to answer this question by using a specific, and very simple, example and then give a more general interpretation of the MaxEnt choice.

2.1 The kangaroo problem

MaxEnt is not the only regularising function used in image reconstruction: several

have been recommended. We will follow Gull & Skilling (1983) in using the *kangaroo problem* to demonstrate our preference for the MaxEnt choice over the alternatives. It is a physicists' perversion of a mathematical argument given by Shore & Johnson (1980), where they formally show that MaxEnt is the only regularising function that yields self-consistent results when the same information is used in different ways. The *kangaroo problem* is as follows:

- Information:** (1) One third of kangaroos have blue eyes.
 (2) One third of kangaroos are left-handed.
Question: On the basis of this information alone, estimate the proportion of kangaroos that are both blue-eyed and left-handed.

Clearly, we do not have enough information to know the correct answer: all solutions of the type shown in the 2x2 contingency table of Fig. 1 (a) fit the data—these constitute the *feasible set* of solutions, each of which is equally likely. Figs. 1(b)-(d) show three of the myriads of feasible solutions: namely, the one with no correlation and the ones with the maximum positive and negative correlations, respectively. Although the data do not allow us to say which is the *correct* solution, our common sense compels us to the uncorrelated solution if we are forced to make a choice—no other single choice is defensible.

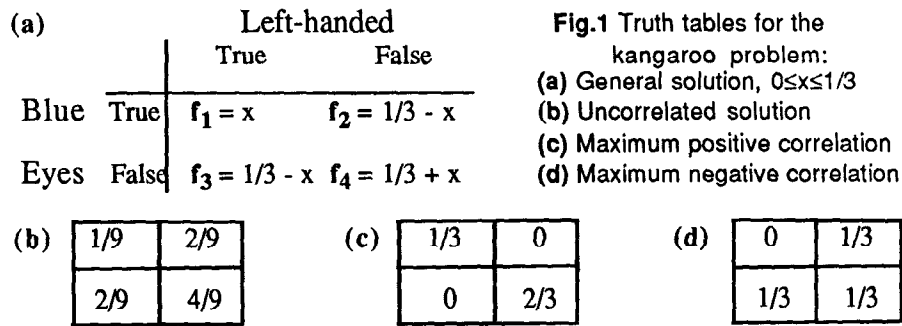


Table 1 shows the result of selecting the solution by maximising four commonly used regularising functions. For this very simple example, where common sense tells us the "best" answer when faced with insufficient (but noise-free) data, it is only the Shannon-Jaynes entropy that yields a sensible answer!

Table 1

Regularisation function	Proportion blue-eyed and left-handed (x)	Correlation
$-\sum f_j \log(f_j)$	1/9	Uncorrelated
$-\sum f_j^2$	1/12	Negative
$\sum \log(f_j)$	0.13013	Positive
$\sum f_j^{1/2}$	0.12176	Positive

Can we interpret the MaxEnt choice more generally?

2.2 The monkey argument

Our common sense recommended the uncorrelated solution because, intuitively, we knew that this was the least committal choice. The data itself did not rule out correlation but, without actual evidence, it was (*a priori*) more likely that the genes controlling handedness and eye-colour were on different chromosomes than on the same one. Although we cannot usually appeal to specific knowledge like genes and chromosomes, we can use the *monkey argument* of Gull & Daniell (1978) to see more generally that the MaxEnt choice is the one that is *maximally non-committal* about the information we do not have. The monkey argument can (again) be thought of as a physicists' perversion of the formal work of Shannon (1948) showing that entropy was a unique measure of information content. The monkey argument is as follows:

Imagine a large team of monkeys who make images ($\hat{\mathbf{f}}$), at random, by throwing small balls of flux at a (rectangular) grid. Eventually, they will generate all possible images. If we have some data relating to an object (\mathbf{f}), we can reject most of the monkey images because they will not give data consistent with the experimental measurements. Those images that are not rejected constitute the *feasible set*. If we are to select just one image from this *feasible set*, the image that the monkeys generate most often would be a sensible choice. This is because our hypothetical team of monkeys have no particular bias, and so this choice represents that image which is consistent with the measured data but, at the same time, is *least committal* about the data we do not have. This *preferred* image is the MaxEnt solution.

3. Model-fitting and least squares

The quantity of interest \mathbf{f} is usually a continuous quantity. For computational purposes, however, we digitize it into a discrete set of pixels $\{f_j\}$. This is not a limitation because we can digitize as finely as we like, but it does result in us having to estimate a large number of parameters (flux in each pixel) from a relatively small number of data. The problem tends to be grossly under-determined and, hence, we use MaxEnt to help us.

Sometimes we are more fortunate in that we have a functional model for \mathbf{f} —the sum of six δ -functions, or two Gaussians, for example. In this case \mathbf{f} can be parameterised by a handful of variables. We now have to estimate a small number of parameters from a relatively large number of data—the problem is over-determined. In these cases, and with suitable assumptions, the method of least squares is usually appropriate.

If we have a sound basis for our model, then model-fitting with least squares will give more accurate results than MaxEnt—we are using much more *prior* knowledge in the model-fitting procedure than we are in MaxEnt. If we do not have a functional model, or if our model is *ad hoc* ("try fitting Gaussians"), then we are better off using MaxEnt. It is possible, and perhaps to be recommended, that we combine the use of MaxEnt and model-fitting: use MaxEnt to obtain an initial reconstruction to get an

overall picture; if the MaxEnt reconstruction and our prior physical knowledge suggest a functional model, then use this in a least squares sense for further quantitative analysis.

4. General examples

We refer the reader to a comprehensive review by Gull & Skilling (1984) for numerous examples of the applications of MaxEnt. With their kind permission, a small selection of these are reproduced in Fig. 2. They illustrate the wide range of problems to which MaxEnt is now applied—forensic imaging, radio astronomy, plasma diagnostics, medical tomography, and blind deconvolution.

5. Difficult problems

The principle of MaxEnt is quite general and can be applied to any problem where the object of interest is a positive and additive quantity. Actually, finding the MaxEnt solution can be very difficult for non-linear problems because there are many local minima of χ^2 in a large parameter-space (typically 10^5 pixels). A particularly well-known example is the notorious Fourier *phase problem* in crystallography, the gravity of the situation being graphically illustrated in Fig. 3. We will not pursue this topic any further here except to state that, in general, the use of additional prior knowledge is essential for these problems (see, for example, Sivia 1987).

6. The Filter Difference Spectrometer

We now give an explicit example from neutron scattering—the deconvolution of data from a filter difference spectrometer (FDS). For the experimental and spectrometer details, the reader is referred to Taylor, *et al.*, (1984). The essential point for our purposes is that the FDS has a resolution function with a fairly sharp edge and a long decaying tail. The Cambridge algorithm was used throughout to maximise the entropy (Skilling & Bryan, 1984).

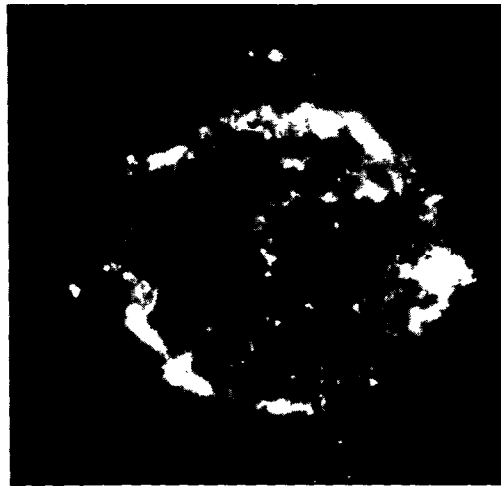
We start with simple simulations to highlight the differences between MaxEnt and a conventional direct inverse under "controlled" conditions. They do not mimic the FDS exactly but capture its salient features. For these simulations, the true spectrum $f(x)$ (scattering law) is shown in Fig. 4(a): it consists of two spikes separated by a low plateau on the left and a much broader peak on the right. This "truth" was generated on a grid of 128 pixels and convolved with a sharp-edged exponential $e^{-x/\tau}$, where $\tau = 15$ pixels, shown in Fig. 4(b), to create a noiseless data set of 128 points. A constant ("known") background equal to 10% of the peak datum was used and Gaussian random noise with a standard deviation equal to the square root of each datum was added. Fig. 4(b) shows this simulated data set when the peak datum was 10^8 (counts)—essentially noiseless. For this case, both MaxEnt and the direct inverse ($O^{-1}.d$) gave reconstructions indistinguishable from the truth (Fig. 4a). Figures 5 and 6 show the corresponding results when the data were made more noisy (fewer counts). The quality of the reconstructions deteriorates for both methods. Since the direct inverse does not take into account the fact that the data are noisy (Section 1), it produces numerous artifacts and deteriorates much more rapidly than MaxEnt.



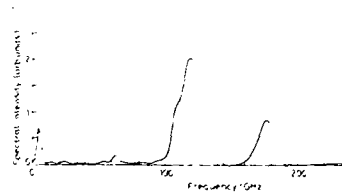
before after
Maximum entropy deconvolution
(UK Home Office)



ME X-ray tomography
(skull in perspex, FMI Ltd)



SNR Cas A at 5 GHz - 1024² ME image
(5-km telescope MRAO, Cambridge)



mm-wave Michelson interferometer spectrum of cyclotron emission from DITE tokamak
(Culham Laboratory)

BLIND	BLIND	BLIND
ORIGINAL	BLURRED	RECONSTRUCTED
r	MAXENT	r
TRUE P. S. F.		RECONSTRUCTED

"Blind" deconvolution of unknown blurring.
(left) true image & blurring
(middle) data as given to ME program
(right) reconstructions
(T.J. Newton)

Fig. 2 General examples of MaxEnt image reconstruction. Reproduced by courtesy of Drs. Gull and Skilling.



Fig. 3 A graphic illustration of the phase problem: (a) and (b) are the original images. (c) is the (Fourier) reconstruction which has the Fourier phases of (a) and Fourier amplitudes of (b); (d) is the reconstruction with the phases of (b) and the amplitudes of (a).

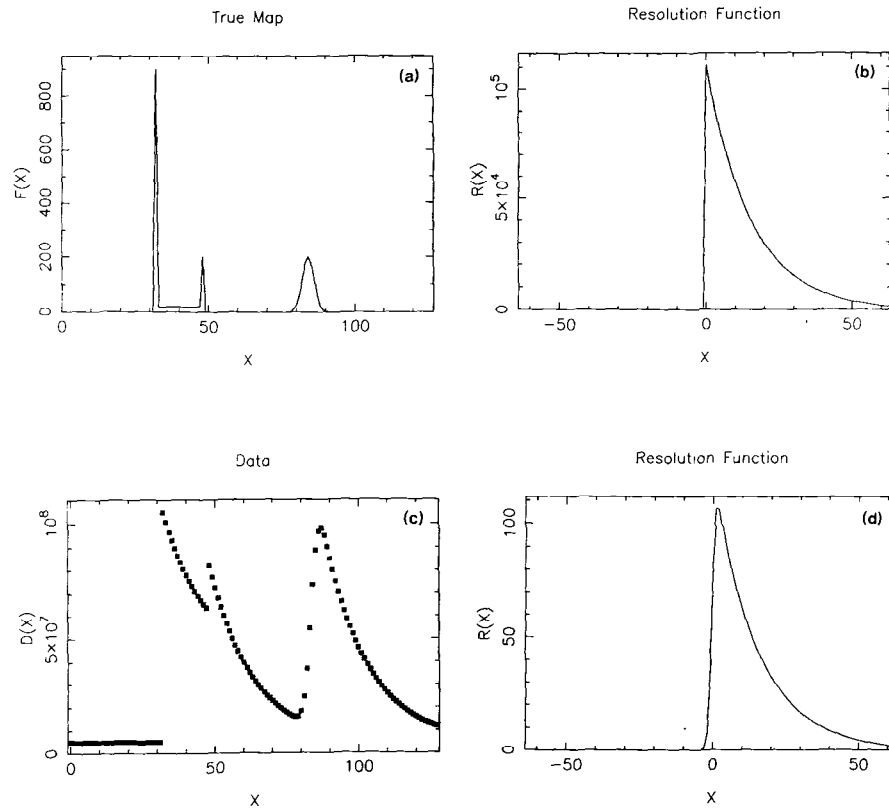


Fig. 4 (a) Spectrum or idealised scattering law, used in the FDS simulations. (b) A first approximation to the FDS resolution function: a sharp-edged exponential. (c) simulated data set with very good statistics. (d) A better approximation to the FDS resolution function: sharp-edged exponential convolved with a narrow Gaussian.

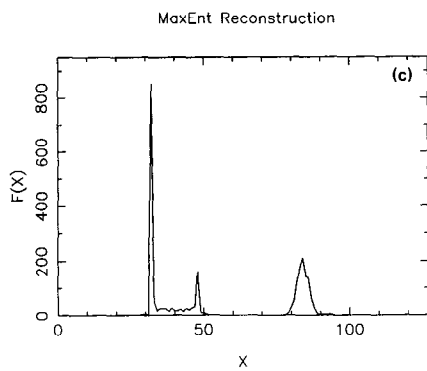
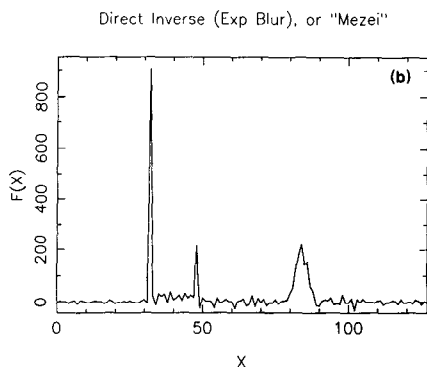
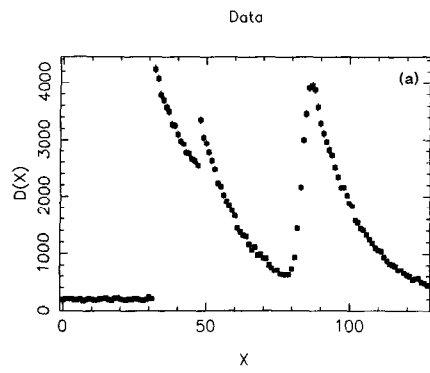


Fig. 5 (a) Simulated FDS data with some noise. (b) Direct inverse. (c) MaxEnt reconstruction.

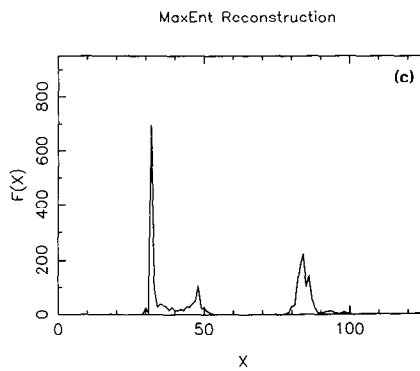
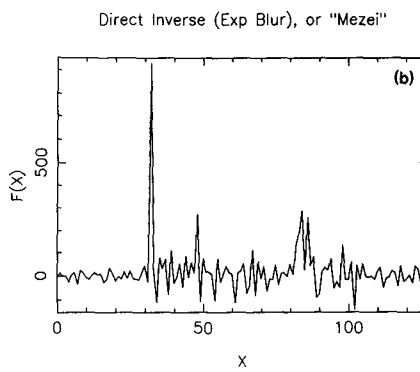
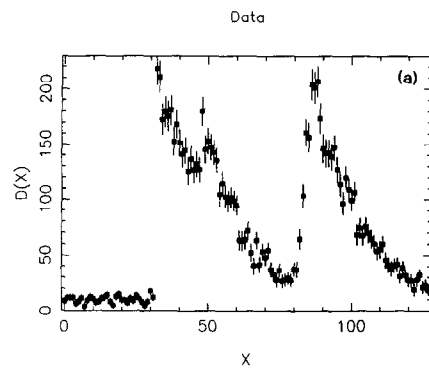


Fig. 6 (a) Simulated FDS data with a lot of noise. (b) Direct inverse. (c) MaxEnt reconstruction.

Rather than use the sharp-edged exponential above (Fig. 4b), we obtain a better approximation to the FDS resolution function if we convolve it with a narrow Gaussian (standard deviation of one pixel). The resulting simulated data, with good statistics, is shown in Fig. 7(a). Although we can deconvolve this new resolution function with the direct method in principle, we will only deconvolve the exponential component as is done in practice (Mezei & Vorderwisch 1989). This is because the inverse is easy to calculate if there is a sharp edge (by direct substitution), but more so because the inverse becomes badly conditioned (very sensitive to noise in the data) when the Gaussian component is included. With MaxEnt, however, we can safely deconvolve the "smoothed" resolution function. The inverse and MaxEnt reconstructions are shown in Figs. 7(b) & (c), respectively. The MaxEnt reconstruction shows much improved resolution and some noise suppression over the direct inverse.

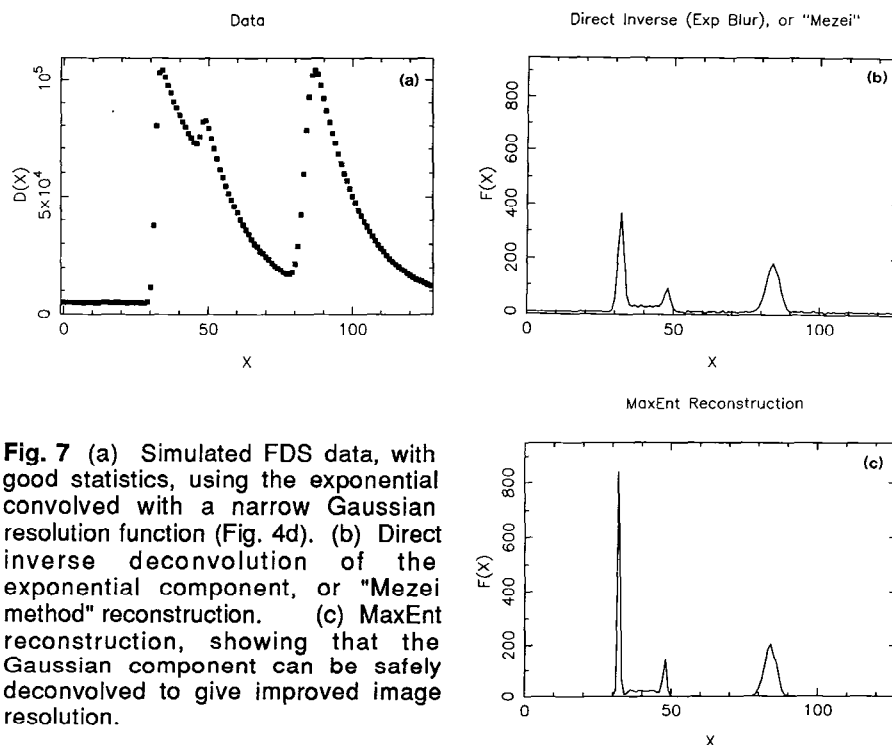


Fig. 7 (a) Simulated FDS data, with good statistics, using the exponential convolved with a narrow Gaussian resolution function (Fig. 4d). (b) Direct inverse deconvolution of the exponential component, or "Mezei method" reconstruction. (c) MaxEnt reconstruction, showing that the Gaussian component can be safely deconvolved to give improved image resolution.

Finally, we show the result of using MaxEnt on a real FDS data set. The data and resolution function were provided by Vorderwisch, experimental and analysis details being given in forthcoming papers (Vorderwisch 1989, and Sivia *et al.*, 1989). Fig. 8(a) shows the Be data for hexamethylene-tetramine (HMT) at 15 K taken at the Los Alamos Neutron Scattering Center (LANSCE). Fig. 8(b) shows the conventional *Filter Difference* spectrum: a crude hardware deconvolution obtained by subtracting the data obtained with Be and BeO filters. Fig. 8(c) shows the MaxEnt reconstruction, and Fig. 8(d) shows this overlaid on the direct inverse reconstruction

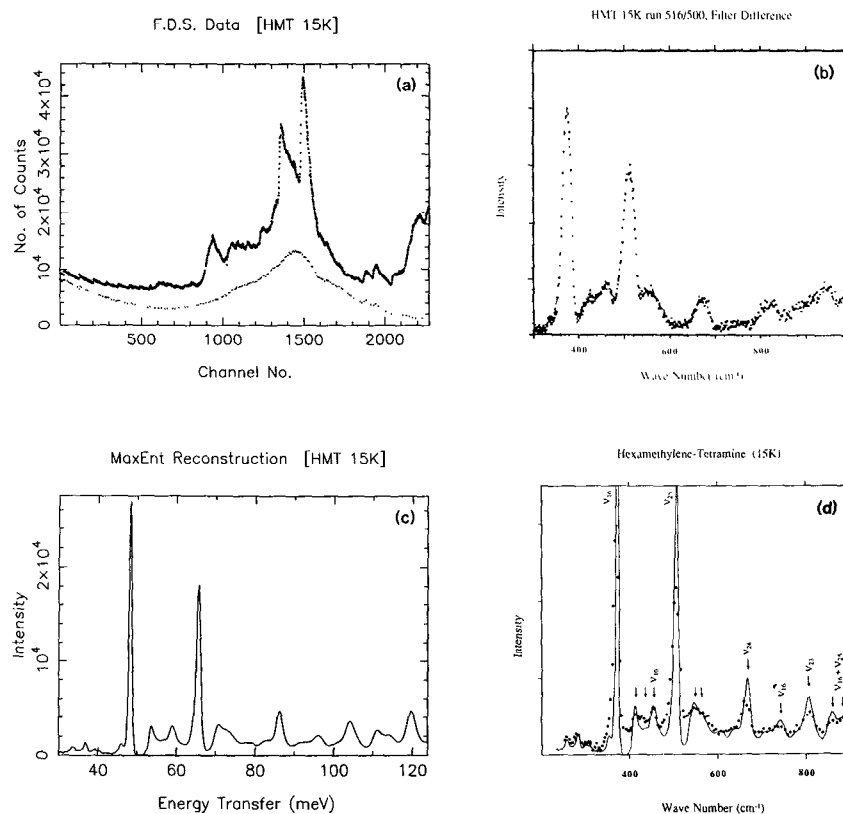


Fig. 8 (a) FDS data for Hexamethylene-Tetramine, at 15 K, taken with the Be filter at LANSCE. The channels are in increasing time-of-flight, or decreasing energy transfer. (b) The Filter Difference spectrum, or a crude hardware deconvolution obtained by subtracting data obtained with Be and BeO filters. (c) The MaxEnt reconstruction. (d) The direct inverse, or "Mezei method", reconstruction (dots) overlaid on the MaxEnt reconstruction.

mentioned above. As expected, we find that MaxEnt has improved the resolution and reduced the noise. The improvement is obvious, but not dramatic, in this particular example, because we had good statistics and the intrinsic Gaussian-like contribution to the resolution function is very narrow with little effect.

7. Concluding remarks

We have shown that MaxEnt provides an optimal criterion for selecting a positive image when faced with incomplete and noisy data. The MaxEnt choice can be interpreted as the maximally non-committal solution that is consistent with the data. As such, it tends to be less noisy and has fewer artifacts than conventional methods, thus making it easier to interpret the results.

We mention in passing that a unified approach to all data analysis (MaxEnt, model-fitting, or whatever) can be achieved by casting all such problems in the probabilistic framework of a Bayesian analysis. This not only gives us the way to select the optimal answer to any given problem, but it also tells us how to estimate the reliability of that solution; unfortunately, however, the error analysis is usually impossible to implement in practice except for the smallest of problems. The difficulty does not arise because we are using MaxEnt, but because we are trying to estimate a large number of parameters.

Acknowledgements

The approach to MaxEnt provided here is due almost entirely to Steve Gull and John Skilling, who brought me up on a good dose of MaxEnt and Bayes' theorem. Many thanks to Richard Silver and Roger Pynn for their support and encouragement in applying MaxEnt to neutron scattering problems in general, and to Peter Vorderwisch for working with me on the FDS example.

References

- Gull, S. F. & Daniell, G. J., 1978, *Nature*, 272, 686-690.
 Gull, S. F. & Skilling, J., 1984, *IEE Proc.*, 131F, 646-659.
 Jaynes, E. T., 1983, (Collected Works) *Papers on Probability, Statistics and Statistical Physics*, ed. by R.D. Rosenkrantz, Dordrecht Holland, Reidel.
 Mezei, F. & Vorderwisch, P., *Physica B*, in press.
 Shannon, C. E., 1948, *Bell System Tech. J.*, 27, 379-423 and 623-656.
 Shore, J. E. & Johnson, R. W., 1980, *IEEE Trans., Information Theory*, Vol. 1, IT-26, No. 1, 26-37.
 Sivia, D. S., 1987, *Phase Extension Methods*, Ph.D. Thesis, Cambridge University.
 Sivia, D. S., Vorderwisch, P., Silver, R. N., 1989, "Maximum Entropy Deconvolution of the Filter Difference Spectrometer," in preparation.
 Skilling, J., 1988, "The Axioms of Maximum Entropy," *in* *Maximum Entropy & Bayesian Methods in Science and Engineering* (Vol. 1), ed. by G. Erickson and C. R. Smith.
 Skilling, J. and Bryan, R.K., 1984, *Mon. Not. R. Astr. Soc.*, 211, 114-124.
 Taylor, A. D., Wood, E. J., Goldstone, J. A., and Eckert, J., 1984, *Nucl. Instr. and Meth.*, 221, 408-418.
 Vorderwisch, P., 1989, "The Resolution Function of the Filter Difference Spectrometer," in preparation.