Mon. Not. R. astr. Soc. (1980) **191**, 69–79

Deconvolution by maximum entropy, as illustrated by application to the jet of M87

R. K. Bryan and J. Skilling *Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge CB3 9EW*

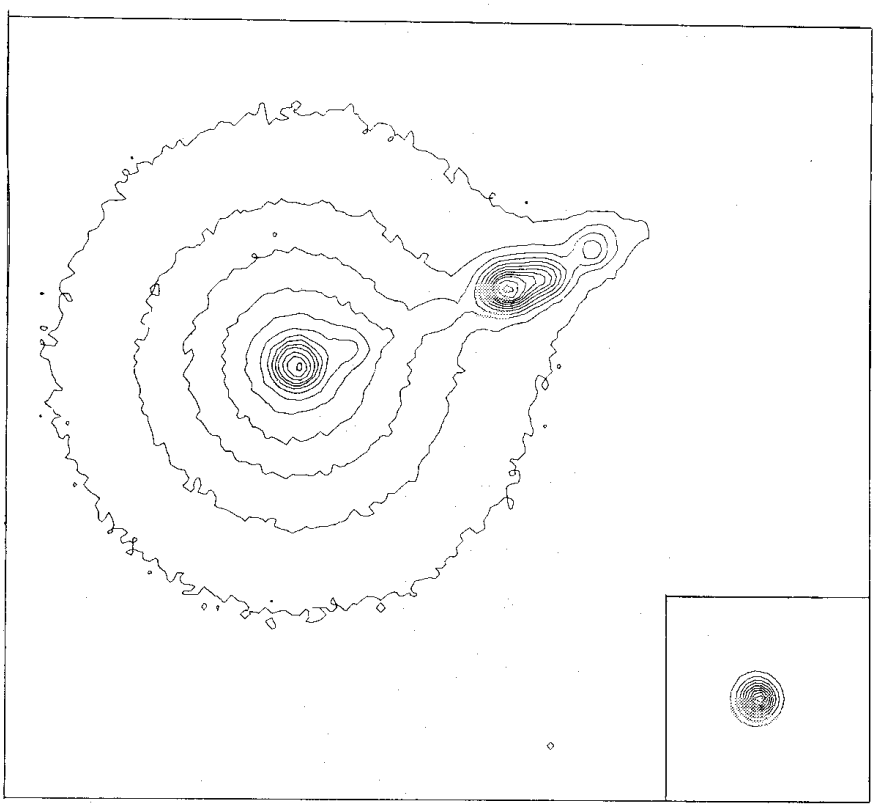
Received 1979 August 24

Summary. We present an improved method of deconvolving blurred and noisy data, appropriate for pictures of the sky taken with astronomical telescopes. The maximum entropy criterion gives the smoothest possible structure of the sky consistent with the observed image. Our improvement lies in the consistency test; we force the noise to have its correct statistical distribution. This provides greater resolution and more accurate fitting. The method is illustrated by deconvolving an optical photograph of the nuclear regions of M87.

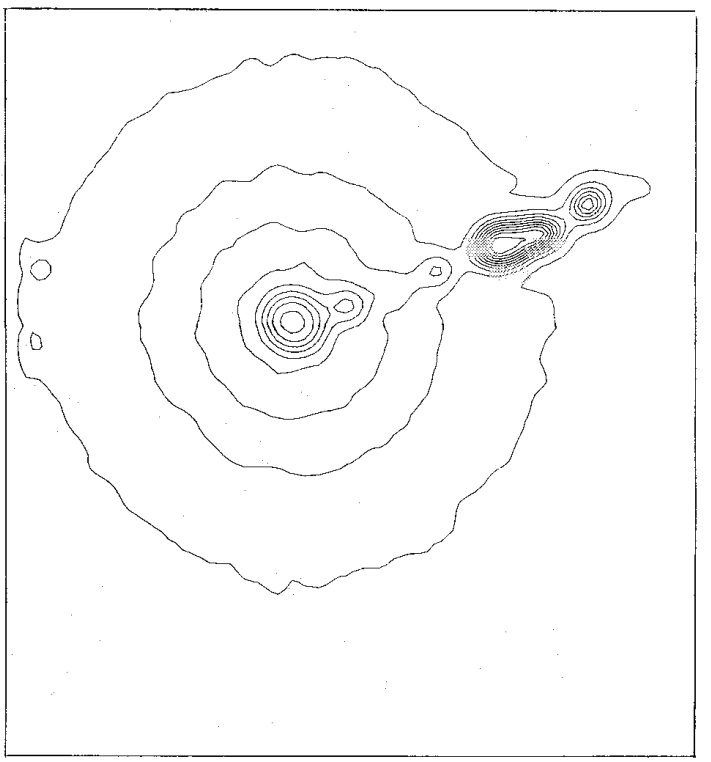
1 Introduction

There are many cases where data from experiments are the result of convolving the physical quantities of interest with the response of an instrument. Unfortunately, because most imaging problems are ill-conditioned (Andrews & Hunt 1977), a direct deconvolution can only be performed for the ideal case of complete and noise-free data, and in practice any noise on the data can be magnified by very large factors. Constraining the deconvolution by the maximum entropy criterion surmounts this difficulty, and gives the most uniform solution consistent with the data. We describe the maximum entropy method in Section 2. As applied in the past (Abels 1974; Wernecke & d'Addario 1977; Gull & Daniell 1978), however, the χ^2 statistic used to test for consistency with the data did not necessarily produce a restored noise distribution appropriate to the assumed noise model, as shown in Section 3. We propose, in Section 4, the use of a consistency test designed to give restored noise values with the correct distribution, and apply it to the maximum entropy method in Section 5.

To illustrate the method we used an optical photograph of the nucleus and jet of M87 (Plate 1 and Fig. 1a) taken with the 200-in telescope at Mount Palomar and digitized on a 128×128 pixel grid. It is appropriate to use a Gaussian noise model for a photographic plate (see, e.g. Andrews & Hunt 1977), with the standard deviation of the noise dependent on the image density. Fig. 2 shows the noise characteristics of the plate used: the noise is greatest where the image is densest. For the point-spread function of the telescope we used its response to a star, symmetrized to reduce noise (inset, Fig. 1a). All of these data were



(a)



(b)

Figure 1. (a) Contour map of photograph (Plate 1) of M87. Contour levels are: 10, 20, 30,..., 100, 120, 140, 160.... Inset: Contour map of the telescope point-spread function, to the same scale. Contour levels are 10, 20, 30,..., 90 per cent of the central maximum. (b) Deconvolution by the χ^2 statistic. Contour levels as in (a). (c) Deconvolution by the E statistic. Contour levels as in (a) and (b).

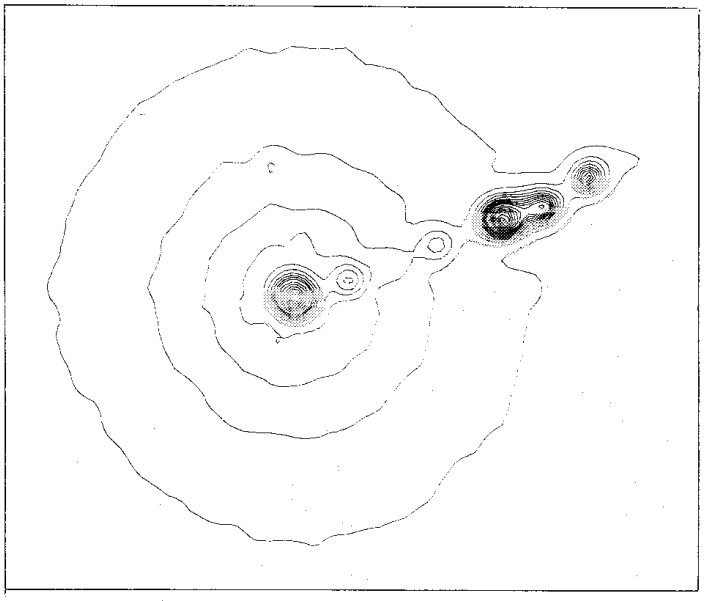


Figure 1 (c)

supplied by J. Lorre of the Jet Propulsion Laboratory. Arp & Lorre (1976) have previously deconvolved this photograph by a Wiener filter technique, and our results can be compared with theirs.

2 Formulation of the maximum entropy criterion

To formulate the deconvolution problem in digital terms we divide the photographic image into some (large) number N of equal-sized cells, the jth cell having density d_j. We wish to find an estimate $\mathbf{f}$ of the original sky intensity distribution which, after being blurred by the point-spread function $\mathbf{b}$, produced the image $\mathbf{d}$. We suppose that the recorded image is noisy, with the noise in each cell independent and Gaussian, of standard deviation σ_j in cell j. Thus the quantities

$$n_k = \frac{1}{\sigma_k}\left(d_k - \sum_{j=1}^{N} f_j b_{k-j}\right) \qquad (k = 1, 2, \ldots, N) \tag{1}$$

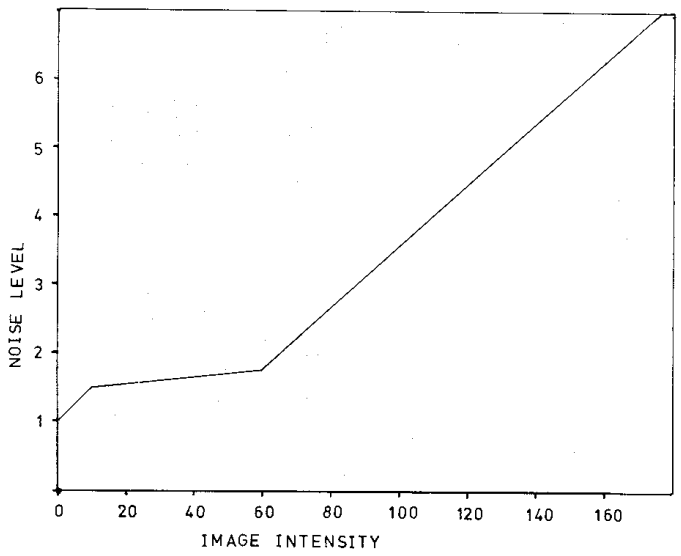


Figure 2. Noise level σ as a function of the image intensity d on the photographic plate.

should form a set of N random samples from the normal distribution $\mathcal{N}(0, 1)$. In the standard two-dimensional problem the suffices j and k stand for two-dimensional integer coordinates. For simplicity we take the point-spread function to be constant over the image, although this assumption, and that of Gaussian noise, could be relaxed if necessary.

Given the quantities **d**, **b** and **σ** as supplied by the experimenter, we construct a probability density $\pi(\mathbf{f})$ for the N-dimensional vector **f**. This is taken to be simply proportional to the likelihood or conditional probability $P(\mathbf{d}|\mathbf{f})$ that the image **d** can be produced from the estimate of the sky **f** (Skilling, Strong & Bennett 1979). We then select some confidence level x ($0 < x < 1$, $x = 0.95$ would give 95 per cent confidence) and construct a confidence domain $\mathscr{D}(x)$ of the most likely maps **f**, drawing the boundary of $\mathscr{D}$ to enclose the selected proportion x of the probability:

$$\mathscr{D}(x) = \{\mathbf{f}: \pi(\mathbf{f}) \geqslant q\} \tag{2}$$

where q is implicitly defined by

$$\int_{\mathscr{D}(x)} \pi(\mathbf{f})\, d^N f = x.$$

In practice one wants a single representative map **g** from this domain. Various ways of selecting a single map have been proposed, all of which involve choosing extreme values of some functional of **f**. Turchin & Turovtseva (1974), in their method of statistical regularization, would minimize the discrete analogue of the curvature 2-norm $[\![\int [f''(\xi)]^2 d\xi]\!]^{1/2}$. Abels (1974), Wernecke & d'Addario (1977) and Wernecke (1977) maximize

$$\sum_j \log f_j,$$

which they call the entropy of **f**. Frieden (1972) and Gull & Daniell (1978) maximize a different form

$$-\sum_j f_j \log f_j$$

of the entropy. We prefer, and shall use, the latter formulation, because (1) the intensity f_j in each cell is automatically positive, (2) this functional smooths noise on both bright and faint regions of the sky, (3) the functional does not discriminate against super-resolution if this is suggested by the data, and (4) when normalized to

$$\sum f = 1$$

it represents (minus) the information content (Shannon 1948) of the configurational structure of **f**. Thus we select our 'maximum entropy' representative by choosing that map **g** for which

$$S = -\sum_{j=1}^{N} p_j \log p_j, \qquad p_j = f_j / \Sigma f \tag{3}$$

takes its greatest value in, say, $\mathscr{D}(0.95)$. We are then 95 per cent certain that the true sky lies in $\mathscr{D}(0.95)$ and contains *more* configurational information than the maximum entropy map. The maximum entropy map is minimally informative, whilst still being consistent with the data, so any structure seen in it probably corresponds to real structure in the true sky – at least there is evidence for such structure in the data.

Plate 1. Original photograph of M87, as digitized.

[*facing page* 72]

Plate 2. Photograph of our deconvolution (Fig. lc).

3 Use of χ^2 statistic

The most straightforward way of constructing the probability density $\pi(\mathbf{f})$, and thence the confidence domain $\mathscr{D}$, is to use the assumption of Gaussian noise in each cell and set

$$\pi(\mathbf{f}) \propto \exp\left(-\tfrac{1}{2}\sum_k n_k^2\right) \tag{4}$$

where the noise coefficients n_k are independent and vary with $\mathbf{f}$ according to equation (1). The statistic

$$\chi^2 \equiv \sum_k n_k^2$$

has a χ^2 distribution with N degrees of freedom. Thus the 95 per cent confidence domain is

$$\mathscr{D}(0.95) = \{\mathbf{f}: \chi^2 \leqslant \chi^2_{0.95}\}$$

where $\chi^2_{0.95}$ is the 95 per cent point of the distribution, given by

$$\chi^2_{0.95} \simeq N + 1.645(2N)^{1/2} \simeq 16682$$

for $N = 128 \times 128 \gg 1$. This domain is an ellipsoid in f space (Fig. 3).

If the flat map (for which S has its unique unconstrained maximum) lies within this domain, then that will be our deconvolution: consistency with the data is attainable without any variation at all in intensity. Except in this unusual event, the solution for $\mathbf{f}$ will be on the surface $\chi^2 = \chi^2_{0.95}$. The resulting map will necessarily be unique, since the surfaces S = constant and χ^2 = constant are both convex. Introducing a Lagrange multiplier λ for the constraint, we maximize

$$Q = S - \lambda\chi^2 \quad \text{under} \quad \chi^2 = \chi^2_{0.95}. \tag{5}$$

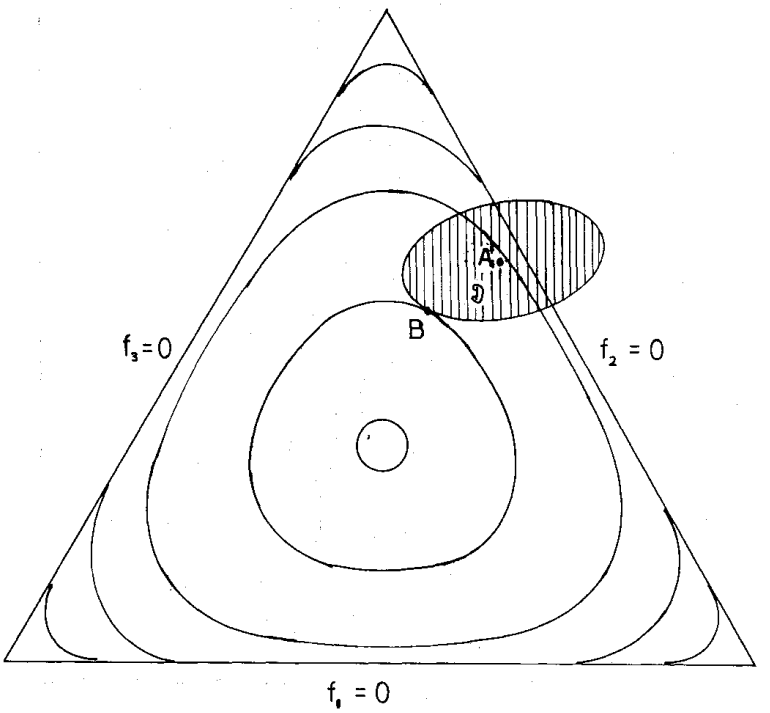


Figure 3. The S criterion and the χ^2 statistic in f space for a 3-cell map normalized to $\Sigma f = 1$. S surfaces are convex and χ^2 surfaces are ellipsoids. A is the map which fits the data exactly: B is the maximum entropy map.

Table 1. Residuals n_k around the central peak of the map as deconvolved using χ^2.

1.13	1.10	0.30	1.08	0.03	-0.32	1.06	0.05	0.97	0.96
0.59	2.22	2.21	3.23	3.23	1.93	0.30	-0.54	-0.22	0.24
0.57	3.42	5.21	6.81	6.38	4.70	2.48	0.71	-0.00	0.20
3.29	5.98	8.39	9.91	9.78	7.45	4.16	1.15	0.58	0.64
3.08	6.79	9.99	10.95	11.26	10.04	5.97	1.99	0.20	0.10
3.95	5.74	9.46	11.54	11.68	10.55	5.84	1.64	1.11	0.33
2.26	4.78	7.15	9.23	10.04	8.39	3.71	1.12	0.29	0.25
0.78	2.10	4.21	5.48	5.59	4.51	2.14	0.41	0.36	0.06
0.21	1.20	2.18	2.79	2.20	2.14	1.08	0.65	1.88	0.46
0.17	0.28	0.76	-0.05	0.34	-0.19	-0.13	0.32	1.05	1.27

This is the same formulation as Gull & Daniell (1978) except that they set $\chi^2 = N = 16384$, corresponding to the boundary of the 50 per cent confidence domain. With $N \gg 1$, there is little difference between the 50 per cent and the 95 per cent domains.

We solved the maximization (5) numerically for the photograph of M87, using a quadratic optimization, and Fig. 1(b) is a contour map of the solution. As expected of maximum

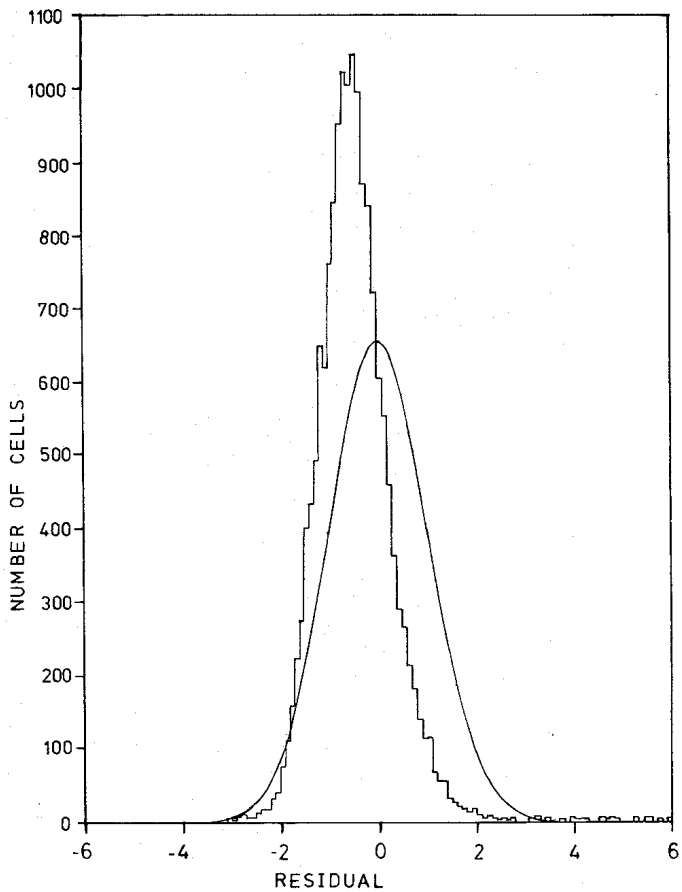


Figure 4. Histogram of χ^2 residuals compared with unit Gaussian.

entropy maps, it is much smoother than the data, yet the resolution has increased to show more structure in the jet. The major peaks are, however, weaker than those in the data, whereas one might have expected a deconvolution to enhance any real peaks. Presumably the normalized residuals n_k are large and positive there; this is confirmed by Table 1, which tabulates the residuals near the central peak. These large residuals, combined with the large noise level σ on highly exposed parts of the photographic plate, allow the peaks of the map to relax a long way down towards the average level. The effect is to make the residuals markedly non-Gaussian. In Fig. 4 we compare their histogram with the expected Gaussian of unit width. There are 45 points not shown beyond six standard deviations. Furthermore, the bulk of the histogram is off-centre and too narrow. This means that the background, which includes most points on the map, is shifted systematically from the data, and also follows the variations in the data too closely.

The χ^2 statistic has fitted the *variance* of the histogram to the expected value (close to unity) but has not constrained the *shape*. Constraining χ^2 to a smaller value could make the peaks of the map match the data more closely, but will lead to spurious resolution elsewhere, with noise on the data being interpreted as true signal (Gull & Daniell 1978). A different statistic is needed.

4 The 'exact error fitting' statistic E

We need to construct a statistical test which fits the noise residuals n_k to their known distribution (in our case $\mathcal{N}(0, 1)$). This could be accomplished by fitting several different moments of the histogram, as well as the variance χ^2, but each moment needs a separate Lagrange multiplier and such an approach is computationally difficult. Instead, we sort the residuals n_k into ascending order to give the 'order-statistics' $n_{(i)}$

$$n_{(1)} < n_{(2)} < \ldots < n_{(N)} \tag{6}$$

(suffixes in brackets denote sorted quantities). Were these to be from an 'exact' normal distribution, the ith sorted residual would be

$$v_{(i)} = \Phi^{-1}\left(\frac{i - \tfrac{1}{2}}{N}\right) \qquad (i = 1, 2, \ldots, N) \tag{7}$$

where

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \exp\left(-\tfrac{1}{2}u^2\right) du$$

is the cumulative normal probability.

We shall use the distance

$$E = \left\{ \sum_{i=1}^{N} (n_{(i)} - v_{(i)})^2 \right\}^{1/2} \tag{8}$$

between $\mathbf{n}$ and its 'exact' form $\boldsymbol{\nu}$ as our statistic, requiring it to be sufficiently small to ensure that the n_k are closely Gaussian. To evaluate the expected value of E, we first write the probability distribution of $n_{(i)}$ as the binomial distribution

$$P(\phi \leqslant \Phi(n_{(i)}) \leqslant \phi + d\phi) = \frac{N!}{(i-1)!\,(N-i)!}\, \phi^{i-1}(1-\phi)^{N-i}\, d\phi$$

(see, e.g. Kendall & Stuart 1977). Then we integrate over this to obtain the expected square difference between $n_{(i)}$ and $\nu_{(i)}$.

$$\langle (n_{(i)} - \nu_{(i)})^2 \rangle \simeq \langle [\Phi(n_{(i)}) - \Phi(\nu_{(i)})]^2 \rangle / [\Phi'(\nu_{(i)})]^2$$

$$\simeq N^{-1} \frac{i - \tfrac{1}{2}}{N} \left(1 - \frac{i - \tfrac{1}{2}}{N}\right) 2\pi \exp(\nu_{(i)}^2) \qquad (N \gg 1).$$

Summing this over i to obtain $\langle E^2 \rangle$, the summation is dominated (albeit weakly) by values of i away from the central fraction of the distribution, so that the asymptotic form

$$\exp(\nu_{(i)}^2) \sim \frac{1}{4\pi} \left(\frac{N}{i - \tfrac{1}{2}}\right)^2 \Big/ \log\left(\frac{N}{i - \tfrac{1}{2}}\right)$$

for $1 \lesssim i \ll N/2$, and the similar form for $N/2 \ll i \lesssim N$, can be used. Hence

$$\langle E^2 \rangle \simeq \sum_{i=1}^{N/2} \frac{1}{(i - \tfrac{1}{2}) \log(N/(i - \tfrac{1}{2}))} \simeq \log \log N. \tag{9}$$

Thus $\langle E \rangle \sim (\log \log N)^{1/2}$, so that our E statistic is always expected to be O(1) for any practical value of N. Direct computer simulations for $N = 16384$ confirm this, and give a 95 per cent confidence limit $E < E_{0.95} = 2.8$.

5 Maximum entropy with exact error fitting

We obtain our maximum entropy map by maximizing S over the domain defined by $E < E_{0.95}$, this procedure being effected by the following numerical algorithm.

(1) Start with initial trial map (flat).

(2) Sort the residuals n_k of the current map.

(3) Perform one (quadratic) iteration towards the maximum of $Q = S - \lambda E^2$, where E is defined on the ordering obtained in step 2, and λ is chosen to aim at $E = E_{0.95}$.

(4) Go to 2, or stop if converged.

The algorithm stops when a map is found which has maximum entropy on a surface $E = E_{0.95}$ and also correctly ordered residuals. This solution does *not*, however, necessarily have the maximum entropy over all the possible $E_{0.95}$ surfaces corresponding to different orderings of the residuals. Despite this apparent non-uniqueness, we show in the Appendix that the expected uncertainty that it allows corresponds to a distance in residual space between our solution **f** and the true maximum entropy map **g** of less than 1, i.e.

$$\sum_k \left(\frac{1}{\sigma_k} \sum_j b_{k-j}(f_j - g_j)\right)^2 \lesssim 1.$$

For large datasets the permitted results will be indistinguishable for all practical purposes, and the algorithm will yield an effectively unique map.

We applied this method to the picture of M87 and the results are displayed in Plate 2 and Fig. 1(c). The two main peaks have clearly increased greatly in magnitude and the subsidiary peaks are resolved more clearly. The increase in resolution is similar to that achieved by Arp & Lorre (1976) using a Wiener filter. The peaks in their restoration are, however, surrounded by dark haloes, and the background appears to contain faint light and dark patches, both of which are artefacts caused by filter cut-off at high spatial frequencies. In comparison, the background of our restoration remains smooth, and there is no sign of

Table 2. Residuals n_k around the central peak of the map as deconvolved using E.

2.32	1.72	0.23	0.50	-0.65	-0.56	1.51	1.07	2.14	2.03
0.73	1.05	-0.33	-0.05	-0.10	-0.76	-1.11	-0.48	0.76	1.43
-0.78	0.14	0.31	1.21	0.55	-0.53	-1.17	-0.76	0.39	1.38
0.83	1.31	2.19	3.08	2.77	0.54	-1.42	-1.90	0.18	1.67
-0.21	1.28	3.28	3.28	3.67	2.97	-0.42	-2.00	-0.77	1.01
0.89	-0.07	2.50	4.27	4.33	3.68	-0.71	-2.52	0.14	1.24
-0.13	-0.02	0.54	2.02	3.05	1.67	-2.33	-2.20	-0.23	1.32
-0.26	-1.08	-0.73	-0.54	-0.68	-1.10	-1.96	-1.38	0.65	1.30
0.57	0.23	-0.12	-0.44	-1.38	-0.77	-0.56	0.53	2.85	1.77
1.27	0.80	0.64	-0.66	-0.34	-0.49	0.29	1.36	2.33	2.51

ringing near the peaks. The residuals near the central peak (Table 2) are still systematically positive, but cover a very much smaller area, and are (of course!) no larger than the maximum expected for 16384 samples from a Gaussian distribution. Positive residuals at peaks are an almost inevitable result of using a smoothness criterion in a restoration method. The histogram of residuals is, as expected, a very close fit to a Gaussian (Fig. 5).

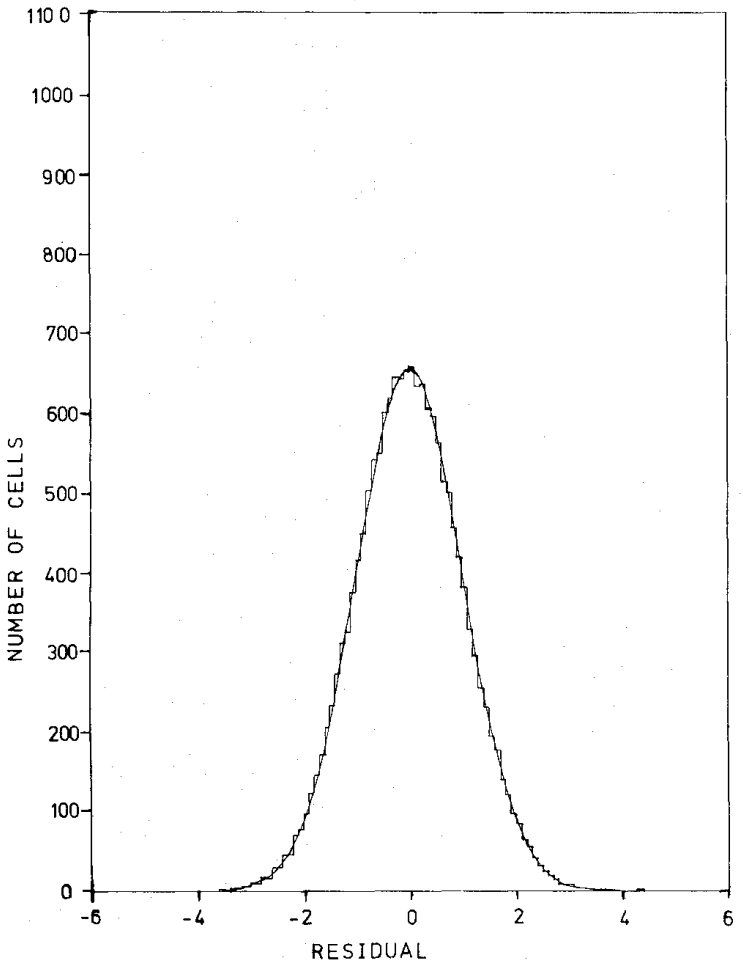


Figure 5. Histogram of E residuals compared with unit Gaussian.

This algorithm has a further useful feature. Suppose an isolated point in the data is corrupted by more than about four standard deviations. It will then be impossible to deconvolve this single point with the telescope point-spread function. No map will be found to satisfy $E < E_{0.95}$. However, such corrupted points are automatically picked out as outliers when the residuals are sorted, and can be inspected individually. Our data for M87 contained two such points, which were effectively eliminated by increasing the corresponding errors σ_k considerably.

6 Conclusions

The maximum entropy algorithm gives good deconvolutions of blurred and noisy pictures. All versions of the algorithm use a statistical test of goodness-of-fit between the actual data and what would be expected on given hypotheses about the brightness distribution on the sky.

For the best deconvolutions, this statistical test should be chosen with care. We suggest fitting the residuals to their correct statistical distribution, as supplied by the experimenter, via our E statistic: certainly this gives results superior to those obtained by other authors using the χ^2 statistic. This idea should also be applicable to other restoration problems, beyond the linear deconvolution case presented here. Furthermore, it may be that some other test, perhaps using the correlation function of the residuals, would be even better. Meanwhile, we have a program which can routinely deconvolve images described by equation (1) on a 128×128 raster in 5 or 10 min of CPU time on an IBM 370/165.

Acknowledgments

We are grateful to Dr J. Lorre for supplying the test data in digital form, and also to Dr S. F. Gull for help with the data and for very many conversations on maximum entropy. The computations were performed on the IBM 370/165 of the University of Cambridge Computer Laboratory. One of us (RKB) is in receipt of financial support from the Science Research Council.

References

Abels, J. G., 1974. *Astr. Astrophys. Suppl.*, **15**, 383.
Andrews, H. C. & Hunt, B. R., 1977. *Digital Image Restoration*, Prentice-Hall, New Jersey.
Arp, H. & Lorre, J., 1976. *Astrophys. J.*, **210**, 58.
Frieden, B. R., 1972. *J. opt. Soc. Am.*, **62**, 511.
Gull, S. F. & Daniell, G. J., 1978. *Nature*, **272**, 686.
Kendall, M. & Stuart, A., 1977. *Advanced Theory of Statistics*, 4th edn, ch. 14, Griffin, London.
Shannon, C. E., 1948. *Bell System Tech. J.*, **27**, 379 and 623.
Skilling, J., Strong, A. W. & Bennett, K., 1979. *Mon. Not. R. astr. Soc.*, **187**, 145.
Turchin, V. F. & Turovtseva, L. S., 1974. *Optics and Spectroscopy*, **36**, 162.
Wernecke, S. J., 1977. *Rad. Sci.*, **12**, 831.
Wernecke, S. J. & d'Addario, L. R., 1977. *IEEE Trans.*, **C26**, 351.

Appendix: quasi-uniqueness of solution using E statistic

Consider the algorithm as it operates in the space of residuals $\mathbf{n}$ (Fig. A1). Corresponding to each ordering, the condition $E \leqslant E_{0.95}$ gives a N-sphere of radius $E_{0.95}$. The centres of these spheres ($\mathbf{n} = \boldsymbol{\nu}$, $E = 0$) have the residuals fitting the Gaussian exactly, and hence lie on the surface $\chi^2 = N$, which is a larger sphere of radius $\sqrt{N}$. It follows that the $N!$ smaller

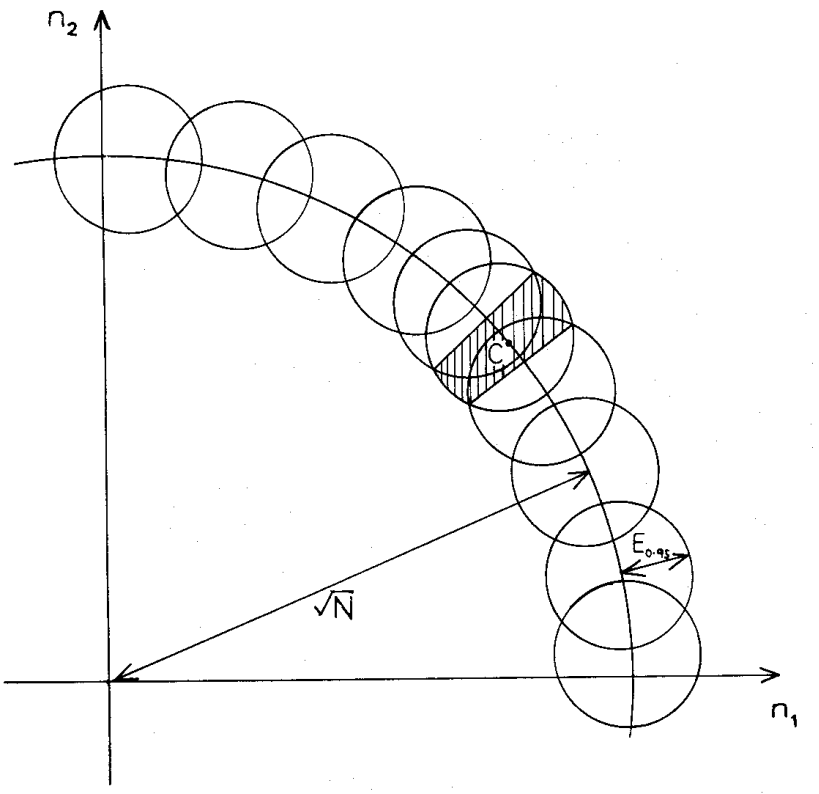


Figure A1. The E statistic in n space. Only the shaded region of the sphere centred at C has the residuals correctly sorted.

spheres overlap by at least a factor $\omega = N!(E_{0.95}/\sqrt{N})^N$. Only at most the small fraction $1/\omega$ of the points on each sphere correspond to residuals $\mathbf{n}$ which are ordered in the same way as the corresponding $\boldsymbol{\nu}$, and only the small fraction $1/\omega$ of the $N!$ conceivable maxima of S are actually permitted. These $N!/\omega$ points (still a large number) will be expected to lie on adjacent (and overlapping) spheres, because both χ^2 and S are convex. The volume filled by these adjacent spheres being at most a fraction $1/\omega$ of the volume filled by the sphere of radius $\sqrt{N}$, it follows that the centres of the spheres will be expected to lie within at most the Nth root of this (i.e. a fraction $\omega^{-1/N}$) of the distance $\sqrt{N}$. The permitted maxima of S will also lie within this distance $\omega^{-1/N}\sqrt{N} \simeq e/E_{0.95} \simeq 0.8$ of each other.