J. R. Statist. Soc. B (1992)
54, *No.* 1, *pp.* 41–81

Maximum Entropy and the Nearly Black Object

By DAVID L. DONOHO, IAIN M. JOHNSTONE†, JEFFREY C. HOCH and ALAN S. STERN

University of California, Berkeley, USA *Stanford University, USA* *Rowland Institute for Science, Cambridge, USA*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, April 17th, 1991,* Dr F. Critchley *in the Chair*]

SUMMARY

Maximum entropy (ME) inversion is a non-linear inversion technique for inverse problems where the object to be recovered is known to be positive. It has been applied in areas ranging from radio astronomy to various forms of spectroscopy, sometimes with dramatic success.

ill-posedness occurs when the effective dimension of $\mathbf{y}$ is considerably smaller than the dimension of $\mathbf{x}$, i.e. when the operator K has few singular values which are significantly different from 0 (Bertero *et al.*, 1985; Bertero and Pike, 1982; Barakat and Newsam, 1985a, b). In one of the most common examples, image deblurring, K would be a smoothing transform, with singular values small at singular vectors corresponding to high frequencies, so that the detailed high frequency information in $\mathbf{x}$ is lost; the inverse problem is to recover $\mathbf{x}$, with high frequencies restored (if possible).

Were it not for the ill-posedness, it would be natural to approach the problem by least squares. After all equation (1) is just a linear model, and an estimate of $\mathbf{x}$ can be obtained from the least squares principle $\hat{\mathbf{x}}_{\mathrm{LS}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - K\mathbf{x}\|_2^2$, giving $\hat{\mathbf{x}}_{\mathrm{LS}} = (K^{\mathrm{T}}K)^{-1}K^{\mathrm{T}}\mathbf{y}$. However, because of ill-posedness, this estimate is either undefined, or else has very poor performance, even if we interpret the matrix inverse as a generalized inverse. It is by now traditional to approach such problems by least squares regularization. Then we estimate $\mathbf{x}$ by the solution to the optimization problem $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - K\mathbf{x}\|_2^2 + 2\lambda\|\mathbf{x}\|_2^2$, which gives the formula $\hat{\mathbf{x}}_{\mathrm{RLS}} = (K^{\mathrm{T}}K + \lambda I)^{-1}K^{\mathrm{T}}\mathbf{y}$. Here λ is a tuning constant specified by the user in some way. This idea has been used in many fields and also goes by many other names, such as ridge regression, penalized likelihood and damped least squares.

In this paper we focus on problems where the object $\mathbf{x}$ to be recovered has non-negative co-ordinates. Think of images, chemical spectra or other measurements of intensities. In this context, ME (Gull and Daniell, 1978) is a regularization method which gives an estimate of $\mathbf{x}$ by the prescription

$$\max_{\mathbf{x}}\Big(-\sum_i x_i \log x_i\Big) \qquad \text{subject to } \|\mathbf{y} - K\mathbf{x}\|_2^2 \leqslant S^2; \tag{2}$$

see also Wernecke and D'Addario (1977) and Frieden (1972) for related definitions of ME. We prefer to define it in the equivalent form

$$\hat{\mathbf{x}}_{\mathrm{ME}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - K\mathbf{x}\|_2^2 + 2\lambda \sum_i x_i \log x_i, \tag{3}$$

which emphasizes the similarity to regularized least squares. There is a (data-dependent) one-to-one correspondence between λ in equation (3) and S in expression (2) which makes the two optimization problems have the same solution.

Although ME has a formal similarity with least squares regularization (we are, after all, just replacing the quadratic penalty Σx_i^2 with $\Sigma x_i \log x_i$) there are important differences. Because of properties of the entropy $H(\mathbf{x}) = -\Sigma x_i \log x_i$, the solution to equation (3) must always have *non-negative* entries. Second, because the objective in equation (3) is not quadratic, the solution is non-linear in the observations vector $\mathbf{y}$. Finally, no closed form expression is known for the solution of equation (3). Instead, equation (3) must be approached as a general convex optimization problem and solved by some variant of gradient descent. However, special optimizers for this problem have been developed (Skilling and Bryan, 1984) which can solve very high dimensional problems.

The implicit claim made by advocates and users of ME inversion is that these differences from least squares regularization matter: that the positivity and non-linearity of the ME process provide benefits in applications which are worth the computational expense of ME.

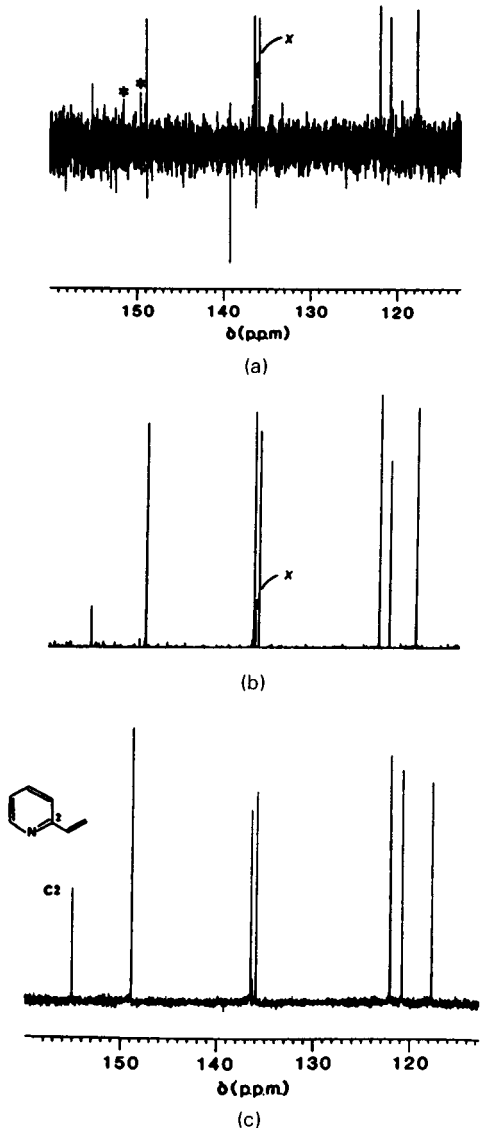


Fig. 1. (a) Conventional Fourier transform NMR reconstruction; (b) ME reconstruction for the same data as (a); (c) Fourier transform NMR reconstruction from a better experiment (from Sibisi *et al.* (1984), Figs 1 and 2)

Many applications of ME have been developed: to problems in NMR spectroscopy (Sibisi *et al.*, 1984), in astronomy (interferometry) (Gull and Daniell, 1978) and in infra-red absorption spectroscopy (Frieden, 1972). Many published reconstructions obtained via ME are excellent, and a few side-by-side comparisons show that ME regularization can, in certain cases, dramatically outperform quadratic regularization. We mention two prototypical examples.

(a) Sibisi *et al.* (1984) compare the ME reconstruction of a nuclear magnetic resonance (NMR) spectrum with reconstruction by conventional (least squares)

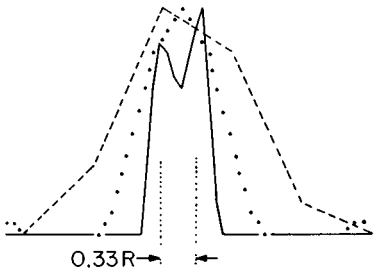


Fig. 2. (Noise-free) data (--------), best linear recovery (· · · · · ·) and ME recovery (———) (the true object consists of two spikes at $0.33R$ spacing) (taken from Frieden (1972), Fig. 4)

methods (Fig. 1). Not only does the ME reconstruction look nicer (fewer noisy oscillations), but ME does a better job, in an objective sense. The ME reconstruction resembles closely the reconstruction which conventional methods could obtain only on data from a much more sensitive experiment, i.e. an experiment with higher signal-to-noise ratio.

(b) Frieden (1972) shows that ME can sometimes *superresolve*. We shall explain the terminology in Section 4; but we illustrate the point with Frieden's diagram. Fig. 2 shows a true, 'spiky' object, a least squares reconstruction and an ME reconstruction. In this case, the true object consists of two closely spaced spikes, and the data are diffraction limited. The reconstruction by ME clearly shows two spikes; the reconstruction by least squares does not. The term 'super-resolution' is used here because ME in this case resolves better than the so-called *Rayleigh limit*, a resolution limit which all linear translation invariant methods must obey. In particular, the two spikes are spaced less than a third of the Rayleigh distance R apart, yet the ME reconstruction resolves them.

These examples illustrate the basic phenomena that sometimes occur with ME reconstructions:

(a) *signal-to-noise* enhancement and
(b) *superresolution*.

The purpose of our paper is to explain how and why these phenomena occur, and particularly *when* (i.e. under what conditions) they occur. We hope to make three main points.

(a) The phenomena are real, and due to the non-linearity of ME. However, they are delicate, and they occur if and only if the image to be recovered is *nearly black*—nearly zero in all but a small fraction of samples.
(b) ME is not the only non-linear inversion technique able to exploit near-blackness and to produce these phenomena. For example, another method, l_1-reconstruction, can do so optimally, from one point of view.
(c) The improvements obtained by such non-linear processing do not fully substitute for improving the sensitivity by doing a better experiment.

The paper is organized as follows. Section 2 discusses a simple estimation problem in which it can be shown how the non-linear behaviour of ME allows for an

improvement in signal-to-noise ratio. Section 3 shows how this surprisingly simple analysis extends to studying the behaviour of ME inversion in NMR spectroscopy. Finally, Section 4 sketches a theory explaining superresolution. For a brief concluding comment see Section 5. Appendix A contains proofs of the theorems.

2. IMPROVING SIGNAL-TO-NOISE RATIO

Consider the simple problem of estimating $\mathbf{x}=(x_i)_{i=1}^n$ from noisy data $\mathbf{y}$:

$$y_i = x_i + z_i, \qquad i=1, \ldots, n \tag{4}$$

where the noise terms z_i are independent and normally distributed with variance σ^2. This is a special case of equation (1), with K the identity operator.

In this model, the ME estimate of equation (3) is the solution to

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}=(x_i)} \sum_i (x_i - y_i)^2 + 2\lambda \sum_i x_i \log x_i \tag{5}$$

where only positive x need be considered in the minimum. Taking partial derivatives, we find that at the solution $(\hat{x}_i)$, say,

$$0 = 2\lambda(1 + \log \hat{x}_i) + 2(\hat{x}_i - y_i), \qquad i=1, \ldots, n$$

so that $\hat{x}_i$ is implicitly given as the solution to

$$y_i = \hat{x}_i + \lambda(1 + \log \hat{x}_i).$$

Let $\delta_{\mathrm{ME},\lambda}(y)$ be the solution to the equation

$$y = \delta + \lambda(1 + \log \delta).$$

Then the solution to equation (5) can be written explicitly in the form

$$\hat{x}_i = \delta_{\mathrm{ME},\lambda}(y_i), \qquad i=1, \ldots, n. \tag{6}$$

In words, the ME estimate is the result of applying the simple non-linearity $\delta_{\mathrm{ME},\lambda}$ co-ordinatewise.

Fig. 3 displays the function $\delta_{\mathrm{ME},\lambda}(\ )$ for three different parameter values $\lambda = \frac{1}{10}, \frac{1}{2}, 2$. The non-linearity is defined for both positive and negative arguments, is always

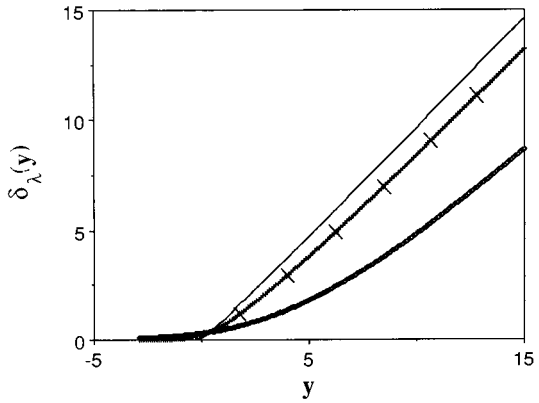


Fig. 3. Non-linearity $\delta_{\mathrm{ME},\lambda}$ for three different parameter values: ——, $\lambda=0.1$; —|—, $\lambda=0.5$; ▬▬, $\lambda=2$

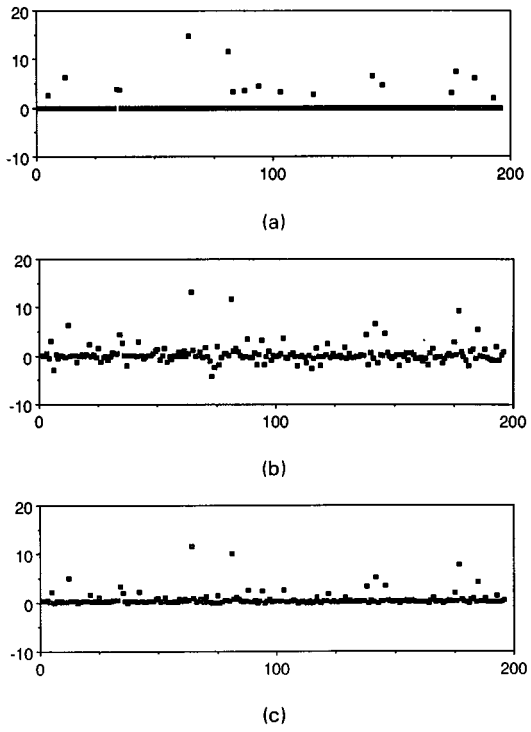


Fig. 4. (a) Object $\mathbf{x}$ to be recovered; (b) noisy data $\mathbf{y}$ from model (4) ($\sigma = 1$, $n = 196$); (c) ME estimate $\hat{\mathbf{x}}$ using $\lambda = \frac{1}{2}$

positive, tends to 0 for extreme negative arguments and tends to ∞ for extreme positive arguments. The ME non-linearity has a fixed point, $\delta_{\mathrm{ME},\lambda}(e^{-1}) = e^{-1}$, towards which the data are always 'shrunk':

$$|\delta_{\mathrm{ME},\lambda}(y) - e^{-1}| < |y - e^{-1}|.$$

The amount of shrinkage, as measured by the left-hand side of this inequality, increases as λ increases.

The effects that this non-linearity can produce are shown in Fig. 4. Fig. 4(a) shows a 'signal' $\mathbf{x}$ consisting mostly of 0s and a few large spikes. Fig. 4(b) shows data $\mathbf{y}$ observed when the normal errors $\mathbf{z}$ have standard deviation 1. Fig. 4(c) shows the ME estimate $\hat{\mathbf{x}}$ obtained with $\lambda = \frac{1}{2}$. The ME estimate has many visual similarities to the 'truth' in Fig. 4(a). There are only a few peaks standing out from a nearly constant background. To some readers, the transition from Fig. 4(b) to Fig. 4(c) will seem a dramatic visual improvement.

There is certainly a quantitative improvement. Define the mean-squared error $\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) = n^{-1} \Sigma_i (\hat{x}_i - x_i)^2$. Then the raw data of Fig. 4(b) have $\mathrm{MSE}(\mathbf{y}, \mathbf{x}) \approx 1$, whereas for the estimate of Fig. 4(c) we have $\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x}) \approx 0.45$, an improvement by a factor of 2.

2.1. *Improvement and Nearly Black Images*

This improvement is due to the special nature of $\mathbf{x}$ used in Fig. 4. Let Y be distributed

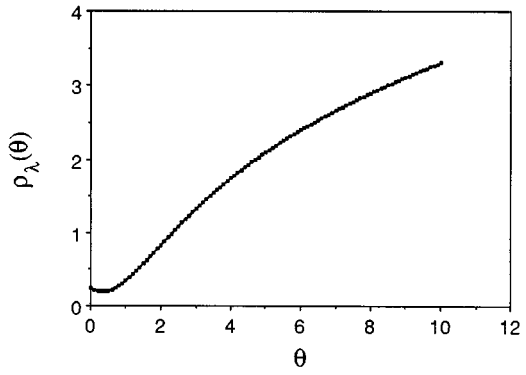


Fig. 5. Risk $\rho(\theta; \lambda, \sigma)$ with $\lambda = \frac{1}{2}$, $\sigma = 1$

$N(\theta, \sigma^2)$, and introduce the *risk* $\rho(\theta; \lambda, \sigma) = E\{\delta_{\mathrm{ME},\lambda}(Y) - \theta\}^2$, the expectation referring to the distribution of Y. Then we have

$$E\{\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x})\} = \frac{1}{n} \sum_{i=1}^{n} \rho(x_i)$$

with expectation referring to model (4).

Fig. 5 plots $\rho(\theta)$; the parameters λ and σ^2 were chosen exactly as in Fig. 4. Evidently, the risk ρ is small if and only if θ is near 0. Hence the expected MSE of $\hat{\mathbf{x}}$ is small compared with σ^2 only if most co-ordinates of $\mathbf{x}$ are nearly 0. Indeed, we can read off the graph that

$$E\{\mathrm{MSE}(\hat{\mathbf{x}}, \mathbf{x})\} \leqslant \epsilon\sigma^2$$

implies (by Markov's inequality)

$$\frac{1}{n} \#\{i : x_i \geqslant 2\sigma\} \leqslant \frac{5}{4}\epsilon$$

etc. (where # denotes 'cardinality of'). The trivial estimate $\mathbf{y}$ has expected mean-squared error σ^2. This shows that if ME improves significantly on the trivial estimate then the true image must be significantly non-zero in only a small fraction of samples. Hence Fig. 4 is in some sense the generic example of ME's ability to improve the mean-squared error in model (4).

2.2. *Optimal Performance with Nearly Black Images*

ME is not the only estimate that can be used in model (4). Consider the optimization problem

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} = (x_i)} \sum_i (x_i - y_i)^2 + 2\lambda \sum_i x_i \tag{7}$$

where now the minimum is over non-negative $\mathbf{x}$. We call this the minimum l_1-rule because it uses a penalty which is the same as the l_1-norm for non-negative $\mathbf{x}$. Repeating the analysis above, we have that the minimum l_1-estimate is obtained by applying a co-ordinatewise non-linearity:

$$\hat{x}_i = \delta_{l_1,\lambda}(y_i), \qquad i = 1, \ldots, n,$$

where $\delta_{l_1,\lambda}(y) \equiv \max(0, y-\lambda)$. Here we obtain $\hat{x}_i$ by pulling down every measured observation y_i by an amount λ, taking care to ensure a non-negative result.

If we were to display the analogue of Fig. 4(c) for this estimator we would obtain a plot visually resembling the 'true' answer for that situation, Fig. 4(a). The reader is invited to imagine this for himself.

It turns out that for nearly black images of the type in Fig. 4(a) the l_1-method does an excellent job quantitatively, and not just visually. Thus for the data of Fig. 4(b) the choice $\lambda = \frac{1}{2}$ gives $\mathrm{MSE}(\delta_{l_1,\lambda}, \mathbf{x}) \approx 0.3$. This is 50% better than ME and more than three times as good as the trivial estimate y.

In fact, the l_1-procedure has a certain optimality in dealing with nearly black images. We formalize this property by using minimax decision theory.

Definition 1. The class of ϵ-black images $X_n(\epsilon)$ is the set of sequences of length n satisfying

(a) $x_i \geqslant 0$ for all i and
(b) $\#\{i : x_i > 0\} \leqslant n\epsilon$.

Suppose that we have a rule $\hat{\mathbf{x}} = \delta_n(\mathbf{y})$ for estimating $\mathbf{x}$ in a problem of size n. If this rule makes excellent use of the nearly black property, then it should have a small expected mean-squared error for any $\mathbf{x} \in X_n(\epsilon)$. Thus, the following worst case mean-squared error should be small:

$$M_n(\delta_n, \epsilon) = \sup_{\mathbf{x} \in X_n(\epsilon)} (E[\mathrm{MSE}\{\delta_n(\mathbf{y}), \mathbf{x}\}]).$$

The smallest that this can possibly be for any rule is

$$M_n(\epsilon) = \inf_{\delta_n} \{M_n(\delta_n, \epsilon)\}.$$

A rule attaining this minimum is called *minimax*.

The class $X_n(\epsilon)$ contains all images which are nearly black: images in which the non-zero pixels can have any conceivable arrangement in space and in amplitude; $M_n(\epsilon)$ therefore measures how accurately it is possible to reconstruct $\mathbf{x}$ from $\mathbf{y}$ just using the information that $\mathbf{x} \geqslant 0$ and that $x_i > 0$ in a small fraction of samples. The following result describes the behaviour of

$$M(\epsilon) \equiv \sup_n \{M_n(\epsilon)\}$$

and shows that the l_1-rule is nearly minimax for small ϵ.

Theorem 1. Let $\mathcal{F}_\epsilon$ be the class of distributions of non-negative random variables which place at least $1-\epsilon$ of their mass at 0. Then

$$M(\epsilon) = \sigma^2 \sup\{1 - I(\Phi * F) : F \in \mathcal{F}_\epsilon\} \tag{8}$$

where $I(G) = \int (g')^2/g$ is the Fisher information and Φ the standard Gaussian distribution. We have the asymptotic result

$$M(\epsilon) = 2\sigma^2 \epsilon \log \epsilon^{-1}\{1 + o(1)\} \qquad \text{as } \epsilon \to 0. \tag{9}$$

Moreover, let $M(l_1, \epsilon) = \inf_\lambda [\sup_n \{M_n(\delta_{l_1,\lambda}, \epsilon)\}]$ denote the minimax performance among l_1-rules over all ϵ-black images. Then the optimal λ satisfies $\lambda^2(\epsilon) \sim 2\sigma^2 \log \epsilon^{-1}$ and

TABLE 1
Maximum risk over ϵ-black objects†

ϵ	$M(\epsilon)$	$M(l_1, \epsilon)$	$M(\delta_{exp}, \epsilon)$
0.01	0.046	0.052	0.20
0.02	0.078	0.087	0.27
0.05	0.153	0.16	0.42
0.10	0.248	0.26	0.56
0.20	0.390	0.41	0.72

†Assumes that $\sigma^2 = 1$. The raw data have worst case MSE = 1. $M(\delta_{exp}, \epsilon)$ denotes the performance of Bayes rule for an exponential prior with mean 3 (for comparison).

$$M(l_1, \epsilon)/M(\epsilon) \to 1 \qquad \text{as } \epsilon \to 0. \tag{10}$$

The proof is in Appendix A. Our analysis has several points of contact with work of Bickel (1983) and Pinsker (1980).

We interpret result (9) as follows. If we knew *a priori* which x_i were non-zero in $\mathbf{x}$, we could always estimate the other x_i as 0 and estimate the non-zero x_i by y_i. This would give $E(\text{MSE}) = \sigma^2\epsilon$. Relation (9) says that even without knowing *a priori* which x_i are non-zero we can obtain a mean-squared error which is worse only by logarithmic terms.

Table 1 illustrates the fact that relation (10) is a good approximation for ϵ as large as 5% or even 10%. Included for illustration is the behaviour of a Bayes rule which assumes that the x_i are random variables, independent and exponentially distributed; this rule does far worse than the l_1-rule.

Incidentally, ME is not competitive with l_1 in this worst case analysis. As Fig. 4 shows, the risk of ME tends to $+\infty$ as any component $x_i \to \infty$. Hence $M(\delta_{\text{ME},\lambda}, \epsilon) = +\infty$. In fact, ME is asymptotically not competitive even in the best case. Because of the fixed point property of ME, $\delta_{\text{ME},\lambda}(y) > e^{-1}$ if $y > e^{-1}$. Therefore, we immediately have

$$\inf_{\lambda}\{\inf_{X_n(\epsilon)}(E[\text{MSE}\{\delta_{\text{ME},\lambda}(y), \mathbf{x}\}])\} > e^{-2}(1-\epsilon)\,\Phi\{-(\sigma e)^{-1}\},$$

which does not go to 0 with ϵ.

In another direction, the minimax risk among linear procedures in this problem is precisely σ^2, for each $\epsilon > 0$: linear procedures are unable to take advantage of sparsity. Finally, the family of 'threshold' (T-) estimators $\delta_{T,\lambda}(y) = y\mathbf{1}_{\{y \geq \lambda\}}$ has a minimax risk $\inf_\lambda\{M(\delta_{T,\lambda}, \epsilon)\}$ which goes to 0 with ϵ but performs quantitatively somewhat worse than does the l_1-method.

2.3. *Bias versus Variance*

The risk improvements attained by the non-linear methods above come at a price: the estimators are biased. This may be seen by comparison of Figs 4(a) and 4(c). All the 0 values in Fig. 4(a) are estimated in Fig. 4(c) by positive values offset from 0 by a (small) nearly constant displacement. The peak values in Fig. 4(a) are estimated in Fig. 4(c) by systematically smaller $\hat{\mathbf{x}}$. Such 'amplitude bias' is present also for the

l_1-estimate. This bias is necessary to obtain the risk savings. The unbiased estimate $\hat{\mathbf{x}} = \mathbf{y}$ has expected mean-squared error σ^2, which is much larger.

2.4. *Relation to Practical Maximum Entropy*

The signal-plus-noise model discussed above is highly idealized. In practice, we generally modify the idealized ME above to deal with three specific issues.

2.4.1. *Scaling*

The benefits of the ME shrinker are most evident where the fixed point e^{-1} is small compared with the peak amplitudes in $\mathbf{x}$. We may always rescale our data to arrange for this. Alternatively, we may modify the entropy term, to $-\Sigma_i x_i \log(\beta x_i)$. The resulting scaled ME non-linearity $\delta_{\mathrm{ME},S,\lambda,\beta}(\ )$ has its fixed point at $A = (\beta e)^{-1}$. A is often called the 'default value', particularly in discussions of the Cambridge ME algorithm. Analysis with the modified objective is conceptually the same as rescaling the data so that A on the old scale corresponds to e^{-1} on the new scale, and then using the ME approach as we introduced it earlier. See also Skilling (1988).

2.4.2. *Normalization*

In certain settings we know that the true object is normalized by $\Sigma_i x_i = 1$. This constraint may be imposed on the ME solution by explicitly adding it to the problem formulation. For a certain data-dependent constant β, the solution of this constrained problem with entropy term $-\Sigma_i x_i \log x_i$ is the unconstrained solution of a modified ME problem, with entropy term $-\Sigma_i x_i \log(\beta x_i)$. This solution amounts, once again, to applying the non-linearity $\delta_{\mathrm{ME},S,\lambda,\beta}(\ )$ co-ordinatewise.

2.4.3. *Choice of regularization parameter*

Statisticians would naturally see a variety of possibilities for choosing λ in the analysis of real data, ranging from cross-validation, to use of Stein's unbiased risk estimates. Researchers in ME have developed still other methods; see, for example, Gull (1989). We do not discuss such methods here, as we are interested in what happens once a good choice of λ has already been provided to us.

The idealized picture that we have developed here continues to apply, with minor modifications and approximations, when practical considerations of scaling and normalization are enforced. For more information about the application of ME to large-scale practical situations, see Gull and Daniell (1978), Skilling and Bryan (1984) and Narayan and Nityananda (1986).

3. APPLICATIONS TO SPECTROSCOPY

The results of the previous section have a broader significance than one might at first suppose. Suppose that instead of observations (4) we have observations according to the original model (1), with the linear operator K an *orthogonal matrix*. Such K arise in Hadamard transform spectroscopy (Harwit and Sloane, 1979) where they are Hadamard matrices.

With K orthogonal, K^{-1} exists, and we can define pseudodata $\tilde{\mathbf{y}} = K^{-1}\mathbf{y}$. As K preserves Euclidean distances, $\|\mathbf{y} - K\mathbf{x}\|_2^2 = \|\tilde{\mathbf{y}} - \mathbf{x}\|_2^2$. The general optimization problem (3) can therefore be rewritten, in this particular case, as

$$\min_{\mathbf{x}} \sum_i (\tilde{y}_i - x_i)^2 + 2\lambda \sum_i x_i \log x_i. \tag{11}$$

This is the same as optimization problem (5) that we encountered in the signal plus noise situation, only with pseudodata $\tilde{\mathbf{y}}$ replacing $\mathbf{y}$. It follows that the solution to the ME problem is given simply by

$$\hat{x}_i = \delta_{\mathrm{ME},\lambda}(\tilde{y}_i) \qquad i = 1, \ldots, n. \tag{12}$$

A related analysis applies in NMR spectroscopy. In that area, when relaxation times and observation times are long, so that 'peak deconvolution' is not required (Freeman, 1988), K may be modelled as the complex $n \times n$ discrete Fourier transform matrix. The data $\mathbf{y}$ and the object $\mathbf{x}$ to be recovered are then, in general, complex. It is possible to define an entropy for complex objects in several ways, and this leads to different properties of estimates; see Hoch *et al.* (1989). We mention here the simplest definition, which leads to

$$\min_{\mathbf{x}} \|\mathbf{y} - K\mathbf{x}\|_2^2 + 2\lambda \sum_i |x_i| \log |x_i|, \tag{13}$$

where, in this equation, $|z|$ denotes the modulus $(z\bar{z})^{1/2}$ of the complex number z. Now the discrete Fourier transform matrix is, up to a constant factor, unitary; defining pseudodata $\tilde{\mathbf{y}} = K^{-1}\mathbf{y}$, it turns out, by repeating earlier arguments, that $\hat{x}_i = \delta_{\mathrm{CME},\lambda}(\tilde{y}_i)$, for a certain 'complex ME' non-linearity closely related to the 'real ME' non-linearity. Thus, to solve for $\mathbf{x}$ in equation (13), we first take the inverse discrete Fourier transform of the observations $\mathbf{y}$, obtaining pseudodata $\tilde{\mathbf{y}}$, then we apply a complex data ME non-linearity co-ordinatewise. In contrast, conventional NMR spectroscopy consists in simply taking the inverse discrete Fourier transform of the data, and using the pseudodata $\tilde{\mathbf{y}}$ to estimate $\mathbf{x}$.

Hence in one area of NMR spectroscopy ('without deconvolution of line widths') the difference between conventional and ME restoration is simply in the application of a co-ordinatewise non-linearity. We have conducted experiments to show this. Fig. 6 presents three versions of the real part of an NMR spectrum of the compound tryptophan in D_2O at 400 MHz, taken on the JEOL GX-400 NMR spectrometer at the Rowland Institute, Cambridge, Massachusetts. Fig. 6(a) was prepared by using standard Fourier transform methodology. Fig. 6(b) was prepared with the Cambridge ME program (Sibisi *et al.*, 1984; Skilling and Bryan, 1984) by using the 'four-channel' method for treating complex spectra, with default value parameter $A = 0.01$ and noise level parameter $S^2 = 5.1 \times 10^4$ (these parameters are called def and C_0 in the software documentation). Fig. 6(c) was prepared by using the idea described in Hoch *et al.* (1989) and Donoho *et al.* (1990): computing the discrete Fourier transform, followed by a co-ordinatewise application of a complex ME non-linearity. In principle, Figs 6(b) and 6(c) must be identical; owing to numerical imprecision, though, they agree only to six digits of accuracy.

It is interesting to review Fig. 1, taken from Sibisi *et al.* (1984), with the developments of this section in mind. Visual comparison between Figs 1(a), 1(b) and 1(c) and the corresponding panels of Figs 4 or 6 leaves little doubt that the same effect which is

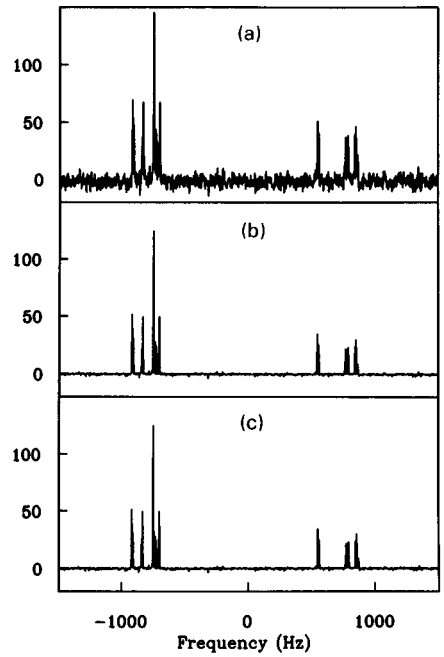


Fig. 6. (a) Fourier transform NMR recovery of the real part of the spectrum of tryptophan; (b) Cambridge ME recovery of the real part; (c) recovery of the real part resulting from applying a non-linearity $\delta_{CME,\lambda}(\mathbf{y})$ elementwise to (a)

achieved in Fig. 1 by ME reconstruction could also be obtained by simply applying the right non-linearity co-ordinatewise to Fig. 6(a). In other words, the qualitative effect of using ME can, in this case, be obtained by a simple non-linearity. The theory of Section 2 therefore gives an explanation of how ME has been able to improve the signal-to-noise ratio in Sibisi *et al.* (1984).

Let us recall our three points.

(a) Non-linearities can be used to improve the mean-squared error of estimation when the true object is near 0 in all but a small fraction of samples. However, as we saw in Section 2.1, if the object to be reconstructed is not nearly black, little improvement will be obtained. In some practical cases the true NMR spectrum has a 'background level' that is significantly higher than the noise level over a long interval; there we expect that ME does little to improve the signal-to-noise ratio. Fig. 7 gives a comparison of an ME reconstruction (Fig. 7(b)) with a conventional Fourier transform NMR reconstruction (Fig. 7(a)). The object to be recovered is well away from 0 in a significant portion of the display. The ME reconstruction is noticeably less noisy than Fourier transform reconstruction in the areas of the figure where the values are small, but the two figures differ negligibly in noisiness in the interval where the object is well away from 0.

(b) If near-blackness is present, it is clear from the discussion of Section 2.2 that other non-linearities can exploit it as well. We could, for example, define a 'complex l_1-method' as the solution to

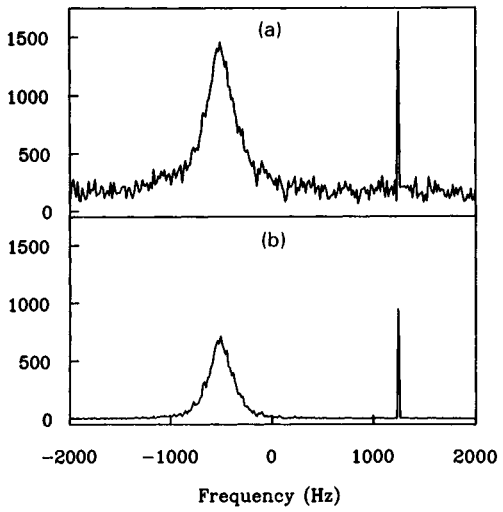


Fig. 7. (a) Real part of the Fourier transform reconstruction from synthetic data (the data consist of two decaying sinusoids with line widths 1.0 and 1000.0 Hz and amplitudes 100, 5000; series length $n=256$); (b) real part of the ME reconstruction for the same data, using the Cambridge method with $C_0=3.6\times10^7$

$$\min_{\mathbf{x}} \|\mathbf{y}-K\mathbf{x}\|_2^2+2\lambda\sum_i |x_i|.$$

with | | again the modulus; compare Newman (1988). Presumably, an analysis similar to Section 2.3 would show this to perform well in the nearly-black, complex-valued case.

(c) We may seriously question the extent to which a gain in mean-squared error leads to a gain in insight. When the comments of this section apply, the difference between ME and conventional reconstruction amounts to presenting the same data on two different plotting scales. Suppose that we wanted to identify peaks that were 'statistically significant', by the simple device of drawing a horizontal line across the plot at the 95th percentile of the null distribution of the plotted quantity (here null refers to the assumption that the true signal value at that sample is 0). The calculation of the height at which such a line should be drawn would differ, depending on whether we were plotting $\tilde{y}_i$ or the ME reconstruction $\delta_\lambda(\tilde{y}_i)$, but the same i-co-ordinates would be identified as significant.

We might therefore maintain that ME improves the signal-to-noise ratio (if the signal is nearly black) but does not improve *sensitivity* (i.e. the ability to discriminate small amplitude signal from noise successfully). Compare Freeman (1988) and Donoho *et al.* (1990).

The bias in amplitude estimates produced by ME should also be mentioned. While the ME reconstruction may be close to the truth in mean square, its peak amplitudes are biased. (The habit, in papers such as Sibisi *et al.* (1984), of suppressing axis labels on plots obscures this fact.) In contrast, by improving the experiment, we obtain a better signal-to-noise ratio without introducing bias.

4. SUPERRESOLUTION

A full discussion of superresolution would require considerable space, so we specialize. We suppose that we have data according to model (1), where the object is a vector of dimension n and the operator K consists of the first m rows of the $n \times n$ discrete Fourier transform matrix:

$$K_{ji} = \begin{cases} \frac{1}{\sqrt{n}} \cos\left\{\frac{\pi(i-1)(j-1)}{n}\right\} & j = 1, 3, \ldots, m, \\ \frac{1}{\sqrt{n}} \sin\left\{\frac{\pi(i-1)j}{n}\right\} & j = 2, 4, \ldots, m-1. \end{cases}$$

Then the observations vector $\mathbf{y}$ is of dimension m, and the elements of $\mathbf{y}$ represent noisy observations of the m low order Fourier coefficients of $\mathbf{x}$. (Our definition requires that m be *odd*.) This is a discrete model of diffraction-limited imaging; compare Bertero and Pike (1982), Frieden (1972), Pike *et al.* (1984) and Barakat and Newsam (1985a, b).

As in many other inverse problems, here the operator K is of less than full rank. It has m non-zero singular values and a null space of dimension $n-m$. Consequently, analysis by least squares or regularized least squares faces certain limitations. The formula $\hat{\mathbf{x}}_{\mathrm{RLS}} = (K^{\mathrm{T}}K + \lambda I)^{-1}K^{\mathrm{T}}\mathbf{y}$ gives an estimate $\hat{\mathbf{x}}$ which must lie in a subspace of dimension m, consisting of those vectors $\mathbf{x}$ whose last $n-m$ Fourier coefficients vanish. Vectors whose high order Fourier coefficients vanish are representable as sums of low frequency sinusoids and are therefore 'smooth'.

Although there certainly are applications where the object to be recovered is smooth, in areas like astronomy or spectroscopy the object to be recovered is nearly a set of scattered spikes. The smoothing effect of regularized least squares can be to lump two closely spaced spikes together into a single bump. Therefore, in such areas, regularized least squares can hide important structure. The terminology 'Rayleigh distance' has arisen to explain this; this is the minimal distance that two spikes must be spaced apart so that they can still be visually recognized as separate features in the conventionally reconstructed image. The Rayleigh distance in this discrete model may be taken as $R = n/m$. This is the reciprocal of the *incompleteness ratio* $\epsilon = m/n$.

Frieden (1972) demonstrated convincingly that ME can, sometimes, resolve structures closer together than the Rayleigh distance. This was illustrated in Fig. 2.

4.1. *Theory of Superresolution*

Frieden's is not the only example of superresolution. There is by now a considerable literature documenting many different non-linear algorithms which may be employed to obtain superresolution. Jansson (1984) contains several articles detailing different approaches.

However, superresolution is not well understood theoretically. To our knowledge, no theoretical treatment has emerged to answer questions such as, if the Rayleigh limit can be circumvented by non-linear procedures, what is the true limit of resolution? The absence of a theory of superresolution makes it easy for sceptics to invoke general principles which, in their view, cast doubt on the whole superresolution phenomenon.

We have developed a theory which shows when superresolution is possible, and what its limits are. From empirical work reported in Donoho *et al.* (1991), we believe

that the theory adequately models the application of superresolving methods to real data. In this paper, we present only a limited selection of results, just enough to make our main points. Other aspects of this theory will be reported elsewhere (Donoho, 1990; Donoho *et al.*, 1991). Here and later we adopt the usual conventions $\|\mathbf{v}\|_1 = \Sigma_i |v_i|$, $\|\mathbf{v}\|_2 = \sqrt{(\Sigma_i v_i^2)}$ and $\|\mathbf{v}\|_\infty = \max_i |v_i|$.

Definition 2. Let

$$\omega(\Delta;\ \mathbf{x}) = \sup\{\|\mathbf{x}' - \mathbf{x}\|_1 : \|K\mathbf{x}' - K\mathbf{x}\|_2 \leqslant \Delta \text{ and } \mathbf{x}' \geqslant 0\}. \tag{14}$$

We say that $\mathbf{x}$ admits of superresolution if

$$\omega(\Delta;\ \mathbf{x}) \to 0 \qquad \text{as } \Delta \to 0. \tag{15}$$

The definition makes sense. Suppose that $\mathbf{x}$ admits superresolution according to our definition. When we observe data $\mathbf{y} = K\mathbf{x} + \mathbf{z}$, and the noise $\mathbf{z}$ satisfies $\|\mathbf{z}\|_2 \leqslant \Delta$, then if $\hat{\mathbf{x}}$ is any purported reconstruction satisfying

$$\|\mathbf{y} - K\hat{\mathbf{x}}\|_2 \leqslant \gamma\Delta \tag{16}$$

and

$$\hat{x}_i \geqslant 0 \qquad i = 1, \ldots, n, \tag{17}$$

we must have, by the triangle inequality,

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_1 \leqslant \omega\{(1+\gamma)\Delta;\ \mathbf{x}\}. \tag{18}$$

If the noise level Δ is sufficiently small, this means that $\hat{\mathbf{x}}$ accurately reconstructs $\mathbf{x}$ from the partial information $\mathbf{y} = K\mathbf{x} + \mathbf{z}$.

Examples of methods satisfying inequalities (16) and (17) are ME in constrained form (2), with $S = \gamma\Delta$, or the minimum l_1-variant, defined by

$$\min\Big(\sum_i x_i\Big) \qquad \text{subject to } \|\mathbf{y} - K\mathbf{x}\|_2 \leqslant S \text{ and } \mathbf{x} \geqslant \mathbf{0}$$

also with $S = \gamma\Delta$. We could also mention the positive-constrained least squares estimate, defined by

$$\min \|\mathbf{y} - K\mathbf{x}\|_2 \qquad \text{subject to } \mathbf{x} \geqslant \mathbf{0},$$

which satisfies inequalities (17) and (16) with $\gamma = 1$.

In contrast, suppose that $\mathbf{x}$ does not admit superresolution, as we have defined it. Then, there exists $\mathbf{x}'$ which is non-negative and unequal to $\mathbf{x}$, yet $K\mathbf{x}' = K\mathbf{x}$. Even with noiseless data, we cannot say whether $\mathbf{x}$ or $\mathbf{x}'$ is the true object. Other pathologies occur in this case; we can show that neither the minimum l_1-estimate nor the positivity-constrained least squares estimate is uniquely defined for small noise levels, etc.

Our definition is particularly strong, intended to convince sceptics rather than to reassure advocates. In fact, the definition is so strong that it may be surprising that there is any $\mathbf{x}$ which admits of superresolution. Under this definition, $\mathbf{x}$ must be so special that any method (ME, or any of those described in Jansson (1984)) must give a good restoration of the full object $\mathbf{x}$ from incomplete, noisy data $\mathbf{y}$—if the noise level Δ is small, and if the restoration method obeys inequalities (16) and (17).

In fact, all sufficiently nearly black objects admit of superresolution, as we see in theorem 3 later.

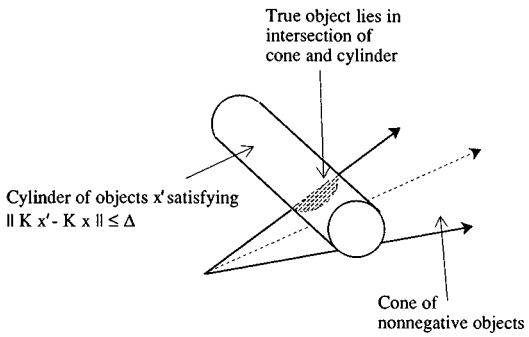


Fig. 8. Geometry of superresolution: the set of non-negative objects is a cone; the set of objects satisfying $\|K\mathbf{x}-K\hat{\mathbf{x}}\| \leqslant \Delta$ is a cylinder; when the intersection of the cone and cylinder becomes small as $\Delta \to 0$, we say that $\mathbf{x}$ admits superresolution

A certain geometry underlies the definition: see Fig. 8. When inequalities (16) and (17) hold, we know from the triangle inequality that $\|K(\mathbf{x}'-\mathbf{x})\|_2 \leqslant (1+\gamma)\Delta$. For the particular K that we are using, the set of all possible reconstructions $\mathbf{x}'$ obeying this inequality is a cylinder, the product of an m-dimensional sphere with an $(n-m)$-dimensional affine subspace. This set of reconstructions is unbounded. However, when we include the constraint that we are interested only in positive reconstructions, attention focuses on the small shaded area in the figure. When we are sufficiently lucky that the geometry of the situation is as in that figure, the set of all possible reconstructions is small, and also its diameter tends to 0 as $\Delta \to 0$.

The issue is to find for which $\mathbf{x}$ the geometry is of the favourable kind indicated by Fig. 8. The following (basically technical) result permits a reduction.

Theorem 2. $\mathbf{x}$ admits of superresolution if and only if there is a finite, positive constant C so that

$$\|\mathbf{x}'-\mathbf{x}\|_1 < C\|K\mathbf{x}'-K\mathbf{x}\|_2 \tag{19}$$

holds whenever

$$x_i' \geqslant 0 \qquad i=1, \ldots, n. \tag{20}$$

Let $C(K, \mathbf{x})$ denote the smallest constant for which inequalities (19) and (20) hold, and $C(K, \mathbf{x})=\infty$ if no such relations hold. If $C<\infty$, $\mathbf{x}$ admits superresolution, and

$$\omega(\Delta; \mathbf{x}) \sim C(K, \mathbf{x})\Delta \qquad \text{as } \Delta \to 0. \tag{21}$$

We therefore turn attention to the coefficient $C(K, \mathbf{x})$. It is clear from the proof of theorem 2 that $C(K, \mathbf{x})$ does not depend on $\mathbf{x}$ except through the number and arrangement of non-zero elements, i.e. the amplitudes of the non-zero elements of $\mathbf{x}$ do not matter. Our main result involves just the number of non-zero elements.

Theorem 3.

(a) If $\mathbf{x}$ has $\frac{1}{2}(m-1)$ or fewer non-zero elements then $C(K, \mathbf{x})<\infty$.
(b) If $\frac{1}{2}(m+1)$ divides n, there exists $\mathbf{x}$ with $\frac{1}{2}(m+1)$ non-zero elements yet $C(K, \mathbf{x})=\infty$.
(c) If $\mathbf{x}$ has more than m non-zero elements then $C(K, \mathbf{x})=\infty$.

The proof, in Appendix A, revolves around a lower bound on the number of negative values taken on by high frequency sequences. That bound, in lemma 2, may be viewed as an analogue, for high frequency sequences, of Logan's (1965) results on the number of 0s of high frequency functions in continuous time.

We restate the result in the language of our title. If the incompleteness ratio is $\epsilon = m/n$, $\mathbf{x}$ must admit superresolution if $\mathbf{x}$ is $\frac{1}{2}\epsilon$-black. Moreover, $\mathbf{x}$ might not admit superresolution if $\mathbf{x}$ is not $\frac{1}{2}\epsilon$-black and $\mathbf{x}$ cannot admit superresolution if $\mathbf{x}$ is not ϵ-black.

Thus, near-blackness is both necessary and sufficient for superresolution.

We now briefly turn to consider the size of C. When C happens to be very large, say 10^{12}, superresolution is largely a theoretical curiosity, since subquantum noise levels would be required to make $C(1+\gamma)\Delta$ sufficiently small to exert useful control on the reconstruction error $\|\hat{\mathbf{x}} - \mathbf{x}\|_1$.

It turns out that C is strongly correlated with the *spacing* of non-zero elements in $\mathbf{x}$. If all non-zero elements in $\mathbf{x}$ are well spaced, then C can be moderately small, but if it can happen that as many as r are bunched together within a Rayleigh interval then C can be very large, growing roughly exponentially in r.

Theorem 4. There exists $\mathbf{x}$ having only r non-zero elements, and a non-negative $\mathbf{x}'$ such that

$$\|\mathbf{x}' - \mathbf{x}\|_1 = \Gamma(r, m, n)\,\|K(\hat{\mathbf{x}} - \mathbf{x})\|_2 \tag{22}$$

where, if $m, n \to \infty$ with r fixed, and $m/n \to \epsilon$,

$$\Gamma(r, m, n)^{-2} \sim \frac{1}{2\pi}\int_0^{\pi\epsilon} |P_r(\theta)|^2\,d\theta, \tag{23}$$

where P_r is a certain trigonometric polynomial having $P_r(\theta) \sim (\theta/2)^{2r}$ as $\theta \to 0$. In particular, for small ϵ,

$$\frac{1}{2\pi}\int_0^{\pi\epsilon} |P_r(\theta)|^2 \approx (4r+2)^{-1}\left(\frac{\pi}{2}\right)^{2r}\epsilon^{2r+1}. \tag{24}$$

4.2. *Interpretation*

Our three claims in Section 1 apply to superresolution also.

(a) Superresolution is a real, delicate, non-linear effect. It is *real*, because we have proved that under certain conditions ME accurately reconstructs the unknown object, despite massive incompleteness. It is *delicate*, because it depends on the near-blackness of the object to be reconstructed (by theorem 3). It is *non-linear*, because it depends on the two properties (16)–(17); these cannot both be guaranteed by linear methods.

(b) ME is not the only method which exhibits superresolution. As mentioned earlier, many different methods have been shown to exhibit it in published examples. In our theory, any method with the two properties (16)–(17) exhibits superresolution.

(c) Superresolution produced by these non-linear methods is not, in general, a substitute for superresolution produced by developing better instrumentation (i.e. increasing m). If the object to be recovered consists of a few spikes spaced well apart, the coefficient C can be moderate in size, and ME and like methods

might conceivably give a highly accurate reconstruction. But if there are several spikes close together theorem 4 shows that the coefficient C can be very large, and the prospects for accurate reconstruction are doubtful. (See also the examples in Donoho *et al.* (1991).) In contrast, developing better instrumentation (where possible!) would increase the resolution of the experiment for both the easy (well-spaced) and the difficult cases.

5. COMMENT

This paper is written against a background of some controversy: Skilling (1984), Redfearn (1984) and Titterington (1984). Some ME proponents have made, in open forums, claims that ME is the one and only method to use for solving inverse problems where the answer is known to be positive. We quite naturally feel an affinity for Titterington's objections to such fundamentalism. In fact, we believe that many statisticians feel some distrust towards the fundamentalist school of ME. It sometimes seems that the feeling is reciprocated.

The ME literature encompasses many points of view, and not just fundamentalist ones. For example, Frieden, Komesaroff, Narayan and Nityananda have all evidenced concern for analysing why ME can sometimes improve significantly on conventional methods, and for describing conditions under which ME *fails* to improve on conventional methods; see Frieden (1972, 1985), Komesaroff *et al.* (1981) and Narayan and Nityananda (1982, 1986).

Our most balanced assessment of the situation is this. Proponents of ME have performed a service to inverse theorists by demonstrating the possibility of signal-to-noise enhancement and superresolution, and to practitioners by making available efficient ME software which makes it possible to exploit these phenomena. Perhaps both statisticians and ME proponents can find common ground in the recognition of these contributions and in building on these achievements.

ACKNOWLEDGEMENTS

The authors would like to thank M. Burns, B. R. Frieden, E. Gassiat, F. J. Gilbert, R. L. Parker, P. B. Stark and J. W. Tukey for interesting discussions and correspondence. B. F. Logan kindly provided a bound copy of his doctoral thesis. Lorenzo Sadun, Moxiu Mo and Cha-Yong Koo (in chronological order) provided computing assistance. We also thank the referees for helpful comments.

D. L. Donoho was supported by National Science Foundation grant DMS 84-51753 and by grants from Apple Computer, Schlumberger–Doll Research and Western Geophysical. I. M. Johnstone was supported by National Science Foundation grant DMS 84-51750, by the Science and Engineering Research Council and by the Sloan Foundation.

APPENDIX A: PROOFS

A.1. *Proof of Theorem 1*

The arguments presented here are related to those in Donoho and Johnstone (1989). The reader may also be helped for Section A.1.1 by consulting Pinsker (1980) and for Section A.1.2 by consulting Bickel (1983).

A.1.1. *Proof of formula for M(ϵ)*

Let Π_G denote the set of exchangeable probability measures on $\mathbf{R}^n$ which put mass 1 on $X_n(\epsilon)$. For a (prior) measure π on $\mathbf{R}^n$, let μ denote the joint distribution of $(\mathbf{x}, \mathbf{y})$ when $\mathbf{x} \sim \pi$ and $\mathbf{y} \sim N(\mathbf{x}, \sigma^2 I)$. We define the Bayes risk

$$\rho(\pi) = n^{-1} E_\pi [\, \| E_\mu\{x \mid y\} - x \|_2^2 \,].$$

As in Donoho and Johnstone (1989), the minimax theorem of decision theory implies that

$$M_n(\epsilon) = \sup\{\rho(\pi) : \pi \in \Pi_G\}. \tag{A.1}$$

Given $\pi \in \Pi_G$ let π_0 be the product measure with the same marginal. As in Donoho and Johnstone (1989) we have $\rho(\pi) \leqslant \rho(\pi_0)$. Now if we define the rescaled marginal $F_{1,\sigma}(t) = \pi\{\mathbf{x} : x_1/\sigma \leqslant t\}$, the Bayes risk $\rho(\pi_0) = \sigma^2\{1 - I(F_{1,\sigma} * \Phi)\}$, where Φ denotes the standard univariate Gaussian distribution and I denotes the Fisher information (this is an identity due to L. D. Brown; see for example Bickel (1983)). Finally observe that $\{F_{1,\sigma} : \pi \in \Pi_G\} \subset \mathscr{F}_\epsilon$. Combining these facts,

$$M_n(\epsilon) \leqslant \sigma^2 [\, 1 - \inf\{I(F * \Phi) : F \in \mathscr{F}_\epsilon\} \,]$$

for every n, so that

$$M(\epsilon) \leqslant \sigma^2 [\, 1 - \inf\{I(F * \Phi) : F \in \mathscr{F}_\epsilon\} \,]. \tag{A.2}$$

Let $\alpha < \epsilon$. There exists a sequence $(F_{k,\alpha},\ k = 1, 2, \ldots)$ of distributions in $\mathscr{F}_\alpha$ with

$$I(F_{k,\alpha} * \Phi) \to \inf\{I(F * \Phi) : F \in \mathscr{F}_\alpha\}$$

and, additionally, $\operatorname{supp}(F_{k,\alpha}) \subset [\, 0, k \,]$.

Put $A_n = \{\mathbf{x} \in X_n(\epsilon)\}$. Let π_0 be the product measure on $\mathbf{R}^n$ with marginal $F_{k,\alpha}$. Define the conditional measure $\pi_n(B) = \pi_0(B \mid A_n)$. Then $\pi_n \in \Pi_G$, and so by equation (A.1) $M_n(\epsilon) \geqslant \rho(\pi_n)$. Let $\operatorname{bin}(n, p)$ denote a random variable with binomial distribution having parameters n and p. Then as $n \to \infty$

$$\pi_0(A_n) = P\{\operatorname{bin}(n, \alpha) \leqslant n\epsilon\} \to 1. \tag{A.3}$$

Hence π_0 and π_n are very close, in variation distance, for large n. Hence π_n and π_0 give almost the same expectations to bounded measurable functions. The boundedness $\operatorname{supp}(\pi_0) \subset [\, 0, k \,]^n$ can therefore be used to show the equivalence of Bayes risks:

$$\rho(\pi_n)/\rho(\pi_0) \to 1 \tag{A.4}$$

as $n \to \infty$. As $\rho(\pi_0) = \sigma^2\{1 - I(F_{k,\alpha} * \Phi)\}$, we then have

$$\liminf_{n \to \infty} \{M_n(\epsilon)\} \geqslant \sigma^2\{1 - I(F_{k,\alpha} * \Phi)\}.$$

As $M(\epsilon) \geqslant M_n(\epsilon)$, it follows on letting $k \to \infty$ that

$$M(\epsilon) \geqslant \sigma^2 [\, 1 - \inf\{I(F * \Phi), F \in \mathscr{F}_\alpha\} \,].$$

Now $I(F * \Phi)$ is a continuous functional of F in supremum norm (compare, for example, Donoho (1988)); letting $\alpha \to \epsilon$ and invoking this continuity, we obtain the reverse inequality to inequality (A.2), and equation (8) follows.

A.1.2. *Asymptotic formula for M(ϵ)*

In this section, we establish the lower bound

$$M(\epsilon) \geqslant 2\sigma^2 \epsilon \log \epsilon^{-1}\{1 + o(1)\}. \tag{A.5}$$

The upper bound on behaviour of the l_1-rule of the next subsection shows that equality holds.

Here and in the next section we take $\sigma = 1$, without any loss of generality. Inequality (A.5) follows from equation (8), taking a large in expression (A.7) below.

Proposition 1. Let $F_{\epsilon,\mu} = (1-\epsilon)\nu_0 + \epsilon\nu_\mu$, where ν_x denotes Dirac mass at x. Let $a > 0$, and, for all sufficiently small ϵ, define μ implicitly as a function of ϵ by

$$\mu^2 + 2a\mu = 2\log\epsilon^{-1}. \tag{A.6}$$

Then

$$1 - I(\Phi * F_{\epsilon,\mu}) \sim \epsilon\mu^2\,\Phi(a) \qquad \text{as } \epsilon \to 0. \tag{A.7}$$

Proof. By L. D. Brown's identity, $1 - I(\Phi * F_{\epsilon,\mu})$ is the Bayes risk for estimation of the one-dimensional parameter θ from data which is $N(\theta, 1)$, when $\theta = 0$ with probability $1-\epsilon$, and $\theta = \mu$ with probability ϵ.

This risk may be written

$$\rho = (1-\epsilon)\int (E\{\theta \mid y\} - 0)^2\,\phi(y)\,dy + \epsilon\int (E\{\theta \mid y\} - \mu)^2\,\phi(y-\mu)\,dy.$$

Define $p(y) = P\{\theta = \mu \mid y\}$, and note that

$$E\{\theta \mid y\} = \mu\,p(y).$$

Lemma 1 below shows that, as $\epsilon \to 0$, p tends to 0 or 1 depending on whether y is smaller than or bigger than $\mu + a$. Applying this, and carefully bounding remainder terms,

$$\rho = (1-\epsilon)\mu^2\int_{\mu+a}^{\infty}\phi(y)\,dy + \epsilon\mu^2\int_{\infty}^{\mu+a}\phi(y-\mu)\,dy + o(\epsilon\mu^2). \tag{A.8}$$

Now using assumption (A.6), we have $\phi(\mu + a) = \phi(0)\,\epsilon\exp(-a^2/2)$; combined with the standard inequality

$$1 - \Phi(t) \leqslant \phi(t)/t \qquad \text{for all } t \geqslant 1 \tag{A.9}$$

we obtain

$$(1-\epsilon)\mu^2\int_{\mu+a}^{\infty}\phi(y)\,dy = O(\epsilon\mu) = o(\epsilon\mu^2),$$

and so the first term in equation (A.8) is negligible compared with the second term, leaving

$$\rho \sim \epsilon\mu^2\,\Phi(a)$$

as required.

Lemma 1. With the assumptions and notation of proposition 1,

$$p(\mu + z) \to \begin{cases} 1 & z > a, \\ 0 & z < a, \end{cases} \tag{A.10}$$

as $\epsilon \to 0$, uniformly in $z > a + \delta$ and in $z < a - \delta$, $\delta > 0$.

Proof.

$$p(y) = \frac{\epsilon\,\phi(y-\mu)}{(1-\epsilon)\,\phi(y) + \epsilon\,\phi(y-\mu)}$$

so that $p(\mu + z) = \{(1-\epsilon)\exp(-\mu z - \mu^2/2 + \log\epsilon^{-1}) + 1\}^{-1}$. Now by equation (A.6) $-\mu^2/2 + \log\epsilon^{-1} = a\mu$, so

$$p(\mu + z) = [\,(1-\epsilon)\exp\{\mu(a-z)\} + 1\,]^{-1}.$$

From this, expression (A.10) is immediate.

A.1.3. *Asymptotic minimaxity of l_1-rule*

We complete the proof of equations (9) and (10) by showing that if we put

$$\lambda^2 = 2 \log \epsilon^{-1} \tag{A.11}$$

then, for the l_1-rule δ_n based on this choice of λ, we have

$$\sup_n \{M_n(\delta_n, \epsilon)\} \leqslant \lambda^2 \epsilon \{1 + o(1)\} \qquad \text{as } \epsilon \to 0. \tag{A.12}$$

Since the μ of the last section and λ of this section are asymptotically equivalent, this shows that the inequality in expression (A.5) can be replaced by equality. It also shows that choice (A.11) is asymptotically optimal. Letting $\delta(y) = \max(0, y - \lambda)$, we have

$$E\{\mathrm{MSE}(\delta_n, \mathbf{x})\} = \int\int \{\delta(\theta + z) - \theta\}^2 \phi(z)\, dz\, dF_n(\theta)$$

where F_n is the empirical distribution of the x_i. As this functional is linear in F_n, and as $F_n(0) \geqslant 1 - \epsilon$ whenever $\mathbf{x} \in X_n(\epsilon)$, we have

$$E\{\mathrm{MSE}(\delta_n, \mathbf{x})\} \leqslant (1 - \epsilon)\, r(0) + \epsilon \sup_\mu \{r(\lambda, \mu)\}$$

where

$$r(\lambda, \mu) \equiv E\{\delta(\mu + z) - \mu\}^2.$$

A calculation gives

$$r(\lambda, \mu) = \mu^2 \Phi(\lambda - \mu) - (\lambda + \mu)\phi(\lambda - \mu) + (\lambda^2 + 1)\{1 - \Phi(\lambda - \mu)\}.$$

This implies that $\partial r(\lambda, \mu)/\partial \mu = 2\mu \Phi(\lambda - \mu)$, and so

$$\sup_\mu \{r(\lambda, \mu)\} = r(\lambda, +\infty) = \lambda^2 + 1.$$

Moreover,

$$r(\lambda, 0) = -\lambda \phi(\lambda) + (\lambda^2 + 1)\{1 - \Phi(\lambda)\}.$$

By equation (A.11) we have $\phi(\lambda) = \epsilon \phi(0)$; using again inequality (A.9), we obtain $1 - \Phi(\lambda) \leqslant \epsilon \phi(0)/\lambda$ and so

$$r(\lambda, 0) \leqslant \epsilon \phi(0)/\lambda.$$

We conclude that for $\epsilon < e^{-1}$

$$(1 - \epsilon)\, r(\lambda, 0) + \epsilon\, r(\lambda, \mu) \leqslant \epsilon(\lambda^2 + 2)$$

and inequality (A.12) follows.

A.2. *Proof of Theorem 2*

Let $\mathcal{Z} = \{i : x_i = 0\}$. Let $\mathcal{V} = \{\mathbf{v} : \|\mathbf{v}\|_1 = 1 \text{ and } v_i \geqslant 0,\ i \in \mathcal{Z}\}$. Then $\mathcal{V}$ is closed and compact. Let v^* be any minimizer of $\|K\mathbf{v}\|_2$ on $\mathcal{V}$.

If $\|K\mathbf{v}^*\|_2 > 0$, we claim that $C = 1/\|K\mathbf{v}^*\|_2$ and that $\omega(\Delta; \mathbf{x}) = C\Delta$ for sufficiently small Δ. If $\|K\mathbf{v}^*\|_2 = 0$, we claim that $C = \infty$ and that $\omega \not\to 0$. These two claims together prove the theorem.

Assume that $\|K\mathbf{v}^*\|_2 > 0$. Put $\mathbf{v} = \mathbf{x}' - \mathbf{x}$, $\mathbf{x}' \geqslant \mathbf{0}$. Then $\mathbf{v}/\|\mathbf{v}\|_1 \in \mathcal{V}$, so

$$\|K\mathbf{v}\|_2 \geqslant \|K\mathbf{v}^*\|_2 \|\mathbf{v}\|_1;$$

hence $C \leqslant 1/\|K\mathbf{v}^*\|_2$.

Now, if we pick α very small, then $\mathbf{x}' = \mathbf{x} + \alpha\mathbf{v}^*$ defines a non-negative vector, with

$$\|K(\mathbf{x}' - \mathbf{x})\|_2 = \alpha \|K\mathbf{v}^*\|_2 = \|K\mathbf{v}^*\|_2 \|\mathbf{x}' - \mathbf{x}\|_1$$

so

$$\omega(\Delta;\ \mathbf{x}) \geqslant \alpha \| \mathbf{v}^* \|_1 \qquad \text{for } \Delta > \alpha \| K\mathbf{v}^* \|_2 \tag{A.13}$$

for all $\alpha \leqslant \alpha_0$, say. As $\| K\mathbf{v}^* \|_2 > 0$, this says

$$\omega(\Delta;\ \mathbf{x}) \geqslant \Delta / \| K\mathbf{v}^* \|_2$$

for $\Delta < \Delta_0$. By definition of ω and C,

$$\omega(\Delta;\ \mathbf{x}) \leqslant C(K,\ \mathbf{x})\Delta, \qquad \Delta > 0. \tag{A.14}$$

It follows that $C \geqslant 1/\| K\mathbf{v}^* \|_2$, and hence $C = 1/\| K\mathbf{v}^* \|_2$. We may also conclude that

$$\omega(\Delta;\ \mathbf{x}) = C(K,\ \mathbf{x})\Delta, \qquad \Delta < \Delta_0,$$

completing our first claim.

For the second claim, suppose that $\| K\mathbf{v}^* \|_2 = 0$. Let α_0 be any positive value for α for which inequality (A.13) holds. Then

$$\liminf_{\Delta \to 0} \{\omega(\Delta;\ \mathbf{x})\} \geqslant \alpha_0,$$

i.e. $\omega \not\to 0$ as $\Delta \to 0$. From this and inequality (A.14), $C = \infty$, completing the proof of the second claim.

A.3. *Proof of Theorem 3*

Assertion (c) of the theorem is an exercise in parameter counting and linear algebra. We omit the argument.

Let $0 \leqslant k < n/2$. $\mathscr{B}(k)$ is the set of *discrete band-limited sequences* of length n and bandwidth k, i.e. the real sequences (b_i) satisfying

$$\sum_{i=1}^{n} b_i \exp\left\{ \sqrt{(-1)} \frac{2\pi(i-1)j}{n} \right\} = 0 \qquad j = k+1,\ \ldots,\ n-1-k.$$

Let $1 < l < n/2$. $\mathscr{H}(l)$ is the set of *discrete high pass sequences* of length n, i.e. the real sequences (h_i) satisfying

$$\sum_{i=1}^{n} h_i \exp\left\{ \sqrt{(-1)} \frac{2\pi(i-1)j}{n} \right\} = 0 \qquad j = 0,\ \ldots,\ l-1,\ n-l+1,\ \ldots,\ n-1.$$

It follows from Parseval's relation for the finite discrete Fourier transform that if $\mathbf{b} \in \mathscr{B}(k)$ and $\mathbf{h} \in \mathscr{H}(l)$, with $k < l$, then

$$\sum_i b_i h_i = 0. \tag{A.15}$$

The importance of these two vector spaces is that

$$\mathscr{B}\left(\frac{m-1}{2}\right) = \text{range}(K^{\mathrm{T}}K), \qquad \mathscr{H}\left(\frac{m+1}{2}\right) = \text{kernel}(K^{\mathrm{T}}K).$$

Let $\mathscr{J} = \{i : x_i > 0\}$. From the proof of theorem 2 we know that $C(K,\ \mathbf{x}) = \infty$ if and only if there exists $\mathbf{v} \neq \mathbf{0}$ such that $K\mathbf{v} = \mathbf{0}$ yet $v_i \geqslant 0$, $i \in \mathscr{J}$. From now on we fix $l = (m+1)/2$. We can rephrase the condition of theorem 2 as

$$C(K,\ \mathbf{x}) = \infty \qquad \text{if and only if for some non-zero } h \in H(l),\ h_i \geqslant 0,\ i \in \mathscr{J}. \tag{A.16}$$

We use this to prove the two halves of the theorem.

A.3.1. *Fewer than m/2 non-zero elements*

Lemma 2 shows that a non-zero $\mathbf{h} \in \mathscr{H}(l)$ has at least l negative elements. For such an $\mathbf{h}$ to satisfy $h_i \geqslant 0$, $i \in \mathscr{J}$, we must have $\text{card}(\mathscr{J}) \leqslant n-l$, i.e. $\mathbf{x}$ has at least l non-zero elements. As $l > m/2$ this cannot happen. The proof of the first half is complete.

Lemma 2. Let $\mathbf{h} \in \mathscr{H}(l)$. If $\mathbf{h} \neq \mathbf{0}$, then $\mathbf{h}$ has at least l negative elements.

Proof. Put $\mathscr{P} = \{i : h_i \geqslant 0\}$. Put $k = \text{card}(\mathscr{P}^c)$. We claim that there is a sequence $\mathbf{b} = (b_i)$ so that

$$\min_i |b_i| > 0, \tag{A.17}$$

$$b_i h_i \geqslant 0, \qquad i = 1, \ldots, n, \tag{A.18}$$

and

$$\mathbf{b} \in \mathscr{B}(k). \tag{A.19}$$

It follows from equations (A.17) and (A.18) that

$$\sum_i b_i h_i \geqslant \min_i |b_i| \sum_i |h_i|.$$

Now if $k < l$, $\mathscr{H}(k+1) \supseteq \mathscr{H}(l)$. Therefore $\mathscr{B}(k)$ and $\mathscr{H}(l)$ are orthogonal. It follows that $\Sigma_i b_i h_i = 0$, which forces $\mathbf{h} = \mathbf{0}$. Consequently, if $k < l$ then $\mathbf{h} = \mathbf{0}$.

The construction of such a vector $\mathbf{b}$ is made by adapting a construction of Logan, who used it to show that continuous time high pass functions must change sign frequently (Logan (1965), theorem 5.3.1).

Let $(i_u)_{u=1}^k$ be an enumeration of the elements of $\mathscr{P}^c$. Define

$$s_u = \frac{2\pi(i_u - \frac{1}{2})}{n}, \tag{A.20}$$

$$t_u = \frac{2\pi(i_u + \frac{1}{2})}{n}, \tag{A.21}$$

for $u = 1, \ldots, k$. Define the sequence

$$b_i^{(u)} = \sin\left\{\frac{1}{2}\left(\frac{2\pi i}{n} - s_u\right)\right\} \sin\left\{\frac{1}{2}\left(\frac{2\pi i}{n} - t_u\right)\right\}. \tag{A.22}$$

The reader will want to check at this point that

$$\begin{aligned} b_i^{(u)} &> 0 \qquad i \in \{1, \ldots, n\} - \{i_u\}, \\ b_i^{(u)} &< 0 \qquad i = i_u. \end{aligned} \tag{A.23}$$

Now putting equation (A.22) in the equivalent form

$$b_i^{(u)} = \left[\cos\left(\frac{\pi}{n}\right) - \cos\left\{\frac{2\pi}{n}(i - i_u)\right\}\right] \Big/ 2$$

shows that, for certain constants $e_0^{(u)}$, $e_1^{(u)}$ and $f_1^{(u)}$, we have

$$b_i^{(u)} = e_0^{(u)} + e_1^{(u)} \cos\left\{\frac{2\pi(i-1)}{n}\right\} + f_1^{(u)} \sin\left\{\frac{2\pi(i-1)}{n}\right\}, \tag{A.24}$$

i.e. $\mathbf{b}^{(u)} \in \mathscr{B}(1)$. Define now

$$b_i = (-1) \prod_{u=1}^{k} b_i^{(u)}; \tag{A.25}$$

this gives a sequence (b_i) such that

$$b_i > 0 \qquad i \in \mathscr{P},$$
$$b_i < 0 \qquad i \in \mathscr{P}^c.$$

Properties (A.17)–(A.18) follow. By the convolution theorem for the finite discrete Fourier transform, we may show that if $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(k)}$ are all in $\mathscr{B}(a)$ then $\Pi_{u=1}^{k} \mathbf{b}^{(u)}$ is in $\mathscr{B}(ka)$. As equation (A.24) shows that the $\mathbf{b}^{(u)}$ are all in $\mathscr{B}(1)$ we conclude that expression (A.19) holds.

We could also check this directly by combining equation (A.24) with equation (A.25), giving explicitly the representation

$$b_i = e_0 + \sum_{u=1}^{k} e_u \cos\left\{\frac{2\pi(i-1)u}{n}\right\} + f_u \sin\left\{\frac{2\pi(i-1)u}{n}\right\},$$

which implies expression (A.19).

A.3.2. *More than m/2 non-zero elements*

By hypothesis, l divides n. Pick $0 < t < n/l$, and define

$$h_i = \begin{cases} 1 & i = 0,\ n/l,\ 2n/l,\ \ldots, \\ -1 & i = t,\ t+n/l,\ t+2n/l,\ \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Now (h_i) is periodic with period n/l. Thus $\mathbf{h}$ is representable as a Fourier sum using only sinusoids also of period n/l. Hence

$$\sum_{i=1}^{n} h_i \exp\left\{\sqrt{(-1)}\frac{2\pi(i-1)j}{n}\right\} = 0, \qquad j \notin \{0,\ l,\ 2l,\ \ldots,\ n-l\}.$$

Now by construction $\Sigma_i h_i = 0$. Hence,

$$\sum_{i=1}^{n} h_i \exp\left\{\sqrt{(-1)}\frac{2\pi(i-1)j}{n}\right\} = 0, \qquad j \notin \{l,\ 2l,\ \ldots,\ n-l\},$$

which implies $\mathbf{h} \in \mathscr{H}(l)$.

If we define

$$x_i = \begin{cases} 1 & i = t,\ t+n/l,\ t+2n/l,\ \ldots, \\ 0 & \text{otherwise,} \end{cases}$$

then x_i has only l non-zero elements. Putting $\mathscr{J} = \{i : x_i = 0\}$, we see that $h_i \geqslant 0$, $i \in \mathscr{J}$. Hence, condition (A.16) is satisfied for this $\mathbf{h}$ and this $\mathbf{x}$, and $C(K, \mathbf{x}) = \infty$.

A.4. *Proof of Theorem 4*

We treat indices circularly, so that 0 is identified with n, -1 with $n-1$, etc. Put $u = r-1$.

Suppose that r is odd. Let $x_i = 1$ if $i = -u,\ -u+2,\ \ldots,\ -2,\ 0,\ 2,\ \ldots,\ u-2,\ u$, and $x_i = 0$ otherwise. If, instead, r is even, let $x_i = 1$ if $i = -u,\ -u+2,\ \ldots,\ -1,\ 1,\ \ldots,\ u-2,\ u$, and $x_i = 0$ otherwise.

Define the sequence $\mathbf{c} = (c_i)_{i=1}^{n}$ via

$$c_i = \begin{cases} (-1)^i \binom{2r}{r+i} 2^{-2r} & -r \leqslant i \leqslant r, \\ 0 \quad \text{otherwise.} \end{cases}$$

Pick $\alpha > 0$ with $\alpha < \min_{-r \leqslant i \leqslant r}(1/|c_i|)$. Define $\mathbf{x}'$ by

$$x_i' = x_i + (-1)^r \alpha c_i \qquad i = 1, \ldots, n;$$

then $\mathbf{x}' \geqslant \mathbf{0}$. Now $\|\mathbf{x}' - \mathbf{x}\|_1 = \alpha \Sigma_i |c_i| = \alpha$ and $\|K(\mathbf{x}' - \mathbf{x})\|_2 = \alpha\{\Sigma_{j=1}^m (\hat{c}_j)^2\}^{1/2}$, where

$$\hat{c}_j = \begin{cases} \dfrac{1}{\sqrt{n}} \sum_i c_i \cos\left\{\dfrac{\pi i(j-1)}{n}\right\} & j = 1, 3, 5, \ldots, m, \\ \dfrac{1}{\sqrt{n}} \sum_i c_i \sin\left\{\dfrac{\pi ij}{n}\right\} & j = 2, 4, 6, \ldots, m-1 \end{cases}$$

are the discrete Fourier coefficients of $\mathbf{c}$. $\mathbf{c}$ is an even sequence, so that the sine coefficients vanish: $\hat{c}_j = 0$, $j = 2, 4, 6, \ldots$.

Define now $P_r(\theta) = \Sigma_{i=-r}^{r} c_i \cos(\theta i)$, so that $\hat{c}_j = n^{-1/2} P_r\{(\pi j - 1)/n\}$ for $j = 1, 3, 5, \ldots$. With this notation, equation (22) holds, with

$$\Gamma(r, m, n)^{-2} = \sum \hat{c}_j^2 = n^{-1} \sum_{0 \leqslant j < m/2} \left| P_r\left(\frac{2\pi j}{n}\right) \right|^2. \tag{A.26}$$

As P_r is Riemann integrable, if we let $n, m \to \infty$ with $m/n \to \epsilon \in (0, 1)$,

$$n^{-1} \sum_{0 \leqslant j < m/2} \left| P_r\left(\frac{2\pi j}{n}\right) \right|^2 \to \frac{1}{2\pi} \int_0^{\pi\epsilon} \left| P_r(\theta) \right|^2 \mathrm{d}\theta.$$

To estimate the behaviour of P_r, let f denote a function defined on the real line. Define the $2r$th-order differencing operator

$$(\Delta_h^{2r} f)(x) = \sum_{i=-r}^{r} (-1)^i \binom{2r}{r+i} f(x + hi).$$

Then, if f is integrable, we have for the Fourier transforms

$$2^{2r} P_r(\theta) \hat{f}(\theta) = \widehat{(\Delta_1^{2r} f)}(\theta).$$

Now if f is C^∞ and of compact support,

$$h^{-2r} \Delta_h^{2r} f \to D^{2r} f \qquad \text{as } h \to 0$$

in L_1, where D^{2r} denotes the differentiation operator of order $2r$. Hence for each fixed θ we have that as $h \to 0$

$$2^{2r} h^{-2r} P_r(h\theta) \hat{f}(\theta) \to \theta^{2r} \hat{f}(\theta)$$

for all f in C^∞ of compact support. Consequently

$$P_r(\theta) = (\theta/2)^{2r}\{1 + o(1)\} \qquad \text{as } \theta \to 0$$

and, applying equation (A.26), approximation (24) follows.

REFERENCES

Barakat, R. L. and Newsam, G. N. (1985a) Algorithms for reconstruction of partially known, bandlimited Fourier transform pairs from noisy data: I, The prototypical linear problem. *J. Integr. Eqn.*, **9**, 49–76.

——— (1985b) Algorithms for reconstruction of partially known, bandlimited Fourier transform pairs from noisy data: II, The nonlinear problem of phase retrieval. *J. Integr. Eqn.*, **9**, 77–108.

Bertero, M., de Mol, C. and Pike, E. R. (1985) Linear inverse problems with discrete data: I, General formulation and singular system analysis. *Inv. Prob.*, **1**, 301–330.

Bertero, M. and Pike, E. R. (1982) Resolution in diffraction-limited imaging, a singular-value analysis: I, The case of coherent illumination. *Opt. Acta*, **29**, 727–746.

Bickel, P. J. (1983) Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics* (eds M. H. Rizvi, J. S. Rustagi and D. Siegmund), pp. 511–528. New York: Academic Press.

Donoho, D. L. (1988) One-sided inference about functionals of a density. *Ann. Statist.*, **16**, 1390–1420.

——— (1990) Super-resolution via sparsity constraints. *Technical Report*. Department of Statistics, University of California, Berkeley.

Donoho, D. L. and Johnstone, I. M. (1989) Minimax risk over l_p balls. *Technical Report 201*. Department of Statistics, University of California, Berkeley.

Donoho, D. L., Gassiat, E. and Stark, P. B. (1991) Superresolution and positivity constraints: theory and experiment. *Technical Report*. Department of Statistics, University of California, Berkeley.

Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1990) Does maximum entropy improve sensitivity? *Proc. Natn. Acad. Sci. USA*, **87**, 5066–5068.

Freeman, R. (1988) *A Handbook of Nuclear Magnetic Resonance*. London: Longman.

Frieden, B. R. (1972) Restoring with Maximum Entropy: II, Superresolution of photographs with diffraction-blurred impulses. *J. Opt. Soc. Am.*, **62**, 1202–1210.

——— (1985) Dice, entropy, and likelihood. *Proc. IEEE*, **73**, 1764–1770.

Gull, S. F. (1989) Developments in Maximum Entropy data analysis. In *Maximum Entropy and Bayesian Methods* (ed. J. Skilling) Boston: Kluwer.

Gull, S. F. and Daniell, G. J. (1978) Image reconstruction from incomplete and noisy data. *Nature*, **272**, 686–690.

Harwit, M. and Sloane, N. J. A. (1979) *Hadamard Transform Optics*. New York: Academic Press.

Hoch, J. C., Stern, A. S., Donoho, D. L. and Johnstone, I. M. (1989) Reconstruction of complex, phase-sensitive spectra. *J. Magn. Resnce*, **86**, 236–246.

Jansson, P. A. (ed.) (1984) *Deconvolution, with Applications in Spectroscopy*. New York: Academic Press.

Komesaroff, M. M., Narayan, R. and Nityananda, R. (1981) The Maximum Entropy method of image restoration—properties and limitations. *Astron. Astrophys.*, **83**, 419–450.

Logan, B. F. (1965) Properties of high-pass signals. *PhD Thesis*. Columbia University, New York.

Narayan, R. and Nityananda, R. (1982) Maximum entropy image reconstruction—a practical, non-information theoretical approach. *J. Astrophys.*, **3**, 419–450.

——— (1986) Maximum entropy image restoration in astronomy. *Ann. Rev. Astron. Astrophys.*, **24**, 127–170.

Newman, R. H. (1988) Maximization of entropy and minimization of area as criteria for NMR signal processing. *J. Magn. Resnce*, **79**, 448–460.

Pike, E. R., McWhirter, J. G., Bertero, M. and de Mol, C. (1984) Generalized information theory for inverse problems in signal processing. *IEEE Proc.*, **131**, 660–667.

Pinsker, M. S. (1980) Optimal filtration of square-integrable signals in Gaussian white noise. *Prob. Inform. Transmissn*, **16**, 120–133.

Redfearn, J. (1984) *The Times*, Oct. 17th, 16C.

Sanz, J. L. C. (1985) Mathematical considerations for the problem of Fourier transform phase retrieval from magnitude. *SIAM J. Appl. Math.*, **45**, 651–664.

Sibisi, S., Skilling, J., Brereton, R. G., Laue, E. D. and Staunton, J. (1984) Maximum entropy signal processing in practical NMR spectroscopy. *Nature*, **311**, 446–447.

Skilling, J. (1984) The maximum entropy method. *Nature*, **309**, 748–749.

——— (1988) The axioms of maximum entropy. In *Maximum Entropy and Bayesian Methods in Science and Engineering* (eds G. J. Erickson and C. R. Smith), pp. 173–188. Boston: Kluwer.

Skilling, J. and Bryan, R. K. (1984) Maximum Entropy image reconstruction: general algorithm. *Mthly Not. R. Astron. Soc.*, **211**, 111–124.

Titterington, D. M. (1984) The maximum entropy method for data analysis. *Nature*, **312**, 381–382.
Wernecke, S. J. and D'Addario, L. R. (1977) Maximum entropy image reconstruction. *IEEE Trans. Comput.*, **26**, 351–364.

DISCUSSION OF THE PAPER BY DONOHO, JOHNSTONE, HOCH AND STERN

B. D. Ripley (University of Oxford): My interest in maximum entropy (ME) methods arose from work on image enhancement in astronomy. It may surprise many statisticians that ME has the entrenched position as the received wisdom in several areas of physical science. Rafael Molina and I encountered this view quite forcefully recently when attempting to publish an account of our own methods (Molina and Ripley, 1989) in a major astronomical journal. We had made some comparisons with ME but the referee wanted to see us demonstrate whether ME could match *every* insight we extracted from our examples.

How has ME reached this position? First it seems to work well and looks impressive; its proponents have been tenacious in developing adequate algorithms, for which they deserve considerable credit. Second, it is flexible and appealing, as the following quote from a 'guru' shows (Jaynes, 1978):

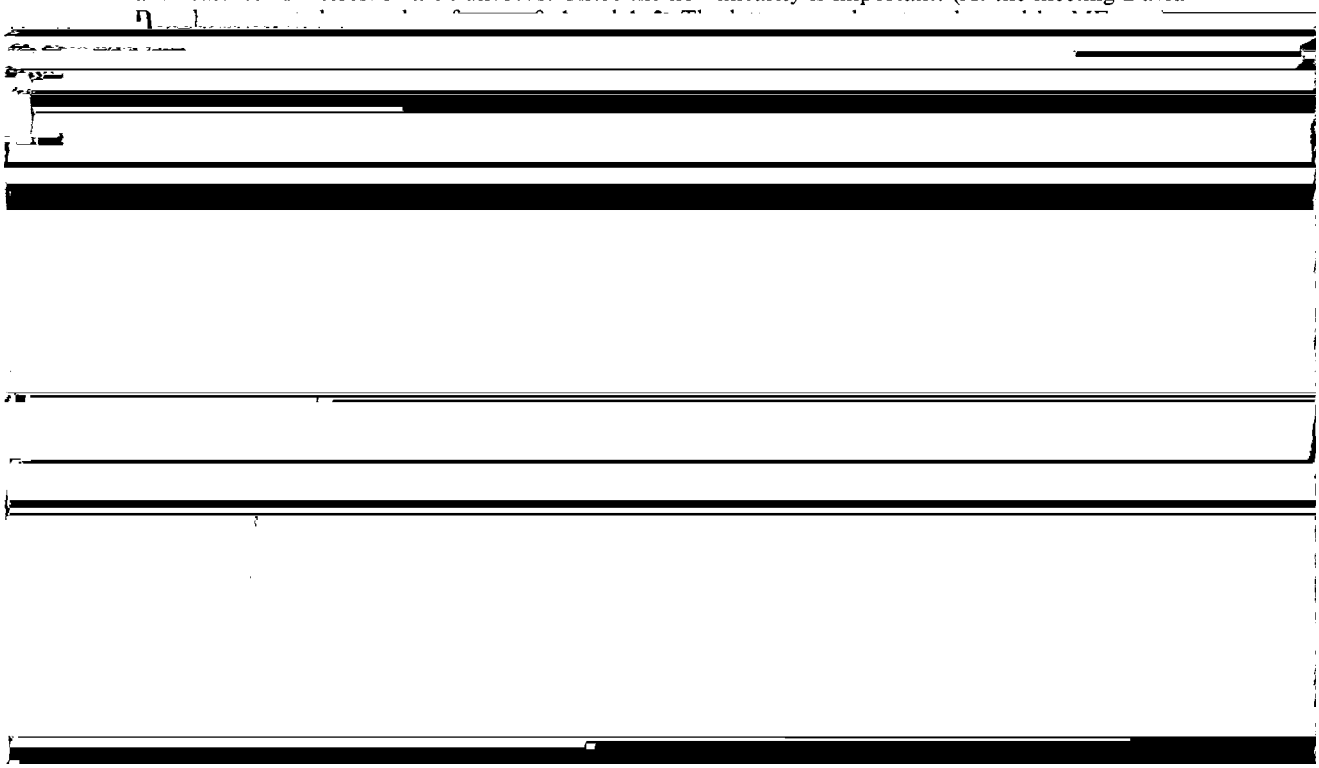
> 'It is the obvious importance of Shannon's theorems that first commands our attention and respect; but as I only realised later, it is just his vagueness on these conceptual issues—allowing every reader to interpret the work in his own way—that made Shannon's writings, like those of Neils Bohr, so eminently suited to become the Scriptures of a new Religion, as they so quickly did in both cases.'

The evangelists have clearly been influential!

We have to be careful with arguments that methods 'look good'. Our human image interpretation system is good at some things, but not others, such as removing one-dimensional blurs. Thus some impressive ME results can be matched by many other algorithms. The more I learn about research on human cognition, the less I believe we do know about our own cognitive system!

One of the major advantages of ME is that it is non-linear and respects non-negativity. The authors are cavalier in the use of 'positive' and 'non-negative'; ME gives strictly positive solutions in most problems, including those they describe, even when zero may be desirable. The use of linear methods with a non-negativity constraint is perhaps the appropriate bench-mark comparison, and Roy Frieden states that it works well in many deblurring problems.

Many astronomical problems have large dynamic ranges, with peak photon counts of say 100000 and features of interest in the hundreds. There the non-linearity is important. (At the meeting David



What maximum entropy is not

Despite the claims in this paper, it should be said that ME is *not*

(a) a regularization method,
(b) a superresolution technique or
(c) a means for noise suppression.

It is certainly true that our ME programs have all these properties, so that the authors' analyses of the ability of ME to increase resolution and to suppress noise are a valuable exercise. To settle on those aspects, however, is to misunderstand the motivation for ME and to ignore its role as a fundamental technique. This is crucially important because, by misunderstanding the basic rationale of ME, we are led to inappropriate and misleading generalizations. These include

(a) attempts to generalize the functional form of the entropy expression (e.g. to use an L_1-norm) and
(b) a plethora of *ad hoc* choices for the regularization constant.

At the same time, useful generalization paths are ignored, including those which enable us to side-step the main conclusions of this paper, namely that the usefulness of ME is limited to objects that are 'nearly black'.

What maximum entropy is

Modern ME data analysis (Gull and Daniell, 1978; Skilling, 1989; Gull, 1989) is a fully quantitative tool for *inference*. We 'fundamentalists' believe that *any* inference must be based on strict adherence to the laws of probability theory, because any deviation automatically leads to inconsistency (Cox, 1946). Consequently, we are Bayesians. We set up our hypothesis space for the image processing problem as follows: f represents the object being reconstructed and D denotes our data set;

$$\underset{\text{joint}}{\Pr(f, D)} = \underset{\text{prior}}{\Pr(f)} \times \underset{\text{likelihood}}{\Pr(D|f)}$$

$$= \underset{\text{evidence}}{\Pr(D)} \times \underset{\text{posterior}}{\Pr(f|D)}.$$

The names for the various terms are familiar, except for the term that we now call the 'evidence'. We find that this neglected term is the most useful of all, because it enables us to compare the posterior probability of alternative hypotheses. For Gaussian errors the likelihood is proportional to $\exp(-\chi^2/2)$ and there are good reasons to take the prior $\Pr(f) \propto \exp(\alpha S)$, where S is essentially the same entropy defined by the authors. The maximum of the posterior distribution, which corresponds to our best inference about f, can therefore be found by maximizing $\alpha S - \chi^2/2$. Note that this derivation shows that there is a good reason to have the Lagrange multiplier on the prior, rather than the likelihood.

The following are some advantages of the Bayesian view of ME.

(a) There is a consistent choice of regularization constant α: use Bayes's theorem to find what it should be (Gull, 1989). This prescription can also be found in the literature: Davies and Anderssen (1986) are 'right'; all other prescriptions are, therefore, 'wrong'.
(b) We can quantify the reliability of our reconstructions and place error bars on our results.
(c) We can estimate the noise level of our data set if it is not known *a priori*.
(d) We can make a series of typical samples from the posterior distribution, and display them as a probabilistic film (Skilling *et al.*, 1991).
(e) Spatial correlations between pixels can be incorporated, and any resulting improvements assessed quantitatively.
(f) There is an enormous variety of applications.

Once we see ME processing in these terms, we can generalize our hypothesis space to overcome its apparent restriction to nearly black objects. Fig. 9 shows an example of a Raman spectrum supplied by P. Graves at Harwell Laboratories. The spectroscopist was interested in a quantitative assessment of the amount of signal in the sharp lines, but these lines are confused by a background of scattered laser light. Our ME reconstruction shows *two* separate channels: a diffuse, background, channel and a sharper, signal, channel that is now 'nearly black'.

An example of modern ME processing that can incorporate spatial correlations in a picture and thereby

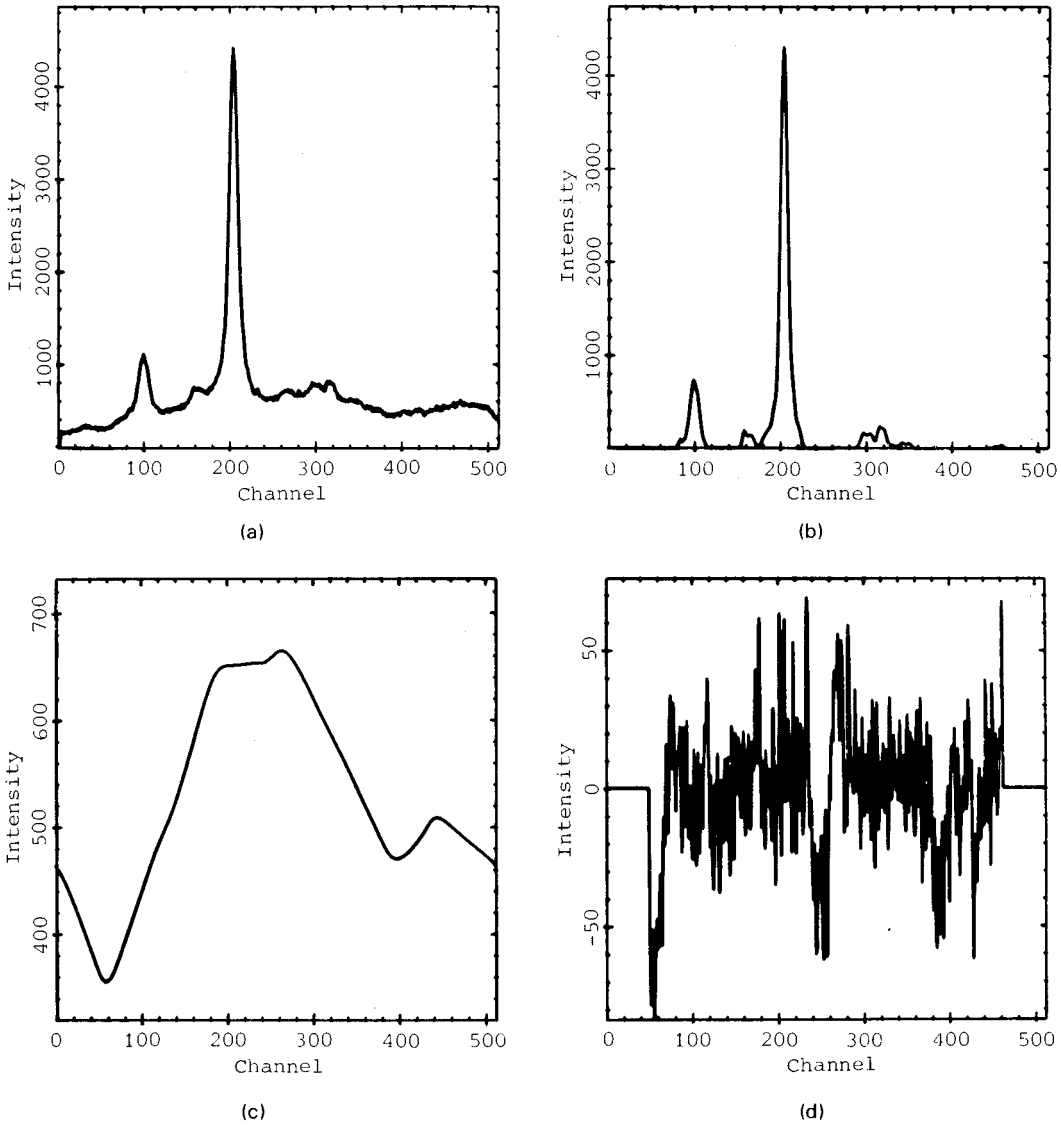


Fig. 9. Raman spectrum: (a) data; (b) maximum entropy—spectrum; (c) maximum entropy—background; (d) maximum entropy—noise

reduce noise is given by Gull (1989). Once again, 'near-blackness' is not the issue—we must merely have a sensible hypothesis space. Finally, Charter (1991) gives a state-of-the-art example that shows how an orally administered drug enters the bloodstream via the stomach and the liver. ME is used to measure the proportion of the dose that was effective (for this case it was 80% ± 5%) and the median time for entry of the drug into the systemic circulation. In all cases it is the quantitative power of Bayesian ME to make these inferences that is crucial. It is far more than a means of producing pretty pictures.

On the lighter side, I note that the authors are wary of fundamentalists, and they suggest that 'this feeling is reciprocated'. I wonder, therefore, whether, in seconding the vote of thanks to the authors for their stimulating contribution, I may be permitted to end by saying 'amen' to these sentiments.

The vote of thanks was passed by acclamation.

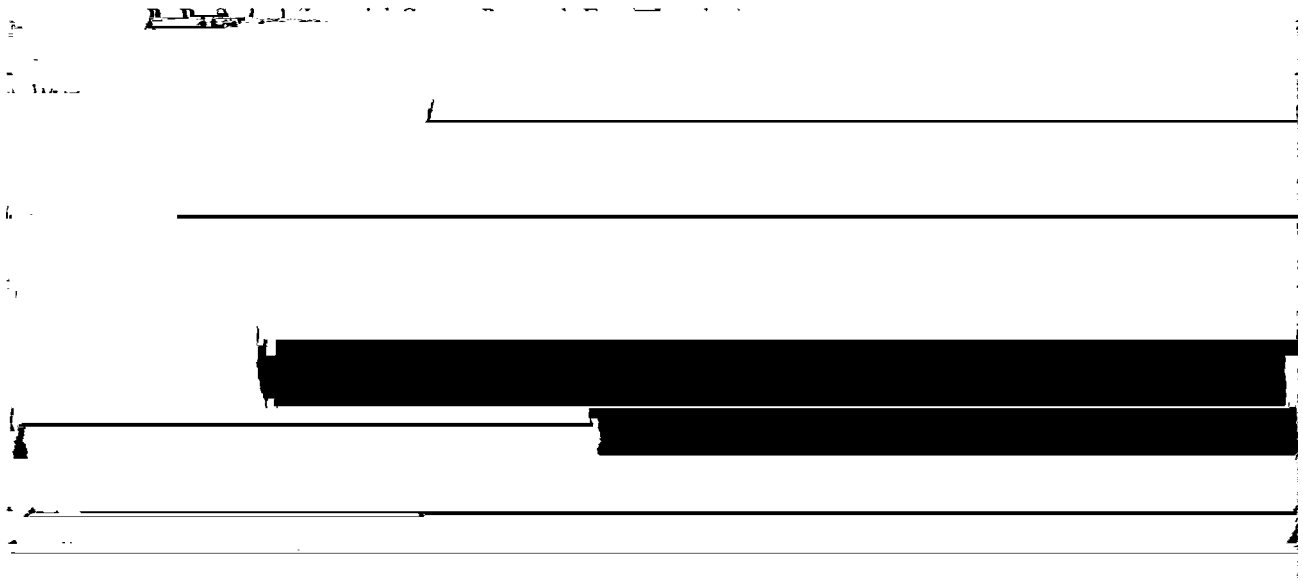
J. A. Jones (University of Oxford): I would like to take a completely different tack on this problem and to say a little about how maximum entropy processing looks to a practising nuclear magnetic resonance (NMR) spectroscopist (this is based on work done with my supervisor, Dr P. J. Hore, in the Physical Chemistry Laboratory, University of Oxford).

What is important to an NMR spectroscopist? The answer is not signal to noise on the base-line, but how reliably various parameters may be extracted from the data, in particular peak areas on which we have decided to concentrate.

We have abandoned the theoretical approach in favour of a crude, but robust, Monte Carlo simulation method. This works as follows: a large number of data sets which contain the same signal but different pseudorandom noise are synthesized; each data set is then processed by the method under investigation and the parameter of interest is extracted. The mean and variance of the parameter are then calculated.

For the peak areas the simple maximum entropy processing described gives both a larger variance and a larger negative bias than conventional processing (direct Fourier transformation). This is obviously completely useless to NMR spectroscopists. If, however, information on the line shape is included, i.e. if we attempt to deconvolve the line shape, then maximum entropy processing can give areas which are both less variable and less biased than those obtained by conventional processing. We feel that this is a genuine improvement. However, maximum entropy processing is not unique in this. It is always possible to obtain a better estimate of areas, in terms of both variability and bias, if an attempt is made to fit a model function (curve fitting).

We would argue therefore that, although maximum entropy processing can give spectra which are in some sense better than conventional spectra, it is not unique in this. Other methods can do better still, particularly if there is a known model function.



Ad hoc estimator (when $K=I$)

Figs 1 and 4 suggest a threshold estimator. The authors state that the l_1-method is better. We can understand this by considering small, medium and large values of y. When y is very small (or negative) it is probably all noise: both estimators have $\delta(y)=0$ for $y<\lambda$. When y is large there is a signal and, assuming symmetric noise, $\delta(y)=y$ seems reasonable; however, $\delta_{l_1}(y)=y-\lambda$. For intermediate values of y we may be observing a large amount of noise or a small signal plus some noise: a good estimator will shrink towards 0, $0<\delta(y)<y$. This is what δ_{l_1} does, but $\delta_T(y)$ is either 0 or y (dependent on $y<\lambda$). Since intermediate values of y are far more common than large values, δ_{l_1} is better on average.

This analysis suggests a new estimator:

$$\delta_A(y)=\begin{cases} 0, & \text{if } y<\lambda_1, \\ (y-\lambda_1)\lambda_2/(\lambda_2-\lambda_1), & \text{if } \lambda_1 \leqslant y \leqslant \lambda_2, \\ y, & \text{if } y>\lambda_2. \end{cases}$$

We may even argue (*à la* James–Stein) that for large y we should 'shrink' towards $\mu=E[x\mathbf{1}_{(x>0)}]$.

Bayes estimator (when $K=I$)

Our model is $Y_i=x_i+z_i$ where $z_i \sim N(0, \sigma^2)$ and $x_i=0$ with probability $1-\epsilon$, and is otherwise from a distribution on $[0, \infty)$ with density f. If σ, ϵ and f are known, the Bayes estimator is the mean of

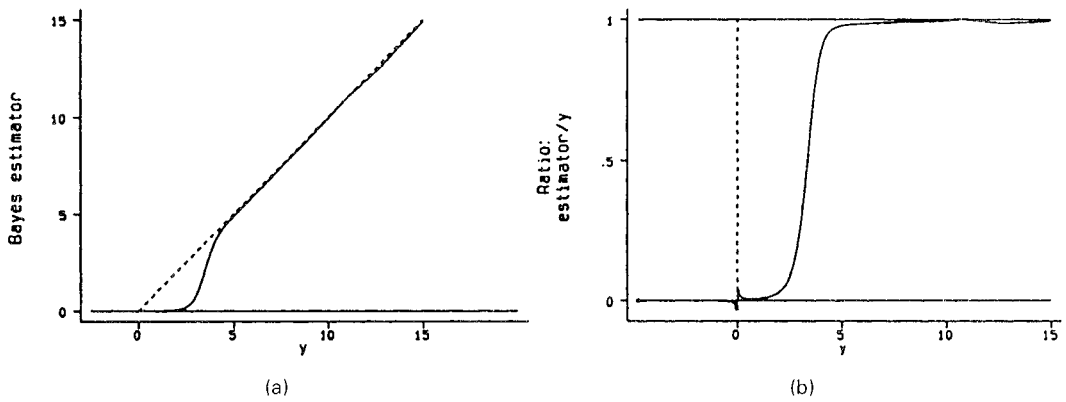


Fig. 10. (a) Bayes estimator compared with $y=x$; (b) ratio of Bayes estimator to y

The problem is then to estimate the vector $\boldsymbol{\beta}$ for a given design matrix X with the knowledge that $\beta_i \geqslant 0$ and all but p are 0. When the total number of covariates is large (n) best subset and backward elimination are impractical. Instead we start with forward selection. Proceeding stepwise we may also allow deletions. Thus, for instance, we grow large regression trees and then prune them. Non-linear estimators are quite familiar to statisticians but have names such as subset selection.

When X is orthogonal there is no need for backward steps. Instead we may build a big model and then shrink those coefficients that are of marginal significance: $\hat{\beta}=(X_0'X_0+\lambda G)^{-1}X_0'Y$, where $X_0=(X_1, X_2, \mathbf{0})$ $(n\times n)$ and

$$G=\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & I\end{pmatrix}$$

$(n\times n)$ with $X_1(n\times p_1)$ the matrix of important covariates, $X_2(n\times p_2)$ a matrix of marginal covariates and I the $(n-p_1)$-identity. (λ is a ridge parameter.) $\hat{\beta}$ has the intuitively appealing properties of the *ad hoc* estimator. (Classically we wish to estimate $X\boldsymbol{\beta}$; here we concentrate on $\boldsymbol{\beta}$ *per se*.)

John T. Kent (University of Leeds): I greatly enjoyed this paper. It has substantially enhanced my understanding of the way that the maximum entropy method works. The paper is concerned with the reconstruction of a signal $\mathbf{x}=(x_i)$ from indirect noisy measurements $\mathbf{y}$. There seem to be two key assumptions about the signal $\mathbf{x}$:

(a) the signal $\mathbf{x}$ is nearly black, i.e. $x_i=0$ for nearly all i, and
(b) the components of $\mathbf{x}$ are non-negative, i.e. $x_i\geqslant 0$ for all i.

Maximum entropy and related methods are effective at reconstructing signals satisfying assumption (a) because they shrink a naïve reconstruction of $\mathbf{x}$ towards 0 or a nearby value. Similarly they are effective for signals satisfying assumption (b) because the estimated reconstruction is required to have non-negative components.

Could the authors give us some insight into the relative importance of assumptions (a) and (b) for explaining the effectiveness of the maximum entropy method, both for the examples in the paper and more widely? In particular how effective is maximum entropy if the signal is only partly black? Also non-negativity seems to be a key assumption in Section 4 on superresolution (see especially Fig. 8) but is relaxed to allow negative and even complex signals in Section 3. Finally, another example where maximum entropy has been used effectively is in deblurring problems, where it seems to be the non-negativity constraint which helps to avoid the 'ringing' artefacts of linear deblurring methods.

William E. Strawderman (Rutgers University, New Brunswick): I congratulate the authors for an interesting and well-written paper. It provides a particularly nice example of the insight and potentially practical benefits that can be obtained by applying statistical optimality concepts to a difficult and important practical problem. I am not an expert in inverse problems and no doubt my comments will reflect my naïvety. In the context of Section 2 at least an empirical Bayes approach might preserve

(at least nearly) the optimality properties of the minimum l_1-rule in the nearly black case and perhaps allow as good or better behaviour when the object is not nearly black. One possible class of priors would be a mixture of a point mass at 0 with probability to be estimated from the data and a gamma prior with parameters also to be estimated. Here the probability of the point mass at 0 would play the role of the tuning parameter λ for the maximum entropy and minimum l_1-methods and would be adaptively determined. The potential advantage of such a method should come in its ability to adapt at least somewhat to the non-black part of the image. As such empirical Bayes methods are well known to the authors I presume that there are computational or other reasons for not using some variation on the empirical Bayes theme.

Didier Dacunha-Castelle (Université Paris Sud, Orsay): I am pleased to congratulate the authors for their fascinating paper. For their model (1) $Y_t = (Kx)_t + \epsilon_t$, x is chosen by maximizing $S(x)$ subject to the relaxed constraint. Model (2) is $\| Y - Kx \|_2 < \epsilon$; of course the procedure can be thought of as a deterministic fitting procedure. There is a certain kind of optimality if we choose $S(x) = \| x \|_{l_1}$. This optimality depends on

(a) the normality of (ϵ_t) and
(b) on the asymptotics chosen, i.e. a level of noise tending to 0.

What happens if we change condition (a)? What is the link between this result with the classical asymptotic optimality in robustness theory (as in P. Huber's work)? From this deterministic point of view, instead of model (2) we can regularize a problem maximizing $S(x)$ under constraint (3), $m_k(x) = m_k(y)$, $1 \leqslant k \leqslant m$, when

$$m_k(x) = \frac{1}{n} \sum_{k=1}^{n} x^k.$$

This point of view is linked to the use of a generalization of the maximum entropy principle as in Dacunha-Castelle and Gamboa (1990) and Gamboa and Gassiat (1990) and in some statistical problems. It seems well adapted when we work with the asymptotics of discretization, with (x_t) thought of as a finite number of values of a continuous function x_t. In this asymptotic, there is a Bayesian interpretation of the choice of S (Gamboa and Gassiat, 1991a, b).

Elisabeth Gassiat (Université Paris-Sud, Orsay): I would first like to congratulate the authors for their clear exposition of the properties of the so-called 'maximum entropy' methods in general, underlying the role of positiveness and 'nearly blackness' of the object to be recovered. I shall focus my contribution on the second part of the paper, where the operator K is not easily invertible. The authors show a superresolution property in the particular case where K is the finite Fourier transformation: $\mathbf{y} = K\mathbf{x} + \boldsymbol{\epsilon}$, $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} \in \mathbf{R}^m$, $m << n$. Here $\boldsymbol{\epsilon}$ is a deterministic error.

In this approach, and looking at their proofs, the role of the discretization and of the geometric structure induced by K does not demonstrate clearly why superresolution occurs.

The problem may be rephrased in a more general form: if U is a compact metrizable space, $\mathbf{x}$ a positive measure, K an m-dimensional continuous function on U, the observation is

$$\mathbf{y} = \int_U K \, d\mathbf{x} + \boldsymbol{\epsilon}.$$

The problem is now a generalized moment problem. In Gamboa and Gassiat (1991b), it is shown that superresolution occurs if and only if $\int_U K \, d\mathbf{x}$ is a determinate point, i.e. a point for which the set of positive measures solutions to the moment problem is reduced to a singleton. In that paper criteria are also given to show whether a point is determinate or not. All the results rely on developments on 'maximum entropy methods on the mean' developed in Gamboa and Gassiat (1991a). Another point of view of near-blackness has also been studied in Gassiat (1990), relating it to some concentration property of the support of any positive measure solution of the moment problem.

Andrew Gelman (University of California, Berkeley): The authors present an enlightening discussion of where and why maximum entropy methods are effective. Many of their results become even clearer from a Bayesian perspective: better prior distributions yield better estimates. Jaynes (1987) makes a similar point in his discussion of the non-linear shrinkage that results from Bayesian spectral estimation using a nearly black model (without maximum entropy).

For simplicity, I shall analyse the example of Section 2. Each of the estimates considered by the authors

TABLE 2
Shrinkage estimates and their corresponding prior densities

Estimate $\hat{x}_i$	*Prior density on* x_i
y_i	Uniform$(-\infty, \infty)$
$\max(y_i, 0)$	Uniform$[0, \infty)$
$\hat{x}_{\mathrm{RLS}}$	Normal$(0, \sigma^2/\lambda)$
$\max(\hat{x}_{\mathrm{RLS}}, 0)$	Normal$(0, \sigma^2/\lambda)$, truncated to be non-negative
δ_{ME}	Density proportional to $\exp\{(-\lambda/\sigma^2)x_i \log x_i\}$
δ_{l_1}	Exponential(σ^2/λ)
δ_{Thresh}	Mixture of point mass at 0 and uniform$[0, \infty]$

may be interpreted as a posterior mode under the appropriate family of proper or improper prior distributions on $x_1, \ldots, x_n$ (Table 2).

The regularization parameter λ is assumed known. In practice, the normal and maximum entropy prior distributions typically each take another parameter, fitting location for the normal and scale for the entropy distribution. In either case, fitting the additional parameter allows the fixed point of shrinkage to be fitted to the data, rather than be fixed at 0 for quadratic regularization and e^{-1} for maximum entropy. (Incidentally, fitting the second parameter eliminates the problem with maximum entropy described in the penultimate paragraph of Section 2.2.)

As the authors report, maximum entropy reconstruction shrinks large data values proportionately less than small data values, thus preserving peaks while suppressing noise. Least squares regularization performs worse for the nearly black object, even when restricted to positivity, because it shrinks all positive data by a common factor.

The shrinkage behaviour of the estimates makes perfect sense in light of the corresponding prior distributions. The prior density for the trivial estimate is constant, so the data are not pulled at all. The priors for least squares regularization have rapidly decaying $\exp(-x^2)$ tails, so large data points will be pulled strongly towards the fixed point. The priors for maximum entropy—and also for the l_1-estimate—decay only like $\exp(-x)$ and so large data values are shrunk less strongly. The threshold estimate δ_{Thresh} pulls points to the point mass at 0, but its uniform component fails by not shrinking the positive components x_i to their common mean.

In summary: under the 'nearly black' model, the normal prior is terrible, the entropy prior is better and the exponential prior is slightly better still. (An even better prior distribution for the nearly black model would combine the threshold and regularization ideas by mixing a point mass at 0 with a proper distribution on $[0, \infty]$.) Knowledge that an image is nearly black is strong prior information that is not included in the basic maximum entropy estimate.

The following contributions were received in writing after the meeting.

Bob Anderssen (CSIRO, Canberra): In making this important contribution, the authors, not surprisingly, find that maximum entropy has its limitations. The impact which this paper has will depend on the extent to which it removes the belief that maximum entropy is a panacea. The shortcomings of maximum entropy have already been noted by others (e.g. Engl and Landl (1991), Koch and Anderssen (1987) and Titterington (1984)). What the present authors have achieved is to have placed this earlier understanding on a much firmer foundation.

Certainly, strong support for maximum entropy can be based on physical (thermodynamical) and philosophical (minimum assumptions) considerations. But, it is one among many such principles like minimum energy in elasticity and minimum action in optics and seismology. The choice of methodology should be driven by the information flowing from the context in which the data have been collected rather than other considerations. The context also has a bearing on the statistical interpretation of the approach (O'Sullivan, 1986; Wahba, 1990). For example, Wahba (1990), section 1.5, gives a Bayesian estimation interpretation for regularization with a quadratic regularizor.

The success (when it occurs) of maximum entropy relates to its implementation via a regularization framework such as

$$\min_{\mathbf{x}}\{\|\mathbf{y} - K\mathbf{x}\|_2^2 + 2\lambda\,\Lambda(\mathbf{x})\}, \tag{25}$$

and to the ramifications associated with the use of this variational formulation. The present authors understand and exploit this point. In fact, their conclusions fit the general picture about (discrete) regularization which theoretical studies have so far uncovered. From the variational framework (25), it is clear that, in the minimization, the regularizor $\Lambda(\mathbf{x})$ controls the smoothness of $\mathbf{x}$. For quadratic regularizors, where $\Lambda(\mathbf{x}) = \|\mathbf{T}\mathbf{x}\|^2$ with $\mathbf{T}$ a linear differential operator, we know from theoretical investigations (Lukas, 1980; Wahba, 1990) that

(a) If $K = I$ (data smoothing), then the structure of the solution of expression (25) (the regularized approximation), x_λ, is piecewise with respect to the ordinates of the data and, between the data points, is a function from the null space of $\mathbf{T}^*\mathbf{T}$. If K is a linear operator, then x_λ is not necessarily piecewise, but its structure is known explicitly (Lukas, 1980).
(b) As $\lambda \to 0$, the regularized approximation x_λ tends to a corresponding $\bar{x}$ such that $K\bar{x}$ interpolates the data.
(c) As $\lambda \to \infty$, the regularized approximation x_λ tends to the least squares function which best fits the data from among the functions which form the null space of $\mathbf{T}$.

It is therefore natural to conclude that, because the null space of the maximum entropy regularizor lacks the structure associated with that of quadratic regularizors, the utility of maximum entropy regularization will be limited to specialized situations.

C. A. Glasbey and G. W. Horgan (Scottish Agricultural Statistics Service, Edinburgh): This paper sheds welcome light on what to us was a nearly black object, maximum entropy. We have three comments.

(a) Although Table 1 gives minimax mean-squared errors for various shrinkage operators, in practice, with a particular probability density function for $\mathbf{x}$, relative performances could be quite different. For example, in many cases a threshold operator will do better than the uniform shrinkage of the l_1-estimator.
(b) In applications such as nuclear magnetic resonance spectroscopy and image analysis, $\mathbf{x}$ is usually a function over continuous one- or two-dimensional space, although $\mathbf{y}$ is only observed on a grid of points. Could the authors comment on the recovery of $\mathbf{x}$ in this case?
(c) In image analysis, near-blackness is but one of many forms of prior knowledge. The human vision system exploits features such as spatial continuity and straightness of edges in images to achieve greater resolution than the limits predicted from the optics of the eye. In computer vision it remains a largely open question which assumptions lead to tractable algorithms, and how much can be gained in resolving power.

I. J. Good (Virginia Polytechnic Institute and State University, Blacksburg): In the method of maximum penalized likelihood (MPL) a roughness penalty is subtracted from a log-likelihood, and the difference is then maximized. (For numerous references see, for example, MPL in the first index of Good (1983a), especially Good and Gaskins (1980) and Good and Deaton (1981). See also Tapia and Thompson (1978), Good (1983b) and Good (1989).) Without the penalty, i.e. if pure maximum likelihood (ML) is used, the result is usually too rough if the number of parameters exceeds the number of observations, e.g. if the number is in principle infinite (the 'nonparametric' case). The penalty can be multiplied by a factor λ which acts as a smoothing parameter, or, in the Bayesian interpretation, as a hyperparameter, and can also be interpreted as a Lagrange multiplier. For categorical data the negative entropy was suggested as a roughness penalty (Good (1963), p. 931) thus providing the natural compromise between ML (appropriate for large samples) and 'maximum entropy' (appropriate for zero-size samples). This suggestion was developed in detail for contingency tables by Pelz (1977). We have found, in work not yet published, that the method appears to be most appropriate for the estimation of the physical probabilities corresponding to cells containing small or zero frequencies. Thus the MPL method should be useful for multidimensional contingency tables because, putting it roughly, the larger the dimensionality the larger the fraction of empty cells. This state of affairs is analogous to the requirement, for continuous problems, that the data be 'nearly black', as in the work of the present authors.

For continuous problems, involving the estimation of probability densities, where there is always a natural ordering, I have preferred, from a theoretical point of view, to use measures of roughness that depend on derivatives of the density function (or differences in some numerical approaches). Even for contingency tables, if the rows have a natural ordering (or the columns etc.), the use of sums of squares of first or second differences might be theoretically preferred to entropy. But in practice the

entropic penalty might be just as good. In particular, I hope that the authors' impressive theorems can be extended to categorical data.

My main point is that the analogy between continuous problems on the one hand, such as those dealing with radio astronomy and nuclear magnetic resonance, and discrete problems, on the other hand, should be held in mind.

Jim Kay (University of Glasgow): The authors are to be congratulated for producing an interesting paper which presents the advantages of the maximum entropy method in an illuminating and non-fundamentalist manner. The paper refers to ill-posed problems and yet in the theory developed in Sections 2 and 3 the contexts considered are not ill posed. Clearly, in general the stationarity equations are

$$\mathbf{x} = \hat{\mathbf{x}}_{\mathrm{LS}} - \lambda (K^{\mathrm{T}}K)^{-1} \operatorname{grad}(\Phi),$$

where Φ is a non-linear penalty function, and so the simple co-ordinatewise non-linearity is lost. To what extent does this theory extend to (real) ill-posed problems?

With regard to computation, Green (1990) has introduced an algorithm based on penalized likelihood estimation. It is simple to programme and yields the simple recursion

$$\mathbf{x}^{\mathrm{new}} = \mathbf{x}^{\mathrm{old}} - n^{-1} D^{-1} [\, K^{\mathrm{T}}(\mathbf{y} - K\mathbf{x}^{\mathrm{old}}) - \lambda \operatorname{grad}\{\Phi(\mathbf{x}^{\mathrm{old}})\}\,],$$

where $D = \operatorname{diag}(K^{\mathrm{T}}K)$.

Some other penalty functions that might be of interest are

$$\Phi_1(\mathbf{x}) = \sum_i x_i \beta \qquad (0 < \beta < 1),$$

$$\Phi_2(\mathbf{x}) = \sum_i x_i (x_i^{\alpha - 1}/\alpha) \qquad (0 < \alpha < 1),$$

$$\Phi_3(\mathbf{x}) = \sum_{i \sim j} x_i \log(x_i / x_j)$$

and

$$\Phi_4 = \sum x_i \log(x_i / x_j)(1 - e_{ij}) + \gamma \sum_{i-j} e_{ij},$$

where $i \sim j$ means that pixels i and j are neighbours. In the special case considered in Section 2, the use of penalties Φ_1 and Φ_2 would involve simple co-ordinatewise non-linearities and might also exhibit signal-to-noise enhancement. Reconstruction with Φ_2 has a fixed point at $(\alpha + 1)^{-1/\alpha}$ and a corresponding shrinkage property; as $\alpha \to 0$ this penalty yields maximum entropy whereas, as $\alpha \to 1$, it gives a quadratic regularization. In general, penalties Φ_1 and Φ_2 would exhibit superresolution as conditions (16) and (17) would be satisfied. The use of penalty Φ_3 would invoke a local smoothing, similar in spirit to first-order quadratic regularization: generalizations are possible; see Thompson (1988). In penalty Φ_4, $e_{ij} = 1$ when pixels i and j are connected by an edge and $e_{ij} = 0$ otherwise. The $\{e_{ij}\}$ may be 'estimated' from the data. Such an approach might be helpful in reducing the negative bias of spikes.

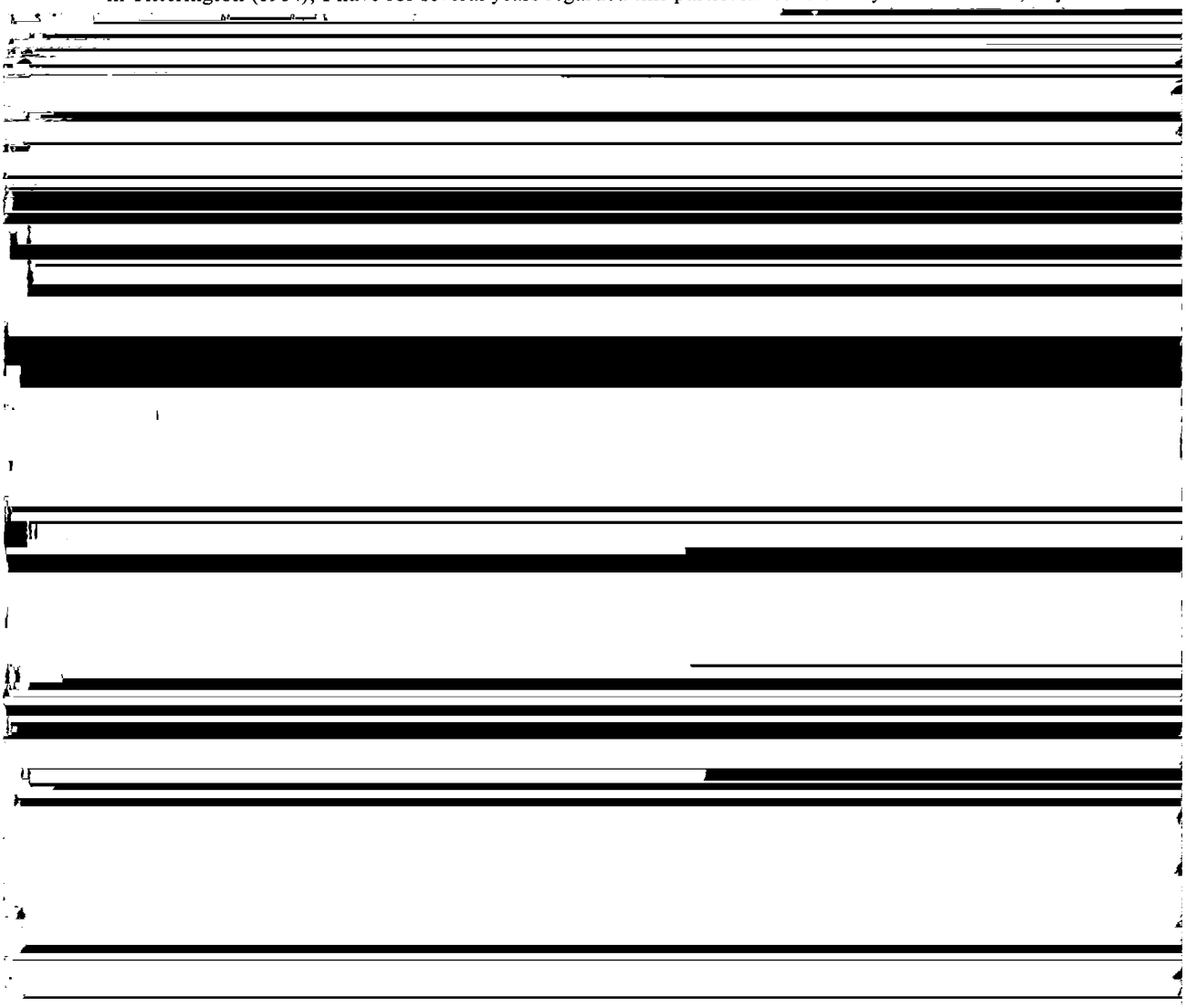
Alan M. Thompson (University of Glasgow): I would like to congratulate the authors for their paper which goes a long way towards putting the importance of the maximum entropy method into perspective. In light of their work it seems ironic that the reason that the maximum entropy method is used in radio astronomy is because it reconstructs images of spatially distributed sources better than its main rival CLEAN (Cornwell and Evans, 1985).

There has also been some recent work on a comparison of some methods of selecting the smoothing parameter λ (e.g. Thompson *et al.* (1991)) which shows that the choice of smoothing parameter can have a significant effect on the quality of the reconstructed image. In particular, the new methods developed by the maximum entropy school (Gull, 1989) which are mentioned by the authors have been examined by Thompson and Kay (1991) who have obtained results which suggest that these methods produce reconstructed images which are undersmoothed (i.e. produce reconstructed images which retain more noise than the best fit solution).

D. M. Titterington (University of Glasgow): I was very sorry not to be able to attend the meeting at which this important paper was presented. The paper provides a major service in identifying

concrete comparative characteristics of a range of non-linear inversion techniques that include maximum entropy (ME) as a special case. I owe the ME method a personal debt of gratitude in that my initial encounter with it was a major factor in my being led from a comparatively narrow interest in density estimation to an awareness of a general structure underlying many statistical smoothing exercises (Titterington, 1985a, b). I had my differences with the originators of the ME method in that I could not accept that the approach is uniquely special among regularization procedures; I could not fit the wide range of data analytic applications of the method into the axiomatic ME formalism for probability measures (Jaynes, 1968; Shore and Johnson, 1980; Johnson and Shore, 1983; Tikochinsky *et al.*, 1984a, b). At the time, I worried that I was just missing the key link, and I was greatly reassured when I discovered that Terry Speed, for one, was of a similar mind.

Although I would now probably wish to put my objections in a way different from that expressed in Titterington (1984), I have for several years regarded this particular controversy as a non-issue; any



quantitatively as one of a wide class of comparable prescriptions for smoothing. There are clearly many further aspects to be examined, some of them alluded to in the paper. Among them are the following.

(a) In spite of what is said in Section 2.4.3, some systematic study of the influence, in practical terms, of the choice of λ should be carried out. An oft-repeated, throwaway line in the context of smoothing procedures is that the choice of method (e.g. choice of kernel function in kernel-based density estimation) is much less important than the choice of smoothing parameter. Is the converse being postulated here, partially, at least?

(b) If the 'object' is believed to be nearly black, should this be accommodated at the outset, through appropriate modelling and suitable, possibly semiparametric, analysis, rather than by an essentially nonparametric approach? In this context, algorithms such as CLEAN, referred to by Koch and

We again thank the authors for examining carefully the role of positivity and l_1-estimation in the case of the nearly black object.

The **authors** replied later, in writing, as follows.

The discussants approached their task from many different viewpoints, each with its own set of characteristic concerns and reactions. Below we paraphrase these viewpoints and compare them with our own. (Our grouping is arbitrary; the discussants may not agree with it.)

Scientists

Ripley claims that the squared error loss may be an inadequate measure of how scientists would judge the quality of fit; he asks how to quantify the visual quality of reconstruction. Anyone who has proved theorems about mean-squared-error optimality, and then compared the theorems with pictures of actual reconstructions, has probably lost sleep over this question. Now Ripley states the question publicly. It is important for our profession, and not just this discussion, that someone finds an answer!

Jones points out that scientists are often interested in the recovery of certain linear functionals (peak areas) rather than the whole curve. Except for John Rice, statisticians have not looked at the recovery of peak areas. We hope that Jones's question will prompt renewed interest. We suspect that Jones is correct: that non-linear estimates of peak areas can improve on linear estimates.

Decision theorists

Decision theorists focus on quantifying and improving performance under a given statistical model, and on understanding how variations of the model affect the results.

Accordingly, both Sasieni and Strawderman suggest improvements to the l_1-rule of Section 2.3. With good choices for λ_1 and λ_2, Sasieni's δ_A will certainly perform at least as well as the l_1-rule, though of course we now have two parameters to choose rather than one. The class of Bayes rules δ_B is ideally suited to precise prior information that the image is ϵ-black. A more detailed study of the minimax risk $M(\epsilon)$ of theorem 1 (Johnstone, 1991) leads to a three-term asymptotic expansion of which equation (9) is the leading term. From this, it follows that, although the l_1-rule is first order minimax as shown in theorem 1, it is not even second-order minimax. It would be interesting to know whether Sasieni's

of a parallel to 'superresolution' in the model selection literature. Superresolution occurs in the underdetermined case, when the unknown parameter vector is sparse and positive; what is the model selection analogue?

Good's suggestion to keep sparse contingency tables in mind is well taken. The details of a corresponding theoretical development would differ somewhat from the Gaussian case, if only for the following reason. In theorem 1, the phenomenon underlying the value for the minimax risk is the threshold λ_ϵ at which an observation is equally likely (in the marginal distribution) to be a (rare) large error imposed on a (likely) zero signal or a negligible error on a (rare) signal of magnitude about λ_ϵ. In a contingency table with, say, Poisson counts, the corresponding situation cannot arise for a cell with zero mean because positive counts cannot occur by definition. Perhaps this could be handled by assuming a small positive base-line intensity on which signal is added.

Bayesians

Bayesians seem most interested in quantifying prior information about near-blackness probabilistically, and in interpreting the resulting Bayes rules.

For the signal-plus-noise model, Gelman interprets various estimators as maximum *a posteriori* for various prior distributions. This is a natural way to obtain some quick intuition into the structure of a given estimator. We had done this also, and found it curious that, whereas the greatest successes of ME *vis-à-vis* linear methods occurred in nearly black settings, the imputed maximum *a priori* prior suggested that ME was somehow optimized for independent and identically distributed 'snowy' images with the $\exp\{-(\lambda/\sigma^2)x_i \log x_i\}$ pixelwise distribution!

Accordingly, the success of ME derives from its *non-linear* behaviour at those objects which are near-black, rather than from the astuteness of its prior.

Naturally, however, better priors give better results. Sasieni and Strawderman both suggest the use of empirical Bayes methods, which in effect derive a prior from the data. There is little doubt that these methods would outperform both ME and l_1-methods. However, optimality *per se* has not been our main goal.

Maximum entropy

As Gull's discussion shows, the Cambridge rationale for ME is undergoing changes, so that the 'consistency' argument of years past, which leads to entropy penalization of the likelihood ('classical ME'), has been replaced by a 'coherence' argument, which leads to 'modern ME'.

The example Gull (1989) of modern ME processing is, in Bayesian language, a hierarchical Bayes model based on Gaussian priors at two levels, with hyperparameters chosen to maximize the appropriate marginal distribution (see Berger (1985) and Good (1983a) for further exposition of this standard 'type II maximum likelihood' approach). The specific model used seems quite close in spirit to the conditionally autoregressive models of Besag (1974) or the integrated Brownian motion priors of Grace Wahba.

The cited advantages of Gull's Bayesian approach to ME are largely characteristic of the use of the Bayesian paradigm, and by no means specific to the use of entropic priors on appropriately chosen 'hypothesis spaces'. In the colourful words of L. J. Savage, if one breaks the Bayesian egg, one gets to enjoy the Bayesian omelette. Of course, one had better use good eggs!

Frequently the *non-linearity* of ME, rather than its Bayesian character, or other philosophical underpinnings, is responsible for the method's relative successes. In Gull's Raman spectrum example, the separation into 'sparse spectrum' plus 'rolling base-line' is remarkable. This separation is due to a non-linear effect called *Logan's phenomenon* by Donoho and Stark (1989) and Donoho and Logan (1991), in a study of minimum l_1-norm reconstruction. Understanding this successful example requires ideas of Ben Logan or Werner Heisenberg (i.e. the uncertainty principle) rather than Thomas Bayes or Edwin Jaynes.

Quadratic regularization

Quadratic regularization is the most widely employed method for linear inverse problems. Grace Wahba, a distinguished developer of techniques based on quadratic regularization, makes interesting historical comments about the improvements which she has discovered when supplementing quadratic regularization with positivity constraints. Her volumetric heuristic may persuade many of the importance of positivity constraints.

Kay and Anderssen discuss the structure of quadratic regularizers and ask whether similar structure persists with non-quadratic regularizers. Analysis of quadratic regularizers is based on methods of calculus,

but superresolution occurs where the regularizer and associated optimum solution operator are non-differentiable. Superresolution seems intrinsically tied up with the failure of traditional linearizations.

Recent research associated with quadratic regularization has focused on the choice of the regularization parameter λ. Both Titterington and Thompson express some surprise at our relative lack of discussion of this topic. Titterington's (1985a) excellent survey was concerned mostly with families of *linear* estimators indexed by a smoothing parameter. When the choice is between linear estimators, we would agree that the main practical issue is often the choice of the smoothing parameter rather than the specific kernel, filter or whatever. However, when we consider situations with strong prior information about sparsity, the advantages of appropriate non-linear methods over linear methods may be so large that they warrant separate study, and this was our focus. The appropriate choice of smoothing or regularization parameter for non-linear superresolving methods is the natural next question, and we thank the discussants for their references to ongoing work on this topic.

Applied mathematicians

Gassiat and Gamboa show that superresolution is quite general in an abstract qualitative sense. Our approach is more specialized, but it yields explicit quantitative information on the stability of superresolution; compare Donoho (1990) and Donoho *et al.* (1991).

John Kent asked us to compare the superresolution available from sparsity constraints alone, and from sparsity in conjunction with positivity constraints. Donoho (1990) and Donoho *et al.* (1991) develop a quantitative theory of superresolution and provide explicit stability estimates under sparsity constraints which are parallel to explicit stability estimates under positivity constraints. From the point of view of theoretical performance, the combination of sparsity and positivity constraints is not intrinsically more powerful than sparsity constraints alone.

From the point of view of computation, however, the two types of constraint are very different. Whereas imposition of positivity constraints in least squares fitting is accomplished with standard quadratic programming software, or through entropy penalization, imposing sparsity constraints is a job of seemingly combinatorial complexity requiring a best sparse subset search through an enormous number of linear least squares fits. Our understanding of sparsity constraints is at present largely theoretical, whereas many published reconstructions with positivity constraints exist.

Signal/image processors

Glasbey and Horgan ask how our conclusions generalize to continuous objects, objects with spatial continuity and objects with edges. Superficially, there is little applicability, as we discuss objects such as discretized star maps and molecular spectra, which do not have such features.

The recently developed wavelet transform (Meyer, 1990; Mallat, 1989; Daubechies, 1989) opens up several possible applications of our ideas in image and signal processing. Work of Coifman *et al.* (1990) suggests that many images and signals, when wavelet transformed, become sparse objects (i.e. objects most of whose co-ordinates are essentially 0).

We may regard such wavelet transforms as nearly black objects. Donoho and Johnstone (1991) exploit this point of view, applying simple threshold non-linearities to wavelet transforms to suppress noise in a theoretically optimal way. When the wavelet transform is inverted, we obtain a curve or surface reconstruction. Despite the apparent simplicity of this 'wavelet shrinkage' method, the resulting curve or surface reconstruction adapts automatically to discontinuities, singularities and spatial smoothness of the unknown object and adapts in a way which is near optimal.

Conclusion

The diverse and thoughtful discussion has slighted, so far, the concept of sparsity. We ask explicitly: how generally useful is the concept of the nearly black object? The optimizing impulses of decision theorists, the philosophizing of Bayesians and ME advocates, the analytical skills of mathematicians—all can be well exercised by an ample subject. Is sparsity ample in this sense?

We believe so. Not only molecular spectra and sky maps, but also images and voice recordings (after suitable transformation) may be seen as nearly black objects. We look forward to work, from the many points of view represented here, identifying and exploiting cases where the object is nearly black.

We thank the Society and discussants for a most stimulating occasion.

REFERENCES IN THE DISCUSSION

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis.* New York: Springer.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.
Bickel, P. J. (1983) Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics* (eds M. H. Rizvi, J. S. Rustagi and D. Siegmund). New York: Academic Press.
Brown, L. D. (1971) Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Ann. Math. Statist.*, **42**, 855–903; correction, *Ann. Statist.*, **1** (1973), 594–596.
Charter, M. K. (1991) Quantifying drug absorption. In *Maximum Entropy and Bayesian Methods* (eds W. T. Grandy, Jr, and L. H. Schick), pp. 245–252. Dordrecht: Kluwer.
Coifman, R., Meyer, Y., Quake, S. and Wickerhauser, V. (1990) Signal processing and compression with wavelet packets. *Technical Report*. Department of Mathematics, Yale University, New Haven.
Cornwell, T. J. and Evans, K. E. (1985) *Astron. Astrophys.*, **143**, 77.
Cox, R. T. (1946) Probability, frequency and reasonable expectation. *Am. J. Phys.*, **14**, 1–13.
Dacunha-Castelle, D. and Gamboa, F. (1990) Maximum d'entropie et problème des moments. *Ann. Inst. Henri Poincaré*, **26**, 567–596.
Daubechies, I. (1989) Orthonormal bases of compactly supported wavelets. *Communs Pure Appl. Math.*, **41**, 909–996.
Davies, A. R. and Anderssen, R. S. (1986) Optimisation in the regularisation of ill-posed problems. *J. Aust. Math. Soc.* B, **28**, 114–133.
Donoho, D. L. (1990) Super-resolution via sparsity constraints. *Technical Report*. Department of Statistics, University of California, Berkeley.
Donoho, D. L., Gassiat, E. and Stark, P. B. (1991) Superresolution and positivity constraints: theory and experiment. *Technical Report*. Department of Statistics, University of California, Berkeley.
Donoho, D. L. and Johnstone, I. M. (1991) Smoothing by wavelet shrinkage: I, The minimax nonlinearity. *Technical Report*. Department of Statistics, Stanford University.
Donoho, D. L. and Logan, B. F. (1991) Signal recovery and the large sieve. *SIAM J. Appl. Math.*, to be published.
Donoho, D. L. and Stark, P. B. (1989) Uncertainty principles and signal recovery. *SIAM J. Appl. Math.*, **49**, 906–931.
Engl, H. W. and Landl, G. (1991) Convergence rates for maximum entropy regularization. *Institutsbericht*. Institut für Mathematik, Universität Linz.
Gamboa, F. and Gassiat, E. (1990) Maximum d'entropie et problème des moments: cas multidimensionnel. *Probab. Math. Statist.*, to be published.
——— (1991a) Extension of the maximum entropy method on the mean and a Bayesian interpretation of the method. Submitted to *Probab. Theory Reltd Flds*.
——— (1991b) MEM techniques for solving moment problems. Submitted to *Probab. Theory Reltd Flds*.
Gassiat, E. (1990) Problème des moments et concentration de mesure. *Compte Rend. Acad. Sci.*, **310**, 41–44.
Good, I. J. (1963) Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.*, **34**, 911–934.
——— (1983a) *Good Thinking*, p. 296. Minneapolis: University of Minnesota Press.
——— (1983b) Review of *The Maximum Entropy Formalism*. *J. Am. Statist. Ass.*, **78**, 987–989.
Good, I. J. and Deaton, M. L. (1981) Recent advances in bump-hunting. In *Computer Science and Statistics: Proc. 13th Symp. Interface* (ed. W. F. Eddy), pp. 92–104. New York: Springer.
Good, I. J. and Gaskins, R. A. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data (with discussion). *J. Am. Statist. Ass.*, **75**, 42–73.
Good, I. J., Holtzman, G. I., Deaton, M. L. and Bernstein, L. H. (1989) Diagnosis of heart attack from two enzyme measurements by means of bivariate probability density estimation: statistical details. *J. Statist. Computn Simuln*, **32**, 68–76.
Green, P. J. (1990) On use of the EM algorithm for penalized likelihood estimation. *J. R. Statist. Soc.* B, **52**, 443–452.
Gull, S. F. (1989) Developments in Maximum Entropy data analysis. In *Maximum Entropy and Bayesian Methods* (ed. J. Skilling), pp. 53–71. Dordrecht: Kluwer.
Gull, S. F. and Daniell, G. J. (1978) Image reconstruction from incomplete and noisy data. *Nature*, **272**, 686–690.
Jaynes, E. T. (1968) Prior probabilities. *IEEE Trans. Syst. Sci. Cybernet.*, **4**, 227–241.
——— (1978) Where do we stand on maximum entropy? In *The Maximum Entropy Formalism* (eds R. D. Levine and M. Tribus). Cambridge: Massachusetts Institute of Technology Press.
——— (1987) Bayesian spectrum and chirp analysis. In *Maximum-entropy and Bayesian Spectral Analysis and Estimation Problems* (eds C. R. Smith and G. J. Erickson), pp. 1–37. Dordrecht: Reidel.
Johnson, R. W. and Shore, J. E. (1983) Comments on and correction to Shore and Johnson (1980). *IEEE Trans. Inf. Theory*, **29**, 942–943.
Johnstone, I. M. (1991) On minimax estimation of nearly-black signals in Gaussian white noise. *Technical Report*. Stanford University.
Koch, I. and Anderssen, R. S. (1987) A direct surface smoothing procedure for Fourier image reconstruction in radiophysics. *Astron. Astrophys.*, **183**, 170–176.
Lukas, M. A. (1980) Regularization. In *The Application and Numerical Solution of Integral Equations* (eds R. S. Anderssen, F. R. de Hoog and M. A. Lukas). Alphen aan den Rijn: Sijthoff and Noordhoff.
Luttrell, S. P. (1990) A Bayesian derivation of an iterative autofocus/super-resolution algorithm. *Inv. Prob.*, **6**, 975–996.
Mallat, S. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.

Meyer, Y. (1990) *Ondelettes*, vol. I. Paris: Hermann.
Molina, R. and Ripley, B. D. (1989) Using spatial models as priors in astronomical image analysis. *J. Appl. Statist.*, **16**, 193–206.
O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, **1**, 502–527.
Pelz, W. (1977) Topics on the estimation of small probabilities. *Doctoral Thesis*. Virginia Tech, Blacksburg.
Ripley, B. D. (1991) The use of spatial models as image priors. *IMS Lect. Notes*, 309–340.
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. Inf. Theory*, **26**, 26–37.
Skilling, J. (1989) Classic maximum entropy. In *Maximum Entropy and Bayesian Methods* (ed. J. Skilling), pp. 45–52. Dordrecht: Kluwer.
Skilling, J., Robinson, D. R. T. and Gull, S. F. (1991) Probabilistic displays. In *Maximum Entropy and Bayesian Methods* (eds W. T. Grandy, Jr, and L. H. Schick), pp. 365–368. Dordrecht: Kluwer.
Tapia, R. A. and Thompson, J. R. (1978) *Nonparametric Probability Density Estimation*. Baltimore: Johns Hopkins University Press.
Thompson, A. M. (1988) On the use of quadratic regularisation within maximum entropy restoration. In *Maximum Entropy and Bayesian Methods* (ed. J. Skilling). Boston: Kluwer.
Thompson, A. M., Brown, J. C., Kay, J. W. and Titterington, D. M. (1990) *IEEE Trans. Pattn Anal. Mach. Intell.*, to be published.
Thompson, A. M. and Kay, J. W. (1990) On hyperparametric inversion strategies. To be published.
Tikochinsky, Y., Tishby, N. Z. and Levine, R. D. (1984a) Consistent inference of probabilities for reproducible experiments. *Phys. Rev. Lett.*, **52**, 1357–1360.
——— (1984b) Alternative approach to maximum-entropy inference. *Phys. Rev.* A, **30**, 2638–2644.
Titterington, D. M. (1984) The maximum entropy method for data analysis. *Nature*, **312**, 381–382.
——— (1985a) Common structure of smoothing techniques in Statistics. *Int. Statist. Rev.*, **53**, 141–170.
——— (1985b) General structure of regularization procedures in image processing. *Astron. Astrophys.*, **144**, 381–387.
Wahba, G. (1980) Ill posed problems: numerical and statistical methods for mildly, moderately and severely ill posed problems with noisy data. *Proc. Int. Conf. Ill Posed Problems* (ed. M. Z. Nashed).
——— (1982) Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine. In *Statistical Decision Theory and Related Topics, III* (eds S. Gupta and J. Berger), vol. 2, pp. 383–418. New York: Academic Press.
——— (1990) Spline models for observational data. *Reg. Conf. Ser. Appl. Math.*, **59**.