Eur Biophys J (1990) 18:165–174

European Biophysics Journal

Maximum entropy analysis of oversampled data problems

R. K. Bryan

Europäisches Laboratorium für Molekularbiologie, Meyerhofstrasse 1, D-6900 Heidelberg, Federal Republic of Germany

Received September 28, 1989/Accepted in revised form January 2, 1990

Abstract. An algorithm for the solution of the Maximum Entropy problem is presented, for use when the data are considerably oversampled, so that the amount of independent information they contain is very much less than the actual number of data points. The application of general purpose entropy maximisation methods is then comparatively inefficient. In this algorithm the independent variables are in the singular space of the transform between map (or image or spectrum) and data. These variables are much fewer in number than either the data or the reconstructed map, resulting in a fast and accurate algorithm. The speed of this algorithm makes feasible the incorporation of recent ideas in maximum entropy theory (Skilling 1989 a; Gull 1989). This algorithm is particularly appropriate for the exponential decay problem, solution scattering, fibre diffraction, and similar applications.

Key words: Maximum entropy – Inverse problem – Dynamic light scattering

1. Introduction

The use of Maximum Entropy for the analysis of data in many applications is rapidly increasing. Some problems, such as dynamic light scattering (DLS), have data of high accuracy, but oversampled, and for this sort of application a previous algorithm (Bryan 1980; Skilling and Bryan, 1984), designed for general image processing problems, such as deconvolution, converges disappointingly slowly. The original purpose of this work was to develop a numerical algorithm to solve such problems more efficiently, and the new algorithm is the subject of §2. The advantages are more rapidly convergence, and the ability to deal with data with a much greater signal to noise ratio (SNR) than hitherto. Recent work by Skilling (1989 a) and Gull (1989) has also embedded Maximum Entropy in a more general Bayesian framework. Bricogne (1988) develops a similar theory in the more limited context of noiseless crystallographic Fourier data. The advantages of the Bayesian approach include a firm axiomatic foundation for the entropy as the prior of the distribution of intensity over a positive, additive density (image, map or spectrum); a Bayesian estimate of the value of the Lagrange multiplier α, which balances the relative weights of the entropy and data constraint; and an estimate of the covariance of the solution. We start by briefly reviewing the parts of this theory relevant for the current work, and later discuss the extension of the theory required for our application. Unlike the image processing problem considered by Gull (1989), which has a large number of independent observations and a precisely-determined α, our problem results in a range of reasonably probable α. This distribution must be averaged over to estimate the spectrum.

Present address and address for offprint requests: Laboratory of Molecular Biophysics, South Parks Road, Oxford OX1 3QV, United Kingdom

1.1. Theory

An experimental dataset, represented by an N_d dimensional vector $\boldsymbol{D}$, is inevitably noisy, so the result of a data-processing method should be the probability distribution over possible images, represented by an N dimensional vector $\boldsymbol{f}$, which could give rise to the data, rather than a single 'solution'. Thus we wish to find $p_r(\boldsymbol{f}\,|\,\boldsymbol{D})$, which may be done by applying the product rule for probabilities in two ways to the joint distributon of $\boldsymbol{D}$ and $\boldsymbol{f}$,

$$p_r(\boldsymbol{f}, \boldsymbol{D}\,|\,I) = p_r(\boldsymbol{f}\,|\,\boldsymbol{D}, I)\, p_r(\boldsymbol{D}\,|\,I),$$
$$= p_r(\boldsymbol{D}\,|\,\boldsymbol{f}, I)\, p_r(\boldsymbol{f}\,|\,I),$$

which can be rearranged to give Bayes' theorem

$$p_r(\boldsymbol{f}\,|\,\boldsymbol{D}, I) = p_r(\boldsymbol{D}\,|\,\boldsymbol{f}, I)\, p_r(\boldsymbol{f}\,|\,I)/p_r(\boldsymbol{D}\,|\,I), \quad (1)$$

showing that the posterior distribution of $\boldsymbol{f}$ is the product of the prior of $\boldsymbol{f}$, $p_r(\boldsymbol{f}\,|\,I)$, and the probability of the data given the image, $p_r(\boldsymbol{D}\,|\,\boldsymbol{f}, I)$. I refers to all other information relevant to the problem, such as the response func-

tion of the experiment and the specification of coordinate ranges, and is implicit in all probabilities, even when not explicitly included. $\boldsymbol{D}$ is a measured dataset, so $p_r(\boldsymbol{D}\,|\,I)$ is simply a normalisation constant, and will be omitted from further equations.

When $p_r(\boldsymbol{D}\,|\,\boldsymbol{f}, I)$ is considered as a function of $\boldsymbol{f}$ for fixed $\boldsymbol{D}$ it is known as the likelihood, and is constructed to define the relation of the image to the measured data, by expressing the probability that noise makes up the difference between the 'mock data' that would be observed if $\boldsymbol{f}$ were the true image, and the actual data $\boldsymbol{D}$. In this paper, this relation is restricted to the functional form $p_r(\boldsymbol{D}\,|\,\boldsymbol{f}, I) = \exp - L(\boldsymbol{F}, \boldsymbol{D})$, where $\boldsymbol{F}$ is linearly related to $\boldsymbol{f}$, $\boldsymbol{F} = T\boldsymbol{f}$, but the relation between $\boldsymbol{F}$ and $\boldsymbol{D}$ is not further restricted. In fact, many experimental problems, including virtually all those to which maximum entropy has already been applied, allow the likelihood to be cast in this form. In particular, a linear problem with uncorrelated Gaussian noise has $L(\boldsymbol{F}, \boldsymbol{D}) = \frac{1}{2}\sum_k (F_k - D_k)^2/\sigma_k^2$, which may be recognised as 'χ^2'. In this case, $\boldsymbol{F}$ itself represents the 'mock data', but, for example in the cases of Fourier intensity or DLS data (§3.2), they depend on the square of $\boldsymbol{F}$. The common feature of the problems our numerical algorithm is intended to solve is that the rows of T are not linearly independent, and in this sense the data is called *oversampled*. That is not to say they are redundant; they all will contribute to the statistical accuracy of the solution.

Using the same axioms of coordinate invariance, subset independence and system independence that Shore and Johnson (1980) used to derive the principle of maximum entropy for a probability density, Skilling (1989a) shows that the prior $p_r(\boldsymbol{f}\,|\,\alpha, \boldsymbol{m})$ for an unnormalised positive additive density $\boldsymbol{f}$ is proportional to $\exp \alpha S$, where the entropy S of the required map $\boldsymbol{f}$ relative to an initial map $\boldsymbol{m}$ is given by

$$S = \sum_i f_i - m_i - f_i \log f_i/m_i .$$

α is initially undetermined, so the solution is conditional on two more quantities, α and $\boldsymbol{m}$. Defining the normalisation constants $N_f = \sum f$, $N_m = \sum m$, we note that this unnormalised form may be rewritten as

$$S = -N_f \sum_i p_i \log p_i/q_i + N_f - N_m - N_f \log(N_f/N_m),$$

where $\boldsymbol{p} = \boldsymbol{f}/N_f$ and $\boldsymbol{q} = \boldsymbol{m}/N_m$. The first term is a multiple of the entropy of the normalised densities as used before (e.g., Skilling and Bryan 1984; Livesey et al. 1986), contains all the information on their shapes, and is not influenced by the other terms, which simply measure the discrepancy in normalisation. If the normalisation N_f is known exactly, we must set $N_m = N_f$, and the multiplier N_f may be absorbed into α, giving exactly the same result as the normalised form. However, if the normalisation is to be estimated from the data, it will at best be subject to noise, or even unavailable should the necessary datum or combination of data be missing. Since, for consistency, there should be no dramatic change in the form of the solution for these various cases, the use of the unnormalised form is preferable. A further consequence of Skilling's derivation is that a measure $d^N f/\prod f^{1/2}$ should be applied to integrals of the probability over $\boldsymbol{f}$ space.

Inserting in (1) gives the posterior distribution

$$p_r(\boldsymbol{f}\,|\,\boldsymbol{D}, \alpha, \boldsymbol{m}) = \exp(\alpha S - L)/Z_S Z_L, \tag{2}$$

where the Z functions normalise the prior and likelihood respectively. Z_L depends only on the data, and $Z_S(\alpha) \propto \alpha^{-N/2}$ (Gull 1989). Hence, for given α, (2) attains its maximum at $\hat{\boldsymbol{f}}$ which maximises $Q = \alpha S - L$, or in other words, maximises the entropy subject to the conditioning provided by the data through L.

Previous implementations have assumed that α should be chosen such that $L = N_d/2$, equivalent to '$\chi^2 = N_d$' for Gaussian noise (Gull and Daniell 1978). This criterion generally tends to underfit the data. An improved way of finding α is to use Bayes' theorem on the joint distribution of $\boldsymbol{f}$ and α, which has been applied to quadratic regularisation problems (Turchin and Nozik 1969; Turchin et al. 1971), and more recently by Sibisi (1989). Gull (1989) has applied this method to the estimate of α in the entropic case, as follows Bayes' theorem gives

$$p_r(\boldsymbol{f}, \alpha\,|\,\boldsymbol{D}, \boldsymbol{m}) \propto p_r(\boldsymbol{D}\,|\,\boldsymbol{f}, \alpha, \boldsymbol{m})\, p_r(\boldsymbol{f}\,|\,\alpha, \boldsymbol{m})\, p_r(\alpha), \tag{3}$$

thus introducing $p_r(\alpha)$, the prior for α, which assume to be independent of $\boldsymbol{m}$. The posterior distribution of α is obtained by integrating the epression (3) over $\boldsymbol{f}$. First make the Gaussian approximation

$$\exp(\alpha S - L) \approx \exp(Q(\hat{\boldsymbol{f}}) + \tfrac{1}{2}\,\delta\boldsymbol{f}^T\,\nabla\nabla Q\,\delta\boldsymbol{f}),$$

where $\delta\boldsymbol{f} = \boldsymbol{f} - \hat{\boldsymbol{f}}$,

so that

$$p_r(\boldsymbol{f}, \alpha\,|\,\boldsymbol{D}, \boldsymbol{m}) \propto \alpha^{N/2} \exp(Q(\hat{\boldsymbol{f}}) + \tfrac{1}{2}\,\delta\boldsymbol{f}^T\,\nabla\nabla Q\,\delta\boldsymbol{f})\, p_r(\alpha), \tag{4}$$

then integrate over $\boldsymbol{f}$ using the $d^N f/\prod f^{1/2}$ measure to obtain (Gull 1989)

$$p_r(\alpha\,|\,\boldsymbol{D}, \boldsymbol{m}) \propto \prod_i \left(\frac{\alpha}{\alpha + \lambda_i}\right)^{1/2} \mathrm{e}^{\hat{Q}}\, p_r(\alpha), \tag{5}$$

where the $\{\lambda_i\}$ are the eigenvalues of $\mathrm{diag}\{f^{1/2}\}\,\nabla\nabla L\,\mathrm{diag}\{f^{1/2}\}$ evaluated at $\hat{\boldsymbol{f}}(\alpha)$. Also, the distribution of $\boldsymbol{f}$ at fixed α is, up to an unimportant normalisation constant,

$$p_r(\boldsymbol{f}, |\,\alpha, \boldsymbol{D}, \boldsymbol{m}) = \prod_i (\alpha + \lambda_i)^{1/2} \exp \tfrac{1}{2}\,\delta\boldsymbol{f}^T\,\nabla\nabla Q\,\delta\boldsymbol{f}. \tag{6}$$

Gull then shows that when the number of observations is large, (5) gives a sharp optimum for α, at $\hat{\alpha}$, say. Assuming that $d\lambda/d\alpha$ is negligible, and that any reasonable prior for α will be overwhelmed by the data, this leads, after rearrangement of the derivative $dp_r(\alpha\,|\,\boldsymbol{D})/d\alpha = 0$, to the condition

$$-2\hat{\alpha} S = \sum_i \frac{\lambda_i}{\hat{\alpha} + \lambda_i}.$$

Each eigenvalue λ_i which is significantly larger than $\hat{\alpha}$ contributes one to the right hand side, which is therefore a count of the number of good observations, N_g. The approximation $p_r(\alpha\,|\,\boldsymbol{D}, \boldsymbol{m}) = \delta(\alpha - \hat{\alpha})$ is made, and the posterior distribution of $\boldsymbol{f}$ is obtained as a Gaussian (6)

about $\hat{\boldsymbol{f}}(\hat{\alpha})$. As will be shown in a later section, neither assumption holds for the problems of interest here: – there is a wide distribution of reasonably probable α, and the $\{\lambda_i\}$ vary significantly over this range.

The Bayesian analysis gives rise to a set of maximum entropy maps $\hat{\boldsymbol{f}}$, parameterised by α, which are the modes of the probability distributions (6). We no longer interpret entropy maximisation as a way of selecting a single preferred map, as previously argued (Skilling and Bryan 1984; Livesey et al. 1986). Quantitative results must depend on the whole probability distribution, so displaying a single maximum entropy map should only be thought of as a way of representing its mode. Expectation values of $\boldsymbol{f}$ and other quantities are obtained by integration over the distribution (4) of $\boldsymbol{f}$ and α. However, we see that this distribution is characterised by the maximum entropy maps $\hat{\boldsymbol{f}}(\alpha)$, the eigenvalues $\{\lambda_i\}$, and the posterior probability (5) of α. Thus the calculation of a maximum entropy map is still central to the problem. In §2.1 and 2.2 the required numerical methods will be developed, and in §2.5 the integration over α will be discussed.

2. Numerical algorithm

The theory described in §1 requires $Q=\alpha S-L$ to be maximised. Skilling and Bryan (1984) presented an algorithm which has been used successfully in many applications (reviewed in Gull and Skilling 1984; some newer applications are contained in the volume edited by Skilling 1989). This algorithm works by finding an increment $\delta\boldsymbol{f}$ to $\boldsymbol{f}$ at each iteration, which is an approximate solution to the Newton equations $\delta\boldsymbol{f}=-(\nabla\nabla Q)^{-1}\nabla Q$. For problems of any useful size, this $N\times N$ set of equations is too big to solve directly. In Skilling and Bryan (1984) the approximate solution is found by (conceptually) using the binomial theorem on $(\nabla\nabla Q)^{-1}=(\alpha\nabla\nabla S-\nabla\nabla L)^{-1}$. Since $\nabla\nabla S^{-1}=-\mathrm{diag}\{f\}$, this expansion gives $(\nabla\nabla Q)^{-1}$ as a power series in the operator $\mathrm{diag}\{f\}\nabla\nabla L$, so that $\delta\boldsymbol{f}$ may be approximated by a linear combination of the 'search directions' $\mathrm{diag}\{f\}\nabla S$, $\mathrm{diag}\{f\}\nabla L$, and powers of $\mathrm{diag}\{f\}\nabla\nabla L$ acting on these two initial vectors. Provided that the noise on the data is uncorrelated, acting on a vector by $\nabla\nabla L$ may be performed by a combination of vector operations and map-data space transformations, T, so all these operations are possible, even on very large data and map arrays. This algorithm works well when the Hessian matrix $\nabla\nabla Q$ is roughly isotropic, so the directions of $\delta\boldsymbol{f}$ and $\mathrm{diag}\{f\}\nabla Q$ are not too different, and only a few applications of $\mathrm{diag}\{f\}\nabla\nabla L$ give good search directions. Depending on the size and complexity of the problem, between three and ten directions have been used.

Since

$$\nabla L=\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{f}}\frac{\partial L(\boldsymbol{F},\boldsymbol{D})}{\partial \boldsymbol{F}}=T^T\frac{\partial L(\boldsymbol{F},\boldsymbol{D})}{\partial \boldsymbol{F}}, \tag{7}$$

it is seen that ∇L lies in the column space ot T^T. Suppose the singular value decomposition (SVD) of T is $T=V\Sigma U^T$, where V is an $N_d\times N_d$ orthogonal matrix, U an $N\times N$ orthogonal matrix, and the $N_d\times N$ matrix Σ is zero except for the elements $\Sigma_{ii}=\sigma_i$, $i=1,\ldots,s$. The σ_i, conventionally ordered $\sigma_1\geq\sigma_2\geq\ldots\geq\sigma_s>0$, are the singular values of T, and $s=\mathrm{rank}\,(T)$. The column space of T^T is therefore the same as the space spanned by the columns of U associated with the non-zero singular values. This s-dimensional space will be referred to as the *singular space* for convenience. Let U_s denote the $s\times N$ matrix formed by the first s columns of U. Thus $\mathrm{diag}\{f\}\nabla L$ and all the search directions formed by acting with $\mathrm{diag}\{f\}\nabla\nabla L$ will lie in the space spanned by the columns of $\mathrm{diag}\{f\}U_s$. Indeed, $\mathrm{diag}\{f\}\nabla S$ is the only search direction *not* in that space. Thus, irrespective of the number of directions generated, this algorithm searches in an at most $(s+1)$-dimensional subspace of the N-dimensional $\boldsymbol{f}$ space. Now $\nabla\nabla Q$ is reasonably isotropic if both $\nabla\nabla S$ and $\nabla\nabla L$ are, which mean respectively that the dynamic range of the reconstruction must not be too large, and that the non-zero singular values of T must all be of the same order of magnitude. An example of an applications with these characteristics is critically sampled Fourier transform data, where $\Sigma=V=I$, $U=\mathfrak{F}$. On the other hand, problems such as the exponential-decay problem have singular values which decrease exponentially with index. Thus the generated search directions are dominated by $\mathrm{diag}\{f\}U_1$, the search is confined to a two-dimensional space, and convergence is correspondingly slow. A high SNR exacerbates this problem.

2.1. *Development of the algorithm*

To begin development the new algorithm, the condition for the maximum, $\alpha\nabla S-\nabla L=0$, is written as

$$-\alpha\log f_i/m_i=\sum_j T_{ji}\frac{\partial L(\boldsymbol{F},\boldsymbol{D})}{\partial F_j}. \tag{8}$$

The solution could therefore be represented in terms of a new variable $\boldsymbol{u}$, where $\log f/m=T^T\boldsymbol{u}$ (Pollard 1988). However, unless T is of full rank, the components of $\boldsymbol{u}$ will not be independent, and care must be taken to achieve a stable algorithm. Exploiting the fact that ∇L lies in the singular space, it is better to use the representation $\log f/m=U_s\boldsymbol{u}$ for some s-dimensional vector $\boldsymbol{u}$. Then

$$f_i=m_i\exp\sum_{t=1}^{s}U_{it}u_t, \tag{9}$$

and we see that all possible $\boldsymbol{f}$ which may arise from datasets produced by the response function T may be represented in this way. Therefore we can reduce the problem from an N-dimensional to an s-dimensional optimisation. If $s<N_d$, then the dataset is oversampled; without noise, some items of data could be predicted from the rest. The algorithm presented here is particularly efficient if $s\ll N_d$. In §3 example results are shown where the map and data spaces may each have a dimension of hundreds, but the singular space a few tens at most.

In the following, all relevant matrix and vector quantities will be assumed to be restricted to the s-dimensional singular space, so that the diagonal components of Σ are

all non-zero, unless otherwise stated. Writing ∇S in terms of $U\,\boldsymbol{u}$, (8) becomes

$$-\alpha\, U\,\boldsymbol{u}=U\,\Sigma\,V^T\frac{\partial L(\boldsymbol{F},\boldsymbol{D})}{\partial \boldsymbol{F}},$$

and hence, since the columns of U are orthogonal,

$$-\alpha\,\boldsymbol{u}=\Sigma\,V^T\frac{\partial L(\boldsymbol{F},\boldsymbol{D})}{\partial \boldsymbol{F}}=\boldsymbol{g},\quad \text{say,} \tag{10}$$

thus defining the entropy maximum for given α in terms of $\boldsymbol{u}$, where $\boldsymbol{g}$ is a function of $\boldsymbol{u}$ through (9) and $\boldsymbol{F}=T\boldsymbol{f}$. $\boldsymbol{F}$ may be computed efficiently by successive application of U^T, Σ and V to $\boldsymbol{f}$, which takes $s(N+1+N_d)$ operations, as opposed to the $N\,N_d$ needed for the direct application of T.

Although it is tempting to try to solve (10) by fixed-point iteration (i.e., starting from some trial value of $\boldsymbol{u}$, successively calculating $\boldsymbol{g}$ from the current value of $\boldsymbol{u}$, and resetting $\boldsymbol{u}=-\boldsymbol{g}/\alpha$), analogous to the method of Gull and Daniell (1978), to do so is doomed to failure, as such a procedure will only converge if the modulus of the largest singular value of $\partial\boldsymbol{g}/\partial\boldsymbol{u}$ is less than α. Straightforward damping helps only if these non-zero singular values are all of the same magnitude, which is certainly not the case considered here. However, a Newton method can be applied to (10), the increment at each iteration being given by $J\,\delta\boldsymbol{u}=-\alpha\,\boldsymbol{u}-\boldsymbol{g}$, where $J=\alpha I+\partial\boldsymbol{g}/\partial\boldsymbol{u}$ is the Jacobian of the system. Now

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}=\Sigma\,V^T\frac{\partial^2 L(\boldsymbol{F},\boldsymbol{D})}{\partial \boldsymbol{F}^2}\frac{\partial \boldsymbol{F}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}},$$

and

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}=\operatorname{diag}\{f\}\,U,$$

so

$$\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}=\Sigma\,V^T\,WV\Sigma\,U^T\operatorname{diag}\{f\}\,U=M\,K,$$

where for convenience we have made the definitions $W=\partial^2 L(\boldsymbol{F},\boldsymbol{D})/\partial\boldsymbol{F}^2$, $M=\Sigma\,V^T\,WV\Sigma$ and $K=U^T\operatorname{diag}\cdot\{f\}\,U$. M and K are both symmetric $s\times s$ matrices. If the noise on the data samples is independent, W is diagonal. $\partial\boldsymbol{u}$ thus solves

$$(\alpha I+M\,K)\,\delta\boldsymbol{u}=-\alpha\,\boldsymbol{u}-\boldsymbol{g}. \tag{11}$$

At each iteration the size of the increment $\delta\boldsymbol{u}$ must be restricted so that the second-order approximation used in (11) remains accurate. Previously, (Skilling and Bryan 1984), $-\nabla\nabla S=\operatorname{diag}\{1/f\}$ was used as a metric in $\boldsymbol{f}$ space, a better justification being given by Skilling (1989 a). Now

$$\delta\boldsymbol{f}^T\operatorname{diag}\{1/f\}\,\delta\boldsymbol{f}=\delta\boldsymbol{u}^T\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\operatorname{diag}\{1/f\}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{u}}\delta\boldsymbol{u},$$

$$=\delta\boldsymbol{u}^T\,K\,\delta\boldsymbol{u},$$

so K is the equivalent metric in $\boldsymbol{u}$ space. The step-length restriction is achieved by augmenting J with a multiple of the unit matrix (Levenberg 1944; Marquardt 1963), giving

$$((\alpha+\mu)\,I+M\,K)\,\delta\boldsymbol{u}=-\alpha\,\boldsymbol{u}-\boldsymbol{g}, \tag{12}$$

where μ is chosen so that $\delta\boldsymbol{u}^T\,K\,\delta\boldsymbol{u}\le O(\sum m)$.

As in Skilling and Bryan (1984), the values of α and μ are adjusted so that the iteration proceeds, depending on requirements, either towards the maximum probability (5) or towards the maximum of Q for a specific value of α, whilst imposing the step-length constraint. This search may be made efficiently if (12) is diagonalised, so that only $O(s)$ operations are required for each trial α-μ pair, rather than $O(s^3)$ if (12) is solved directly. First diagonalise K by solving the eigenproblem

$$KP=P\Xi,\quad \Xi=\operatorname{diag}\{\xi\},\quad P^T\,P=I \tag{13}$$

and then define

$$A=\operatorname{diag}\{\xi^{1/2}\}\,P^T\,M\,P\,\operatorname{diag}\{\xi^{1/2}\}.$$

Solving the further eigenproblem

$$AR=R\Lambda,\quad \Lambda=\operatorname{diag}\{\lambda\},\quad R^T\,R=I, \tag{14}$$

and defining

$$Y=P\operatorname{diag}\{\xi^{-1/2}\}\,R, \tag{15}$$

the following relations hold:

$$K=Y^{-T}\,Y^{-1},$$

$$\Lambda=Y^{-1}\,M\,Y^{-T},$$

and the columns of Y are the eigenvectors of $K\,M\,K$ with respect to K. Hence (12) becomes

$$((\alpha+\mu)\,I+\Lambda)\,Y^{-1}\,\delta\boldsymbol{u}=-\alpha\,Y^{-1}\,\boldsymbol{u}-Y^{-1}\,\boldsymbol{g}, \tag{16}$$

giving s *independent* equations for the components of $Y^{-1}\,\delta\boldsymbol{u}$, and the step length

$$\delta\boldsymbol{u}^T\,K\,\delta\boldsymbol{u}=|Y^{-1}\,\delta\boldsymbol{u}|^2.$$

In practical calculations, if $\boldsymbol{f}$ has a high dynamic range, K may be close to singularity, so that some ξ are approximately zero, and $\delta\boldsymbol{u}$ cannot be found by applying Y to the $Y^{-1}\,\delta\boldsymbol{u}$ calculated by (16), since (15) shows that Y contains a $\xi^{-1/2}$ term. Nevertheless, $Y^{-1}\,\delta\boldsymbol{u}$ itself is correctly determined. To investigate how $\delta\boldsymbol{u}$ can be found, P, the matrix of eigenvectors of K, may be partitioned into those vectors associated with zero eigenvalue ξ (the null space), designated by a subscript 0, and the non-null space (subscript 1). Hence

$$K=P\,\Xi\,P^T=(P_0\quad P_1)\begin{pmatrix}0 & 0\\ 0 & \Xi_1\end{pmatrix}\begin{pmatrix}P_0^T\\ P_1^T\end{pmatrix}.$$

Writing $\boldsymbol{p}=P^T\boldsymbol{u}$, multiplying (12) on the left by P^T and partitioning, we obtain

$$(\alpha+\mu)\begin{pmatrix}\delta\boldsymbol{p}_0\\ \delta\boldsymbol{p}_1\end{pmatrix}+\begin{pmatrix}P_0^T\\ P_1^T\end{pmatrix}M(P_0\quad P_1)\begin{pmatrix}0 & 0\\ 0 & \Xi_1\end{pmatrix}\begin{pmatrix}\delta\boldsymbol{p}_0\\ \delta\boldsymbol{p}_1\end{pmatrix}$$
$$=-\alpha\boldsymbol{p}-P^T\boldsymbol{g},$$

so that

$$(\alpha+\mu)\,\delta\boldsymbol{p}_0+P_0^T\,M\,P_1\,\Xi_1\,\delta\boldsymbol{p}_1=-\alpha\boldsymbol{p}_0-P_0^T\boldsymbol{g}, \tag{17}$$

$$(\alpha+\mu)\,\delta\boldsymbol{p}_1+P_1^T\,M\,P_1\,\Xi_1\,\delta\boldsymbol{p}_1=-\alpha\boldsymbol{p}_1-P_1^T\boldsymbol{g}, \tag{18}$$

and

$$\delta \boldsymbol{u}^T K\, \delta \boldsymbol{u} = \delta \boldsymbol{p}^T \Xi\, \delta \boldsymbol{p} = \delta \boldsymbol{p}_1^T \Xi_1\, \delta \boldsymbol{p}_1 . \tag{19}$$

The P_1 space Eqs. (18) are of the same form as (12), and may be solved in the same way. The null-space increment, $\delta \boldsymbol{p}_0$, is now directly given by (17) in terms of $\delta \boldsymbol{p}_1$, and does not contribute to the step-length estimate (19). However, this still leaves the numerical problem of deciding when an eigenvalue of K is sufficiently small for it to be treated as zero. Fortunately, (12) can be rewritten as

$$\begin{aligned}(\alpha+\mu)\,\delta \boldsymbol{u} &= -\alpha \boldsymbol{u} - \boldsymbol{g} - M K\, \delta \boldsymbol{u}, \\ &= -\alpha \boldsymbol{u} - \boldsymbol{g} - M Y^{-T} Y^{-1}\, \delta \boldsymbol{u},\end{aligned} \tag{20}$$

so finally, (16) is solved for $Y^{-1}\, \delta \boldsymbol{u}$, the result inserted in (20), and the true $\delta \boldsymbol{u}$ recovered. For the $\delta \boldsymbol{p}_1$ components, (20) is an identity; for the $\delta \boldsymbol{p}_0$, it is the same as the partitioned equation (17); the Y^{-T} on the right eliminates any contribution from its null space. The two spaces merge smoothly. Only multiplication by Y^{-1} is required, never by Y itself, and since $Y^{-1} = R^T \operatorname{diag}\{\xi^{1/2}\} P^T$, the calculations are always stable, requiring no division by small eigenvalues. Furthermore, the exponential representation (9) of $\boldsymbol{f}$ avoids the need to protect the algorithm against points going negative during the calculation.

The cutoff on singular values is not critical, provided sufficient vectors are included to span the row space of T; including additional columns of U which correspond to zero singular values simply results in the extra components of $\boldsymbol{u}$ being calculated as zero. Similar algorithms may be developed for any representation in terms of a matrix U, whose columns form a set of linearly independent basis functions. Orthogonality, as produced here by the SVD, is not strictly necessary; linear independence is sufficient, as used in a specialised application by Skilling (1989 b).

Linear inverse problems have been analysed in terms of SVDs before (e.g., Hanson 1971). The restored map is represented as a linear combination of the singular vectors associated with large singular values, and hence well determined by the data. In the simplest form, making no allowance for noise, this method uses the Moore-Penrose inverse of Σ, defined by $\Sigma^+ = \operatorname{diag}\{\sigma_1^{-1}, \ldots \sigma_s^{-1}, 0, \ldots, 0\}$, so that $\boldsymbol{f} = \mathrm{U} \Sigma^+ V^T \boldsymbol{F}$. More generally, an SVD analysis has been used in combination with quadratic regularisation (e.g., Strand 1974), providing a smooth roll-off of the singular vectors, of the typical form

$$\boldsymbol{f} = U \operatorname{diag}\left\{\frac{\sigma_i}{\sigma_i^2+\alpha}\right\} V^T \boldsymbol{F}.$$

Indeed, such a method has been applied to the exponential decay problem (Provencher 1982), although additional inequality constraints are also needed to ensure a positive spectrum, leading to increased algebraic complexity. Maximum entropy may be considered to be a method using a non-quadratic regularising functional, and the solution (9) is determined by the same singular vectors; but it is more convenient to think of their coefficients, $\boldsymbol{u}$, as Lagrange multipliers constraining $\boldsymbol{f}$ to fit the data. Furthermore there is no restriction to problems with a completely linear relation between data and reconstruction.

2.2. Eigenvalues

The non-zero eigenvalues of $\operatorname{diag}\{f^{1/2}\} \nabla\nabla L \operatorname{diag}\{f^{1/2}\}$ are required for the computation of the probabilites (5, 6). This is an eigenproblem in N-dimensional $\boldsymbol{f}$-space, but we show now by a careful change of basis that the eigenvalues are the same as for the singular space eigenproblem (14). Assume for this section only that U, K, etc, stand for the quantities in the full space, and those with subscript s for those restricted to the singular space (i.e., the unsubscripted ones of the previous section). The eigenproblem is

$$\operatorname{diag}\{f^{1/2}\}\, U M U^T \operatorname{diag}\{f^{1/2}\}\, X = X \Lambda,$$

which, after pre-multiplying by $U^T \operatorname{diag}\{f^{1/2}\}$, can be manipulated to

$$K M K U^T \operatorname{diag}\{f^{-1/2}\}\, X = K U^T \operatorname{diag}\{f^{-1/2}\}\, X \Lambda. \tag{21}$$

K is positive definite, so its Cholesky decomposition $K = C C^T$, where C is lower triangular and non-singular, exists. Premultiplying (21) by C^{-1} gives the eigenproblem

$$C^T M C Z = Z \Lambda, \quad \text{where} \quad Z = C^T U^T \operatorname{diag}\{f^{-1/2}\}\, X.$$

Since the only non-zero elements of Σ are the first s elements of the diagonal, it is clear that only the leading $s \times s$ submatrix of M is non-zero, which we denote by M_s. Together with the pattern of zeros in C, this means that

$$C^T M C = \begin{pmatrix} C_s^T M_s C_s & 0 \\ 0 & 0 \end{pmatrix}$$

where C_s is the leading $s \times s$ submatrix of C, and the eigenproblem may be written as

$$\begin{pmatrix} C_s^T M_s C_s & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Z_s & 0 \\ 0 & \Phi \end{pmatrix} = \begin{pmatrix} Z_s & 0 \\ 0 & \Phi \end{pmatrix} \begin{pmatrix} \Lambda_s & 0 \\ 0 & 0 \end{pmatrix},$$

where new s-subscripted quantities have obvious meanings, and Φ, which forms a basis for the null space, is any $(N-s)$-dimensional orthogonal matrix. C performs a change of basis which diagonalises K whilst preserving the singular space. Since, in this new basis, all the eigenvectors except those in the singular space have zero eigenvalues, all the non-zero eigenvalues may be obtained from the eigenproblem restricted to the singular space. Clearly

$$C_s C_s^T = K_s = U_s^T \operatorname{diag}\{f\}\, U_s,$$

and the non-zero eigenvalues are given by the equivalent problem

$$K_s M_s K_s Y_s = K_s Y_s \Lambda_s, \qquad Y_s = C_s^{-T} Z_s, \tag{22}$$

where K_s and M_s are identical with K and M of the previous section. The solution of this s-dimensional eigenproblem thus provides all the information for estimates of both the increment (16, 20) and $p_r(\boldsymbol{f}, \alpha \mid \boldsymbol{D}, \boldsymbol{m})$, (4).

2.3. Convergence test

To test for convergence, the magnitude of the vector difference of $\alpha\, \partial S/\delta \boldsymbol{u}$ and $\partial L/\partial \boldsymbol{u}$ is compared with the sum of

their magnitudes, using

$$t=2\left|\alpha\frac{\partial S}{\partial \boldsymbol{u}}-\frac{\partial L}{\partial \boldsymbol{u}}\right|^2\Big/\left(\left|\alpha\frac{\partial S}{\partial \boldsymbol{u}}\right|+\left|\frac{\partial L}{\partial \boldsymbol{u}}\right|\right)^2,$$

again evaluated using the K metric (in fact, K^{-1}, as the gradients are covariant). Since $\partial S/\partial \boldsymbol{u}=-K\boldsymbol{u}$ and $\partial L/\partial \boldsymbol{u}=K\boldsymbol{g}$, these quantities are easily evaluated. This test checks for equal lengths of the gradients, and not only for parallelism, so is stricter than one used previously (Skilling and Bryan 1984), and, when zero, confirms that the maximum of $Q(\boldsymbol{f}\,|\,\alpha)$ is attained. All the results in §3 have $t\leq 10^{-4}$.

2.4. *Covariance matrix*

A further result of the analysis of Gull (1989) and Skilling (1989 b) is the covariance matrix of $\boldsymbol{f}(\alpha)$, which in the Gaussian approximation (6) is $-(\nabla\nabla Q)^{-1}$. Similar algebra to that of Skilling (1989 b) enables one to obtain an expression in terms of singular-space quantities. Thus, after some manipulation,

$$\begin{aligned}-\nabla\nabla Q&=-\alpha\nabla\nabla S+\nabla\nabla L,\\&=\operatorname{diag}\{1/f\}\,U\,Y^{-T}(\alpha I+\Lambda)\,Y^{-1}\,U^T\operatorname{diag}\{1/f\},\end{aligned}$$

so

$$\begin{aligned}-(\nabla\nabla Q)^{-1}&=\operatorname{diag}\{f\}\,U\,Y\operatorname{diag}\left\{\frac{1}{\alpha+\lambda}\right\}Y^T\,U^T\operatorname{diag}\{f\},\\&=\frac{1}{\alpha}\operatorname{diag}\{f\}-\operatorname{diag}\{f\}\,U\,Y\operatorname{diag}\left\{\frac{\lambda}{\alpha(\alpha+\lambda)}\right\}Y^T\,U^T\operatorname{diag}\{f\},\end{aligned}\tag{23}$$

and, using a similar triangularisation argument as in §2.2, the matrices in the second term can again be restricted to the singular space.

2.5. *Alpha*

The expression for the posterior probability of α derived by Gull (1989),

$$\log p_r(\alpha\,|\,\boldsymbol{D},\boldsymbol{m})=\frac{1}{2}\sum_i\log\left(\frac{\alpha}{\alpha+\lambda_i}\right)+\hat{Q}+\log p_r(\alpha),\tag{24}$$

has its maximum where the derivative (conveniently written in $\log\alpha$ space)

$$\frac{d\log p_r(\alpha\,|\,\boldsymbol{D},\boldsymbol{m})}{d\log\alpha}=\frac{1}{2}N_g-\frac{\alpha}{2}\sum_i\left(\frac{d\lambda_i/d\alpha}{\alpha+\lambda_i}\right)+\alpha\hat{S}+\frac{d\log p_r(\alpha)}{d\log\alpha}\tag{25}$$

is zero. Gull's '$-2\alpha S=N_g$' criterion is obtained if $d\lambda_i/d\alpha$ and the prior may be ignored in comparison with N_g and $\alpha\hat{S}$. For notational convenience, we denote the posterior of α without the prior by $p_q(\alpha)$. If N_g is low, $p_q(\alpha)$ can have a broad peak, and the position of the optimum can be affected both by the prior on α and by fluctuations in the values of the $\{\lambda_i\}$. Since α is a scale factor, the Jeffreys prior $p_r(\alpha)=1/\alpha$ is appropriate (Jaynes 1968; Gull 1989), which effectively reduces N_g by two, and thus causes the data to be fitted rather more closely. The examples of §3 demonstrate these effects. In general, the derivative is dominated by the $\frac{1}{2}N_g$ term when $\alpha\ll\hat{\alpha}$, and by the $\alpha\hat{S}$ term when $\alpha\gg\hat{\alpha}$. It can be seen that the error introduced is very much less than that of assuming $2L=N_d$ to be the correct criterion, although the $d\lambda/d\alpha$ terms shift the peak significantly. However, the width of the peak means that for a correct estimate of the expectation of $\boldsymbol{f}$, the average over the distribution of α must be taken. The strategy we have adopted is

1. Locate $\hat{\alpha}$ approximately, using the $-2\alpha S=N_g$ criterion. If required, $\hat{\alpha}$ may be obtained more accurately by performing a search for the true maximum of (24).
2. Estimate $d^2\log p_q(\alpha)/d\log\alpha^2\approx\alpha S$ to get the approximate width of the peak of $p_q(\alpha)$.
3. Using the above approximation as a first step, perform a crude search in α for the upper limit of integration, $\alpha_{\max}$, such that $p_q(\alpha_{\max})<c\,p_q(\hat{\alpha})$. Typically $c=\exp(-8)$.
4. Perform $\int\phi(\alpha)\,p_q(\alpha)\,d\log\alpha$ numerically, using equal steps in $\log\alpha$ until a similar lower cutoff, $\alpha_{\min}$ is attained.

The integrated functions $\phi(\alpha)$ are those required for normalisation of the probability, estimation of $\bar{\alpha}$, $\bar{\boldsymbol{f}}$, etc. Note that the prior on α now comes in via the measure in the integral, so that the integrand contains $p_q(\alpha)$. Also, using (5), (6) and the product rule,

$$\begin{aligned}\bar{\boldsymbol{f}}&=\int\boldsymbol{f}\,p_r(\boldsymbol{f},\alpha\,|\,\boldsymbol{D},\boldsymbol{m})\,\mathrm{d}^N f\prod f^{-1/2}\,d\alpha,\\&\propto\int\hat{\boldsymbol{f}}(\alpha)\,p_q(\alpha)\,d\log\alpha,\end{aligned}$$

so $\bar{\boldsymbol{f}}$ is found by averaging the maximum entropy maps $\hat{\boldsymbol{f}}$ found for each α. $\bar{\boldsymbol{f}}$ itself is not a maximum entropy map for any α. For α significantly less than $\hat{\alpha}$, $d\log p(\alpha)/d\log\alpha\approx\frac{1}{2}N_g$, so $p_q(\alpha)$ decreases with a power law as $\alpha\to 0$, and the truncation of the integral can be estimated as

$$\begin{aligned}\int_0^{\alpha_{\min}}p_r(\alpha\,|\,\boldsymbol{D},\boldsymbol{m})\,d\alpha&\approx\int_{-\infty}^{\log\alpha_{\min}}p_q(\alpha_{\min})\,(\alpha/\alpha_{\min})^{N_g/2}\,d\log\alpha,\\&=\frac{2}{N_g}\,p_q(\alpha_{\min}),\end{aligned}$$

which is negligible with the chosen $\alpha_{\min}$. The integral for $\alpha\to\infty$ is improper (Gull 1989), but in order to take the data into account at all some upper limit for α must be imposed, and then the contribution to the integral is almost entirely from the range of α around $\hat{\alpha}$.

3. Applications

A key aspect of this algorithm is the SVD of the matrix T, which is defined by the geometry of the problem: the coordinates of the data-points; the domain of $\boldsymbol{f}$-space selected for reconstruction; and the functional form of the relation between map and data, which is assumed known.

All this information is implicitly included in the 'other information', which all the probabilities are conditional on, and is fixed for the problem. The decomposition can therefore be performed once and for all the beginning of the calculation, and thereafter only the components of U and V associated with non-zero singular values are required. In some cases, such as deconvolution, an analytic approach is appropriate, which is also possible for the exponential decay problem if the data points are at exponentially increasing times (Provencher 1976). However, practical instruments rarely collect data in such an ideal pattern, so in general the SVD must be done numerically. We have found that the Lanczos algorithm to reduce T to bidiagonal form, followed by the implicit QR algorithm (Golub and van Loan 1983) is very reliable, provided that the Lanczos vectors are re-orthogonalised, and that care is taken if there is a multiple singular value. Calculating the 18 significant singular vectors for the 100×140 exponential decay problem of § 3.2 takes only 15 s cpu time on a VAXstation 2000, compared with about 3 min if the bidiagonalisation is done by Householder transforms. Otherwise, the main computational cost is in the evaluation of the M and K matrices each iteration, the other scalar and singular space calculations being small in comparison. Thus the computational cost per iteration scales as $s^2(N+N_d)$. Moreover, if L is quadratic in $\boldsymbol{F}$, M is constant, and a further computational saving may be made.

The number of iterations required depends mostly on the SNR; better data require more iterations. The example of § 3.2 has a high SNR of approximtely 4000, and 41 iterations are required to obtain the $-2\alpha S = N_g$ solution, using 113 seconds of VAXstation 2000 time. When integrating over α, if the intervals in α are suitably small, only 2–3 iterations are needed to update the previous $\hat{\boldsymbol{f}}(\alpha)$. Thus, evaluating the probabilities at intervals of 0.025 in $\log_{10}\alpha$ to produce the final result takes some 450 iterations in all, or 22 min total. The previous algorithm (Skilling and Bryan 1984), required 172 iterations in 846 s just to reach the $2L = N_d$ solution, and was insufficiently powerful to converge when a closer fit was desired.

The number of singular values depends strongly on the coverage of data space by the data samples, and the range selected in map space for the reconstruction. The selection of the map space range might be achieved from other experimental information on the system being studied, or by considering the regime for which the data could possibly give information. For example, in the exponential decay problem it would be senseless to try to reconstruct the spectrum at such short decay times that the signal would have decayed completely before the first time sample, or at such large decay times that no significant decay would have taken place over the entire data record. The grid interval in map space must also be chosen to be sufficiently small to represent the spectrum; for Fourier problems this is easily calculated from the maximum frequency measured. Otherwise, $\boldsymbol{f}$ is used purely as an intermediate in the calculation, and using a finer grid does not alter the number of significant singular values. To demonstration the use of this algorithm, two problems are solved, one where the non-zero singular values are all nearly equal, and one where they decay rapidly.

3.1. *Oversampled Fourier problem*

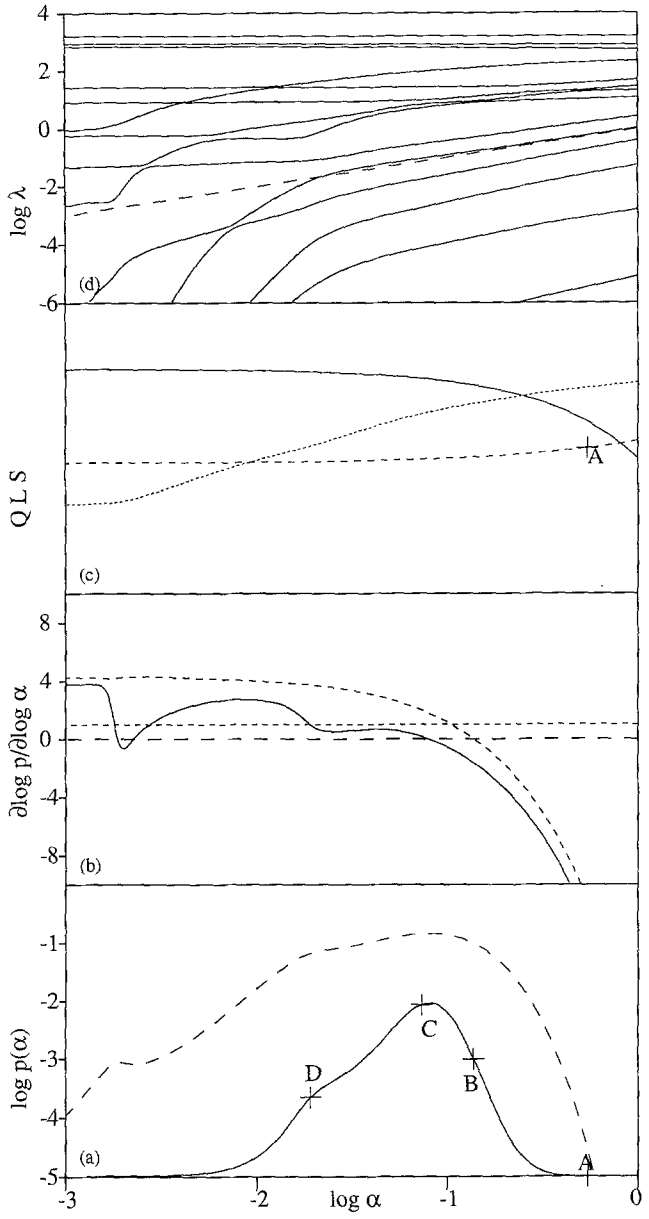
Many scattering experiments produce data as a function of scattering angle, and are thus related to the object of interest by a Fourier transform. Special geometries, such as cylindrical or spherical symmetry, give rise to analogous problems, as in fibre diffraction and solution scattering. A continuous distribution of scattered intensity may be measured, as distinct from the crystallographic problem where discrete reflections are recorded (and $V = \Sigma = I$). The SVD of the Fourier transform relating an image of restricted range to the data may be analysed in terms of prolate spheroidal wavefunctions (PSWs) (Slepian and Pollak 1961; Slepian 1964), but it is usually just as convenient to perform the SVD by linear algebra techniques as to evaluate the PSWs themselves. The results obtained in numerical tests agree closely, as do the singular value spectra. In fact, the SVD may be interpreted as a sampling formula: U^T transforms the map to the sample points in Fourier space; Σ is a set of weights; V interpolates from the sample points to the Fourier-space data points. Shannon sampling is the special case when the range in Fourier space is infinite. Practical sampling schemes, with a finite range, have a non-uniform distribution of sample points, with a rather greater density of samples near the domain boundaries, and always more than the number of Shannon samples on the same interval.

As a simple illustration, a reconstruction from cosine transform data is presented, with $T_{jk} = \cos \pi j \, \delta k \, \Delta$, $j = 1, \ldots, N$, $k = 0, \ldots, N_d - 1$. The reconstruction is thus made on the interval $0 \le x \le X = N\delta$, and data are provided at frequencies $k\Delta$; the algorithm is not, however, restricted to use on equal-interval data sampling like this. The real-space sampling, δ, must be selected to give adequate sampling for the highest frequency signal in the data, $N_d \Delta \delta \ll 1$. For comparison, the Shannon sampling limit is $\Delta = 1/X$. In our example, we take $X = 1$, oversample five times, so $\Delta = 1/5$, and provide data for $N_d = 50$ samples. Hence we require $\delta \ll 1/10$, so we take $N = 100$, $\delta = 1/100$. The SVD analysis gives the singular values in Table 1; the first 10 are of the same magnitude, and then there is a rapid drop off to zero, as expected (Slepian and Pollak 1961).

The results are shown in Fig. 1. Data are constructed from a simulated object consisting of three δ-functions, heights 10, 5 and 5, with noise added, standard deviation 2% of the maximum data value. The fluctuation of the eigenvalues (particularly those with values near α) cause the true derivative of $\log p_q(\alpha)$ to depart considerably from the $\frac{1}{2} N_g + \alpha S$ approximation, shifting the position of

Table 1. Singular values of the example cosine transform

No.	Value.	No.	Value.	No.	Value.
1	17.3682	7	15.8114	13	0.1993
2	16.4758	8	15.8110	14	0.0196
3	15.8114	9	15.7796	15	0.0016
4	15.8114	10	14.7091	16	0.0001
5	15.8114	11	7.6513	>17	Negligible
6	15.8114	12	1.5621		



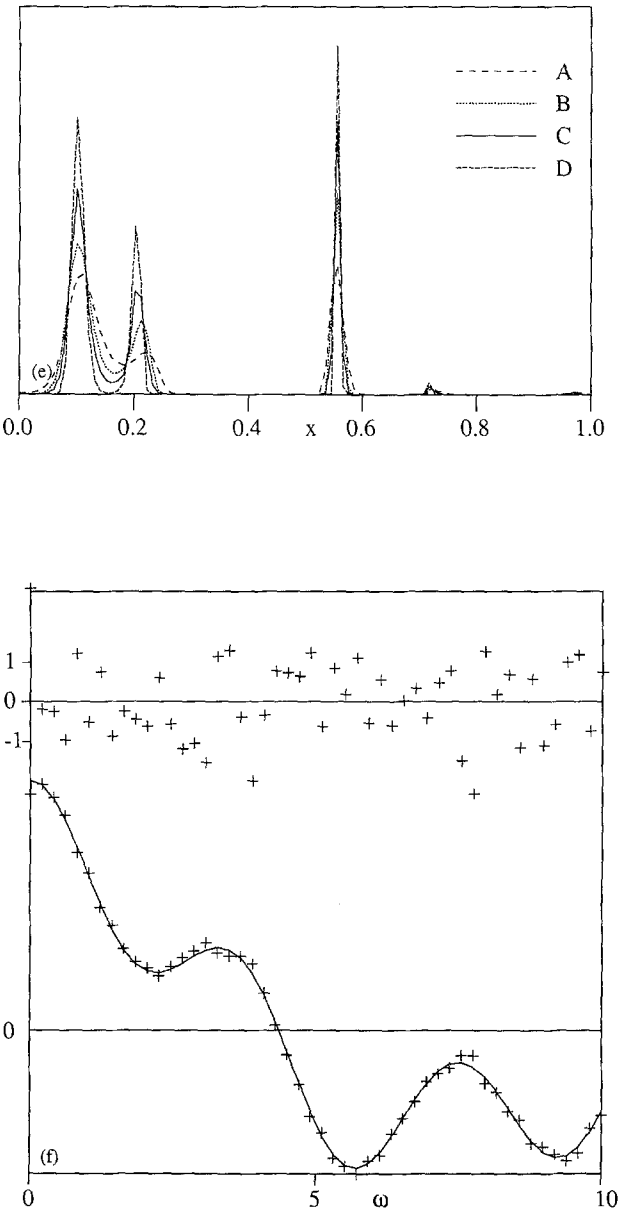


Fig. 1 a–f. Plots of quantities calculated for the cosine problem. **a** Posterior probability of α (continuous) and its log (base 10) (dashed), as functions of $\log_{10}\alpha$, evaluated pointwise with a flat prior (the integrated area is thus correct as if a Jeffreys prior were used). The crosses are at the positions given by A: $2L=N_{\text{data}}$, B: $-2\alpha S=N_g$, C: $\bar{\alpha}$, D: maximum of probability evaluated pointwise with a Jeffreys prior. **b** The numerical derivative of $\log p_q(\alpha)$ (solid), $0.5\,N_g+\alpha S$ (dashed). The horizontal lines are at 1 and 0, so the intersections define the solutions for α with and without the Jeffreys prior. **c** Q (continuous), S (dotted), and L (dashed). Vertical scale is 0–50 for L, and -100–0 for Q and S. **d** Eigenvalues λ. Dashed line is $\alpha=\lambda$. **e** Restorations f, labelling same as (a), except that C is $\bar{f}$. **f** Data (crosses) and transform of restoration (continuous), with residuals (in standard deviations) at top

the maximum, and creating a subsidiary maximum removed by two orders of magnitude in α, whose probability is somewhat lower. The probability of the $2L=N_d$ solution is very low, and shows much less structure. The $\alpha \to 0$ behaviour is confirmed, the log–log plot becoming a straight line. The maximum pointwise probability when the Jeffreys prior on α is used leads to an interesting result. The solution using the approximate derivative moves to a nearby point, as would be expected. However, the full form of the derivative has no zero there; α skips an order of magnitude, and gives a considerably sharper solution. $\bar{\alpha}$, $\hat{\alpha}$, and the approximate $\hat{\alpha}$ with the Jeffreys prior are all closely clustered, and the maps similar, although $\bar{f}$ shows sharper peaks but little difference in the wings. The integrals of f over the peaks are essentially the same for all the maps with significant probabilities. For $f(\hat{\alpha})$ they are

	peak$_1$	peak$_2$	peak$_3$	peak$_4$	background
integrated intensity	10.4682	4.6514	5.0293	0.1712	0.0825
standard deviation	1.5061	1.4860	0.1005	0.2102	0.5526

where the background consists of all points not included in the peaks, and the standard deviations are calculated

as the square roots of the diagonal of the covariance matrix found from (23):

$$\mathrm{cov}(p_1, p_2, p_3, p_4, bg)$$

$$= \begin{pmatrix} 2.2685 & -2.1026 & -0.0008 & 0.0015 & -0.1498 \\ -2.1026 & 2.2081 & -0.0008 & -0.0041 & -0.1053 \\ -0.0008 & -0.0008 & 0.0101 & 0.0010 & -0.0004 \\ 0.0015 & -0.0041 & 0.0010 & 0.0442 & -0.0377 \\ -0.1498 & -0.1053 & -0.0004 & -0.0377 & 0.3054 \end{pmatrix}.$$

The fourth peak, intensity 0.1712 ± 0.2102, is not significant. The third peak is extremely well determined, whereas the neighbouring first and second peaks are strongly negatively correlated. Although their individual variances are large, their sum, 15.1196 ± 0.5210 is much better determined. 74 points are included in the background, so the average background value is 0.0011 ± 0.0642.

3.2. *Dynamic light scattering*

This example illustrates a likelihood expression $L(\boldsymbol{F}, \boldsymbol{D})$ which is non-quadratic and contains an unknown parameter which is also estimated. Livesey et al. (1986) describe an earlier application of maximum entropy to this problem. The counts obtained in a DLS experiment are proportional to $G(t) = F(t)^2 + A$, where $F(t)$ is a multiexponential decay curve, related to the spectrum of decay times $f(\tau)$ by $F(t) = \int f(\tau) \exp -t/\tau \, \mathrm{d}\tau$, and A an unknown baseline count. The counts may be very large (10^6 or more), so the Poisson distribution of noise is approximated by Gaussian, giving the discretised form

$$L = \frac{1}{2} \sum_k \frac{1}{\sigma_k^2} (G_k - D_k)^2,$$

with $G_k = G(t_k)$, $\sigma_k = \sqrt{D_k}$. Poisson noise statistics could also be used directly, with

$$L = \sum_k G_k - D_k \log G_k + \log D_k!,$$

which is also in a functional form acceptable for our algorithm. Following other authors (Provencher 1979; Livesey et al. 1986), $\boldsymbol{f}$ is represented on a grid of points distributed uniformly in $\log \tau$, with a uniform prior $\boldsymbol{m}$ on this grid, whose value is calculated as the best-fit constant. The baseline level may be estimated from measurements at large t, when all components have decayed, but like other 'nuisance parameters', it may be integrated out of the likelihood (Jaynes 1987) to give the 'quasi-likelihood'

$$p_r(\boldsymbol{D} \mid \boldsymbol{f}) = \int p_r(\boldsymbol{D} \mid \boldsymbol{f}, A) \, \mathrm{d}A,$$

and we can then define $\tilde{L} = -\log p_r(\boldsymbol{D} \mid \boldsymbol{f})$, analogously to the original definition of L. L in the Gaussian noise aproximation is quadratic in A, so the integral is easily performed. $\tilde{L}$ and $\partial \tilde{L}/\partial \boldsymbol{F}$ have the same functional form as L and $\partial L/\partial \boldsymbol{F} = 2F_k(G_k - D_k)/\sigma_k^2$, except that A is replaced by $\tilde{A}$, its 'best estimate' at the current $\boldsymbol{F}$,

$$\tilde{A} = -\sum_k \frac{1}{\sigma_k^2} (F_k^2 - D_k) \Big/ \sum_k \frac{1}{\sigma_k^2}, \tag{26}$$

which is the value which minimises the difference between the 'mock' and actual data. $\partial^2 L/\partial F_j \, \partial F_k$ undergoes a rank-1 modification to give

$$\frac{\partial^2 \tilde{L}}{\partial F_j \, \partial F_k} = \mathrm{diag}\{\boldsymbol{W}\} - \boldsymbol{w}\boldsymbol{w}^T \Big/ \sum_k \sigma_k^{-2}$$

where

$$W_k = \frac{2}{\sigma_k^2}(3F_k^2 + \tilde{A} - D_k) \quad \text{and} \quad w_k = \frac{2}{\sigma_k^2} F_k.$$

Computationally, this modification is more economically performed in the singular space, rather than data space. $\partial^2 \tilde{L}/\partial F_j \, \partial F_k$ has at most one extra negative eigenvalue; otherwise the calculation is unchanged.

The posterior distribution of A may also be calculated, by a further application of Bayes' theorem. Assuming a flat prior for A, the result is

$$p_r(A \mid \boldsymbol{D}) = \int p_r(A, \boldsymbol{f}, \alpha \mid \boldsymbol{D}) \, \mathrm{d}\alpha \, \mathrm{d}^N f^{-1/2},$$
$$= \int p_r(A \mid \boldsymbol{f}, \alpha, \boldsymbol{D}) \, p_r(\boldsymbol{f} \mid \alpha, \boldsymbol{D}) \, p_r(\alpha \mid \boldsymbol{D}) \, \mathrm{d}\alpha \, \mathrm{d}^N f \prod f^{-1/2}.$$

Now $p_r(A \mid \boldsymbol{f}, \alpha, \boldsymbol{D}) \propto \exp -\frac{1}{2} \sum_k \sigma_k^{-2} (A - \tilde{A})^2$, and again using the Gaussian approximation for $p_r(\boldsymbol{f} \mid \alpha, \boldsymbol{D})$, the covariance of $\boldsymbol{f}$ (23), and linearising $\tilde{A}(\boldsymbol{f})$ (26), the expectation of A at fixed α is $\tilde{A}$, with variance

$$\left(\sum_k \sigma_k^{-2} + \boldsymbol{w}^T V \Sigma \, Y^{-T} \mathrm{diag}\left\{ \frac{1}{\alpha + \lambda} \right\} Y^{-1} \Sigma \, V^T \boldsymbol{w} \right) \Big/ \left(\sum_k \sigma_k^{-2} \right)^2. \tag{27}$$

An example is shown here, using data (kindly supplied by J. Langowski, EMBL, Grenoble) from a mixture of two sizes of polystyrene spheres in solution. The data are collected at correlation times which are switched in interval, so that $\log t$ space (3–800 μs) is roughly evenly covered, plus six counts at large t (13 ms), to help establish the baseline. The correlator scales the counts by a factor of 16, so the SNR is actually four times that expected from Poisson statistics. The spectrum is reconstructed on $\log_{10} \tau \in [1.3, 3.3]$. The singular values (Table 2) drop off exponentially, compared with the level spectrum for the Fourier problem.

N_g is found to be approximately 4, and again a wide range of probable α is found (Fig. 2a). The $2L = N_d$ solution (Fig. 2b) again has low probability, and fails to resolve the two peaks. The expectation baseline count, $\langle A \rangle = 868\,763$, with variance (27) 7700. $1/\sum_k \sigma_k^{-2} = 400$, so the variance is dominated by the uncertainty in $\hat{\boldsymbol{f}}$, not by the distribution of A about $\tilde{A}$.

Table 2. Singular values of the exponential decay transform

No.	Value.	No.	Value.	No.	Value.
1	65.8017	7	3.875×10^{-2}	13	1.490×10^{-5}
2	17.3744	8	9.231×10^{-3}	14	2.735×10^{-6}
3	5.9095	9	2.994×10^{-3}	15	1.375×10^{-6}
4	1.9106	10	1.815×10^{-3}	16	4.009×10^{-7}
5	0.5665	11	4.050×10^{-4}	17	6.276×10^{-8}
6	0.1480	12	7.971×10^{-5}	18	9.360×10^{-9}

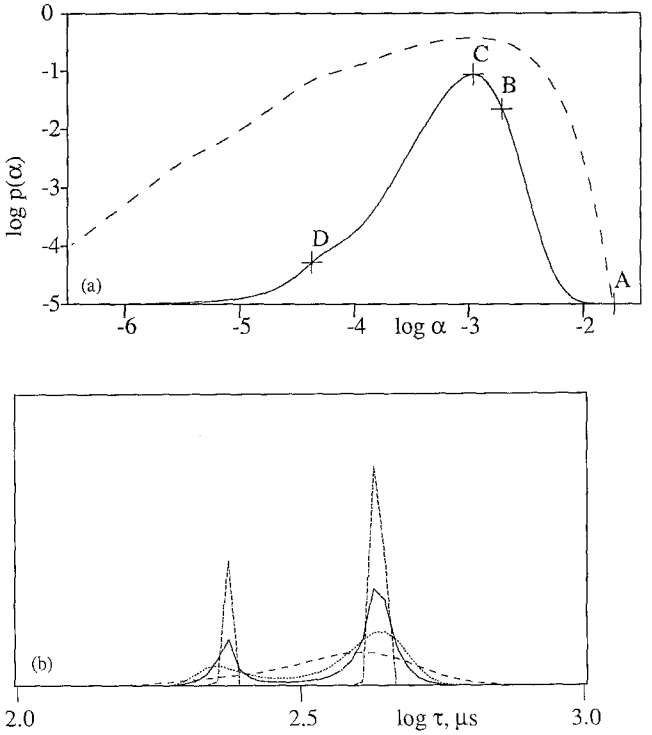


Fig. 2a, b. Plots of quantities calculated for the exponential decay problem. **a** Posterior probability of α (continuous) and its log (dashed). Other details as for Fig. 1a. **b** Part of restorations f, labelling as before. f is flat over the rest of the reconstructed spectrum

4. Conclusion

With the notable exception of crystallographic data, many datasets consist of samples from a continuous distribution, and are thus suitable for analysis with the algorithm presented here. However, when the geometry of the problem is such that the number of significant singular values of the matrix T becomes larger than the size of matrix that can reasonably be handled numerically, other methods must be used. These include approximations to $\sum\lambda/(\alpha+\lambda)$, which could be performed along the lines suggested by Skilling (1989 b), and to the increment $\delta\boldsymbol{u}$, which could be done by calculating $\delta\boldsymbol{u}$ as a linear combination of a (small) number of basis functions, similar to the search directions in $\boldsymbol{f}$ space used by Skilling and Bryan (1984). The analogous search directions would be based on $\boldsymbol{u}$ and $\boldsymbol{g}$, plus powers of MK acting on them.

Two examples have been given here. Both of the spectra were simple line spectra, and could have been analysed by direct fitting of the data by parameters defining the positions and intensities of the lines. However, the method presented here is more general, as it is not restricted to line spectra. In a forthcoming paper (Langowski and Bryan, in preparation) the specific application to DLS data will be examined in greater detail, and applied to both line and continuum spectra.

References

Bricogne G (1988) A Bayesian statistical theory of the phase problem. I. Acta Crystallogr A44:517–545

Bryan RK (1980) Maximum entropy image processing. PhD Thesis, University of Cambridge

Golub GH, van Loan CF (1983) Matrix computations. Johns Hopkins, Baltimore

Gull SF (1989) Developments in maximum entropy data analysis. In: Skilling J (ed) Maximum entropy and Bayesian methods. Kluwer, Dordrecht, pp 53–71

Gull SF, Daniell GJ (1978) Image reconstruction from incomplete and noisy data. Nature 272:686–690

Gull SF, Skilling J (1984) Maximum entropy method in image processing. IEE Proc 131(F):646–659

Hanson RJ (1971) A numerical method for solving Fredholm integral equations of the first kind. SIAM J Numer Anal 8:616–622

Jaynes ET (1968) Prior probabilities. IEEE Trans SCC-4:227–241

Jaynes ET (1987) Bayesian spectrum and chirp analysis. In: Smith CRay, Erickson GJ (ed) Maximum entropy and Bayesian spectral analysis and estimation poblems. Kluwer, Dordrecht, pp 1–37

Levenberg K (1944) A method for the solution of certain non-linear problems in least squares. Q Appl Math 2:164–168

Livesey AK, Licinio P, Delaye M (1986) Maximum entropy analysis of quasielastic light scattering from colloidal dispersions. J Chem Phys 84:5102–5107

Marquardt DW (1963) An algorithm for least squares estimation of non-linear parameters. SIAM J Appl Math 11:431–441

Pollard KOB (1988) Thesis, University of Cambridge

Provencher SW (1976) An eigenfunction expansion method for the analysis of exponential decay curves. J Chem Phys 64:2772–2777

Provencher SW (1979) Inverse problems in polymer characterization: Direct analysis of polydispersity with photon correlation spectroscopy. Makromol Chem 180:201–209

Provencher SW (1982) A constrained regularization method for inverting data represented by linear algebraic or integral equations. Comput Phys Commun 27:213–227

Shore JE, Johnson RW (1980) Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. IEEE Trans IT-26:26–37. Comments and corrections. IEEE Trans IT-29:942–943

Sibisi S (1989) Regularization and inverse problems. In: Skilling J (ed) Maximum entropy and Bayesian methods. Kluwer, Dordrecht, pp 389–396

Skilling J (1989a) Classic maximum entropy. In: Skilling J (ed) Maximum entropy and Bayesian methods. Kluwer, Dordrecht, pp 45–52

Skilling J (1989b) The eigenvalues of mega-dimensional matrices. In: Skilling J (ed) Maximum entropy and Bayesian methods. Kluwer, Dordrecht, pp 455–466

Skilling J, Bryan RK (1984) Maximum entropy image reconstruction: general algorithm. Mon Not R Astr Soc 211:111–124

Slepian D (1964) Prolate spheroidal wave functions, Fourier analysis and uncertainty, IV: Extensions to many dimensions; Generalized prolate spheroidal functions. Bell Syst Tech J 43:3009–3058

Slepian D, Pollak HO (1961) Prolate spheroidal wave functions, Fourier analysis and uncertainty, I. Bell Syst Tech J 40:43–64

Strand ON (1974) Theory and methods related to the singular-function expansion and Landweber's iteration for integral equations of the first kind. SIAM J Numer Anal 11:798–825

Turchin VF, Nozik VZ (1969) Statistical regularization of the solution of incorrectly posed problems. Atmos Oceanic Phys 5:14–18

Turchin VF, Kozlov VP, Malkevich MS (1971) The use of the mathematical statistics method in the solution of incorrectly posed problems. Sov Phys Uspekhi 13:681–703