Communications to the Editor

Maximum Entropy Analysis of Photon Correlation Spectroscopy Data Using a Bayesian Estimate for the Regularization Parameter

The analysis of data collected by dynamic light scattering (DLS) through photon correlation spectroscopy (PCS) has been a matter of discussion since almost the beginning of the technique. The problem consists of finding $f(\tau)$ in

$$G^{(2)}(t) = A + [\int f(\tau)\, e^{-t/\tau}\, d\tau]^2 \qquad (1)$$

where $G^{(2)}(t)$ is the measured second-order PCS autocorrelation function and $f(\tau)$ a distribution of relaxation times. A is a base line which is close to 1.0. The problem is known to be ill-posed,[11] and various regularization techniques have been applied to solve it. Recently, the maximum entropy (ME) technique has been given considerable attention by several authors, such as Livesey et al.[1] or Nyeo and Chu.[2]

One of the features of a typical DLS data set is that it is oversampled; that is, the amount of independent information contained is much less than the number of measured data points. The Cambridge MAXENT algorithm that was used in refs 1 and 2 did not take this property into account explicitly and, therefore, converged rather slowly. With a new algorithm especially adapted to oversampled data,[3] which converges by an order of magnitude faster, it is now possible to include some recent developments in ME theory: automatic determination of the regularizer by a Bayesian estimate, calculation of the covariance of the solution, and Bayesian estimation of so-called "nuisance parameters", like the base line in the case of DLS data. We have reported some preliminary results[3,5] obtained by our program (MEXDLS[12]) for the treatment of DLS and other oversampled data. The details of the algorithm have been described in ref 3. It is applicable to fitting problems where the measured data set (given by an N-dimensional vector $\mathbf{D}$) is related to a model (given by an M-dimensional vector $\mathbf{f}$) through the combination of a linear transformation by a matrix $\mathbf{T}$, followed by some nonlinear function. For example, the second-order correlation function in DLS would be represented by

$$\mathbf{D} = A + b(\mathbf{Tf})^2 \qquad (2)$$

where $\mathbf{D}$ is the second-order autocorrelation function $G^{(2)}(t)$. The matrix $\mathbf{T}$ contains the values of the exponential $e^{-\tau_j/t_k}$ for decay time τ_j and measurement time t_k, A is the experimental base line, and b is the total decay amplitude. The fitting problem is transformed into the space of the singular vectors of $\mathbf{T}$, which greatly reduces the dimensionality of the problem; typically, $\mathbf{T}$ is a 150 × 150 matrix but has only 15-25 significant singular values due to the oversampled nature of the problem.

The main new features of our program are as follows:

(i) Previous studies[1,2] used first-order correlations $g^{(1)}(t)$ as constraints in the calculation. These have to be calculated from the experimental readings of $G^{(2)}(t)$. Consequently, as mentioned in ref 1 and also shown in ref 10, the data points with small $g^{(1)}(t)$ are very sensitive to the value estimated for the base line. Our method uses $G^{(2)}(t)$ directly in a nonlinear likelihood function. As described in ref 3, either the base line A can be integrated out of the likelihood or its most probable value can be found using an iterative method. Thus, we can fit a PCS data set without relying on a base-line estimate.

(ii) Most current PCS data-fitting methods still use $\chi^2 = N_{\mathrm{data}}$ (χ^2 being the variance-weighted sum of squared deviations between the data set and the theoretical curve and N_{data} being the number of data points) as a criterion of fit to the data. The regularizer α, which determines the balance between likelihood (best fit) and smoothness of the solution (maximum entropy), is determined by this criterion. This procedure has been shown in general to underfit the data.[6] Using the Bayesian estimate for α suggested by Gull,[6] considerably better spectra can be achieved, as we show in the following. (This estimate results in a fit with $\chi^2 < N_{\mathrm{data}}$.)

(iii) The numerical algorithm for entropy maximization has been changed to achieve much better performance with oversampled data by transforming the fitting problem into the space of singular vectors of the matrix which transforms from solution space to data space.[3,4] The convergence problems that existed with the old algorithm of Skilling and Bryan[7] when the signal-to-noise ratio of the data is very high, or when the closer fit to the data demanded by the Bayesian fit criterion must be achieved, are considerably reduced by this new algorithm.

(iv) The Bayesian formulation[6] allows the variances and covariances of the sample values of the spectrum to be calculated. Our results on DLS data show that these quantities are essential for a careful interpretation of the final spectrum, due to very large values in the covariance matrix.

One of the main conclusions of ref 2 is that a maximum entropy analysis of DLS data gives equal or lower resolution than the widely used CONTIN program for exponential decay distributions. We show here that the MEXDLS algorithm, using the Bayesian estimate for α, gives better resolution than CONTIN. In particular, we are able to resolve two exponentials whose relaxation times differ by a factor of 2, at a noise level of 0.1%.[5] This resolution is slightly better than that reported in ref 1 for a similar problem, using the $\chi^2 = N_{\mathrm{data}}$ criterion for the fit.

To compare our algorithm with the results given in ref 2, we tested the simulated bimodal distribution data given in Figure 2a of ref 2, using the MEXDLS algorithm. The fit was then done directly to the simulated second-order autocorrelation function $G^{(2)}(t)$, which was obtained from the model $f(\tau)$ by the discretized form of eq 1

$$G^{(2)}(t) = A + [\sum_i f_i e^{-t/\tau_i}]^2 \qquad (3)$$

Five data sets were generated, following exactly the description given in ref 2, with different seeds for the random-number generator and a simulated noise level of 0.1%. The result of the MEXDLS analysis of these data sets is displayed in Figure 1 together with the original distribution. Two peaks are resolved in all cases; while the peak heights and widths scatter widely, their mean positions and areas are very close to those of the original distribution (Table I). If, as a comparison, we use the χ^2

0024-9297/91/2224-6346$02.50/0

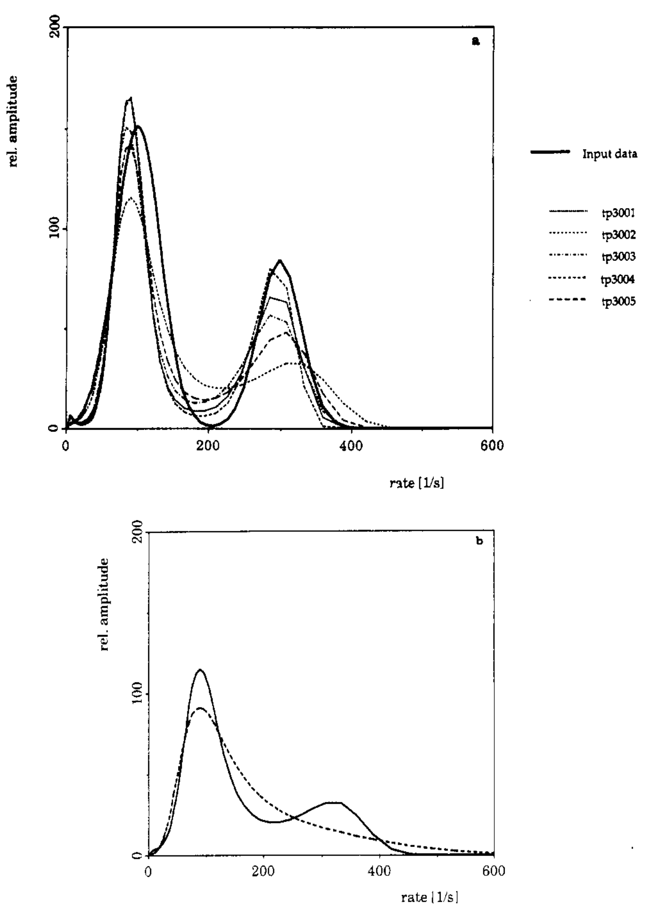

Figure 1. (a) MEXDLS analysis of a simulated autocorrelation function corresponding to a distribution of decay rates. The original distribution is indicated by the solid line; the differently dashed lines show the results of the analysis of five data sets, derived from the same distribution, with the same noise amplitude but different seeds for the random-number generator. The data sets were generated exactly as described in ref 2, adding 0.1% Gaussian noise to the data. The data sets are identified by their file names: (···) TP3001, (- - -) TP3002, (-··-) TP3003, (- -) TP3004, (— —) TP3005. (b) Influence of the choice of the regularizer α on the result of the MEXDLS analysis. The plots show the results of analyzing the TP3002 data set of Figure 1a using the most probable value of α as discussed in ref 3 (solid line) or using the $\chi^2 = N_{data}$ criterion (dashed line).

Table I
Integration of Peak Areas from the Results Displayed in Figure 1[a]

data set no.	peak 1	peak 2	remaining spectrum
TP3001	5.645	1.344	0.07
TP3002	5.288	1.393	0.049
TP3003	5.684	1.356	0.061
TP3004	5.623	1.347	0.069
TP3005	5.414	1.365	0.064
mean	5.531	1.361	0.063
SD	0.17	0.20	0.0084
SD from a covariance matrix	0.76	0.49	0.018

[a] Areas are given in relative units.

$= N_{data}$ criterion with our algorithm, we obtain essentially the same smooth curve as given in ref 2 for the maximum entropy solution (Figure 1a).

It is important to note that the data, although derived originally from a continuous bimodal distribution, only allow the determination of a limited number of independent parameters. This means that the sample values of the calculated distribution are not independent. We find that the "number of good observations", N_g, as defined in refs 5 and 7, is 7.4 ± 0.9 for the spectra presented here. A qualitative interpretation is that parameters such as peak positions, areas, and widths are well-determined but peak shapes are not.

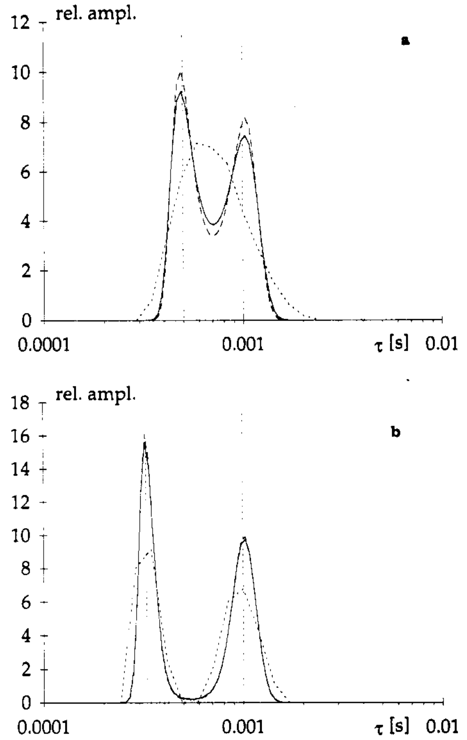

Figure 2. (a) MEXDLS analysis of a simulated autocorrelation function correspond to two exponential decays with relaxation times 10^{-3} and 5×10^{-4} s and equal decay amplitudes (positions indicated by vertical dashed lines). (—) MEXDLS solution using an estimated regularizer as described in ref 3; (- -) MEXDLS solution using the most probable value of α; (- - -) CONTIN's "chosen solution". (b) As a, with relaxation times 10^{-3} and 3.333×10^{-4} s.

Parts a and b of Figure 2 show MEXDLS analyses of DLS data consisting of two discrete components of equal amplitude spaced by a factor of 2 and 3, respectively, with a noise level corresponding to 10^6 counts/channel and an amplitude/base line ratio of 1. MEXDLS resolves the two components in both cases. CONTIN's "chosen solutions"[11] are plotted for comparison; here the factor 2 spacing is not sufficient to achieve resolution. In both cases, the peak centers of the MEXDLS solution are positioned very close to the original data. The solutions were obtained either by using the approximate criterion for the regularizer α, $N_g = -2\alpha S$, described by Gull,[6] where N_g is the number of good observations and S the entropy, or by calculating the most probable value of α by a search method. Both solutions are very close to each other (solid and dashed lines in Figure 2a,b).

The properties of the solution $f(\tau)$ can be better visualized by plotting its covariance matrix $\mathrm{cov}(f(\tau_i) f(\tau_j))$. Figure 3 shows such a plot for one of the five data sets of Figure 1. The diagonal of the plot represents the variance of each point of the spectrum. It can be seen that closely neighboring points always have strong negative covariance. This is plausible since an arbitrary increase of the distribution $f(\tau)$ at τ_i balanced by a similar decrease at $\tau_i + \epsilon$ will result in a theoretical autocorrelation function that fits the data equally well. The two resolved peak centers have positive covariance and correlate negatively with the minimum between. Thus we see that, for different noise realizations, the amplitude ratio of the two peaks stays constant (see Table I) while the valley between the two peaks may be more or less pronounced.

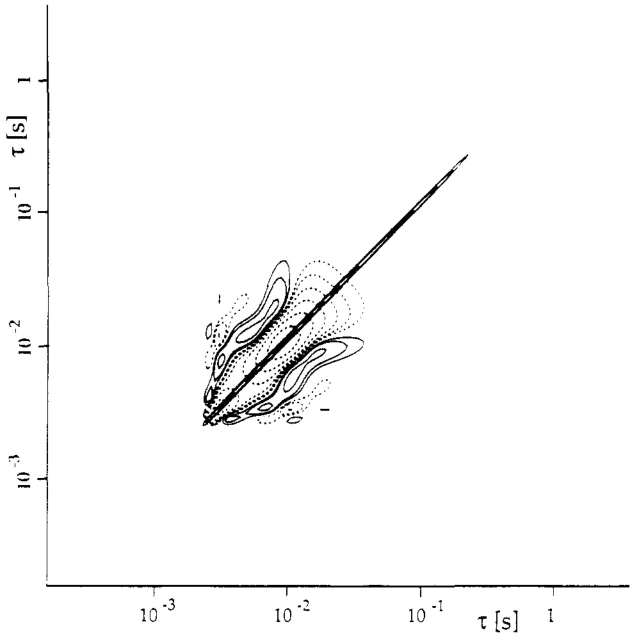

Figure 3. Graphical representation of the covariance matrix for the MEXDLS analysis of the data set TP3002 from Figure 1b. The two axes of the plot correspond to decay time on a logarithmic scale: (solid lines) positive elements; (broken lines) negative elements.

Although the distribution function is in principle continuous, it is always represented on a discrete grid. The value at a grid point therefore represents the integral of the continuous spectrum over a finite interval. Therefore, the values in the covariance matrix of **f** also depend on the grid spacing. To give an example, going from an N-point spectrum to an $N/4$-point spectrum means that the new spectrum is given by summing over four adjacent points of the old spectrum, and each element of the new covariance matrix is obtained by summing over 4 × 4 squares of the old one. For the diagonal elements that determine the variance of the spectrum, the negative off-diagonal contributions partly cancel the positive contributions from the diagonal; thus, the net variance of the diagonal elements of the coarser grid will be smaller.

Hence, integration over larger ranges of the spectrum will give more accurately determined quantities. For example, in Table I we have divided the spectrum into three parts: the two major peaks and the remaining spectrum. The standard deviations (SD) for each integral as obtained from the diagonal elements of the covariance matrix are on the order of 5–10%, whereas the variations between individual data sets due to different noise realizations are somewhat smaller. This result is related to the fact that a change in the noise realization only changes the data, which determines $N_g = 7.4 \pm 0.9$ independent parameters. Changing the noise will therefore sample only a 7.4-dimensional subset of the space of all possible spectra, while the covariance matrix contains contributions due to uncertainties in the data and the width of the prior probability distribution of $f(\tau)$.

In conclusion, we state that a correct application of Bayesian statistics to the analysis of PCS data can lead to multiexponential decay spectra of good resolution in a rather consistent way, with estimation of variances and covariances of the result of the analysis. A cautious interpretation of these results, using averages over suitable parts of the covariance matrix, is however required.

References and Notes

(1) Livesey, A. K.; Licinio, P.; Delaye, M. *J. Chem. Phys.* **1986**, *84*, 5102.
(2) Nyeo, S.-L.; Chu, B. *Macromolecules* **1989**, *22*, 3998.
(3) Bryan, R. *Eur. Biophys. J.* **1990**, *18*, 165.
(4) Bryan, R. In *Maximum Entropy and Bayesian Methods*; Fougere, P. F., Ed.; Kluwer Academic: Dordrecht: The Netherlands, 1990; pp 221–232.
(5) Langowski, J.; Bryan, R. *Prog. Colloid Polym. Sci.* **1990**, *81*, 269.
(6) Gull, S. F. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Kluwer Academic: Dordrecht: The Netherlands, 1989; pp 53–71.
(7) Skilling, J.; Bryan, R. K. *Mon. Not. R. Astron. Soc.* **1984**, *211*, 111.
(8) Livesey, A. K.; Skilling, J. *Acta Crystallogr.* **1985**, *A41*, 113.
(9) Skilling, J. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Kluwer Academic: Dordrecht: The Netherlands, 1989; pp 45–52.
(10) Ruf, H. *Biophys. J.* **1989**, *56*, 67.
(11) Provencher, S. W. *Comput. Phys. Commun.* **1982**, *27*, 213.
(12) The MEXDLS program is available from J. Langowski as a VAX executable version or Macintosh application. We prefer to send the program via electronic mail; contact the author at LANGOWSKI@FREMBL51.bitnet.
(13) EMBL.
(14) Laboratory of Molecular Biophysics.

J. Langowski*,[13] and R. Bryan[14]

EMBL, c/o ILL, 156X
F-38042 Grenoble Cedex, France, and
Laboratory of Molecular Biophysics
South Parks Road, Oxford OX1 3QU, England

Received January 3, 1991
Revised Manuscript Received September 10, 1991