Molecular spectroscopy and Bayesian spectral analysis—how many lines are there?

D. S. Sivia and C. J. Carlile
ISIS Pulsed Neutron Facility, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, OX11 0QX, United Kingdom

(Received 24 July 1991; accepted 24 September 1991)

We demonstrate the Bayesian spectral analysis approach for analyzing neutron scattering molecular tunneling data. It is a generalized form of model fitting, which is appropriate when the number of parameters to be optimized is not known. Specifically, it addresses the question of how many excitation lines there is evidence for in the data. We review the theory of Bayesian spectral analysis relevant to our particular application, describe an efficient algorithm for its implementation, and illustrate its use with both simulated and real data. We believe that this powerful method of analysis will be a very useful tool in experimental molecular spectroscopy.

I. INTRODUCTION

The experimental measurement of molecular excitation lines is obscured by the effects of both the instrumental resolution and finite statistics. The resulting data are essentially a blurred and noisy version of the spectrum we seek, and are usually analyzed by the least-squares fitting of a functional model. A model-independent estimate of the underlying spectrum can also be obtained, by using the maximum entropy (MaxEnt) method. In a sense, both traditional model fitting and MaxEnt are a little extreme for this problem. Traditional model-fitting tends to be too restrictive, by requiring a fully-defined functional model, and implicitly assumes more than we usually know. MaxEnt is somewhat too liberal because the free-form solution often fails to take into account all our prior knowledge about the physics of the situation. Bayesian spectral analysis adopts a middle path. It is a generalized form of model fitting, which is appropriate when the number of parameters to be optimized is not known; the solutions are restricted to a broad class of models rather than to particular model. The choice of the type of model encodes more prior knowledge about the nature of the molecular excitation spectrum than does MaxEnt, while still allowing considerably more flexibility than does conventional model refinement.

We should emphasize, however, that the *logic* behind Bayesian spectral analysis is no different to that of MaxEnt: both rely on the direct use of probability theory for inferring the spectrum from the data and are Bayesian, therefore. Indeed, traditional least-squares model fitting can also be justified within the Bayesian framework. What is different about the three approaches is the precise question that is being asked of the data, and the implicit assumptions which underlie them. In traditional model fitting, for example, we are asking a question of the type: *Given that the excitation spectrum consists of five δ functions, what is the best estimate of their positions and amplitudes*? In MaxEnt, the corresponding question is: *Given that all we know about the excitations is that they constitute a "positive and additive distribution," what is the best estimate of the spectrum*? In Bayesian spectral analysis, we use our prior knowledge about the discrete nature of the excitations but do not restrict their number: *Given that the spectrum consists of a "few" δ functions, how many excitations is there most evidence for in the data and what are their positions & amplitudes*?

In Sec. II we formulate the tunneling spectroscopy problem from a data analysis point of view and outline the theory of the Bayesian spectral analysis approach. In Sec. III we describe an algorithm for the practical implementation of the theoretical results. We illustrate its use with both simulated examples and real neutron data in Secs. IV and V, respectively, and conclude with Sec. VI.

II. THE METHOD

A. Probability theory and data analysis

In science we are often faced with the task of making inferences about some object of interest given incomplete and noisy experimental data. For the case of molecular spectroscopy, we are primarily concerned with inferring the spectrum $F(\epsilon)$, where ϵ is the energy of the excitation and F is proportional to the number of molecules in that state, given a set of data $\{D(\epsilon_k)\}$ $(k = 1,2,3,...,M)$. Given the experimental measurements, what is our best estimate of $F(\epsilon)$ and how confident are we in our prediction? The answer to this question is not clear cut since it depends on both the data and our prior knowledge about $F(\epsilon)$. For example, if physics told us that the spectrum must have a particular functional form, then we need to consider only a very limited set of possibilities for $F(\epsilon)$ defined by a handful of parameters. Alternatively, if we did not have any good *a priori* reason to assume a functional form, then we must consider a much larger set of possibilities for $F(\epsilon)$ described in a suitable free-form fashion. Even without a functional form we may know about the positivity of the spectrum, or its zeroth moment (normalization), or an asymptotic solution, and so on, which will restrict the set of allowed possibilities for $F(\epsilon)$. How, then, should we combine our prior knowledge with the evidence of the data to obtain our best estimate of the spectrum and a measure of its reliability?

 0021-9606/92/010170-09$06.00 © 1991 American Institute of Physics

Cox[1] has shown that any method of inference which satisfies simple rules for logical and consistent reasoning must be equivalent to the use of ordinary probability theory, as originally formulated by Bernoulli,[2] Bayes,[3] and Laplace.[4] Accordingly, the conditional probability distribution function (PDF) prob$[F|D,I]$ summarizes our inference about the spectrum given the data and our prior knowledge I about $F(\epsilon)$ and the experimental setup. Since the numerical value of the probability assigned to any particular $F(\epsilon)$ is a measure of how much we believe that it is the true spectrum, our best estimate is given by that $F(\epsilon)$ which maximizes prob$[F|D,I]$. The width, or spread, of this PDF about the maximum tells us the reliability of the estimate: if the PDF is sharply peaked then we are confident of our prediction, but if it is broad then we are fairly uncertain about the true spectrum.

In order to compute prob$[F|D,I]$ we need to use an important result from probability theory, called Bayes' theorem, which relates the PDF we require to one which we can calculate and to another which encodes our prior knowledge,

$$\text{prob}[F|D,I] \propto \text{prob}[D|F,I] \times \text{prob}[F|I], \tag{1}$$

where we have omitted the normalization constant, $1/(\text{prob}[D|I])$, from the right-hand side, for simplicity. The term on the far right, prob$[F|I]$, is called the *prior* PDF and represents our state of knowledge (or the lack thereof) about $F(\epsilon)$ before we have analyzed the experimental data. Our prior state of knowledge is modified by the data through the so-called *likelihood function*, prob$[D|F,I]$, which encodes details about the experimental setup. The product of the prior PDF and the likelihood function yields the *posterior* PDF we require and represents our state of knowledge about $F(\epsilon)$ after we have analyzed the data.

Let us first consider the likelihood function in more detail. The likelihood function tells us how likely it is that we would have measured the data that we actually did, if we were given an $F(\epsilon)$. In order to compute the likelihood function, therefore, it is essential (but not sufficient) that we should be able to compute an ideal data set $\{\hat{D}(\epsilon_k)\}$ given a spectrum. For our case, the data are given by a convolution of the spectrum with the resolution function of the instrument $R(\epsilon)$ plus a background signal $B(\epsilon)$,

$$\hat{D}(\epsilon_k) = \int_{-\infty}^{+\infty} R(\epsilon_k - x)F(x)dx + B(\epsilon_k). \tag{2}$$

For the moment, let us assume that we have a reasonably good estimate of the resolution function and the background signal; later we will see how these conditions can be relaxed.

The other information we need in order to calculate the likelihood function is some knowledge about the statistical properties of the errors in the experimental data. If we make the simplifying assumptions that the data are independent (so that one measurement does not affect another) and subject to additive Gaussian noise with a root-mean-square-error $\{\sigma_k\}$, then the likelihood function takes the familiar form,

$$\text{prob}[D|F,I] = \frac{e^{-\chi^2/2}}{\prod_k [2\pi\sigma_k^2]^{1/2}}, \tag{3}$$

where χ^2 is the usual sum-of-squared residuals misfit statistic,

$$\chi^2 = \sum_k \frac{[\hat{D}(\epsilon_k) - D(\epsilon_k)]^2}{\sigma_k^2}. \tag{4}$$

It should be noted that this form of the likelihood function can also be derived very easily by using the principle of maximum entropy to assign prob$[D|F,I]$ subject to a constraint on the (expectation value of the) variance of the data:[5-7] $\langle [\hat{D}(\epsilon_k) - D(\epsilon_k)]^2 \rangle = \sigma_k^2$.

Next, let us consider the prior PDF prob$[F|I]$. Before we can think about assigning the prior PDF, we must decide what we know about the spectrum *a priori*. An irrefutable piece of prior knowledge about $F(\epsilon)$ is that it is a *positive and additive* distribution. It is positive because $F(\epsilon_1)$ is proportional to the number of molecular modes with an excitation state of energy ϵ_1, and it is additive because the number of molecular modes with excitation states of energy ϵ_1 and ϵ_2 is given by the sum $F(\epsilon_1) + F(\epsilon_2)$. The appropriate prior for a positive and additive distribution is not immediately obvious but many different types of arguments,[5,8-12] including logical consistency, information theory, coding theory, and combinatorial arguments, lead us to believe that it is of an entropic form: $\exp(\alpha S)$, where S is the generalized Shannon–Jaynes entropy of the spectrum.[13] This, in turn, leads us to the use of the MaxEnt method for data analysis.[13-15]

In molecular spectroscopy, however, we usually know more than just that $F(\epsilon)$ constitutes a positive and additive distribution. The physics of the problem often tells us that the spectrum consists of a "few" discrete excitations. Assuming, for the sake of argument, that we know that the excitation lines are all Gaussian in shape with width W, the spectrum can be written as

$$F(\epsilon) = \sum_{j=1}^{\text{few}} A_j \cdot \exp\left[-\frac{(\epsilon - \epsilon_j)^2}{2W^2} \right]. \tag{5}$$

The prior PDF for the spectrum is now defined by our prior knowledge about the value of the parameters for the amplitudes and positions of the excitations prob$[\{A_j,\epsilon_j\}|I]$. Since the number of the parameters to be estimated is small compared with number of data, the likelihood function will be much sharper than any PDF expressing reasonable *a priori* ignorance; therefore, the posterior PDF will be dominated by the likelihood function. In particular, if we make the algebraically simple assignment of a uniform prior PDF (prob$[F|I]$ = a constant), the maximum of the posterior PDF becomes coincident with the maximum likelihood solution. As the likelihood function above is of the form $\exp(-\chi^2/2)$, our best estimate for the spectrum is given by the set of parameters $\{A_j,\epsilon_j\}$ which minimize χ^2. This, then, provides us with the justification for using the method of least squares.

An important question which we have glossed over in the previous paragraph, but is dear to the hearts of experimental molecular spectroscopists, is the numerical value of

"few" in terms of the number of excitations. Indeed, we need to *know* the number of excitation lines before we can carry out a least-squares analysis!

B. Bayesian spectral analysis—how many lines are there?

To estimate the number of excitation lines N, given the experimental data, we require the posterior PDF $\text{prob}[N|D,I]$. To compute this posterior PDF we need to use two more results from probability theory. The first is essentially a restatement of Bayes' theorem,

$$\text{prob}(a,b|c) = \text{prob}(a|b,c)\times\text{prob}(b|c), \tag{6}$$

and the second concerns *marginalization*, or the integrating out, of *nuisance* parameters,

$$\text{prob}(a|c) = \int_{-\infty}^{+\infty} \text{prob}(a,b|c)db. \tag{7}$$

The mathematics presented below, to compute the posterior PDF for the number of excitation lines, is essentially the same as that found in Chapter 5 of Jeffreys;[16] it can also be found in Gull[6] and Bretthorst.[17]

We begin by using Bayes' theorem,

$$\text{prob}[N|D] = \frac{\text{prob}[D|N]\times\text{prob}[N]}{\text{prob}[D]}, \tag{8}$$

omitting the background information I, as implicitly given throughout, for simplicity. The denominator is a (normalization) constant, which cancels out when comparing the relative probabilities that N has value n_1 and n_2. The prior probability for N, $\text{prob}[N]$, can be taken as uniform between 0 and a few (say 20) so that *a priori* we have no preference for any particular number of lines. Therefore,

$$\text{prob}[N|D] = K\cdot\text{prob}[D|N], \tag{9}$$

where K is a constant. Using Eq. (7), we can write $\text{prob}[D|N]$ as a marginal distribution over the joint PDF for the data and the parameters of the model,

$$\text{prob}[D|N] = \int\int\cdots\int \text{prob}[D,\{A_j,\epsilon_j\}|N]\,d^N A_j\,d^N\epsilon_j. \tag{10}$$

Equation (6) then allows us to write the joint PDF, $\text{prob}[D,\{A_j,\epsilon_j\}|N]$, as the product of a conditional probability and a prior probability (given N throughout),

$$\text{prob}[D,\{A_j,\epsilon_j\}|N] = \text{prob}[D|\{A_j,\epsilon_j\},N]\times\text{prob}[\{A_j,\epsilon_j\}|N]. \tag{11}$$

The first term on the right-hand side is just the likelihood function, or goodness-of-fit term, which we have approximated by the familiar $\exp(-\chi^2/2)$ form [Eq. (3)]. The second term is a prior probability over the amplitudes and positions of the N lines describing the spectrum. We will make the simple assignment that this prior PDF is uniform (a constant) in the range,

$$\epsilon_{\min}\leqslant\epsilon_j\leqslant\epsilon_{\max},\quad 0\leqslant A_j\leqslant A_{\max}$$

and

$$A_1 + A_2 + \cdots + A_N \leqslant A_{\max}, \tag{12}$$

and zero otherwise. The main reason for using a uniform prior is that it simplifies the subsequent algebra. Although a purist could argue that this is not the optimal assignment (perhaps preferring instead a *Jeffreys' prior*, which is uniform in the logarithm of the amplitudes), our defense is that the analysis depends only *weakly* on the particular form of the prior since we are trying to estimate a single parameter (N) from many data. What we are trying to convey by this assignment is that the positions of the lines lie within the range $\epsilon_{\min}$ to $\epsilon_{\max}$ (presumably defined by the energy range of the data), and that the amplitudes are all positive with no single or combined magnitude exceeding $A_{\max}$ (defined by the integrated intensity of the data).

Putting together the various terms above, we obtain

$$\text{prob}[N|D] = \frac{\text{const}\cdot N!}{[(\epsilon_{\max}-\epsilon_{\min})\cdot A_{\max}]^N}\times\int\int\cdots\int e^{-\chi^2/2}d^N A_j\,d^N\epsilon_j, \tag{13}$$

where the marginalization of Eq. (10) has been reduced to a multiple integral over the likelihood function times a prefactor arising from the normalization of the prior probability for the amplitudes and positions of an N-line model for the spectrum. If we assume that there is only one *significant* maximum in the likelihood function, and make a quadratic Taylor series expansion around it,

$$e^{-\chi^2/2} = e^{-\chi^2_{\min}/2}\times e^{-(\delta\epsilon_j\delta A_j)^T\cdot\nabla\nabla\chi^2\cdot\delta\epsilon_j\delta A_j/4}, \tag{14}$$

then we can do the Gaussian multiple integral analytically and obtain,

$$\text{prob}[N|D] = \frac{\text{const}\cdot N!}{[(\epsilon_{\max}-\epsilon_{\min})\cdot A_{\max}]^N}\times\frac{(4\pi)^N\cdot e^{-\chi^2_{\min}/2}}{[\text{Det}(\nabla\nabla\chi^2)]^{\frac{1}{2}}}, \tag{15}$$

where $\text{Det}(\nabla\nabla\chi^2)$ is the determinant of the Hessian matrix.

Equation (15) tells us how to compute the posterior PDF $\text{prob}[N|D]$, and thereby allows us to estimate the number of lines for which there is most evidence in the data. The closed form of the solution does depend, of course, on the validity of the approximations and assumptions, such as that of the quadratic expansion of Eq. (14), but they are not unusual in the sense that they are always implicit when using the traditional least-squares analysis. An algorithm for the practical implementation of the theoretical results is described in Sec. III and a discussion of the meaning, and performance, of Eq. (15) is given in Sec. IV, following an illustration of its use with simulated data.

C. Dealing with systematic uncertainties

Equation (2) shows how the excitation spectrum is related to the data through the resolution function and the background signal. It is a relationship we require for doing any kind of data analysis and assumes that we know the resolution function and the background. Usually we try to obtain a fairly good estimate of both, perhaps by collecting some data using a standard elastic scatterer and an empty cell, but can we do anything if we are not quite so fortunate?

The answer is that we can, at least in principle.

In order to deal with systematic uncertainties, we must be able to characterize their nature in terms of just a few parameters: a linear background, the width of the resolution function given its shape, and so on. These parameters necessarily enter our analysis, because they are needed to compute the likelihood function [Eqs. (2)–(4)], but we are not really interested in their actual value when estimating the posterior PDF for the number of lines—they are *nuisance* parameters. We eliminate them by using marginalization and Bayes' theorem [Eqs. (1), (6), and (7)], just as we removed the set of parameters $\{A_j,\epsilon_j\}$ from the analysis in the previous section [Eqs. (10)–(15)]. Suppose, for example, that we did not know the value of a constant background B. Then, Eq. (10) would become

$$\text{prob}[D\,|N] = \int\int\cdots\int \text{prob}[D,\{A_j,\epsilon_j\},B\,|N]\,d^N A_j\,d^N\epsilon_j\,dB. \quad (16)$$

The analysis would follow along the same lines as before, leading to Eq. (15), but with B now contributing to both $\chi^2_{\min}$ and the Hessian matrix. Since the prior PDF for B [at Eq. (12)] does not depend on the number of lines, its normalization factor can be incorporated into the existing constants in Eqs. (13) and (15).

Systematic uncertainty does not only arise from the background and resolution functions, but also from the width W of the excitation lines [Eq. (5)] if this is not known. This can be dealt with by optimization and marginalization in exactly the same way as above. Although probability theory provides us with the principles to deal with systematic uncertainties, we should not take this to mean that there is no need or value in trying to obtain an estimate of the resolution function and background. Even when we can use the theoretical apparatus in practice, systematic uncertainties still lead to a (significant) reduction in the reliability of the inferred spectrum.

III. AN ALGORITHM

In order to make use of the analysis of the previous section, we need an algorithm for its practical implementation. Before we describe such a procedure, let us first state some assumptions which are implicit in this implementation. First of all, the background is assumed to be linear so that it can be described by two parameters. The resolution function is taken to be known, and tabulated on a fine-enough grid to allow subsequent linear interpolation as required. Although it is also assumed to be invariant in energy transfer in the examples given in this paper, a known variation in the width of the resolution function (keeping the same shape) can be accommodated with some loss in speed. All the excitation lines are assumed to have the same intrinsic width. The examples in this paper use a Gaussian profile, although a Lorentzian or other lineshape could be used just as well. In principle, probability theory will tell us which intrinsic line shape is to be preferred, on the basis of the data, if we have two or more alternatives. It should be emphasized that none of these assumptions is central to the use of Bayesian spectral analysis, but instead they are made to improve computational speed and robustness; they can be relaxed, but at some cost in algorithmic efficiency. Such simplifications are in keeping with the spirit of Bayesian spectral analysis, where we are trying to estimate the parameters of a *simple* but *adequate* type of model when the number of parameters is not known.

The formula for computing the most probable number of excitation lines, in the light of the data, is given in Eq. (15). It requires us to find the best-fit solution for any specified number of lines, but we also need an initial estimate for the range of the position and amplitude parameters $\epsilon_{\min}$, $\epsilon_{\max}$, and $A_{\max}$. Let us deal with the question of the prior range first. Ideally this should come from the results of previous data or theory but, in practice, a reasonable estimate is provided by the energy range and integrated intensity of the data we wish to analyze. So, we set $\epsilon_{\min}$ to the energy of the datum with the smallest energy and $\epsilon_{\max}$ to the corresponding value for the datum with the highest energy. $A_{\max}$ requires an initial estimate for the linear background (obtained by using the average of the data at either end of the energy range, for example), as well as the integrated intensities of the data and the resolution function (if it is not normalized). Parseval's theorem then tells us that the total intensity of the excitation lines, or $A_{\max}$, is given by the difference between the intensity of the data and background divided by the intensity in the resolution function. As a matter of computational efficiency, we should work with a "resultant" resolution function which combines the broadening effects of the instrumental resolution and the intrinsic shape of the excitation lines; this relies on the associativity of the convolution function: $a*(b*c) = (a*b)*c$.

Next, we must consider a procedure for obtaining the best-fit solution for any specified number of lines. Although it is difficult to give absolute guarantees, because the data are not linearly related to the positions of the excitation lines, we describe a simple algorithm with which we have had considerable success. (i) Start by obtaining an initial estimate for the two parameters describing the linear background (by eye-ball fitting the data with a ruler if necessary). If the background is very small, the initial guess could be zero. (ii) Taking the background as given, do a fairly thorough two-dimensional (2D) search, on a rectangular grid ($0 \leqslant A \leqslant A_{\max}$ and $\epsilon_{\min} \leqslant \epsilon \leqslant \epsilon_{\max}$), for the amplitude and position of the first (and presumably strongest) excitation line. This 2D search actually reduces to just a 1D search for the position of the line, since the amplitude of the line is linearly related to the data (given its position). (iii) Refine the parameters of the background and the excitation line simultaneously. It is often helpful to do this by first using a robust *simplex*like algorithm,[18] to be followed by a gradient *Newton–Raphson*-type algorithm.[19] In any case, the second derivatives need to be computed at the optimal solution (by finite differences, for example) since both $\chi^2_{\min}$ and the determinant of the Hessian matrix are required to calculate the posterior PDF prob$[N=1|D]$. (iv) Taking the background and the first excitation line as given, do a fairly thorough two-dimensional search for the amplitude and position of the second line; again, the linearity of the amplitudes can be exploited to reduce this to a 1D search for the position [as in (i)]. Refine

the parameters of the background and both lines simultaneously, and calculate prob$[N=2|D]$. Continue this cycle, with one additional line in each go, until a maximum in the posterior PDF for the number of lines is evident.

Before we go on to illustrate the use of this algorithm, we should make a couple of additional remarks. It is helpful to use the prior range parameters $A_{\max}$ and $\epsilon_{\max}-\epsilon_{\min}$ (and the initial estimate of the background) to scale the parameters to be optimized. That is to say, if we work in dimensionless units like $A_j/A_{\max}$, the optimal parameters will all be of a similar order. Working in these dimensionless units, we can improve the stability of matrix calculations, such as the inversion of the Hessian matrix to obtain the variance of the inferred parameters, by adding (twice) the identity matrix to the Hessian matrix. Adding the identity matrix does not change the eigenvectors of the Hessian matrix, and so does not change the correlations between the inferred parameters. It does, however, put a lower bound on the eigenvalues, thereby encoding our prior knowledge that the uncertainty in any parameters cannot exceed the order of the prior range assigned to it. Finally, it might have been noted that the intrinsic width W of the excitation lines was not treated as an unknown parameter in the same way as the linear background. In principle it could be, but in practice we have found it more stable to run our program several times over using different given widths. We can then plot the two-dimensional posterior PDF for the number of lines and the intrinsic width, integrating with respect to W to obtain prob$[N|D]$ (or integrating with respect to N to obtain prob$[W|D]$, if desired).

IV. AN EXAMPLE USING SIMULATED DATA

We begin our illustration of the use of the theory and algorithm described in Secs. II and III with the aid of data generated in a computer simulation. These data are shown in Fig. 1(a) and result from the convolution of a spectrum consisting of a "few" excitation lines, having a Gaussian profile, with a Gaussian resolution function of full width at half-maximum (FWHM) 2.0 μeV. As can be seen from Fig. 1(a), the data are corrupted by a linear background and statistical noise. Given this information alone, how many lines is there most evidence for in the data?

Carrying out the analysis, as described in the previous sections, we obtain the (logarithm of the) posterior PDF for the number of excitation lines shown by the continuous line in Fig. 1(b). There is most evidence for two lines, therefore. The FWHM of the intrinsic Gaussian width of the lines is estimated to be 1.03 ± 0.08 μeV, and the best estimate of their positions is 13.98 ± 0.03 and 15.47 ± 0.02 μeV. The spectrum which was used to generate the data in Fig. 1(a) did indeed contain two lines, with a FWHM of 1.0 μeV, centered at 14.0 and 15.5 μeV. The amplitudes of both lines were the same, and were recovered correctly to within 5%.

The example above provides us with a good opportunity to comment on the meaning and performance of the theory presented in Sec. II. The shape of the posterior PDF for the number of lines, shown in Fig. 1(b), is characteristic of this type of Bayesian analysis: (a) there is sharp falloff from the maximum on the left, because there is not enough structure

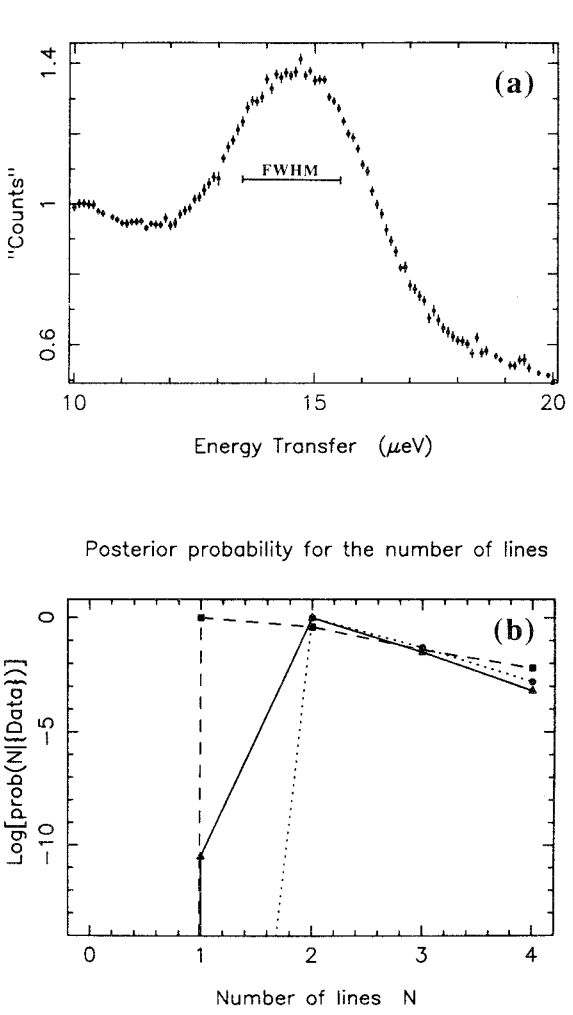

FIG. 1. (a) Data generated by a computer simulation, from a molecular excitation spectrum consisting of a "few" excitation lines (of Gaussian profile). The data are corrupted by statistical noise, in addition to a linear background and an instrumental resolution broadening by a Gaussian of FWHM 2.0 μeV. (b) The continuous line shows the logarithm (to base 10) of the posterior probability for the number of excitation lines, given these data, computed using the formula in Eq. (15). The dashed line shows the results of the corresponding analysis when the data were corrupted by statistical noise whose magnitude was three times larger than in (a); the dotted line is also for the noisier data, but the intrinsic width of the excitation lines (1 μeV) was used as prior knowledge.

in the model to account for the data; (b) there is a slow falloff on the right, as the models become unnecessarily complicated. What we have, in essence, is a *quantitative* statement of *Ockham's Razor*. This is the qualitative statement, named after William of Ockham[20] (who died $\sim$1349), that when several theories exist to describe some phenomenon, we prefer the simplest one consistent with the empirical evidence. In our case, we choose the model with the least number of lines which fits the data. Although this is exactly what we would have done using our "common sense," the value of the analysis is that it sharpens and refines our common sense

far beyond our qualitative intuition. It would be difficult for our common sense to match the formal analysis and state that the two-line model for the spectrum is 10 orders of magnitude more probable than a one-line model, and that the two-line model is 30 times more probable than a three-line model!

We should emphasize, however, that the results are conditional on the particular data which we are analyzing and on our prior knowledge. For example, if the statistical noise on the data of Fig. 1(a) had been three times larger (corresponding to an experiment run for only one tenth of the time), we would have obtained the posterior PDF for the number of lines shown by the dashed line in Fig. 1(b). There is then most evidence for only one line (at $\sim 14.8\,\mu$eV), with an intrinsic width of about $2\,\mu$eV, although the possibility of two lines could not be ruled out at the 90% confidence level. This is because the data of poorer quality can be more simply, but sufficiently, explained by a broader single line than by two narrower ones. Indeed, the maximum of the posterior PDF would occur at zero if the quality of the data was so poor that they could be adequately explained by just a linear background! In that case the maximum would also be very shallow, and so probability theory would be warning us that it was unwise to make too decisive a judgement on the basis of such poor data. Even for the lower quality data, with three times the statistical noise as Fig. 1(a), we would still find most evidence for two lines if we had used the fact that the FWHM of the lines was known to be $1.0\,\mu$eV as prior knowledge; the corresponding posterior PDF is shown by the dotted line in Fig. 1(b).

V. EXAMPLES USING REAL EXPERIMENTAL DATA

We now demonstrate the use of the theory and algorithm described in Secs. II and III for analyzing real experimental neutron tunneling data. The data were taken on the IRIS spectrometer at the pulsed neutron facility ISIS,[21] and on the IN10 spectrometer[22] at the reactor source at the ILL; the sample was *2,6 dimethyl pyridine,* or lutidine. The data from IRIS have been analysed previously by Mukhopad-

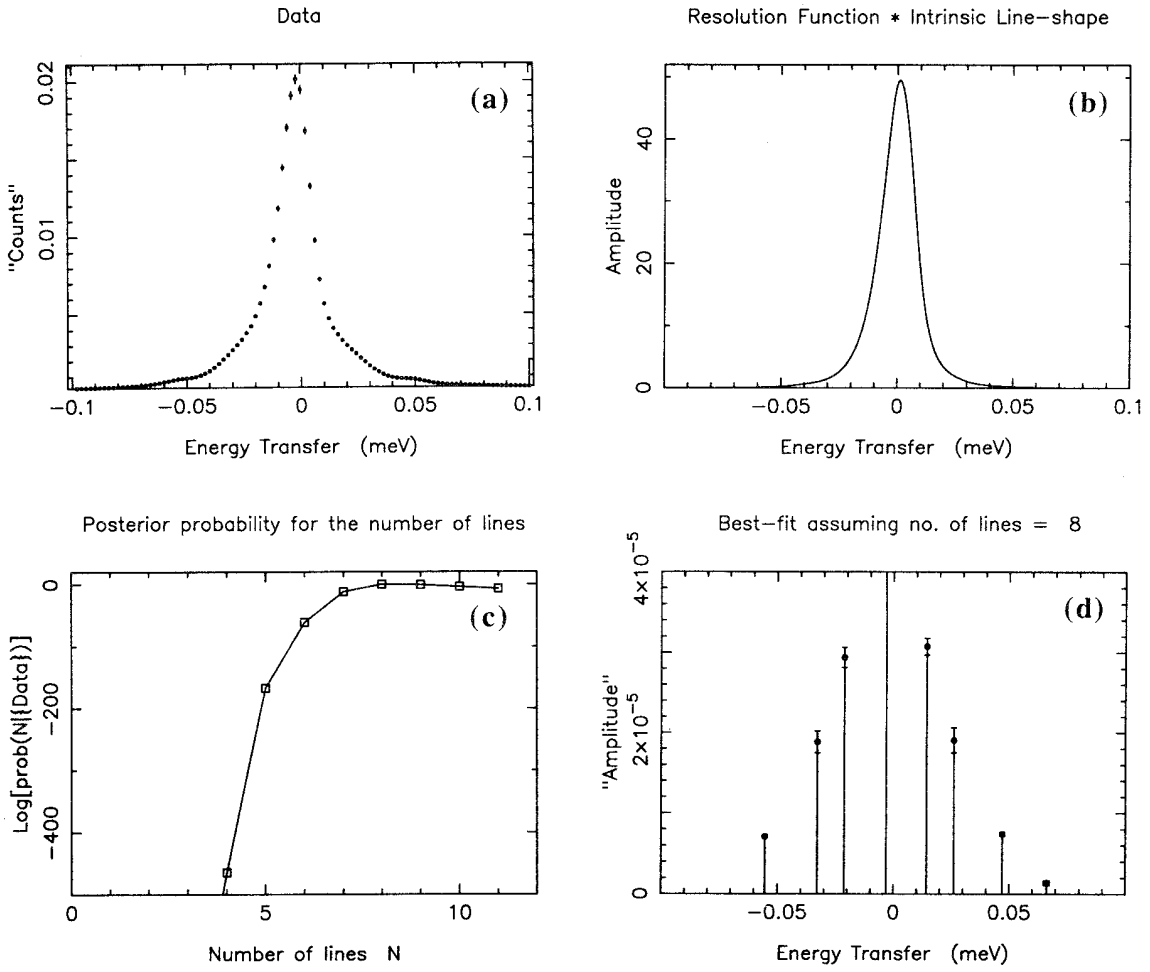

FIG. 2. (a) Data for lutidine taken with the (002) graphite crystal analyser on the IRIS spectrometer at ISIS. (b) The "resultant" resolution function, which includes the broadening contributions of both the instrumental resolution function and (the optimal estimate of) the intrinsic width of the excitation lines. (c) The posterior probability for the number of excitation lines, on a logarithmic scale (to base 10). (d) The best estimate of the amplitudes and positions of the (optimal number of) excitation lines, with their 1-σ error bars.

hyay *et al.*[23] using MaxEnt, who show a threefold improvement in the inferred detail over the instrumental resolution. Our results are consistent with theirs, but have the advantage that they can be presented in a more clean and precise way. This improvement does, of course, stem from the greater prior knowledge about the spectrum which is encoded in the Bayesian spectral analysis formulation of the problem (as compared with MaxEnt).

Figure 2(a) shows the lutidine data, from IRIS, using the pyrolytic graphite (002) crystal analyzer reflection. The resolution function was determined from a measurement of the elastic line from vanadium, and was roughly symmetrical with a FWHM of about 15 μeV. We used MaxEnt, with an *intrinsic correlation function*,[14,24,25] to fit a smooth (but nonparametric) line through the vanadium data, in order to tabulate the resolution function on a fine grid for subsequent numerical manipulation in the Bayesian spectral analysis. The "resultant" resolution function, including the contribution from the optimal estimate of the intrinsic width of the excitation lines, is given in Fig. 2(b). The logarithm of the posterior PDF for the number of lines is shown in Fig. 2(c), and indicates most evidence for eight lines. The best estimate for the amplitudes and positions for these eight lines, with their 1-σ error bars, is shown in Fig. 2(d). It should be noted that the spectrum in Fig. 2(d) is shifted to the left by about 3 μeV; this is because no attempt was made to center the data, or the resolution function, to compensate for any experimental misalignment.

The data in Fig. 3(a), also for lutidine, are from IRIS using the mica (004) analyzer. The resolution function for mica (004) is much narrower than for graphite (002), having a FWHM of only ~ 5 μeV. We can now clearly see evidence for two pairs of excitation lines which were indicated by the analysis of the data in Fig. 2(a), but were not visible to the naked eye. Analyzing the data in Fig. 3(a), as before, we obtain the "resultant" resolution function of Fig. 3(b), the (logarithm of the) posterior PDF for the number of lines in Fig. 3(c) and the best estimate of their amplitudes and positions (with their 1-σ error bars) shown in Fig. 3(d). The substructure of the spectrum indicated by the analysis, over the visible peaks in Fig. 3(a), is again confirmed by the higher-resolution measurements shown in Fig. 4(a). These data were taken on the IN10 spectrometer and are the highest resolution measurements of lutidine to date, with a FWHM

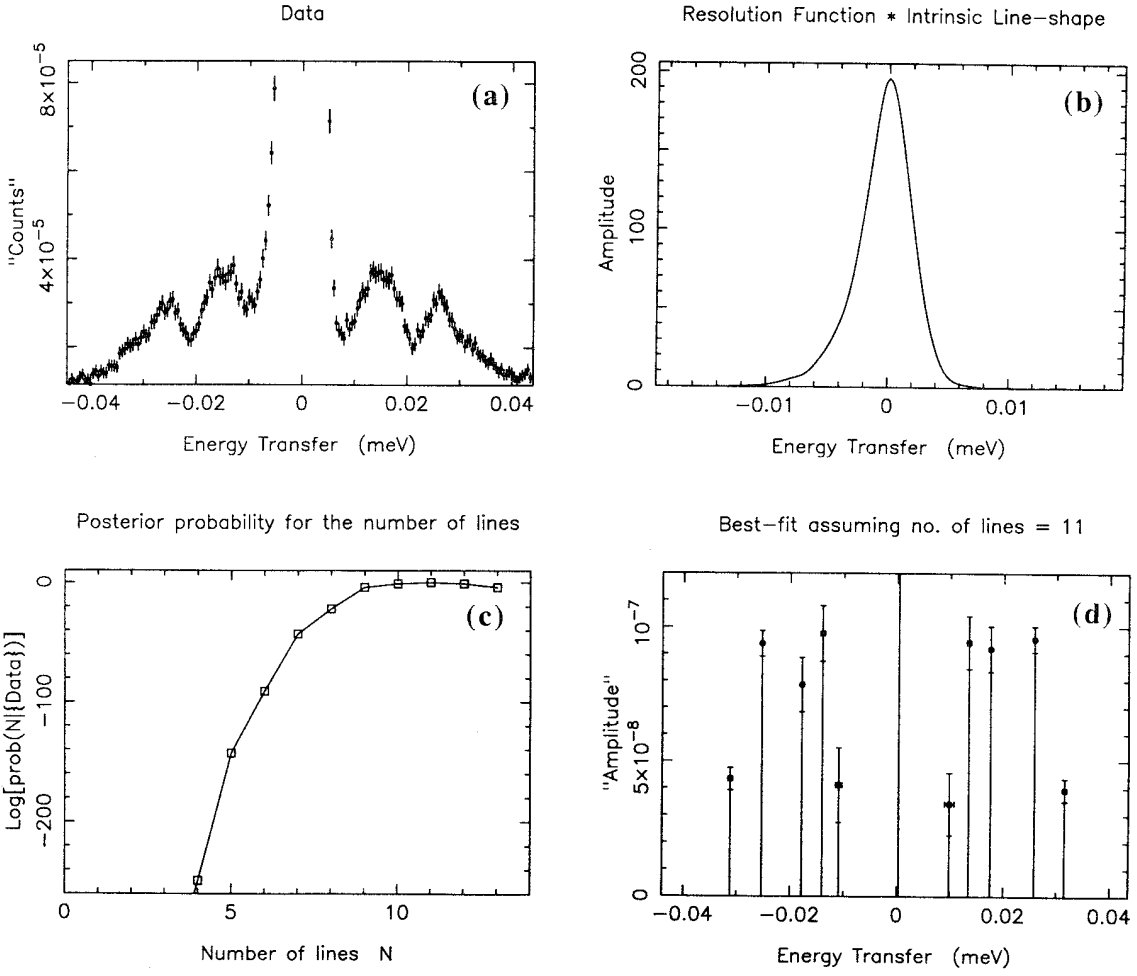

FIG. 3. (a) Data for lutidine taken with the (004) mica crystal analyser on the IRIS spectrometer at ISIS. (b) The "resultant" resolution function. (c) The (logarithm of the) posterior probability for the number of excitation lines. (d) The best estimate of the amplitudes and positions of the excitation lines, with their 1-σ error bars.

of less than 1.5 μeV. Figure 4 shows that an analysis of these (very noisy) data yields evidence for yet more substructure in the tunneling spectrum for lutidine, if we fix the intrinsic width of all the excitation lines to be that which is optimal for the narrow and isolated peak at about 9 μeV.

Before finishing with the conclusions, we should make some additional remarks about systematic uncertainties. In Sec. II C we indicated how these could be dealt with by probability theory, in principle. In practice, however, this is often difficult because we are unable to characterize the systematic uncertainties in terms of a few parameters. For example, the analysis can deal with an unknown linear background in a straightforward manner; the more awkward question is, "how well do we know it's linear?" To start to answer that question in a formal way requires us to present a set of possible alternatives for comparison, and the problem rapidly becomes intractable in a finite amount of time. Our philosophy is to make simplifying, but adequately correct, assumptions to enable us to solve the problem, but to remember that the results are always conditional on the validity of the assumptions. We might, then, include systematic uncertainties into our analysis in a pragmatic way by being conservative in our statements. For the analysis of Fig. 4, for example, we might say that, "there is strong evidence for eight lines, and some evidence for a ninth." The somewhat dubious ninth line is characterized by having the largest absolute error bar for its inferred position and the largest relative uncertainty in its estimated amplitude (and the smallest absolute amplitude). This conservatism can be crudely quantified by using slightly larger error bars for the data to compensate for the systematic errors. An optimal rescaling constant for the error bars can be estimated from the maximum of the posterior PDF prob[*rescaling constant* $|\{Data\}$], computed according to the rules in Sec. II C, and will be given roughly by that value which makes "χ^2 equal to the number of data" (when the number of lines being fit is sufficiently large).

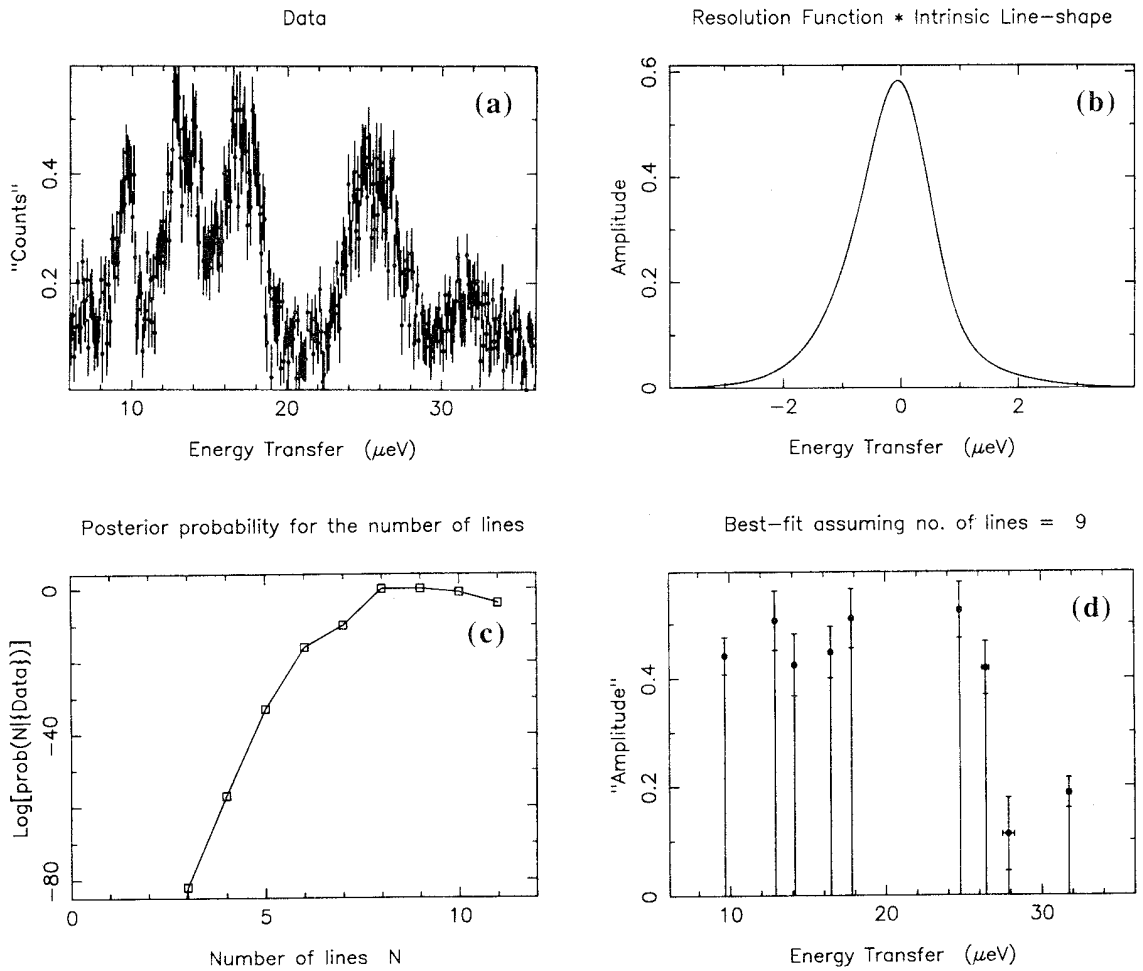

FIG. 4. (a) Data for lutidine taken on the IN10 spectrometer at the ILL, using the temperature scanning monochromator. (b) The "resultant" resolution function, where the intrinsic width of the excitations has been fixed to that value which is optimal for the narrow and isolated peak at $\sim 9\,\mu$eV in (a). (c) The (logarithm of the) posterior probability for the number of excitation lines. (d) The best estimate of the amplitudes and positions of the excitation lines, with their 1-σ error bars.

VI. CONCLUSIONS

We have demonstrated the Bayesian spectral analysis approach for analyzing neutron scattering molecular tunneling data. This method of analysis is most appropriate for molecular spectroscopy because we often know that the spectrum consists of a few discrete excitations, but we do not know how many. Probability theory enables us to answer the question "how many lines is there most evidence for in the data" in a quantitative manner, and allows us to present the results in a very clean and precise way.

We have reviewed the theory of Bayesian spectral analysis, putting it in the broader context of probabilistic data analysis, and have given details relevant to our particular application. We have described an efficient algorithm for the practical implementation of the theoretical results, and illustrated its use with both simulated and real experimental data. We believe that this powerful method of analysis will be a very useful tool in experimental molecular spectroscopy and in crystal field spectroscopy.

ACKNOWLEDGMENTS

The data shown in this paper are part of a program of work on the quantum motions of CH_3 groups carried out by M. Prager, F. Fillaux, G. J. Kearley, and C. J. Carlile. We are grateful for the use of these data prior to publication.

[1] R. T. Cox, Am. J. Phys. **14**, 1 (1946).
[2] J. Bernoulli, Ars. Conjectandi (1713).
[3] T. Bayes, Philos. Trans. R. Soc. London **53**, 370 (1763).
[4] P. S. Laplace, *Theorie Analytique des Probabilités* (Courcier, Paris, 1812).
[5] E. T. Jaynes, *Papers on Probability, Statistics, and Statistical Physics*, edited by R. D. Rosenkrantz (Reidel, Dordrecht, 1983).
[6] S. F. Gull, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, Vol. 1, edited by G. J. Erickson and C. R. Smith (Kluwer, Dordrecht, 1988).
[7] T. J. Loredo, in *Maximum Entropy and Bayesian Methods, Dartmouth 1989*, edited by P. F. Fougère (Kluwer, Dordrecht, 1990).
[8] E. T. Jaynes, in *Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice (Cambridge University, Cambridge, 1986).
[9] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 and 623 (1949).
[10] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **IT-26**, 26 (1980).
[11] Y. Tikochinsky, N. Z. Tishby, and R. D. Levine, Phys. Rev. Lett. **52**, 1357 (1984).
[12] J. Skilling, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, Vol. 1, edited by G. J. Erickson and C. R. Smith, (Kluwer, Dordrecht, 1988).
[13] J. Skilling, in *Maximum Entropy and Bayesian Methods, Cambridge 1988*, edited by J. Skilling (Kluwer, Dordrecht, 1989).
[14] S. F. Gull, in *Maximum Entropy and Bayesian Methods, Cambridge 1988*, edited by J. Skilling (Kluwer, Dordrecht, 1989).
[15] See, for example, S. F. Gull and J. Skilling, IEE Proc. **131** F, 646 (1984), and references therein.
[16] H. Jeffreys, *Theory of Probability* (Oxford University, Oxford, 1939).
[17] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation* (Springer-Verlag, New York, 1988).
[18] J. A. Nelder and R. Mead, Comput. J. **7**, 308 (1965).
[19] See, for example, W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University, Cambridge, 1986).
[20] W. M. Thorburn, Mind **27**, 345 (1918).
[21] B. C. Boland, R. A. L. Report No. 90-041 (1990).
[22] J. C. Cook, W. Petry, A. Heiderman, and B. Frick, I.L.L. Report 91C006T (1991).
[23] R. Mukhopadhyay, C. J. Carlile, and R. N. Silver, Physica B **174**, 546 (1991).
[24] J. Skilling and S. Sibisi, Inst. Phys. Conf. Ser. **107**, 1 (1990).
[25] M. K. Charter, in *Maximum Entropy and Bayesian Methods, Dartmouth 1989*, edited by P. F. Fougère (Kluwer, Dordrecht, 1990).