in Maximum Entropy and Bayesian Methods, pp. 327-335, W. T. Grandy,
and L. H. Schick (eds.) Kluwer Academic Publishers, The Netherlands. (1991)

MOMENT ESTIMATION USING BAYESIAN PROBABILITY THEORY

G. Larry Bretthorst
Washington University
Department of Chemistry
1 Brookings Drive
St. Louis, Missouri 63130

ABSTRACT. In NMR the response of a system of spins is described by a spectral density function $G(\omega)$. Typically, only the moments of this function can be computed from first principles. NMR data are typically sampled in the time domain. In the time domain, these moments are proportional to the derivatives of the Fourier transform of $G(\omega)$, evaluated at time $t = 0$. When comparing theory to experiment, good estimates of the moments are needed. Good estimates are difficult to obtain, because procedures like least-squares and maximum-likelihood do not tell what the data have to say about a particular moment; rather, they give information about all of the moments. In this paper, a Bayesian calculation of the probability for a given moment is presented, and an example of the calculation is given.

1. Introduction

In the moment problem, there is frequency domain spectral density function $G(\omega)$. Some of the moments of $G(\omega)$ have been computed from first principles. The problem is to compare these moments to experiment. One way to do this is to compute the posterior probability density for the moments from the experimental data, and then compare these with the theoretically calculated moments. In NMR the data are typically gathered in the time domain, not the frequency domain. But if $S(t)$, called an autocorrelation function, is the Fourier transform of $G(\omega)$, then the moments of $G(\omega)$ are simply related to the time derivatives of the autocorrelation function:

$$M_k \equiv \left.\frac{d^k S(t)}{dt^k}\right|_{t=0} \propto \int_0^\infty \omega^k G(\omega) d\omega \qquad (k = 0, 1, \cdots, \text{etc.}) \tag{1}$$

where the proportionality constant is the same for all the moments. Estimating the derivatives of $S(t)$ is equivalent to estimating the moments of $G(\omega)$.

To determine the moments, suppose one has a time series which has been sampled at discrete times t_i $(1 \leq i \leq N)$. These discrete samples constitute the data D. The data are known to contains a signal, the autocorrelation function $S(t)$, plus additive noise $e(t)$. The problem is to make the best estimate of the moment, M_k, given the data D and one's prior information I. The question one would like to answer is "What is the best estimate of the

kth moment one can make given the data D and one's prior information I, independent of the functional form of the signal $S(t)$?" Of course, it is not possible to compute this probability density function completely independent of the functional form of the signal; rather one must compute the posterior probability density for the moment, taking into account all of the structure implicit in the data, without expanding the noise.

To make this more specific let $d(t_i)$ be a data item sampled at time t_i, then the data are given formally by the sum of the autocorrelation function $S(t_i)$ plus noise $e(t_i)$:

$$d(t_i) = S(t_i) + e(t_i). \tag{2}$$

Expanding the autocorrelation function, $S(t)$, in a Taylor series around $t = 0$, one obtains

$$S(t) = \sum_{k=0}^{E} \left.\frac{d^k S(t)}{dt^k}\right|_{t=0} \left(\frac{t^k}{k!}\right) = \sum_{k=0}^{E} \frac{M_k t^k}{k!} \tag{3}$$

where E is the expansion order. The expansion order is assumed unknown, but it can be bounded $(0 \leq E \leq N-1)$. The discrete times, t_i, are assumed in the range $(0 \leq t_i \leq 1)$.

When using a Taylor series expansion the moments, M_k, appear in the problem in a very natural way. But any series expansion that constitutes a complete set of functions could be used here. The moments, M_k, may be expressed analytically in this new expansion and the general details of the calculation will remain essentially unchanged. Indeed, a set of functions that are nearly orthogonal on the discrete time values t_i would be preferred, because it would stabilize the resulting numerical calculation. The only reason it is not done here, is that it complicates the mathematics without adding anything new.

Note that in NMR the data are often sampled in such a way that sampling times start at $t_1 > 0$. This is unimportant; whether or not the data are sampled at time $t = 0$ does not affect these equations. The estimated moment will still be the best estimate of the moment one can obtain from the data and one's prior information.

When using probability theory, all of the information in the data relevant to the problem of estimating M_k is summarized in a probability density function. This function is denoted as $P(M_k|D,I)$, where this should be read as the posterior probability for the kth moment, M_k, given the data D and the prior information I. The prior information I is simply a collection of hypotheses that will not be questioned in this calculation. Here this prior information includes the use of a Taylor series expansion, and the separation of the data into a signal plus additive noise. The symbol M_k appearing in $P(M_k|D,I)$ is an index (loosely called a parameter) which denotes a continuum of hypotheses. The problem is to select the most probable hypothesis, corresponding to some numerical value of the index M_k; the problem is typically referred to as a parameter estimation problem.

Two rules from probability theory are used in this calculation: the "sum rule" and the "product rule." The specific form of the "sum rule" needed states that the probability for a series of mutually exclusive hypotheses is just the sum of the probabilities for each of the individual hypotheses. If the joint posterior probability for the moment, M_k, and the expansion order E [denoted as $P(M_k, E|D,I)$] can be computed, then the problem is solved; because the sum rule indicates that the desired probability density function is given by

$$P(M_k|D,I) = \sum_{E=k}^{N-1} P(M_k, E|D,I), \tag{4}$$

where the expansion order, E, has been bounded ($k \leq E \leq N-1$). Probability theory as logic tells one that the probability for moment, M_k, is a sum over the joint probability for the moment and the expansion order. Thus when one computes the probability for the moment, independent of the expansion order, probability theory does not magically get rid of the functional dependence on the expansion order; rather it takes all possible values of the expansion order into account.

The "product rule" of probability theory is next applied to the joint posterior probability for the moment and the expansion order, $P(M_k, E|D, I)$. The product rule states that given three hypothesis, A, B, and C the probability of A and B given C is given by:

$$P(A, B|C) = P(A|C)P(B|A, C) = P(B|C)P(A|B, C). \tag{5}$$

Applying this rule to the joint posterior probability of the moment and the expansion order, $P(M_k, E|D, I)$, one obtains

$$P(M_k, E|D, I) = P(M_k|E, D, I)P(E|D, I), \tag{6}$$

where $P(M_k|E, D, I)$ is the marginal posterior probability for the kth moment given the data D, the expansion order E, and the assumed prior information I. The other term, $P(E|D, I)$, is the marginal posterior probability for the expansion order. Substituting the factored probability density function, Eq. (6), into the probability of the moment, Eq. (4) one obtains:

$$P(M_k|D, I) = \sum_{E=k}^{N-1} P(E|D, I)P(M_k|E, D, I) \tag{7}$$

as the marginal posterior probability density for the moment, M_k, given the data and the prior information I.

The two terms, $P(M_k|E, D, I)$ and $P(E|D, I)$, each represent a straightforward calculation using probability theory. The posterior probability for the moment, $P(M_k|E, D, I)$, is a parameter estimation problem. This probability density function represent the relative ranking for a continuum of hypotheses given that one knows the expansion order. While the posterior probability for the expansion order, $P(E|D, I)$, represents the relative ranking of the Taylor series expansions and selects a "best" model. Thus the problem of estimating the moment, M_k, is a hybrid problem that combines aspects of both parameter estimation and model selection into a single problem and illustrates how probability theory estimates parameters under model uncertainty.

2. The Calculation

The problem, as stated so far, has been formulated in a very general way; but in NMR there is more prior information. For example, for high temperature spin systems only the even moments are nonzero. Additionally, all of the moments must be positive. This prior information could be used in a specialized calculation to obtain results more specific to NMR, but at the cost of restricting the usefulness of the calculation. In this calculation no restriction will be placed on the moments, except that they must be finite. The modifications to include the positive definite nature of the moments is a straightforward modification to the calculation given below. All that must be done is to modify the prior probability density function for the moments to include the lower bound, and then perform the integrals over the nuisance parameters taking into account this bound.

The two factors needed in Eq. (7) are each separately a standard probability theory calculation. The answer to the parameter estimation problem is given in (Jeffreys 1939), and (Bretthorst 1988, 1990). The answer to the model-selection problem is given in (Bretthorst 1988, 1990), and in (Gull 1988). The results from these calculations are used here.

The marginal posterior probability of the kth moment, $P(M_k|E,D,I)$, is given by:

$$P(M_k|E,D,I) = \Gamma(\frac{E}{2})\Gamma(\frac{N-E}{2})\left[\frac{E\overline{h^2}}{2}\right]^{-\frac{E}{2}}\left[\frac{N\overline{d^2}-E\overline{h^2}}{2}\right]^{\frac{E-N}{2}} \tag{8}$$

where

$$\overline{d^2} \equiv \frac{1}{N}\sum_{i=1}^{N}\left[d(t_i)-\frac{M_k t_i^k}{k!}\right]^2, \tag{9}$$

$$\overline{h^2} \equiv \frac{1}{E}\left[\sum_{m=1}^{E}2\hat{B}_m U_m - \sum_{m=1}^{E}\sum_{l=1}^{E}\hat{B}_m\hat{B}_l g_{ml}\right]. \tag{10}$$

Defining the functions $G_l(t)$ as

$$G_l(t) \equiv \begin{cases} \dfrac{t^{l-1}}{(l-1)!} & (1 \le l < k+1) \\ \dfrac{t^l}{l!} & (k+1 \le l \le E) \end{cases} \tag{11}$$

then the matrix g_{ml} is given by

$$g_{ml} = \sum_{i=1}^{N} G_m(t_i)G_l(t_i). \tag{12}$$

The functions $U_m(M_k)$ are given by

$$U_m = \sum_{i=1}^{N}\left[d(t_i)-\frac{M_k t_i^k}{k!}\right]G_m(t_i) \tag{13}$$

and last the $\hat{B}_l$ are given by the solution to

$$\sum_{m=1}^{E}\hat{B}_m g_{ml} = U_l. \tag{14}$$

The formulation given here is equivalent to that given in (Bretthorst 1988); the formulation given in (Bretthorst 1988) is useful in interpreting the equations, while the one given here is computationally more efficient.

The extra terms in Eq. (8), $\Gamma(E/2)$, $\Gamma(N-E/2)$, and $[E\overline{h^2}/2]^{-E/2}$, are due to using fully normalized prior probabilities in the calculation. In the original work (Bretthorst

1988), this was not done because, in parameter estimation problems (when the model is known), these factors always cancel. This is not the case here, and fully normalized prior probabilities must be used.

There are a number of assumptions that must be met in order for these equations to be valid. The major assumption is that, all models contain at least one moment that was removed as a nuisance parameter (a parameter is referred to as a nuisance when the desired probability density function is to be formulated independent of this parameter). For this assumption to be met the Taylor series expansion must contain at least a constant plus a linear term. Making this assumption the smallest value of of the expansion order, E, is one and the Eqs. (8–14) are valid.

The answer to the model-selection problem is given in (Bretthorst 1988 and 1990) for a very general class of models and for linear models in (Gull 1988). Again the results of these calculations are quoted here. The posterior probability for the expansion order, E, given the data and the prior information is given by:

$$P(E|D,I) = \Gamma(\frac{\mathbf{r}}{2})\Gamma(\frac{N-\mathbf{r}}{2}) \left[\frac{\mathbf{r}\overline{\mathbf{h}^2}}{2}\right]^{-\frac{\mathbf{r}}{2}} \left[\frac{N\overline{\mathbf{d}^2}-\mathbf{r}\overline{\mathbf{h}^2}}{2}\right]^{\frac{\mathbf{r}-N}{2}} \tag{15}$$

where

$$\overline{\mathbf{d}^2} \equiv \frac{1}{N}\sum_{i=1}^{N} d(t_i)^2, \tag{16}$$

$$\overline{\mathbf{h}^2} \equiv \frac{1}{\mathbf{r}}\left[\sum_{m=1}^{\mathbf{r}} 2\hat{\mathbf{B}}_m\mathbf{U}_m - \sum_{m=1}^{\mathbf{r}}\sum_{l=1}^{\mathbf{r}} \hat{\mathbf{B}}_m\hat{\mathbf{B}}_l\mathbf{g}_{ml}\right]. \tag{17}$$

Defining the functions $\mathbf{G}_l(t)$ as

$$\mathbf{G}_l(t) \equiv \frac{t^{l-1}}{(l-1)!} \qquad (1 \leq l \leq E) \tag{18}$$

then matrix $\mathbf{g}_{ml}$ is given by

$$\mathbf{g}_{ml} = \sum_{i=1}^{N} \mathbf{G}_m(t_i)\mathbf{G}_l(t_i) \tag{19}$$

and $\mathbf{r} = E + 1$. The numbers $\mathbf{U}_m$ are given by

$$\mathbf{U}_m = \sum_{i=1}^{N} d(t_i)\mathbf{G}_m(t_i) \tag{20}$$

and last the $\hat{\mathbf{B}}_l$ are given by the solution to

$$\sum_{m=1}^{r} \hat{\mathbf{B}}_m\mathbf{g}_{ml} = \mathbf{U}_l. \tag{21}$$

Note the use of boldface type in the posterior probability for the expansion order, E, is just a warning that the definitions of these quantities are slightly different between the parameter-estimation calculation Eqs. (8–14) and the model-selection calculation Eqs. (15–21).

3. Example

In this example, the calculation presented in the previous sections will be applied to computer generated data. Computer generated data will be used for no other reason than, knowing the answer, it allows one to check on the accuracy of the calculation. In this example the data will be generated from

$$d(t_i) = 50 - 40t_i + \frac{5t_i^2}{2} + \frac{t_i^3}{6} - \frac{15t_i^4}{24} + e(t_i) \quad (1 \leq i \leq 1000) \tag{22}$$

where the sampling times, t_i, ranged from zero to one by uniform increments. There are $N = 1000$ data values, and $e(t_i)$ was generated from a random number generator (Press, 1986) with zero mean and standard deviation 0.01. These data are shown in Fig. 1(A). Note that to the eye these data appear to be linear, but this is clearly not the case. Also note that while the data contain noise, it contains very little noise. The reason for this relates to the fact that the width of the probability density function for the moments is inversely related to the signal-to-noise ratio and inversely related to the square root of the number of data values: to get a good estimates for a moment one must have very high signal-to-noise ratio or a large amount of data.

After generating the data, the posterior probability density for moments 0 through 4 were computed. These probability density functions are displayed in Fig. 1(B) through 1(F). The peak value and the widths of each of the probability density functions were then estimated:

$$M_0 = 50.0018 \pm 0.0014, \tag{23}$$

$$M_1 = 40.018 \pm 0.018, \tag{24}$$

$$M_2 = 5.12 \pm 0.16, \tag{25}$$

$$M_3 = 0.6 \pm 0.7, \tag{26}$$

and

$$M_4 = -14 \pm 1.4 \tag{27}$$

at one standard deviation. Note that the higher moments are estimated much more poorly than the lower moments: very roughly the uncertainty increases by an order of magnitude for each higher moment.

To perform this calculation, in principle, one should do all expansion orders out to $N - 1$. This is unnecessary, because eventually $P(E|D, I)$ will begin to decrease. When $P(E|D, I)$ has become negligibly small (compared to the maximum) one can stop the calculation safely. Figure 2(A) is a plot of the base 10 logarithm of $P(E|D, I)$. Note, that this probability distribution rises very rapidly and then, on this logarithmic scale, becomes very flat. The calculation was stoped after the probability for a 7th order Taylor expansion was computed. The probability for the 7th order Taylor series expansion was more than 30 orders of magnitude less probable than the 4th order Taylor expansion, Fig. 2(B).

The Taylor series expansion used in this demonstration is not a suitable way to implement this calculation from a computational point of view. The matrix g_{ml}, represented by

Figure 1: Numerical Example

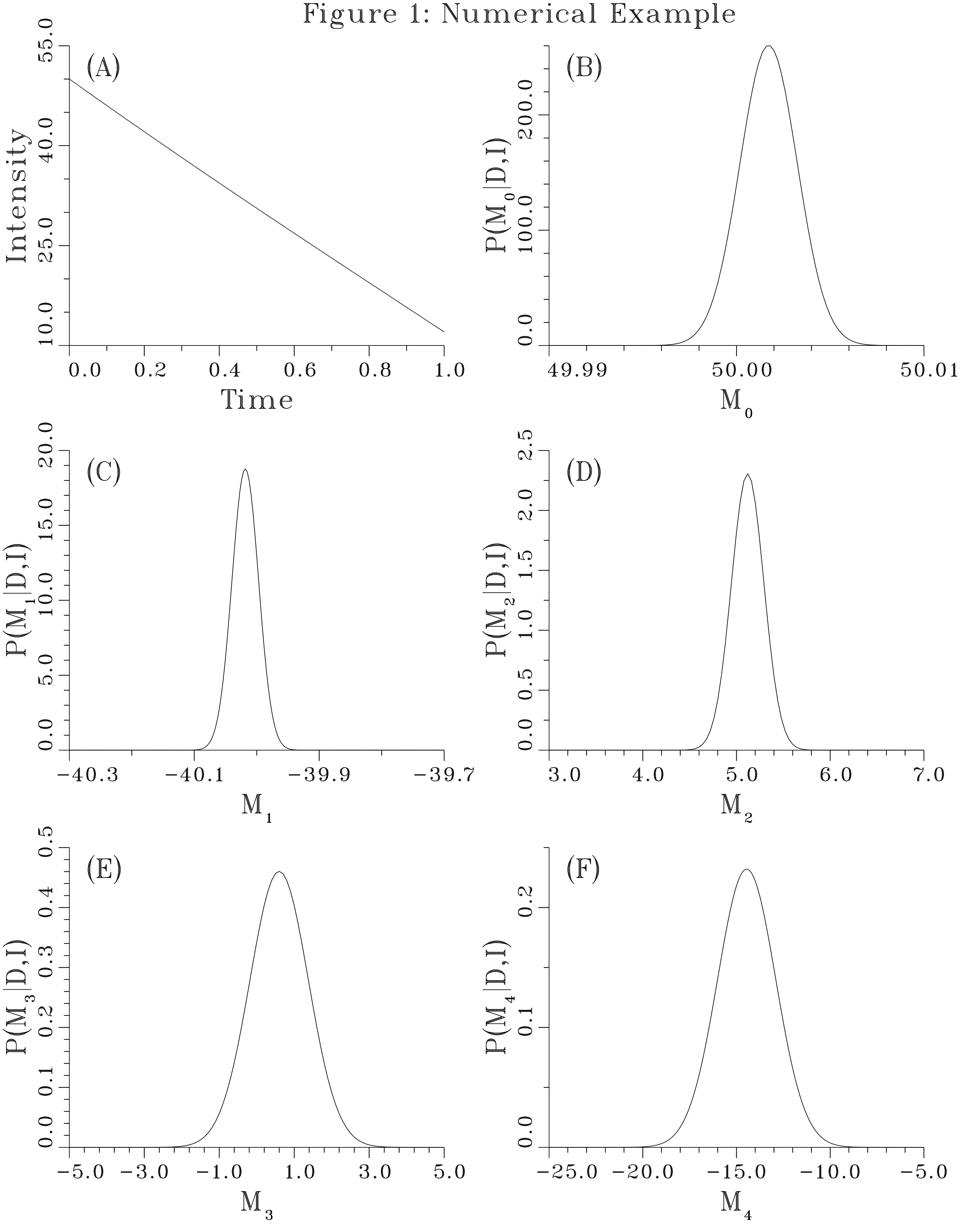
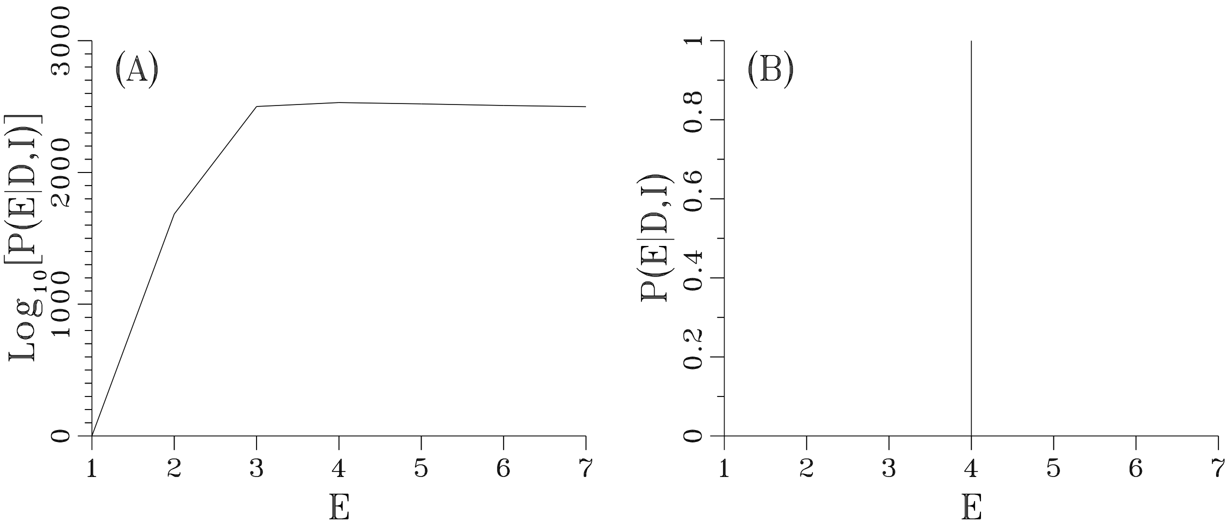


Figure 2: Probability Of The Expansion Order



Panel (A) is a plot of the base 10 logarithm of the posterior probability of the expansion order. The normalization of this plot was set so that the probability for expansion order one was zero. The posterior probability density rises 2530 orders of magnitude and reaches a peak at expansion order 4. The probability of this model is 1 to 10 decimal places, panel (B).

Eq. (19), is computationally nearly singular for large values of E. This problem is easily corrected by taking an expansion that is nearly orthogonal on the discrete sampling times t_i. Many types of expansions have this property, for example a Fourier series is exactly orthogonal, Legendre polynomials are orthogonal on the continuous interval $(0 \leq x \leq 1)$ and the Legendre polynomials evaluated at the discrete time $(0 \leq t_i \leq 1)$ would be nearly orthogonal. These functional representations were not used, because they hide much of the essential simplicity of the problem without adding to one's understanding in any essential way.

4. Conclusions

Probability theory as logic indicates that the answer to the moment estimation problem is a weighted sum. The sum is over the posterior probability for the moment given that one knows the expansion order. The weights are just the posterior probability that this expansion is the "best," i.e., most probable expansion of the data.

ACKNOWLEDGMENTS. The encouragement of Professor J. J. H. Ackerman is greatly appreciated, as extensive conversations with Professor E. T. Jaynes. This work was partially supported by a gift from the Monsanto Company.

REFERENCES

Bretthorst, G. Larry, "Bayesian Spectrum Analysis and Parameter Estimation," in *Lecture Notes in Statistics* **48,** Springer-Verlag, New York, New York, 1988.

Bretthorst, G. Larry, (1990) "Bayesian Analysis I: Parameter Estimation Using Quadrature NMR Models," *J. Magn. Reson.,* **88,** pp. 533-551.

Bretthorst, G. Larry, (1990) "Bayesian Analysis II: Model Selection," *J. Magn. Reson.,* **88,** pp. 552-570.

Gull, S. F., "Bayesian Inductive Inference and Maximum Entropy," in *Maximum Entropy and Bayesian Methods in Science and Engineering"* **1,** pp. 53-75, G. J. Erickson and C. R. Smith *eds.,* Kluwer Academic Publishers, Dordrecht the Netherlands, 1988.

Jaynes, E. T., "How Does the Brain do Plausible Reasoning?" unpublished Stanford University Microwave Laboratory Report No. 421 (1957); reprinted in *Maximum-Entropy and Bayesian Methods in Science and Engineering* **1,** pp. 1-24, G. J. Erickson and C. R. Smith *eds.,* Kluwer Academic Publishers, Dordrecht the Netherlands, 1988.

Jaynes, E. T., "Probability Theory – The Logic of Science," in preparation. Copies of this manuscript are available from E. T. Jaynes, Washington University, Dept. of Physics, St. Louis, MO 63130.

Jeffreys, H., *Theory of Probability,* Oxford University Press, London, 1939; Later editions, 1948, 1961.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes," Cambridge, 1986.