The Maximum Entropy Method Of Moments And Bayesian Probability Theory

G. Larry Bretthorst

Department of Radiology, Washington University, St. Louis, MO 63110,
gbretthorst@wustl.edu,   http://bayes.wustl.edu

Abstract. The problem of density estimation occurs in many disciplines. For example, in MRI it is often necessary to classify the types of tissues in an image. To perform this classification one must first identify the characteristics of the tissues to be classified. These characteristics might be the intensity of a T1 weighted image and in MRI many other types of characteristic weightings (classifiers) may be generated. In a given tissue type there is no single intensity that characterizes the tissue, rather there is a distribution of intensities. Often this distributions can be characterized by a Gaussian, but just as often it is much more complicated. Either way, estimating the distribution of intensities is an inference problem. In the case of a Gaussian distribution, one must estimate the mean and standard deviation. However, in the Non-Gaussian case the shape of the density function itself must be inferred. Three common techniques for estimating density functions are binned histograms [1, 2], kernel density estimation [3, 4], and the maximum entropy method of moments [5, 6]. In following section, the maximum entropy method of moments will be reviewed. Some of its problems and conditions under which it fails will be discussed. Then in later sections, the functional form of the maximum entropy method of moments probability distribution will be incorporated into Bayesian probability theory and it will be shown that all of the technical problems with the maximum entropy method of moments disappear, and that one gets posterior probabilities for the parameters appearing in the problem as well as error bars on the resulting density function.

Keywords: density estimation, maximum entropy method of moments, Bayesian probability theory
PACS: 02.70.Uu

INTRODUCTION

In the problem being formulated, one has a data set consisting of samples drawn from an unknown density function. Figure 1 displays an illustrative set of such data samples, these data samples (gray circles) were generated in a Markov chain Monte Carlo simulation; although the source of the data samples is unimportant for the problem considered here. The horizontal axis is sample number and the vertical axis is the sample value. There are 2500 samples shown in this figure. The problem is to estimate both the density function and the uncertainty in the estimated density function. Often such data samples can be characterized by a Gaussian density function, but just as often the density function is more complicated. Either way, estimating the distribution of intensities is an inference problem. In the case of the Gaussian, one must estimate both the mean and standard deviation. However, in the Non-Gaussian case, the shape of the density function itself must be inferred. Three common techniques for estimating density functions are binned histograms [1, 2], kernel density estimation [3, 4], and the maximum entropy method of moments [5, 6]. In the following section, the maximum entropy method of

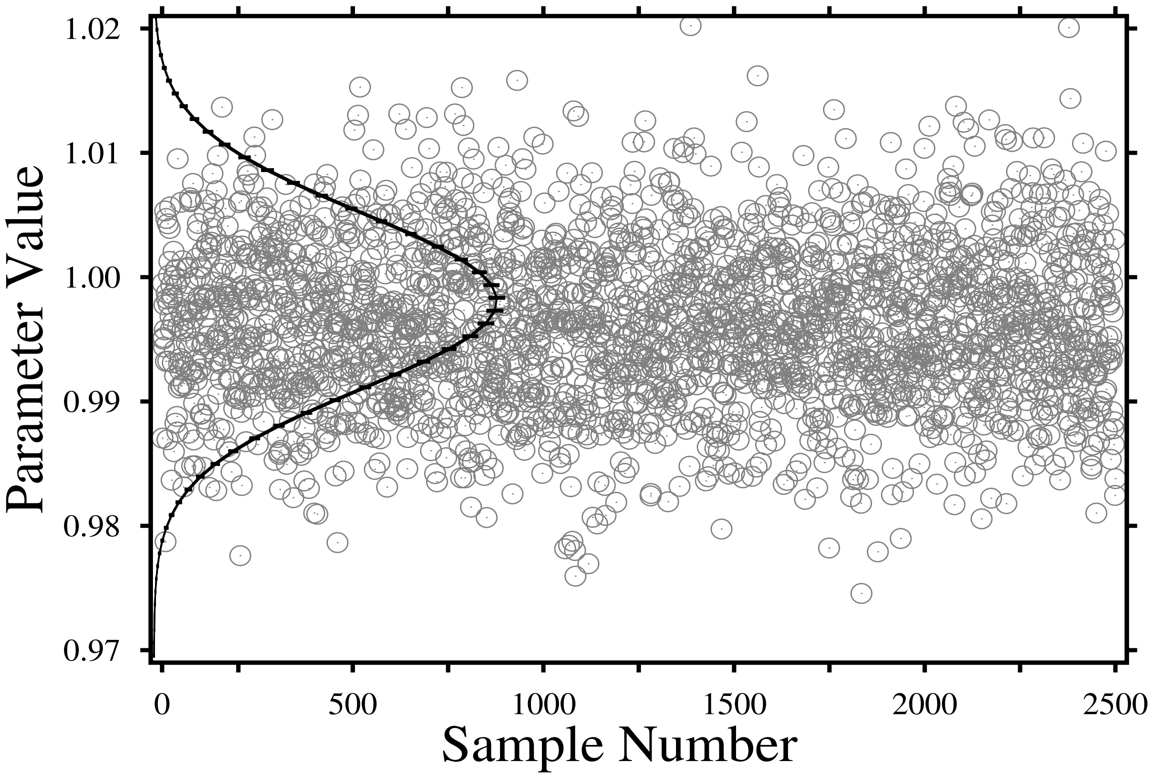

FIGURE 1. In the density estimation problem addressed here, one as a set of samples (open circles) drawn from some unknown density function and one wishes to infer the distribution of the samples (solid line with error bars). This density function was estimated using Bayesian probability theory to determine what probabilities must be assigned. The maximum entropy method of moments was then used to assign the indicated probabilities. Finally, a Markov chain Monte Carlo simulation was used to draw samples from the posterior probability for the density function, see the text for the details.

moments will be reviewed and some of its problems and conditions under which it fails will be discussed. Then in the following sections, the functional form of the maximum entropy method of moments probability distribution will be incorporated into Bayesian probability theory, and it will be shown that all of the technical problems encountered with the maximum entropy method of moments disappear, and that, as a Bayesian calculations, one gets posterior probabilities for the parameters appearing in the problem as well as error bars on the resulting density function. The solid line in Fig. 1 is an example of the estimated density function with error bars generated using the techniques and procedures described in this paper.

THE MAXIMUM ENTROPY METHOD OF MOMENTS

Claude Shannon [5] derived the Shannon entropy as a measure of the information content of a discrete probability distribution. If this discrete probability distribution is represented by f_j, then the Shannon entropy, S, is given by

$$S = -\sum_{j=1}^{n} f_j \log(f_j) \qquad (0 \leq f_j \leq 1) \tag{1}$$

where n is the number of discrete probabilities in the distribution. The entropy S is a measure of the information content of a probability distribution. It reaches its maximum value when all $f_j = 1/n$ and $S = \log(n)$, and it reaches its minimum value when one of the $f_j = 1$, and then $S = 0$. Thus the Shannon entropy maps discrete probability distributions onto the interval $0 \leq S \leq \log(n)$, with $S = \log(n)$ the completely uninformative state, and $S = 0$ the state of certainty. Everything in between represents increasing knowledge for decreasing entropy.

After deriving the entropy function, Shannon proceeded to use the entropy function as a way of assigning maximally uninformative probability distributions that are consistent with some given prior information. In the maximum entropy method of moments, the Shannon entropy is constrained by the power moments. Suppose the probabilities f_j are defined on a set of discrete points x_j, then the expected value of the power moments are given by

$$\langle x^k \rangle = \sum_{j=1}^{n} x_j^k f_j \qquad (k = 0, 1, \ldots, m) \tag{2}$$

where $k = 0$ is the normalization constraint. Because this is an equality, one can move the sum to the left-hand side of the equation, and because this equation is equal to zero one can multiply through by a constant, called a Lagrange multiplier, and the equation will still be zero:

$$\lambda_k \left[\langle x^k \rangle - \sum_{j=1}^{n} x_j^k f_j \right] = 0. \tag{3}$$

Additionally, if one has more than one constraint, one can sum over the constraints and the sum is still zero:

$$\sum_{k=0}^{m} \lambda_k \left[\langle x^k \rangle - \sum_{j=1}^{n} x_j^k f_j \right] = 0. \tag{4}$$

Because this equation is zero, it can be added to the Shannon entropy without changing its value:

$$S = -\sum_{j=1}^{n} f_j \log(f_j) + \sum_{k=0}^{m} \lambda_k \left[\langle x^k \rangle - \sum_{j=1}^{n} x_j^k f_j \right]. \tag{5}$$

To assign numerical values to the f_j, Eq. (5) is maximized with respect to variations in the f_i. The resulting equations can be solved for the functional form of the probability. Taking the derivative with respect to f_i and solving, one obtains:

$$f_i = Z(m, \lambda)^{-1} \exp\left\{ \sum_{k=1}^{m} \lambda_k x_i^k \right\} \tag{6}$$

where $Z(m, \lambda)$ is a normalization constant that is a function of both the number of Lagrange multipliers and their value. This equations gives the functional form of the

Moment	Power	Central	
1	9.96127964E-01	0.00000000E+00	
2	9.92317063E-01	4.61424576E-05	
3	9.88566722E-01	1.95337184E-08	
4	9.84876378E-01	6.64319953E-09	
5	9.81245478E-01	1.06175263E-11	.
6	9.77673479E-01	1.64171891E-12	
7	9.74159848E-01	4.09097233E-15	
8	9.70704063E-01	5.63247606E-16	
9	9.67305611E-01	1.15656817E-18	
10	9.63963991E-01	2.38918537E-19	

FIGURE 2. The first 10 power and central moments computed form the samples shown in Fig. 1

maximum entropy method of moments probability distribution in terms of the Lagrange multipliers λ_j, but one must also satisfy the constraints, namely:

$$\langle x^k \rangle = \sum_{i=1}^{n} x_i^k f_i \qquad (k = 1, \ldots, m). \tag{7}$$

Equations (6) and (7) are a system of coupled nonlinear equations for the Lagrange multipliers. To solve for the values of the Lagrange multipliers that maximizes the entropy, one typically uses a Newton-Raphson [7] searching algorithm. This searching algorithm Taylor expands Eq. (5) about the current estimated values of the Lagrange multipliers to second order, and then solve for the values of the change in Lagrange multipliers that makes the derivatives go to zero. The procedure has to be iterated a few times and, when it converges, it typically converges quadratically.

To make this more concrete, suppose one computes the first 10 moments, of the samples shown in Fig. 1 and use them in a maximum entropy method of moments calculation. What would happen to the maximum entropy distribution as more and more moments are incorporated into the calculation? The first 10 moments are shown in Fig. 2. Two sets of moments are shown, the power moments and the central moments. The power moments are given by

$$\langle \text{Power Moment } k \rangle = \frac{1}{N} \sum_{i=1}^{N} d_i^k \tag{8}$$

and the central moments are given by

$$\langle \text{Central Moment } k \rangle = \frac{1}{N} \sum_{i=1}^{N} (d_i - \overline{d})^k \tag{9}$$

where N is the total number of data values and $\overline{d}$ is the mean data value.

If one incorporate the power moments one at a time into a maximum entropy method of moments calculation, the distributions shown in Fig. 3 result. The flat line marked

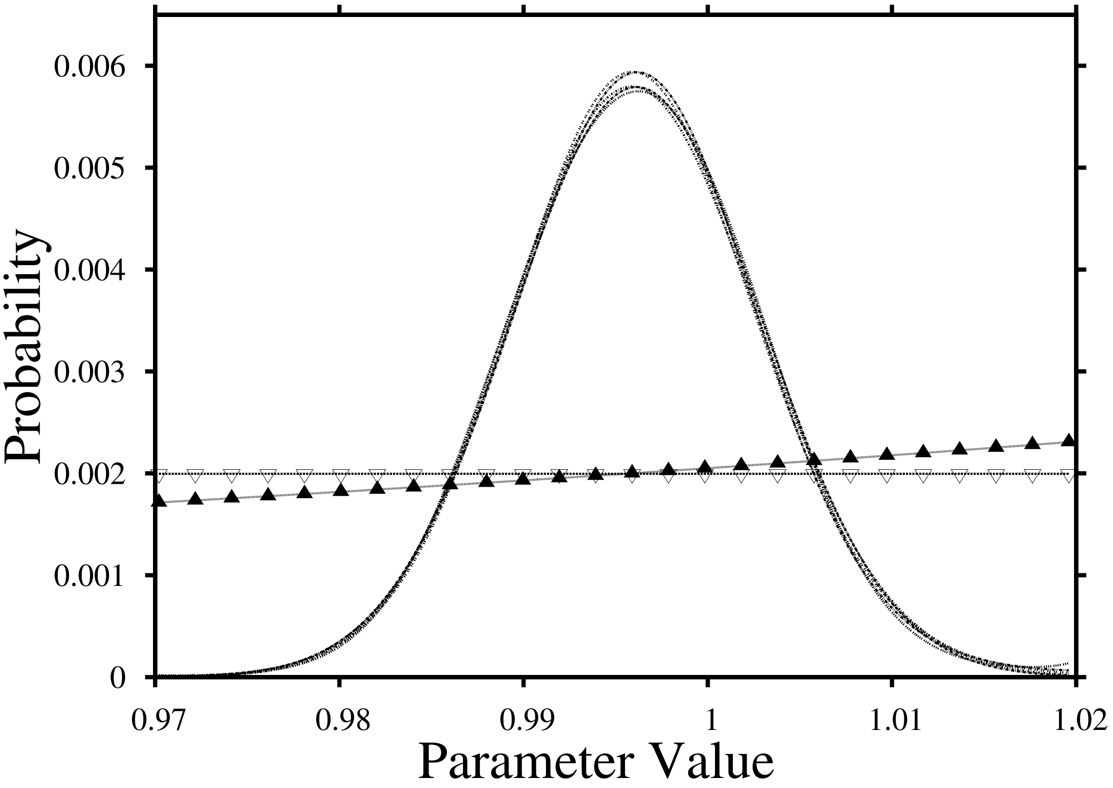

FIGURE 3. The maximum entropy moment distributions as a function of increasing numbers of moments. The flat line with open triangles is a normalized uniform density resulting from using only the zeroth moment. When the first moment is incorporated (line with closed triangles), not much changes because and exponential distribution cannot represent the distribution of samples shown in Fig. 1. However, for second and higher moments (hatched unmarked lines) all of the maximum entropy method of moments distributions closely resemble each other with only minor variations. Only density functions corresponding to moments zero through 7 are shown, the numerical algorithm failed to converge for power moments greater than 7.

with inverted triangles is the zeroth moment, i.e., a uniform distribution. The tilted line marked with closed triangles is an exponential distribution, and because the samples are far from exponentially distributed, this distribution is almost a uniform distribution. Finally, the remaining curves (solid unmarked lines) are the maximum entropy method of moments distributions correspond to power moments two through seven. Note that these distributions are nearly identical, differing only near the peak values.

In Fig. 2 a total of 10 moments were given, however, in Fig. 3 only eight maximum entropy method of moment distributions are shown, corresponding to $k = 0, 1, \ldots, 7$. The reason for this is that the numerical calculation failed to converge when more than the 7 nontrivial moments were incorporated into the calculation. This is typical of the numerical calculations used in solving maximum entropy method of moments problems. Above some number of moments, the searching algorithm fails to converge or the numerical values of the moment were incompatible and no maximum entropy solution exists, see Meed and Papanicolaou [6] for the conditions under which the maximum entropy method of moments can fail.

This completes this review of the maximum entropy method of moments. Here is a short list of some of the problems with this technique:

1. The maximum entropy method of moments did not use the data samples shown in Fig. 1; rather one must compute a number of moments from the samples and use these moments in the calculations. From a Bayesian standpoint this is a rather ad hoc thing to do and has no justification whatsoever. By its very nature, Bayesian probability theory will use the raw data; not the moments, and if the moments are needed, they will show up automatically, they won't have to be artificially forced into the problem.
2. There is no way to determine how many moments, Lagrange multipliers, are needed. That is to say, the maximum entropy method of moments has an arbitrary component to it: one must simply guess the number of moments, or simply continue adding moments until the procedure fails.
3. There is no way to consistently find the maximum entropy method of moments solution. From a finite data sample, the moments can be mutually incompatible and, consequently, no solution may exist [6]. And even if the maximum entropy solution exists, searching algorithms such as Newton-Raphson, which is commonly used on this problem [6, 7], may not be able to find it.
4. There is no way to put error bars on the Lagrange multipliers. Because the maximum entropy method of moments picks out an extremum, the question of putting error bars on the Lagrange multipliers almost does not make sense. After all, maximum entropy picks out a single point. None-the-less from a Bayesian perspective, from a finite amount of data one should be able to put error bars on the multipliers, or better yet compute the posterior probability for the Lagrange multipliers given the data and the prior information.
5. The same comments apply to the assigned density function. Because the maximum entropy method of moments picks out a singe value, there is no way to determine how uncertain one is of the estimated density function. As far as maximum entropy is concerned, there is only a single density function. But from a Bayesian standpoint, this is simply false. From a finite sample, probability theory would never pick out a single density function; rather probability theory will indicate a range of values that the density function could take on that are consistent with the available data and prior information.

When the maximum entropy method of moments works, it gives a good representation of the underlying density function that quickly converges as a function of the number of constraints (moments). However, the maximum entropy method of moments is not a Bayesian techniques: it does not use the raw data, there is no way to determine how uncertain one is of the resulting density function, and it is not uncommon for the maximum entropy method of moments to fail because the set of moments are incompatible. A true Bayesian calculation would never do any of these things. It would always give a results in terms of the calculated posterior probability distributions for the number and value of the Lagrange multipliers and it would do this even if the calculated moments are incompatible.

APPLYING BAYESIAN PROBABILITY THEORY

To resolve these difficulties, Bayesian probability theory will be applied to compute the posterior probability for the number of Lagrange multipliers. The posterior probability for the number of multipliers m given all of the data D is computed using Bayes' theorem [8]:

$$P(m|DI) = \frac{P(m|I)P(D|mI)}{P(D|I)} \tag{10}$$

where $P(m|I)$ is the prior probability for the number of Lagrange multiples, $P(D|mI)$ is a marginal direct probability for the data given the number of multipliers. Finally, $P(D|I)$ is a normalization constant and is computed using the sum and product rules of probability theory:

$$P(D|I) = \sum_{m=1}^{v} P(Dm|I) = \sum_{m=1}^{v} P(m|I)P(D|mI) \tag{11}$$

where v is some given upper limit on the number of Lagrange multipliers.

In Eq. (10), the Lagrange multipliers do not appear. Consequently, Eq. (10) is a marginal posterior probability where the Lagrange multipliers have been removed from the right-hand side using the sum rule of probability theory:

$$P(m|DI) \propto P(m|I) \int P(D\lambda_1 \cdots \lambda_m|mI) d\lambda_1 \cdots d\lambda_m \tag{12}$$

where $P(D\lambda_1 \cdots \lambda_m|mI)$ is the joint probability for all of the data $D \equiv \{d_1, \ldots, d_N\}$ and the Lagrange multipliers given the number of multipliers m and the prior information I. Note that the normalization constant has been dropped, the equal sign has been replaced by a proportionality sign, and this probability distribution will have to be normalized at the end of the calculation. Applying the product rule to the right-hand side of this equation results in:

$$P(m|DI) \propto P(m|I) \int P(\lambda_1 \cdots \lambda_m|mI)P(D|m\lambda_1 \cdots \lambda_m I) d\lambda_1 \cdots d\lambda_m. \tag{13}$$

Assuming logical independence of the data samples, the right-hand side of this equation can be factored:

$$P(m|DI) \propto P(m|I) \int P(\lambda_1 \cdots \lambda_m|mI) \prod_{i=1}^{N} P(d_i|m\lambda_1 \cdots \lambda_m I) d\lambda_1 \cdots d\lambda_m. \tag{14}$$

Finally, assuming logical independence of the Lagrange multipliers, $P(\lambda_1 \cdots \lambda_m|mI)$ may also be factored to obtain

$$P(m|DI) \propto P(m|I) \int \left[\prod_{j=1}^{m} P(\lambda_j|mI)\right] \left[\prod_{i=1}^{N} P(d_i|m\lambda_1 \cdots \lambda_m I)\right] d\lambda_1 \cdots d\lambda_m. \tag{15}$$

The direct probability for the data given the number of Lagrange multipliers and their values, $P(d_i|m\lambda_1\cdots\lambda_m I)$, is the maximum entropy method of moments probability given in Eq. (6). Substituting Eq. (6) into Eq. (15) one obtains

$$P(m|DI)\propto P(m|I)\int\left[\prod_{j=1}^{m}P(\lambda_j|mI)\right]\left[\prod_{i=1}^{N}\frac{1}{Z(m,\lambda)}\exp\left\{\sum_{k=1}^{m}\lambda_k d_i^k\right\}\right]d\lambda_1\cdots d\lambda_m \quad (16)$$

as the posterior probability for the number of Lagrange multipliers given the data and the prior information. Multiplying out the products gives

$$P(m|DI)\propto P(m|I)\int\frac{P(\lambda_1|I)\cdots P(\lambda_m|I)}{Z(m,\lambda)^N}\exp\left\{\sum_{i=1}^{N}\sum_{k=1}^{m}\lambda_k d_i^k\right\}d\lambda_1\cdots d\lambda_m, \quad (17)$$

and evaluating the sum over the data values results in

$$P(m|DI)\propto P(m|I)\int\frac{P(\lambda_1|I)\cdots P(\lambda_m|I)}{Z(m,\lambda)^N}\exp\left\{\sum_{k=1}^{m}\lambda_k N\overline{d^k}\right\}d\lambda_1\cdots d\lambda_m \quad (18)$$

where the $\overline{d^k}$ are the power moments of the samples defined in Eq. (7).

The functional form of Eq. (18) is interesting in several ways. First, the data do not appear in this equation, rather there are m power moments of the data. These power moments are called sufficient statistics. They are sufficient in that they are the only quantities needed for the inference; the data itself are irrelevant. Only maximum entropy distributions have sufficient statistics. In this case the constraint functions are simple polynomials, x^k, so the sufficient statistics are the power moments calculated using the data samples. Second, every term in the sum in Eq. (18) is of the form $\lambda_k N\overline{d^k}$, which can always be driven to infinity by choosing λ_k suitably. So one might think that this could not possibly be a well behaved probability density function. However, this is not the case because this is a fully normalized probability density function and the normalization constant is a function of both the number of Lagrange multipliers and their values. Any attempt to drive the exponent to infinity simply results in a larger normalization constant that keeps everything finite.

The only remaining steps in the calculation are to assign the prior probabilities appearing in Eq. (18) and to perform the indicated calculations. In the numerical calculations that are done, all probability assignments are discretely normalized. This discrete normalization is used to ensure that one has probability distributions, not density functions. Probability density functions can be larger than one, and because of the functional form of the posterior probability, Eq. (18), this cannot be allowed. The prior probabilities were assigned as follows. The prior probability for the number of multipliers, $P(m|I)$, was assigned using an exponential prior probability:

$$P(m|I)\propto\frac{1}{Z_m(\nu)}\exp\{-m\}\qquad(1\leq m\leq\nu) \quad (19)$$

were the ν is the upper limit on the number of moments and expresses a belief that the number of multipliers should be small rather than large. The normalization constant

$Z_m(\nu)$ was computed as

$$Z_m(\nu) = \sum_{m=1}^{\nu} \exp\{-m\} = \frac{1-\mathrm{e}^{-\nu}}{\mathrm{e}-1} \tag{20}$$

and ensures that the prior probability for the number of Lagrange multipliers is normalized and always less than one.

The prior probability for each Lagrange multiplier was assigned using a Gaussian of the form:

$$P(\lambda_j|I) \propto \frac{1}{Z_{\lambda j}} \exp\left\{-\frac{\lambda_j^2}{2\sigma_\lambda^2}\right\} \qquad (\lambda_{\mathrm{Min}} \leq \lambda_j \leq \lambda_{\mathrm{Max}}) \tag{21}$$

where σ_λ is the standard deviation of this Gaussian, λ_{Min} is the smallest value the Lagrange multipliers can take on, λ_{Max} is the largest, and $Z_{\lambda j}$ is the normalization constant for the prior probability for the jth Lagrange multiplier. The standard deviation, σ_λ, was set so that the prior decayed to 7 e-foldings at λ_{Min} and λ_{Max}. This prior probability distribution was normalized discretely. To compute this normalization constant, the prior range was divided into 500 intervals and then summed. In this sum, the kth discrete value of the jth Lagrange multiplier is given by

$$\lambda_{jk} = \lambda_{\mathrm{Min}} + d\lambda(k-1) \qquad (1 \leq k \leq 501) \tag{22}$$

with

$$d\lambda = \frac{(\lambda_{\mathrm{Max}} - \lambda_{\mathrm{Min}})}{500}. \tag{23}$$

The normalization constant was computed as

$$Z_{\lambda j} = \sum_{k=1}^{501} \exp\left\{-\frac{\lambda_{jk}^2}{2\sigma_\lambda^2}\right\}. \tag{24}$$

It is this normalization constant that is used in Eq. (21).

The last normalization constant that must be set is $Z(m,\lambda)$. This is the normalization constant associated with the maximum entropy method of moments probability density function. Again this probability density function was discretely normalize so that a probability distribution was actually used in the numerical calculations. Thus all values computed using Eq. (6) will strictly be probabilities, not probability densities. This normalization constant is computed using the range of the data samples. If the minimum and maximum data value are represented by d_{Min} and d_{Max} respectively, then

$$x_i = d_{\mathrm{Min}} + dx(k-1) \qquad (1 \leq k \leq 501) \tag{25}$$

with

$$dx = \frac{(d_{\mathrm{Max}} - d_{\mathrm{Min}})}{500} \tag{26}$$

and

$$Z(m,\lambda) = \sum_{i=1}^{501} \exp\left\{\sum_{k=1}^{m} \lambda_k x_i^k\right\}. \tag{27}$$

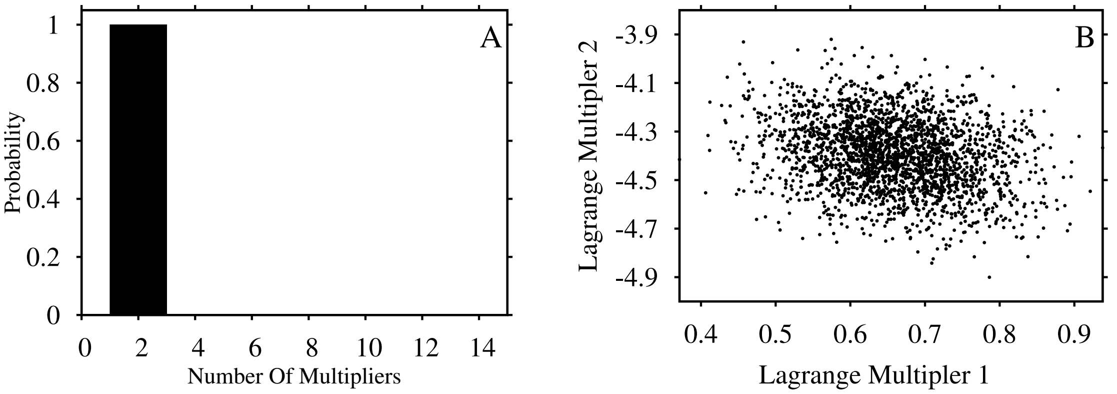

FIGURE 4. The posterior probability for the number of multipliers, Panel A, was computed from Markov chain Monte Carlo samples using the data set shown in Fig. 1. This discrete probability distribution is zero everywhere except when the number of Lagrange multipliers is two, and then the probability is one, indicating that these samples are Gaussian. After determining the number of Lagrange multipliers, the posterior probability for the two multipliers was sampled, Panel B. From these samples one can obtain Markov chain Monte Carlo samples for each Lagrange multiplier.

Again, this normalization constant ensures that Eq. (6) is a probability distribution and sums to one on the x_i. The posterior probability for m is obtained by substituting Eqs. (19, 21 and 27) into Eq (18) one obtains

$$P(m|DI) \propto \int \frac{\exp\{-m\}}{Z_m Z_\lambda^m Z(m,\lambda)^N} \exp\left\{-\sum_{j=1}^{m} \frac{\lambda_j^2}{2\sigma_\lambda^2}\right\} \exp\left\{\sum_{k=1}^{m} \lambda_k N\overline{d^k}\right\} d\lambda \tag{28}$$

where $d\lambda$ means the integral over all m Lagrange multipliers. Equation (28) is the posterior probability for the number of Lagrange multipliers.

In addition to computing the posterior probability for the number of Lagrange multipliers, the posterior probability for λ_j given the number of Lagrange multipliers and the data is also needed. However, this calculation is so similar to the one just given that it will not be repeated. Rather, note that the integrand of Eq. (28) is the joint posterior probability for all of the parameters, $P(m\lambda_1 \cdots \lambda_m|DI)$, and can be used to generate the posterior probability for any one of the Lagrange multipliers by applying the sum rule of probability theory:

$$\begin{aligned} P(\lambda_j|mDI) \quad &\propto \quad \int \frac{1}{Z_\lambda^m Z(m,\lambda)^N} \exp\left\{-\sum_{j=1}^{m} \frac{\lambda_j^2}{2\sigma_\lambda^2}\right\} \\ &\times \quad \exp\left\{\sum_{k=1}^{m} \lambda_k N\overline{d^k}\right\} d\lambda_1 \cdots d\lambda_{j-1} d\lambda_{j+1} \cdots d\lambda_m. \end{aligned} \tag{29}$$

To arrive at this result, we noted that the prior probability for the number of Lagrange is a constant when m is given and dropped it. Additionally, all the Lagrange multipliers, except λ_j, were removed using marginalization. This results in the posterior probability for the single remaining Lagrange multiplier λ_j.

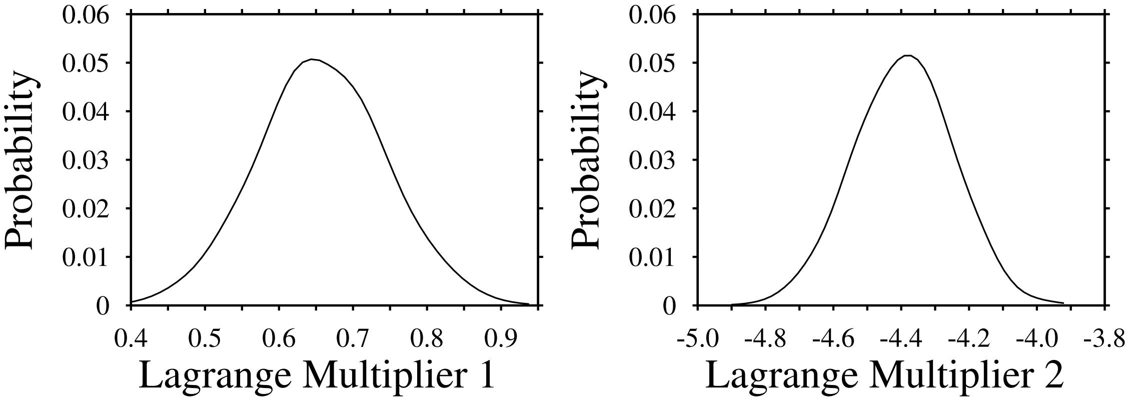

FIGURE 5. These are the posterior probabilities for the first two Lagrange multipliers. Lagrange multiplier 1 is estimated to be about 0.65 ± 0.08 and multiplier number 2 is about -4.4 ± 0.14.

A Markov chain Monte Carlo simulation with simulated annealing was used to draw samples from the integrand of Eq. (28) using the data shown in Fig. 1. In a typical run 50 simulations are run simultaneously and in parallel and 50 samples from each simulation are gathered, so there are 2500 total Markov chain Monte Carlo samples for the number of multipliers and their values. Monte Carlo integration was then used to compute the posterior probability for the number of Lagrange multipliers given the data and the prior information. The posterior probability for the number of multipliers is shown in Fig. 4(A). Note that this posterior probability indicates that only two Lagrange multipliers are needed to represent the density distribution of the data. Consequently, Bayesian probability theory strongly indicates that the data shown in Fig. 1 are Gaussianly distributed.

After determining that the number of Lagrange multipliers was two, the joint posterior probability for the two Lagrange multipliers was sampled. These Markov chain Monte Carlo samples are shown in Fig. 4(B). Each dot in this figure is one sample from one of the 2500 Markov chain Monte Carlo simulations. By using Monte Carlo integration, one can obtain samples from the posterior probability for each Lagrange multiplier, Fig. 5. These one dimensional samples can be used to compute mean and standard deviation estimates of the Lagrange multipliers. However, a means of visually displaying the samples is also desirable. A binned histogram could be used, but even with 2500 samples such histograms are often very rough. Consequently, the program that implements this calculation uses a Gaussian kernel density estimation procedure to generate its histograms.

The 51 bin histograms shown in Fig. 5 were generated using a Gaussian kernel that decays to 3 e-foldings over 6 bins. This kernel was centered on each Markov chain Monte Carlo sample and then added to the histogram by evaluating the kernel at each value of the histogram's x-axis. So each of the 2500 samples were smeared out over a 6 bin interval using the Gaussian kernel. Finally, the normalization is set so that the sum over the 51 bins was one. As can be seen from this figure, Lagrange multiplier 1 is estimated to be about 0.65 ± 0.08 and multiplier number 2 is about -4.4 ± 0.14. Note that these probability density functions for the Lagrange multipliers are not very

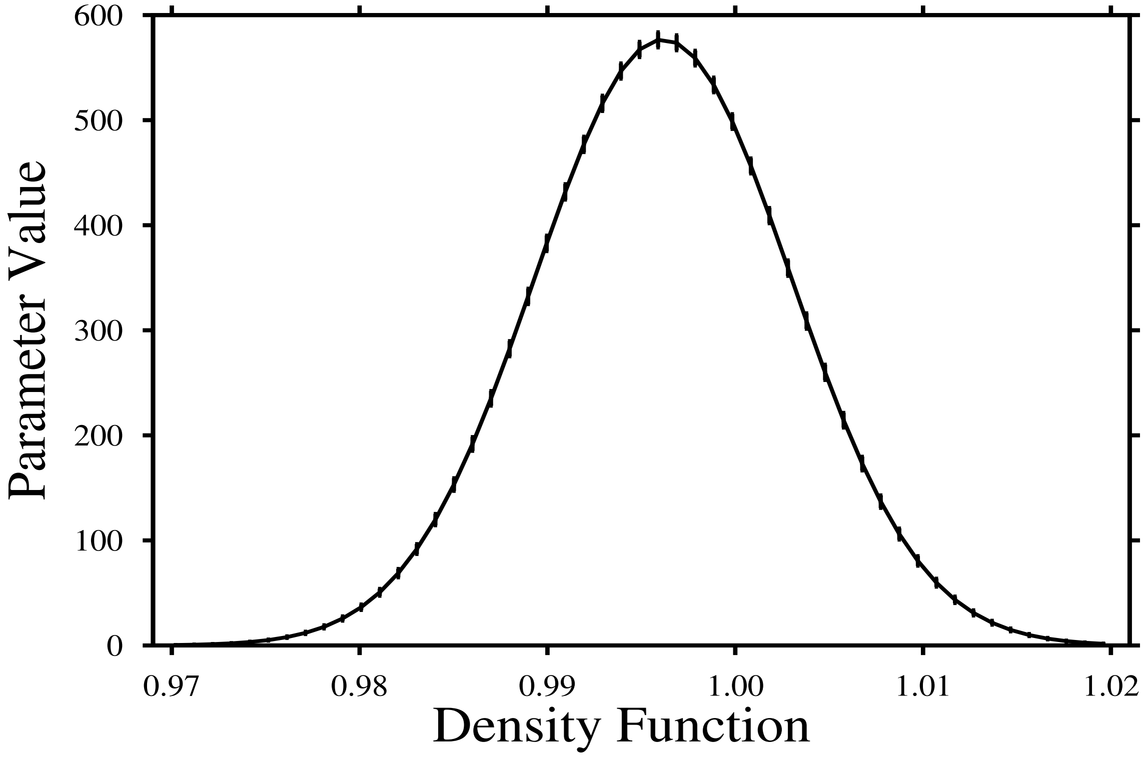

FIGURE 6. This is the model averaged density function with error bars. A Markov chain Monte Carlo simulation was used to draw samples from the joint posterior probability for the number of multipliers and their values. A total of 2500 samples were drawn. Each sample corresponds to a density function that is consistent with the data and the prior information. The solid line in this plot is the mean value of the 2500 density function estimates and the error bars are the standard deviation of the estimates.

compact, and the standard deviation of these probability density functions are 0.08 and 0.14 respectively. Given that there are about 2000 data values and the data are noiseless, this is not a very good determination. Thus while maximum entropy gives one essentially the peak of these probability distributions it does not indicate how uncertain one is of these values and the uncertainty in the value of the Lagrange multipliers is fairly large, having a relative uncertainty of about 20% in multiplier 1 and about 7% for multiplier 2.

Each of the 2500 Markov chain Monte Carlo samples of the Lagrange multipliers shown in Fig. 4(B) corresponds to a density function estimate that is consistent with the given data and prior information. One can use the Lagrange multiplier samples to compute the unknown density function. For example, one could compute the density function at the values specified by Eq. (25). For each x_i there are 2500 samples of the density function. For a given x_i, one can compute the mean and standard deviation. This mean and standard deviation are shown in Fig. 6. At each point in this plot the mean is the solid line and the standard deviation is shown as the error bar. These error bars are a direct measure of the amount of uncertainty in the Lagrange multipliers and thus directly reflect the uncertainty in the underlying density function.

SUMMARY AND CONCLUSIONS

The maximum entropy method of moments is fraught with difficulties. It is computationally unstable. One cannot use the raw data; rather one must compute an unknown number of moments using the data and then use those moments in the maximum entropy method of moments. There is no way to determine how many moments are needed and, finally, there is no way to determine how uncertain one is of the estimated density function.

However, if one uses the maximum entropy formalism to assign the probability for the data given both the number of moments and the value of the Lagrange multipliers, then maximum entropy will assign Eq. (6) as the functional form of the probability distribution. One can then use the rules of Bayesian probability theory to compute the posterior probability for the parameters including the number of Lagrange multipliers. Because the Bayesian calculations are all computed using a forward calculation, i.e., given the values of the parameters compute a probability, and never attempt to solve for the values of the multipliers that satisfy the constraints, Eq. (7), one never runs into computational difficulties. Additionally, the final results are all expressed as probability distributions, so one always knows how uncertain one is of all of the parameters. Finally, because the calculations are implemented using a Markov chain Monte Carlo simulation, one has samples from the joint posterior probability for all of the parameters appearing in Eq. (28) and these samples can be used to form a mean and standard deviation estimate of each point in the unknown density function, thus putting error bars on the unknown density functions value.

ACKNOWLEDGMENTS

I thank Joseph J. H. Ackerman, Jeffrey J. Neil, Joel R. Garbow, for encouragement, support, and helpful comments. This work was supported grants HD062171, HD057098, EB002083, NS055963, CA155365 and HL70037.

REFERENCES

1. Karl Pearson, *Phil. Trans. R. Soc. A* **186**, 343–326 (1895).
2. David Freedman and Persi Diaconis, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**, 453–476 (1981).
3. Murray Rosenblatt, *Annals of Mathematical Statistics* **27**, 832–837 (1956).
4. Emanuel Parzen, *Annals of Mathematical Statistics* **33**, 1065–1076 (1956).
5. Claude E. Shannon, *Bell System Technical Journal* **27**, 379–423, 623–656 (1948).
6. Lawrence R. Mead and Nikos Papanicolaou, *J. Math. Phys.* **25**, 2404–2417 (1984).
7. William Press, Willam T. Vetterling, Saul A. Teukolsky and Brian P. Flannery, *Numerical Recipes, The Art of Scientific Computing*, Cambridge University Press, 2078, third edition.
8. Rev. Thomas Bayes, *Philos. Trans. R. Soc. London* **53**, 370–418 (1763).