ENTROPY, MOMENTS AND PROBABILITY THEORY

Sławomir Kopeć
Jagellonian University
Institute of Physics
Reymonta 4
30-059 Kraków, Poland.

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

ABSTRACT. Bayesian probability theory, using an entropy prior, is applied to the moment problem when the data are noisy. When the noise level tends to zero, the Bayesian solution tends to the classic maximum entropy (MaxEnt) solution. The uncertainty in the estimated function is derived and a numerical example is presented to illustrate the calculation.

1. Introduction

The MaxEnt approach to solving the problem of moments was studied in detail by Mead and Papanicolau (Mead, 1984). Their analysis assumed the availability of exact moments. This is a severe limitation of the technique. An attempt to apply MaxEnt to the noisy moment problem was made by Ciulli *et al.* (Ciulli, 1991). Their solution, although essentially correct, left some important questions unanswered. In particular, an estimate of the uncertainty in the estimated function was not presented. In this paper, the noisy moment problem is addressed using Bayesian probability theory. Many of the details omitted here are contained in (Bretthorst, 1992), where the deconvolution problem is studied.

In the moment problem, there are some known moments or data, d_i, which are related to an unknown function $x(t)$:

$$d_i = \int_0^1 dt\, \omega_i(t)\, x(t) + n_i, \quad i = 1, 2, \ldots N, \tag{1}$$

where $\omega_i(t)$ are linearly independent known functions, and n_i represents the measurement error in the ith moment.

2. Method of Solution

To solve this problem using Bayesian probability theory the continuous function $x(t)$ is first replaced with a set of discrete values $\{x_k\}, k = 1, 2, \ldots, M$, with $x_k := x(t_k)$, and

$dt \rightarrow \Delta t = 1/M$. Using an entropy prior and assigning a gaussian prior probability for the noise, the posterior probability for one of the x_k is then given by

$$P(x_k|D,\beta,\sigma,I) \propto \int dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_M \exp\{S_\beta(x)\} \tag{2}$$

with

$$S_\beta(x) := \frac{2\beta}{M}\sum_{j=1}^{M}\left[x_j(\log x_j - 1) + 1\right] + \sum_{i=1}^{N}\left[d_i - \sum_{j=1}^{M}\frac{\omega_i(t_j)x_j}{M}\right]^2, \tag{3}$$

where D denotes the collection of moments, or data, $D \equiv \{d_1,\ldots,d_N\}$, σ is the standard deviation of the noise, and β is a measure of the relative importance of the prior information.

The integrals are evaluated in the Gaussian approximation. To make this approximation, the values of the x_k that maximize $S_\beta(x)$ are denoted as $\hat{x}_k$. Equation (2) has a maximum when the derivatives of $S_\beta(x)$ vanish. So $\hat{x}_k$ must satisfy

$$\left.\frac{\partial S_\beta(x)}{\partial x_k}\right|_{\hat{x}_k} = \beta\log\hat{x}_k - \sum_{i=1}^{N}\left[d_i - \sum_{j=1}^{M}\frac{\omega_i(t_j)}{M}\hat{x}_j\right]\omega_i(t_k) = 0, \quad k = 1,\ldots,M. \tag{4}$$

Because the $\{\omega_i\}$ are linearly independent, $\log\hat{x}$ is a linear combination of functions ω_i. Thus $\hat{x}$ is given by the solution to

$$\beta\alpha_i = d_i - \sum_{j=1}^{M}\frac{\omega_i(t_j)\hat{x}_j}{M}, \quad i = 1,\ldots,N, \tag{5}$$

$$\hat{x}_j = \exp\left\{\sum_{k=1}^{N}\alpha_k\omega_k(t_j)\right\}, \tag{6}$$

where the α_k play the role of generalized Lagrange multipliers.

Note that vanishing of σ implies vanishing of β. Thus when the noise level tends to zero, the left-hand side of (5) vanishes, and Eqs. (5) and (6) reduce to the maximum entropy solution to the problem. However, for noisy moments the maximum entropy solution is only an approximation to $\hat{x}$.

To complete the gaussian approximation of Eq. (2), a second order Taylor expansion is made about $\hat{x}_j$ to obtain

$$P(x_k|D,\beta,\sigma,I) \propto \int dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_M \exp\left\{-\sum_{jl=1}^{M}\frac{R_{jl}(x_j - \hat{x}_j)(x_l - \hat{x}_l)}{2\sigma^2}\right\} \tag{7}$$

where

$$R_{jl} := \frac{\beta\delta_{jl}}{M\hat{x}_k} + \sum_{i=1}^{N}\frac{\omega_i(t_j)\omega_i(t_l)}{M^2}. \tag{8}$$

Evaluating the integrals one obtains:

$$P(x_k|D,\beta,\sigma,I) \propto \exp\left\{-\frac{(x_k-\hat{x}_k)^2}{2\sigma^2}\left[R_{kk}-\sum_{lmn}\frac{e'_{lm}e'_{ln}R_{km}R_{kn}}{\lambda'_l}\right]\right\}, \tag{9}$$

where λ'_l is the lth eigenvalue of the kth cofactor of the matrix R_{km}, Eq. (8), and e_{lj} is the jth component of the lth eigenvector of this matrix. The kth cofactor of R_{lm} is formed by deleting the kth row and column from R_{lm}. If one adopts the convention that the cofactor's rows and columns are indexed just like R_{lm}, then the sums are over all values of the index except $l=k$ or $m=k$ or $n=k$.

The above result is valid, provided σ is known. In the event σ is unknown, we can eliminate it from the problem by assigning a Jeffrey's prior and integrating. However, note that several terms were dropped from Eq. (9). This could be done because when σ is known, these terms are constants. Recovering these terms and eliminating σ as a nuisance parameter one obtains

$$P(x_k|D,\beta,I) \propto \left[S_\beta(\hat{x})+\left(R_{kk}-\sum_{lmn}\frac{e'_{lm}e'_{ln}R_{km}R_{kn}}{\lambda'_l}\right)(x_k-\hat{x}_k)^2\right]^{-\frac{N+1}{2}}. \tag{10}$$

3. Estimating the Uncertainty

An estimate of the uncertainty in the x_k may be obtained in the (mean ± standard deviation) approximation. To compute this, one must compute both $\langle x_k\rangle$ and $\langle x_j x_k\rangle$, the expectation values of, respectively, x_k and $x_j x_k$. The expectation value of x_k is equal to $\hat{x}_k$, while

$$\langle x_j x_k\rangle = \hat{x}_j\hat{x}_k + \sigma^2\sum_{l=1}^{M}\frac{1}{\lambda_l}e_{lj}e_{lk} \tag{11}$$

where λ_l is the lth eigenvalue of the matrix R_{km}, Eq. (8), and e_{lj} is the jth component of the lth eigenvector of R_{km}. The (mean ± standard deviation) approximation is then given by

$$(x_k)_{\text{est}} = \hat{x}_k \pm \sqrt{\langle\sigma^2\rangle}\left[\sum_{l=1}^{M}\frac{(e_{lk})^2}{\lambda_l}\right]^{1/2}, \quad k=1,\ldots,M. \tag{12}$$

For finite N, this expression tends to infinity when the number of intervals tends to infinity. When M grows, our N moment equations become more and more insufficient to determine the approximate solution. In other words macroscopic data (on moments) cannot affect our knowledge of microscopic structure. This result occurs because neither the prior probability nor the likelihood introduce any point to point correlations in x_k, so on any infinitely small intervals, the function $x(t)$ could be doing wild things, long as the weighted average (represented by the moments) are satisfied.

4. Estimating the Parameters β and σ

In the above, β is assumed known. However, in general, β will not be known. The rules of probability theory tell one how to proceed. One should simply multiply Eq. (9) by an appropriate prior probability and integrate over β. This procedure will yield the posterior probability for x_k, independent of β. Unfortunately, β appears in these equations in a very nonlinear fashion and the integrals are not available in a closed form. However, a good approximation is available, provided the moments are not very noisy. When the moments are not very noisy, the integral over β is well approximated by a delta function, and δ functions just fix the value of the parameter. So if one knew where the integrand peaked as a function of β, one could simply constraint β to its maximum value.

Fortunately, the posterior probability for β can be computed. The value of β for which this probability is maximum is essentially identical to the value of β for which joint probability of x_j and β is maximum. The posterior probability for β, $P(\beta|D,\sigma,I)$ is computed from the joint probability for β and the all $\{x\} := \{x_1 \ldots x_M\}$:

$$\begin{aligned} P(\beta|D,\sigma,I) &= \int dx_1 \cdots dx_M P(\beta,\{x\}|\sigma,DI) \\ &= \int dx_1 \cdots dx_M P(\beta|I) P(\{x\}|\beta,\sigma,D,I). \end{aligned}$$

Assuming Jeffrey's prior, $P(\beta|I) \propto 1/\beta$, the posterior probability for β is given by

$$P(\beta|D,\sigma,I) \propto \beta^{\frac{M}{2}} \exp\left\{-\frac{1}{2\sigma^2} S_\beta(\hat{x})\right\} \prod_{l=1}^{M} (\lambda_l \hat{x}_l)^{-\frac{1}{2}}. \tag{13}$$

The expected value of the noise variance, $\langle\sigma^2\rangle$, should be replaced by its true value in Eq. (12), if σ is known. However, in typical applications the noise level is unknown. We then treat σ as a nuisance parameter and eliminate it from the formalism. In Eq. (12) this resulted in the expected value of σ^2 showing up in the calculation, so to use Eq. (12) the expected value of σ^2 must be computed. This is given by

$$\langle\sigma^2\rangle = \frac{1}{N-2} S_\beta(\hat{x}). \tag{14}$$

5. Numerical Example

To demonstrate that for a fixed M the errors remain finite as the noise goes to zero, a simple example is given. In this example the function $\exp\{-5t^2\}$ was normalized over the interval zero to one and its first five moments, $\omega_i(t) = t^{i-1}$, were computed. These exact moments were then used in the numerical example. The posterior probability for β was first computed and is shown in the logarithmic plot in Fig. 1A. Note that as β goes to zero this posterior probability should go to infinity. However, we were unable to compute the posterior probability for values of β smaller than 10^{-16}. In spite of this, the fully normalized posterior probability, Fig. 1B, is a fair representation to a delta function. Because the probability for β is so sharply peaked, constraining β to its peak value is an extremely good approximation in Eq. 5. Using $\beta = 10^{-16}$ in Eq. 5 one obtains the (mean ± standard deviation) estimates of the function. These are shown in Fig. 2. On this scale there is no observable difference between the estimated function and the (mean ± standard deviation) estimates. Because of this, we have expanded the region around the origin and

Figure 1: Probability for β

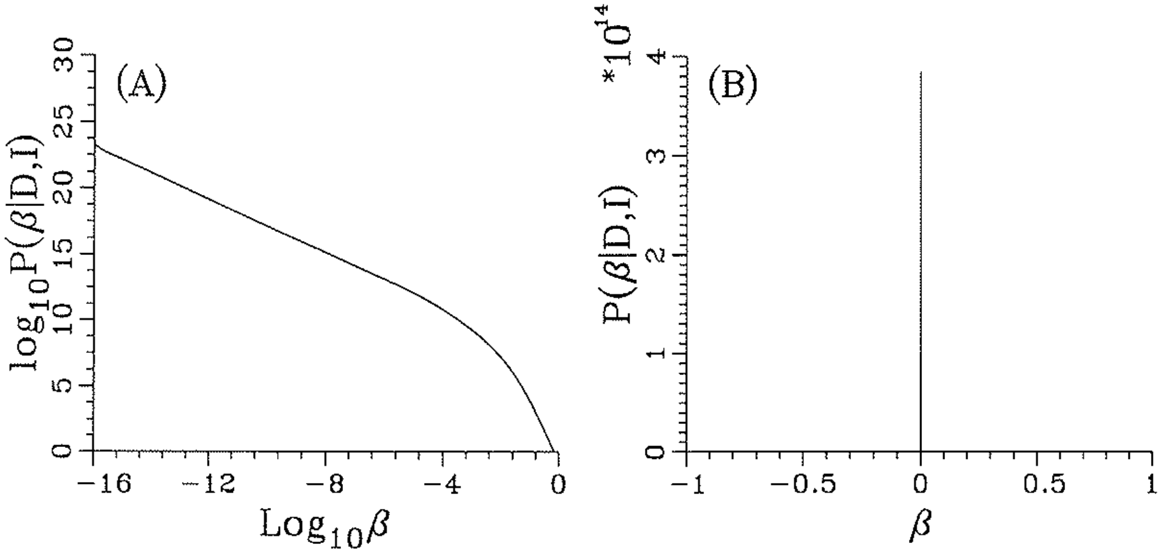

Fig. 1. Panel A is a Log plot of the probability for β. Note that this probability density function drops some 22 orders of magnitude. Panel B is the fully normalized probability density function. For noiseless moments this function should be a delta function. However, computationally we could not compute it for values of β smaller than 10^{-16}, yet the fully normalized density function is still a good approximation to a delta function.

Figure 2: The Expected Function

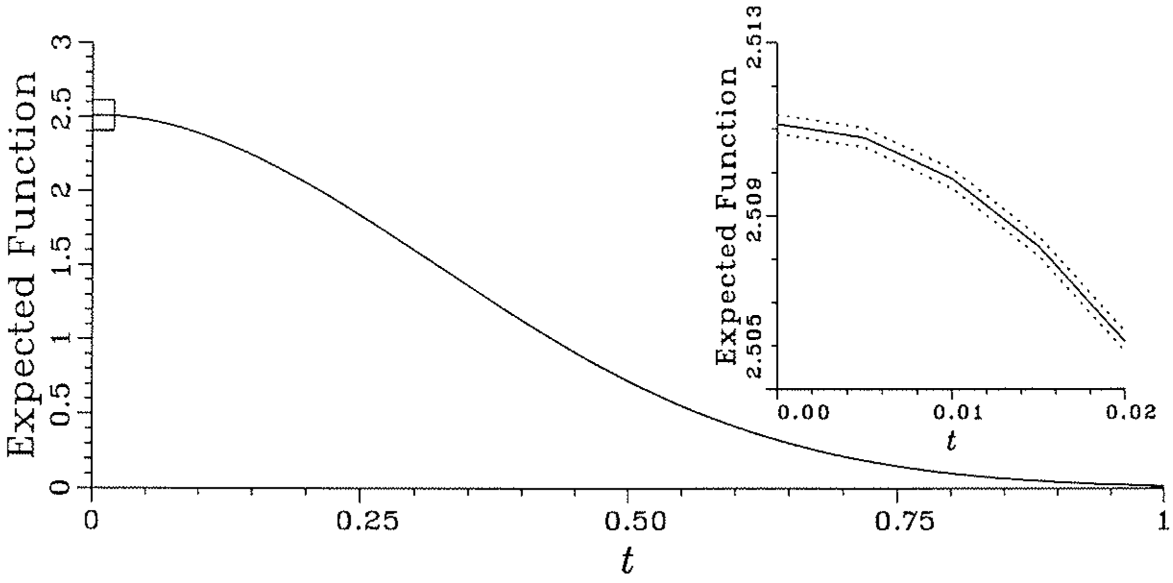

Fig. 2. On this scale no difference is seen between the expected function and the (mean ± standard deviation) estimates. The large box is an expansion of the region around $t = 0$. The expected function is shown as the solid line, the one standard deviation errors are shown as the dotted lines.

have plotted this in the upper right-hand-corner of Fig. 2. Even on this highly expanded scale the uncertainty in the function is barely visible (dotted lines).

6. Final Remarks

Probability theory generalizes the Lagrange multiplier equations from MaxEnt in such a way that a Bayesian solution always exists for some values of β. In the case of very noisy moments, β is estimated to be large, the Lagrange multipliers go to zero and the reconstructed function goes to a uniform function. When the moments are noiseless, σ goes to zero and the Bayesian result is given by the maximum entropy solution to the moment problem. In between there is a kind of minimum–maximum trade off: the result is a maximum entropy solution that has minium chi-squared.

For any given problem the entropy prior may or may not be justified. For example, when the function is known to take on negative values, the entropy prior is not only inappropriate, it simply cannot be used. Additionally, there could be other types of prior information not adequately expressed by entropy. For example, one could know something about the asymptotic form of the function, or one could want an estimate that has minimum curvature. Such information may be incorporated into an appropriate Bayesian prior (Bretthorst, 1992). The resulting Bayesian solution will have the same general characteristics as the one exhibited here: the uncertainty in the function will be a well-behaved quantity, and in the noiseless limit the Bayesian solution will be equal to the solution of some constrained optimization problem.

ACKNOWLEDGMENTS. SK would like to thank the Kosciuszko Foundation for financial support.

REFERENCES

Bretthorst, G.L.: 1992, 'Bayesian Interpolation and Deconvolution',contract number DAAL03-86D-0001, U.S. Army Missile Command.

Ciulli, S., Mounsif, M., Gorman, N., Spearman, T. D.: 1991, 'On the Application of maximum entropy to the moments problem', *J. Math. Phys.* **32**, 1717.

Mead, L.R., and Papanicolau, N.:1984, 'Maximum Entropy in the Problem of Moments, *J. Math. Phys.* **25**, 2404.