BAYESIAN PROBABILITY THEORY AND INVERSE PROBLEMS*

S. Kopeć

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

Bayesian probability theory is applied to approximate solving of the inverse problems. In order to solve the moment problem with the noisy data, the entropic prior is used. The expressions for the solution and its error bounds are presented. When the noise level tends to zero, the Bayesian solution tends to the classic maximum entropy solution in the L_1 norm. The way of using spline prior is also shown.

PACS numbers: 02.50. Cw, 02.60. Cb

1. Introduction

There are numerous physical applications where one encounters the need of solving an inverse problem. The typical examples are: the moment problem, interpolation, approximation, deconvolution. By means of standard procedures [1, 2], the problem of solving integral and differential equations can also be converted to seeking a solution of some particular set of moment equations. Among various techniques of approximate solving inverse problems, the approach based on maximization of entropy (in the sense of information theory) proves successful and becomes increasingly popular (see *e.g.* [2–5], and references therein). The Maximum Entropy (ME) method has, however, several limitations. It applies exclusively to the problems admitting positive solutions. Moreover, the method can be used only in the noiseless case, *i.e.* if available data are known exactly.

The natural formalism of analysing the noisy case is Bayesian probability theory [6–8]. By means of Bayesian methods not only is it possible to find an approximate solution to the problem, but also to give an estimate of the uncertainty in the solution. The basic tool of the Bayesian analysis is

* The paper was supported by the Polish Committee for Scientific Research, grant no. 2 P302 079 04.

the Bayes theorem connecting the posterior probability of the solution with its prior probability and the posterior probability of the data. The explicit form of the prior probability is determined by our additional information about the problem. In particular, one may require that the solution should maximize the information entropy. Other possible forms of the prior probability, potentially easier to handle, can be determined from our knowledge about the nature of the problem [9].

In our paper we show, on an example of the moment problem, how the Bayesian analysis with the entropy prior can be carried out. We focus our attention on the moment problem, because the result can be easily generalized to other inverse problems. We present an expression for the approximate solution and estimate the uncertainty in the solution obtained in the (mean ± standard deviation) approximation. We also show that in the limit of the noiseless moments, the Bayesian result is given by the maximum entropy (MaxEnt) solution to the problem. This result can be proved in a more formal way. We may interpret the Bayesian solution of the noisy problem as a solution to "penalized MaxEnt" and show that in the noiseless limit this solution tends, in a L_1 norm to the MaxEnt solution.

The entropy prior dos not have to be the most appropriate one for a given particular problem. It is also not an easy task to perform computations when the entropy prior is used (due to a nonlinear character of equations it leads to). Therefore other forms of the prior probability are frequently assigned, in accordance to the required properties of the solution. For example, when one expects the solution to be in some sense "smooth", the natural choice is the prior assuring minimalization of the curvature. In this paper, on the interpolation example, we introduce the spline prior. As the cubic spline is the smoothest possible interpolating curve [10], we construct the prior probability such that in the noiseless limit our solution tends to the one given by the spline interpolation. Again, the result can be easily generalized to other inverse problems.

2. The MaxEnt approach to the moment problem

The MaxEnt approach to solving the problem of moments was studied in detail by Mead and Papanicolau [4]. Their analysis, however, implied the availability of exact moments, which is a severe limitation of the technique. An attempt to apply MaxEnt to the noisy moment problem was made by Ciulli *et al.* [11]. 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. In [9], where the deconvolution problem is studied in similar manner, many of the details omitted here are contained.

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

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

where n_i represents a measurement error in the ith moment, and $\omega_i(t)$ are linearly independent known functions. The magnitude of this error may or may not be zero and if nonzero may or may not be known.

Using an entropy prior and the standard likelihood function, Bayes theorem gives

$$P(x|D,\beta,\sigma,I) \propto \exp\left(-\frac{1}{2\sigma^2}\left(2\beta\int_0^1 dt(x(t)(\log x(t)-1)+1)+\sum_{i=1}^{N}(d_i-\int_0^1 dt\omega_i(t)x(t))^2\right)\right) \tag{2}$$

as the posterior probability density for the function $x(t)$, where D denotes the collection of moments, $D \equiv \{d_1, \dots, d_N\}$, σ is the standard deviation of noise, and β is a measure of the relative importance of the prior information.

2.1. Method of solution

To make the (mean ± standard deviation) estimate of the function in a Gaussian approximation, we start from location of the maximum of the posterior probability. Denoting this maximum by $\hat{x}(t)$, we find that

$$\beta\alpha_i = d_i - \int_0^1 dt\, \omega_i(t)\hat{x}(t), \quad i = 1, \dots, N, \tag{3}$$

$$\hat{x}(t) = \exp\left(\sum_{k=1}^{N} \alpha_k \omega_k(t)\right), \tag{4}$$

where the α_k play the role of generalized Lagrange multipliers. Note that this maximum *is the maximum entropy solution to the moment problem,* provided the data are noiseless, *i.e.*, in the limit $\sigma \to 0$ (*cf.* [2]). However, for noisy moments the maximum of the posterior probability and the maximum entropy solution differ.

2.2. Estimating the uncertainty

We start from discretizing the interval $[0,1]$. The continuous function $x(t), t \in [0,1]$ is then replaced with a set of its values $\{x_k\}, k = 1, 2, \ldots, M$, with $x_k := x(t_k)$. The (mean ± standard deviation) estimate of x_k is given by

$$(x_k)_{\mathrm{est}} = \hat{x}_k \pm \sqrt{\langle\sigma^2\rangle}\left(\sum_{l=1}^{M} \frac{(e_{kl})^2}{\lambda_l}\right)^{1/2}, \quad j = 1, \ldots, M, \tag{5}$$

where λ and e are the eigenvalues and eigenvectors of the covariance matrix in the discrete Gaussian approximation. The expected value of the noise variance, $\langle\sigma^2\rangle$, should be replaced by its true value if σ is known, and by

$$\langle\sigma^2\rangle = N^{-1} S_\beta, \tag{6}$$

in the event it is unknown, where

$$S_\beta := 2\beta \int_0^1 dt(\hat{x}(t)(\log \hat{x}(t) - 1) + 1) + \sum_{i=1}^{N} \left(d_i - \int_0^1 dt\omega_i(t)\hat{x}(t) \right)^2. \tag{7}$$

For finite N these "point-like" error bars tend to infinity as the number of intervals tends to infinity. The reason is apparent — 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. Neither the prior probability, nor the likelihood, introduce any point-to-point correlations in x_k, so on small intervals the large variations of $x(t)$ are permitted.

It is still possible, however, to give a *global* expression for the error bounds. The finite and independent of M formula for the global error is

$$\langle X^2\rangle - \langle X\rangle^2 = \sigma^2 \sum_{i,j,l=1}^{M} \frac{e_{li}e_{lj}}{\lambda_l}, \tag{8}$$

where

$$\langle X\rangle = \frac{1}{M} \sum_{i=1}^{M} \hat{x}_i . \tag{9}$$

2.3. Eliminating the nuisance parameter

In the above, β is assumed known. However, in general β will not be known and must be estimated from the data. If one applies the rules of probability theory, one would multiply by a prior for β and integrate. Unfortunately, β appears in these equations in a very nonlinear fashion and these integrals are not available in a closed form. However, a good approximation for β is available, provided the moments are not very noisy. In this case it is justified to constrain β to $\beta_{\max}$, where $\beta_{\max}$ is the value of β maximizing the posterior probability:

$$P(\beta|D,I) \propto \beta^{\frac{M}{2}} \left[S_\beta\right]^{-\frac{N}{2}} \prod_{l=1}^{M} (\lambda_l \hat{x}_l)^{-\frac{1}{2}}. \tag{10}$$

2.4. Penalized MaxEnt

When we seek the moment problem solution belonging to L_1, it is possible to show that the sequence of the "noisy" solutions tend to the MaxEnt solution of the noiseless problem. In this approach we can treat the noisy case as a "penalized version" of the original problem. The proof follows the ideas suggested by Lewis (private communication).

Let S be a finite measure space and the sequence $a_1, a_2, \ldots, a_n$ belong to L_∞. Let I be the functional defined by:

$$I: \quad L_1 \rightarrow (-\infty, +\infty],$$
$$I(x) = \begin{cases} \int x \log x & \text{for } x \geq 0 \text{ a.e.} \\ +\infty & \text{otherwise} \end{cases}$$

Let us define the following condition (P):

$$\inf\{I(x) | \int a_i x = b_i \qquad (i = 1, 2, \ldots, n)\}. \tag{11}$$

Assume that (P) is consistent, that is there exists $\hat{x} \in L_1$ such that $I(\hat{x}) < +\infty, \int a_i \hat{x} = b_i$ for all i. Then, as I has ω-compact level sets, (P) has a unique optimal solution $\hat{x} : I(\hat{x}) = V(P)$.

Let us define also the penalized version of (P), called (P_r):

$$\inf\{I(x) + r \sum_{i=1}^{n} \left(\int a_i x - b_i \right)^2 \}. \tag{12}$$

There exists exactly one x_r optimal for (P_r):

$$V(P_r) = I(x_r) + r \sum_{i=1}^{n} \left(\int a_i x - b_i \right)^2 . \tag{13}$$

We can easily prove the following Lemmas:

Lemma 1

$$V(P_1) \leq V(P_2) \leq \ldots \leq V(P)\,. \tag{14}$$

Proof

By definition

$$V(P_r) \leq I(x_{r+1}) + r\sum_i \left(\int a_i x_{r+1} - b_i\right)^2 I(x_{r+1})$$
$$+(r+1)\sum_i \left(\int a_i x_{r+1} - b_i\right)^2 = V(P_{r+1})\,. \tag{15}$$

As $\hat{x}$ is feasible for (P), we have

$$V(P_r) \leq I(\hat{x}) + r\sum_i \left(\int a_i \hat{x} - b_i\right)^2 = I(\hat{x}) = V(P)\,. \tag{16}$$

Lemma 2

$$\sum_i \left(\int a_i x_r - b_i\right)^2 \to 0\,. \tag{17}$$

Proof

As I has compact level sets, there exists $\alpha > -\infty$ such that $I(x) \geq \alpha$ for all x. Now we have

$$\alpha + r\sum_i \left(\int a_i x_r - b_i\right)^2 \leq I(x_r) + r\sum_i \left(\int a_i x_r - b_i\right)^2 = V(P_r) \leq V(P)\,. \tag{18}$$

Thus

$$0 \leq \sum_i \left(\int a_i x_r - b_i\right)^2 \leq r^{-1}(V(P) - \alpha) \to 0\,. \tag{19}$$

Now we are able to prove the following theorem:

Theorem

$$x_r \rightharpoonup \hat{x} \quad \text{weakly in} \quad L_1\,. \tag{20}$$

Proof

Let us suppose the theorem is not true. Then there exists $z \in L_\infty$ such that for a subsequence (r')

$$\int (x_{r'} - \hat{x})z \geq 1 \quad \text{for all} \quad r'\,. \tag{21}$$

Now

$$I(x_r) \leq I(x_r) + r\sum_i \left(\int a_i x_r - b_i\right)^2 = V(P_r) \leq V(P) = I(\hat{x}) . \tag{22}$$

This means that

$$x_r \in L = \{x \in L_1 | I(x) \leq I(\hat{x})\} , \quad \text{for all} \quad r , \tag{23}$$

which is weakly compact. Thus there exists weakly convergent subsequence $x_{r''} \to$ some $\tilde{x} \in L$. Since

$$\sum_i \left(\int a_i x_r - b_i\right)^2 \to 0 , \quad \sum_i \left(\int a_i \tilde{x} - b_i\right)^2 = 0 .$$

Thus $\tilde{x}$ is feasible for (P) with $I(\tilde{x}) \leq I(\hat{x}) = V(P)$. By uniqueness, $\tilde{x} = \hat{x}$, which contradicts the assumption.

Actually we can show the stronger convergence.

Lemma3

$$I(x_r) \to I(\hat{x}) . \tag{24}$$

Proof

We know that $I(x_r) \leq I(\hat{x})$. Now suppose that $I(x_r)$ is different from $I(\hat{x})$. Then there exists a subsequence $x_{r'}$ and $\epsilon > 0$ with

$$x_{r'} \in L_\epsilon = \{x \in L_1 | I(x) \leq I(\hat{x}) - \epsilon\} , \tag{25}$$

which is again weakly compact. Thus there exists weakly convergent subsequence $x_{r''} \to$ some $\tilde{x} \in L_\epsilon$. As before get $(\int a_i \tilde{x} - b_i) = 0$, so $\tilde{x}$ is feasible for (P). But $I(\tilde{x}) \leq I(\hat{x}) - \epsilon < I(\hat{x}) = V(P)$, which contradicts the assumption.

Now we can prove

Theorem 2

$$||x_r - \hat{x}||_1 = \int | x_r - \hat{x} | \to 0 . \tag{26}$$

Proof

The theorem follows from $x_r \rightharpoonup \hat{x}$, $I(x_r) \to I(\hat{x})$ and strong convexity of I.

3. The spline prior

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, but, since it does not permit negative solutions, 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 a minimal curvature. Such information may be incorporated into an appropriate Bayesian prior [9]. 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.

Another condition, which can be imposed, is the requirement that our solution is the smoothest of all twice-differentiable functions matching the data when $\sigma \to 0$. Then (*cf.* [10]) **the solution must be the cubic polynomial spline fit with second derivatives equal 0 on the boundary.**

When we want to solve the interpolation problem

$$d_i = x(t_i) + n_i\,, \quad i = 1, \beta + 1, \ldots \beta(N-1) + 1\,, \tag{27}$$

the likelihood function is given by

$$P(D|\sigma, x, I) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{l=1}^{N}(d_l - x_{(l-1)\beta+1})^2\right). \tag{28}$$

Now in order to define the cubic polynomial we have to impose $4(N-1)$ conditions. Let us denote by p^k the polynomial at the interval $[t_{(k-1)\beta+1}, t_{k\beta+1}]$, $k = 1, \ldots, N-1$, and by c_k its second derivative at $t_{(k-1)\beta+1}$.

We can choose the following conditions:

1. $$p^k(t_{(k-1)\beta+1}) = x_{(k-1)\beta+1}\,, \qquad k = 1, 2, \ldots, N-1\,, \tag{29}$$

2. $$p^k(t_{k\beta+1}) = x_{k\beta+1}\,, \qquad k = 1, 2, \ldots, N-1\,, \tag{30}$$

3. $$(p^k)'(t_{k\beta+1}) = (p^{k+1})'(t_{k\beta+1})\,, \qquad k = 1, 2, \ldots, N-2\,, \tag{31}$$

4. $$(p^k)''(t_{k\beta+1}) = (p^{k+1})''(t_{k\beta+1})\,, \qquad k = 1, 2, \ldots, N-2\,. \tag{32}$$

As we have $4(N-1) - 2$ equations, we need two more to specify our polynomials uniquely. We choose these two constraints in a form: $c_1 = c_N = 0$. The reason for our choice is that the spline defined this way is the smoothest of all exact matching fits for $\sigma \to 0$. Now our problem is well-defined and the polynomials $p^k(t)$, $k = 1, \ldots, N-1$ can be found in a unique way by solving a recurrence relation.

As the result we obtain the **cubic spline prior**

$$P(x|I,\alpha,\sigma) \propto$$
$$\exp\left(-\frac{\alpha}{2\sigma^2}\sum_{k=1}^{N-1}\sum_{j=0}^{\beta-1}\left(\left(x(t_{(k-1)\beta+j+1}) - p^k(t_{(k-1)\beta+j+1})\right)^2\right.\right.$$
$$\left.\left. + \left(x(t_{(N-1)\beta+1}) - p^{N-1}(t_{(N-1)\beta+1})\right)^2\right)\right) .$$

The coefficient α serves as a measure of the relative importance of our prior information. This expression can be used in *any* inverse problem when we claim that our solution should be the smoothest one.

4. 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 going on. Eq. (3) is the maximum entropy solution to a moment problem when the moments are given by $d_i - \beta\alpha_i$; thus entropy is maximized for every value of β.

I would like to thank Dr G. Larry Bretthorst for stimulating discussions.

REFERENCES

[1] L.R. Mead, *J. Math. Phys.* **27**, 2903 (1986).

[2] S. Kopeć, *J. Math. Phys.* **32**, 1269 (1991).

[3] S. Kopeć, *J. Math. Phys.* **32**, 3312 (1991).

[4] L.R. Mead, N. Papanicolau, *J. Math. Phys.* **25**, 2404 (1984).

[5] M.J. Leaseburg, L.R. Mead, *J. Math. Phys.* **34**, 6009 (1993).

[6] M. Tribus, *Rational Descriptions, Decisions and Designs*, Pergamon Press, Oxford 1969.

[7] A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, New York 1987.

[8] E.T. Jaynes, *Probability Theory — The Logic of Science*, in preparation.

[9] G.L. Bretthorst, contract number DAAL03-86D-0001, U.S. Army Missile Command (1991).

[10] Larry L. Schumaker, *Spline Functions: Basic Theory*, John Wiley and Sons, New York 1981.

[11] S. Ciulli, M. Mounsif, N. Gorman, T.D. Spearman, *J. Math. Phys.* **32**, 1717 (1991).