Bayesian Parameter Estimation: A Monte Carlo Approach

Ray Gallagher

Department of Computer Science, University of Liverpool, Liverpool L69 7ZF, United Kingdom.
Email addresses: rayg@csc.liv.ac.uk

Tony Doran

Department of Computer Science, University of Liverpool, Liverpool L69 7ZF, United Kingdom.
Email addresses: tdustdum@csc.liv.ac.uk

Abstract. This paper presents a Bayesian approach, using parallel Monte Carlo modelling algorithms for combining expert judgements when there is inherent variability amongst these judgements. The proposed model accounts for the situation when the derivative method for finding the maximum likelihood breaks down

Introduction

An expert is deemed to mean a person with specialised knowledge about a given subject area or matter of interest. This paper concerns itself with the situation where we are interested in an uncertain quantity or event and expert opinion is sort out by a decision-maker. The question then arises as to how a decision-maker should then make optimal use of the expert opinion available to them. Moreover, how does a decision-maker make optimal use of expert opinion when several experts are available to them and further resolve conflicting opinions amongst the group of experts. The opinions of an expert may come in many ways: a point estimate, parameters of uncertainty distribution or a "best guess" with upper and lower bounds. The challenge for the decision-maker is to correctly take full advantage of the data provided.

Formally uncertainty can be represented in terms of probability and the ultimate aim is to reach a consensus to arrive at a probability distribution for the uncertain quantity of interest. This distribution should fully reflect the information provided by the experts.

Various consensus procedures for the pooling of experts' opinions and probability distributions have been suggested, encompassing merely the simple averaging of

V.N. Alexandrov et al. (Eds.): ICCS 2001, LNCS 2073, pp. 812-822, 2001.

expert probability distributions through to a formal Bayesian approach. Bayesian methods have been favoured by a number of researchers. Reviews of the available literature being provided by French[1], Cooke[2] together with Genest and Zidek[3]. The models proposed include those by Lindley[4-6],Morris[7,8],Winkler[9,10] and Mosley[11] This paper examines two different methods that allow the decision-maker to make the optimum decision based on available expert opinion. The methods are:

- **Derivative Method**
- **Monte Carlo**

Making the optimal decision based on the derivative method means that the function must be differentiable. We note there are other methods, discussed in Zacks,[12] to address this situation. If the function is not differentiable then we must employ a numerical method (in our case Monte Carlo) to arrive at an estimate of the quantity of interest. We further make use of parallel architectures using MIMD methods to increase the efficiency of the Monte Carlo method in situations where we may have a large body of expert opinion available.

Uncertainty Modelling of Expert Opinion

Suppose we have a parameter $\vec{\theta}$ = (θ_1, θ_2,..... θ_n) and to obtain the best decision about $\vec{\theta}$ we have to use some expert opinion given by $E = \{x_1^*, x_2^*, ..., x_N^*\}$ where x_i is the estimate of the i^{th} expert for an unknown quantity x, with the recognition that the particular value being estimated by that expert may be different from that being estimated by another expert.
The quantity of interest may be a fixed parameter but its exact value is unknown such as the height of a building or it may be an inherently variable quantity such as the IQ's of individual members of a group of people.

The situation arises, for example, when experts provide estimates based on experience with sub-populations of a non-homogeneous population. The objective is to develop an estimate of the distribution representing the variability of *x* in light of the evidence presented.

We attempt to aggregate these expert opinions to reach the "best" decision based on the estimation of $\vec{\theta}$.

For simplification we restrict ourselves to the situation when $\vec{\theta}$ comprises one or two elements. We then provide a general solution for $\vec{\theta}$ dependent on N elements. For formalisation of this discussion we consider the Bayesian approach to probability.

Let us consider the following definition of Bayes's Theorem

$$\pi(\mu \mid E) = k^{-1} L(E \mid \mu)\pi_0(\mu)$$

Where:
$\theta \equiv$ ***The value of interest*** to the decision maker,
$E \equiv$ the set of ***experts' opinions*** about the value of θ, the decision-maker treats this set of opinions as evidence/data,
$\pi_0(\theta) \equiv$ ***the decision maker's prior state*** of knowledge on θ,
$\pi(\mu|E) \equiv$ ***the decision maker's posterior state*** of knowledge on θ,
$L(E|\theta) \equiv$ ***the likelihood of observing*** the evidence E, given that the true value of the unknown quantity is θ,
$k \equiv P(E)$, the normalisation factor that makes $\pi(\theta|E)$ a probability distribution.

The problem of expert opinion is thus reduced to the assessment of the prior, π_0, and the likelihood, L, by the decision-maker. The key element in this approach is the likelihood. The likelihood function is the decision maker's tool to measure the accuracy of the expert's estimate after considering the expert's level of pertinent experience, calibration as an assessor, any known bias, and dependence to other experts.

In this section of the paper we summarise how we can receive $\pi(\hat{\theta} \mid E)$ i.e. with regard to experience, what is the best decision depends on E. Since every x_i^* is just some information concerning x_i we consider $f(x_i|\theta)$ as the actual distribution of the quantity of interest, x. We consider $L(x_i^* \mid \theta)$ is the probability density that the experts' estimate is x_i^* if the decision maker believes that the i^{th} expert is perfect then $L(x_i^* \mid \theta) = f(x_i|\theta)$.

Since the experts are considered independent then we have

$$L(E \mid \theta) = L(x_1^*, x_2^*, ..., x_n^* \mid \theta) = \prod_{i=1}^{n} L(x_i^* \mid \theta) \tag{1}$$

Moreover, $\pi(\theta \mid x_1^*, x_2^*, ..., x_N^*) = k^{-1} L(x_1^*, x_2^*, ...x_n^* \mid \theta)\pi_0(\theta)$. In this method we should first obtain k such that $\pi(\theta \mid x_1^*, x_2^*, ...x_n^*)$ is the conditional distribution. Suppose $P_i = P_i(x_i^* \mid x_i)$, (this P_i is one, if and only if, the expert is considered to be perfect) is the probability that the i^{th} expert says x_i^* when in fact the true value is x_i. The quantity P_i is the decision maker's probability density that the expert's estimate is x_i^* when he is attempting to estimate x_i.

We should note that x_i is one possible value of x and x is distributes according to $f(x|\theta)$. Then

$$L_i(x_i^* \mid \theta) = \begin{cases} \int P_i(x_i^* | xf(x|\theta)dx & if \quad X \quad continuous \\ \sum_j P(x_i^* | x_j) P(x_j | P) & if \quad X \quad discrete \end{cases} \tag{2}$$

For N independent experts we have

$$\pi(\theta \mid x_1^*, x_2^*, \ldots . x_n^*) = \begin{cases} k^{-1}\left\{\int P_i(x_i^* | x) f(x|\theta)dx\right\}\pi_0(\theta) \\ k^{-1}\left\{\prod_{i=1}^{n}\sum_j P_i(x_i^* | x_j) f(x_j|\theta)\right\}\pi_0(\theta) \end{cases} \tag{3}$$

For the best decision based on the evidence, *E,* we can use the derivative method if the derivative exists i.e.

$$\frac{\partial}{\partial \theta_j}\pi(\theta \mid x_1^*, x_2^*, \ldots . x_n^*) = 0 \quad j = 1,2,\ldots,n \tag{4}$$

These systems named normal equations, and receive $\theta_j = \hat{\theta}_j$ and for the maximum of *L* must be

$$\frac{\partial^2}{\partial \theta_j^2}\pi(\theta \mid x_1^*, x_2^*, \ldots . x_n^*)\Big|_{\theta_j = \hat{\theta}_j} < 0 \quad j = 1,2,\ldots,n \tag{5}$$

Example:

Suppose the decision-maker is interested in assessing the probability distribution of a random variable that takes only two values i.e. let

$$X = \{x_1, x_2\}. \tag{6}$$

A discrete distribution of X is completely known if we know P, where
$\theta \equiv \Pr[X = x_1]$ and $1 - \theta \equiv \Pr[X = x_2]$, $0 \le \theta \le 1$.
Suppose now the decision-maker asks the opinion of N experts on whether $X = x_1$ or whether $X = x_2$. Let E, defined as $E = \{x_1^*, x_2^*, ..., x_N^*\}$ be the set of expert responses where x_i^*, the i^{th} response can be either $X = x_1$ or $X = x_2$. Then we have $\pi(\theta \mid E) = k^{-1} L(E \mid \theta)\pi_0(\theta)$ where

$$L(E \mid \theta) = \prod_{i=1}^{n} L_i(x_i^* \mid \theta) \text{ and } L(x_i^* \mid \theta) = \sum_{j=1}^{n} P_r(x_i^* \mid x_j) P_r(x_j \mid \theta). \tag{7}$$

It is trivial that

$$P_r(x_j \mid \theta) = \begin{cases} \theta & if \quad j = 1 \\ 1 - \theta & if \quad j = 2 \end{cases} \tag{8}$$

Where $\Pr(x_i^* \mid x_j)$, is the probability that the ith expert says x_i^* when in fact $X = x_j$. These values represent how good the decision-maker thinks the experts are. For example, let us assume that the decision-maker consults two experts who he believes to be perfect and independent. For simplicity we assume a uniform prior in the closed interval [0,1], i.e. $\pi_0(\theta) = 1$, and consider the following two cases.

Case (i)

The two experts have opposing opinions, e.g. $x_1^* = x_1$ and $x_2^* = x_2$. Then, the likelihood is

$$L = \prod_{i=1}^{n} L_i(x_i \mid \theta) = \theta(1 - \theta) \tag{9}$$

and the posterior will be:

$$\pi(\theta \mid x_1, x_2) = 6\theta(1 - \theta) \tag{10}$$

With regard to equation (9) we have $\pi(\theta \mid E) = k^{-1}\theta(1-\theta)$ since π(θ|E) should be a conditional distribution then

$$\int_0^1 \pi(\theta \mid E)d\theta = k^{-1} \Rightarrow k = 1/6 \tag{11}$$

Then we have

$$\pi(\theta \mid x_1^*, x_2^*, ...x_n^*) = 6\theta(1-\theta) = 6\theta - 6\theta^2 \qquad 0 \le \theta \le 1 \tag{12}$$

Now, with regards to derivative tests for finding the extreme points we have

$$\frac{\partial \pi}{\partial \theta} = 6 - 12\theta = 0 \Rightarrow \hat{\theta} = 1/2 \tag{13}$$

$$\left.\frac{\partial^2 \pi}{\partial \theta^2}\right|_{\hat{\theta}=\frac{1}{2}} < 0 \tag{14}$$

This represents the distribution of all possible distributions of X. The most probable distribution (i.e. the mode of the posterior $\pi(\theta|x_1,x_2)$) is given by θ=1/2. It means that starting from complete lack of knowledge about the distribution of X, the opposing opinions of two independent experts have caused the decision maker to think most probably X= x_1 and X= x_2 are equally likely.

Case (ii)

The two experts have the same opinion; that is, for example, $x_1^* = x_1$ and $x_2^* = x_1$. The posterior in this case will be

$$\pi(\theta \mid x_1, x_2) = 3\theta^2 \tag{15}$$

We leave the proof of the second case as it is essentially the same operation of case i.

The main idea of this paper is when the situation arises when we wish to arrive at the optimum decision when there is no derivative, Zacks,[12]. In this situation we can use the finite difference gradient algorithm.

$$\theta_{i+1} = \theta_i + \alpha_i \hat{\nabla} \pi(\theta \mid E) \tag{16}$$

Therefore

$$\hat{\nabla} \pi = \left(\frac{\hat{\partial} \pi}{\partial \theta_1}, \frac{\hat{\partial} \pi}{\partial \theta_2}, \ldots\ldots\ldots, \frac{\hat{\partial} \pi}{\partial \theta_n} \right) \tag{17}$$

where

$$\frac{\hat{\partial} \pi}{\partial \theta_1} = \frac{g((\theta_1 + \Delta\theta), \theta_2, \ldots\ldots, \theta_n) - g((\theta_1 - \Delta\theta), \theta_2, \ldots\ldots, \theta_n)}{2\Delta\theta} \tag{18}$$

$$i = 1,2,\ldots,n$$

In this case we can consider a Monte Carlo random search algorithm to estimate the optimum decision for θ.

Random Search

We choose the random search double trial algorithm, Rubinstein[14].

$$\theta_{i+1} = \theta_i + \frac{\alpha_i}{2\Delta\theta_i} \left[\pi(\theta_i + \Delta\theta_i t_i) - \pi(\theta_i - \Delta\theta_i t_i) \right] t_i \tag{19}$$

where α_i and $\Delta\theta_1$ are greater than 0. This estimation $\hat{\theta}$ of θ, converges to θ in quadratic mean, in probability, and with probability one, Halton[13].
This algorithm may be performed by generating the random vector t_i continuously distributed on the *n*-dimensional unit sphere.

For this algorithm, if π is a real function which depends on $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_m)$, then we can use $\vec{t_i} = (t_{i1}, t_{i2}, ...t_{in})$, i = 1,...,m, and use *n* random vectors, in this situation we have a lot of random samples and we can try by parallel processing methods in a MIMD environment to obtain θ_i by the i^{th} processor, in a small time interval.

If $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_m)$, we generate (i = 1,...,m) $\vec{t_i} = (t_{i1}, t_{i2}, \ldots, t_{in})$ m random vectors and

$\overrightarrow{t_{01}}$	t_{11}	t_{12}	t_{13}	...	...	...	t_{1n}
$\overrightarrow{t_{02}}$	t_{21}	t_{22}	t_{23}	...	...	...	t_{2n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\overrightarrow{t_{0m}}$	t_{m1}	t_{m2}	t_{m3}	...	...	...	t_{mn}

$\vec{t} = \begin{bmatrix} t_{01} \\ t_{02} \\ t_{03} \\ \vdots \\ t_{0n} \end{bmatrix}$ is a vector with m rows and each element is n-tuples of random

numbers. Then we collect the following random vectors

$$\vec{t}_{01} = \begin{bmatrix} t_{11} \\ t_{21} \\ t_{31} \\ \vdots \\ t_{n1} \end{bmatrix}, \quad \vec{t}_{02} = \begin{bmatrix} t_{12} \\ t_{22} \\ t_{32} \\ \vdots \\ t_{n2} \end{bmatrix} \quad ,..., \quad \vec{t}_{0m} = \begin{bmatrix} t_{1m} \\ t_{2m} \\ t_{3m} \\ \vdots \\ t_{nm} \end{bmatrix},$$

for distribution to processors *1, 2, ..., m* respectively, enabling us to obtain the result $\hat{\theta}_i$ from the i^{th} processor. We will then have $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_n)$ such that $\hat{\theta}$ is an estimate for θ.

Algorithm 1

1. **Set** Sum = 0

2. **Do** N times

3. **Generate** $\vec{t_i} = (t_{i1}, t_{i2}, \ldots, t_{im})$ sampling from $\vec{z_{0i}}$

4. **Do** until convergence

5. **Set**

$$\theta_{i+1} = \theta_i + \frac{\alpha_i}{2\Delta\theta_i}\left[\pi(\theta_i + \Delta\theta_i t_i) - \pi(\theta_i - \Delta\theta_i t_i)\right] t_i$$

6. **Set** Sum = Sum + θ_{i+1}

7. **Goto 3.**

8. **Goto 2.**

9. **Set** Sum = Sum/N

10. **Set** $\hat{\theta} = \theta_n$.

Algorithm 2

1. **Get** the parameter $\vec{\theta} = (\theta_1, \ldots, \theta_m)$, and $\pi(\theta \mid E)$.

2. **Generate** the random vector $\vec{z} = (z_{11}, z_{12}, \ldots, z_{1n})$ for i = 1,2,...,m.

3. **Collect** $\vec{z}_{01} = \begin{bmatrix} z_{11} \\ z_{21} \\ z_{31} \\ \vdots \\ z_{n1} \end{bmatrix}$, $\vec{z}_{02} = \begin{bmatrix} z_{12} \\ z_{22} \\ z_{32} \\ \vdots \\ z_{n2} \end{bmatrix}$, ..., $\vec{z}_{0m} = \begin{bmatrix} z_{1m} \\ z_{2m} \\ z_{3m} \\ \vdots \\ z_{nm} \end{bmatrix}$,

4. **Send** $\vec{z}_{01}$, $\vec{z}_{02}$, ..., $\vec{z}_{0m}$ to processors 1, 2, ...,m.

5. **Algorithm 1.**

6. **Get** $\hat{\theta}_1, \hat{\theta}_2, \ldots\ldots, \hat{\theta}_m$ i.e. the Monte Carlo estimates of θ_i.

7. **Consider** $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m)$, as an estimation of θ such as $\pi(\hat{\theta} \mid E)$ is an optimal estimate of $\pi(\theta \mid E)$ without the need to resort to the derivative method.

Conclusion

This approach allows a solution to be obtained when there is no derivative. Further by virtue of parallel processing it allows complex models containing many experts to be calibrated. It is intended to consider the problems of dependencies amongst experts in a further paper where the computational demands are considered to be excessive.

References

1. French, S., Group Consensus Probability Distribution: A Critical Survey. In *Bayesian Statistics* **2**, ed J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith. North Holland, Amsterdam, (1985), pp. 183-201.
2. Cooke, R. M., Expert Opinion and Subjective Probability in Science. Delft University of Technology Report, Chapter 11, (1990).
3. Genest, C. & Zidek, J. V., Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science*., **1** (1986) 114-48.
4. Lindley, D. V., Reconciliation of Probability Distributions. *Operational Research.*, **31** (1983) 866-80.
5. Lindley, D. V. & Singpurwalla, N., Reliability (and Fault Tree) Analysis Using Expert Opinions. *Journal American Statistical Association*., **81** (393) (1986) 87-90.
6. Lindley, D. V., Tversky, A. & Brown, R. V., On the Reconciliation of Probability Assessments (with discussion). *J. R. Statist. Soc. Ser A*, **142** (1979) 146-80.
7. Morris, P. A. Combining Expert Judgements: A Bayesian Approach. *Management Science*., **23** (1977) 679-93.
8. Morris, P. A., An Axiomatic Approach to Expert Resolution. *Management Science*., **29** (1983) 866-80.
9. Winkler, R. L., The Consensus of Subjective Probability Distributions *Management Science*., **15** (1968) B61-B75).
10. Winkler, R. L., Combining Probability Distributions from Dependent Information *Sources. Management Science*., **27** (1981) 479-88.
11. Mosley, A. Bayesian modeling of expert-to-expert variability and dependence in estimating rare event frequencies. *Reliability Engineering and System Safety* **38** (1992) 47-57.
12. Zacks, S. The Theory of Statistical Inference. John Wiley and Sons. New York (1971)
pages (230-233).
13. Halton, J. H., A retrospective survey of the Monte Carlo method. *Siam Rev*, **12(1)** (1970) 1-63.

14. Rubinstein, R.Y., Simulation and the Monte Carlo Method. John Wiley and Sons. New York (1981) Page 238.

Contact Information: Ray Gallagher. Telephone 44-151-794-3161. Facsimile 44-151-3715