

COMMONWEALTH OF AUSTRALIA

Copyright Regulations 1969

WARNING

This material has been reproduced and communicated to you by or on behalf of Monash University pursuant to Part VB of the *Copyright Act 1968* (**the Act**).

The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

Do not remove this notice.

Digitised by Monash University Library

An Invariant Bayes Method for Point Estimation

C.S. Wallace and D.M. Boulton

Department of Computer Science, Monash University,
Clayton, Victoria, 3168, Australia

INTRODUCTION

We consider the problem of forming a point estimate of an unknown parameter, based on the result of an experiment, and using, via Bayes' theorem, prior information about the parameter. For simplicity of notation, we shall not distinguish between random variables and values assumed by them. Thus if a variable x can take a value x_i , we shall write $\text{prob}(x_i)$ for $\text{prob}(x=x_i)$ etc.

The unknown parameter h is known a priori to lie in a space H , which we suppose to be endowed with a σ -field of subsets. Normally H will be a subset of a Euclidean space. Prior information about h is embodied in a probability measure on this field, which we suppose to be described by a prior probability density function $f(h)$ with respect to a measure denoted by dh . We assume

$$\int_H f(h)dh = 1 \quad \dots 1.1$$

We assume that $f(h)$ is completely known. Thus we are not concerned with the problem of constructing a prior density function to represent some degree of ignorance, nor are we concerned with improving a poorly known $f(h)$ using additional observations or experiments. We take $f(h)$ as given, and are concerned only with the problem of how to derive an estimate of h from $f(h)$ and observed data.

An experiment is designed to improve our knowledge of h . The outcome of the experiment is an observation x known a priori to belong to a set X . We will assume that X is a countable set of possible observations $\{x_i | i=1,2, \dots\}$. It is perhaps more customary to consider that x may range over a continuum. However in practice, even when the observation includes measurements of real-valued quantities such as mass, length etc., the observation as recorded and made available to the statistician is necessarily conveyed by a message of finite length in some finite alphabet. The set of all such messages can be placed in one-to-one correspondence to the positive integers, and hence is countable. Thus the set X of possible observations is always countable, and the more customary treatment of it as a continuum is only an approximation to the practical situation.

We assume that the outcome of the experiment is a random variable whose dependence on the unknown parameter is expressed by a known conditional probability function $p(x|h)$ defined for all $x \in X$ and all $h \in H$ and satisfying

$$\sum_i p(x_i|h) = 1 \quad \text{for all } h \in H \quad \dots 1.2$$

Note that except where otherwise stated, the range of i is to be taken as $\{i|x_i \in X\}$.

An estimation problem is characterised by the quadruple $\{H, X, f(\cdot), p(\cdot|\cdot)\}$. A complete solution to an estimation problem is a mapping of X into H . That is, it is a function or rule which for each possible observation yields an estimate of h .

One route to a solution lies through the posterior probability density given by Bayes' theorem as

$$f^*(h|x) = p(x|h) f(h)/r(x) \quad \dots 1.3$$

where $r(x)$, the prior probability distribution of observations, is given by

$$r(x) = \int_H p(x|h) f(h) dh \quad \dots 1.4$$

We write $r(x_i)$ as r_i for all i .

In particular, if a cost function is known which expresses the cost of making an estimate h' when the true value of the parameter is h'' , a minimum-cost estimate can be found by the standard methods of decision theory. However, there are occasions when the use or range of uses to be made of an estimate is unknown at the time of estimation, so no useful cost function can be defined. In such situations the ideal is perhaps to publish the posterior probability density rather than making an estimate, but this ideal is not generally feasible when h is a complicated vector of parameters. It is certainly the case that the conclusions of most scientific investigations of unknown parameters are expressed and published as point estimates suitably qualified by some indication of their expected error. We therefore consider the problem of point estimation in the absence of a cost function to be a real problem, even if it should ideally not arise.

In the absence of a cost function, one can derive a point estimate from the posterior probability density by choosing that value of h which maximises the posterior density. However, the resulting estimate is not invariant under a change of description of the parameter space. That is, if g is a parameter which is a nonlinear function of h , the value of g which maximises the posterior density of g does not generally correspond to the value of h which maximises the posterior density of h . The same applies to the mean, and the median is unavailable if H has more than one dimension. The lack of invariance of such point estimates makes them suspect in situations where it may be expected that the estimate will be used in a wide range of calculations of varying mathematical form. We therefore in this paper derive an approach to the general point estimation problem which yields estimates which are invariant under 1 to 1 transformations of the parameter space H .

The estimation method developed here turns out to have remarkable generality, being applicable to problems in which H is not a subset of a Euclidean continuum, but rather the set union of several continua of different dimensionality. The method therefore has potential in the analysis of distributions which are mixtures of several simple distributions. Further, the method is shown to be well suited to empirical Bayes estimation, where the prior density $f(h)$ is known only by inference from estimates of the quantities r , defined in (1.4).

2. THE DESCRIPTION OF A SOLUTION

Any proposed solution to a given point estimation problem $\{H, X, f(\cdot), p(\cdot|\cdot)\}$ is fully described by a function $m(x)$ defined on X which for each observation in X yields a parameter value in H . Let H^* be the range of $m(x)$.

$$H^* = \{m(x) | x \in X\} \quad \dots 2.1$$

H^* is no more numerous than X and hence is countable. In fact, we would expect that often, a proposed solution would yield the same estimate for several different observation values, in which case the mapping of X into H^* is many to one. We now construct a description of a proposed solution which makes such many-to-one mappings more explicit.

Since H^* is countable, we may index its members.

$$H^* = \{h_j | j = 1, 2, 3 \dots\} \quad \dots 2.2$$

Each distinct parameter value is included in the indexed sequence only once, even if it may result from several different observations.

For each distinct parameter value in H^* , define the index set

$$c_j = \{i | m(x_i) = h_j\} \quad \dots 2.3$$

Also define

$$C = \{c_j | h_j \in H^*\} \quad \dots 2.4$$

Set c_j contains the indices of those observations which result in an estimated parameter value h_j . Then the indexed sets H^* and C provide an alternative description of the proposed solution. Note that the adoption of this description rather than a description in terms of the mapping function $m(x)$ in no way restricts the proposed solution: any solution may be described in both ways.

Note that we are concerned not with the question of making a single estimate given a single specific observation, but rather with the more general question of devising, perhaps before the experiment or observation is conducted, an estimation function which will yield an estimate from whatever x may later be obtained.

3. CONSTRUCTION OF A SOLUTION

The foregoing sections have merely outlined a notational framework for describing a point estimation problem and solutions to it, and have not suggested any way of constructing a solution. We will now do so.

3.1 Discrete Problems

In problems where one of a discrete set of mutually exclusive hypotheses, each having a known finite prior probability, is to be chosen on the basis of some observation x , it is common to choose that hypothesis having the highest posterior probability, or, equivalently, to choose the hypothesis which maximises the joint probability

$$P(h,x) = P(x|h)P(h) = b(h,x) \quad \text{say} \quad \dots 3.1$$

We do not here advance any argument in favour of this choice, save to note that it in some sense yields the most plausible, or least improbable account of what has been observed.

We shall adopt the same general aim in choosing an estimate out of the continuum H .

Of course, we cannot apply this criterion directly to the point estimation problem, because in general no finite prior probability can be associated with any point parameter value in the continuum H . We instead proceed by substituting for the given point estimation problem a discrete problem chosen to be in some sense a close approximation to the given problem. That is, we replace the continuum H by a discrete subset of values whose members are assigned finite prior probabilities, and then for each $x \in X$, choose that member of the subset which maximises the joint probability of estimate and observation.

Since the discrete subset of H is, by definition, the range of the estimation function, we may identify it with the H^* of section 2, and adopt the notation of section 2 for describing a solution to the discrete problem.

For all j , define q_j as the prior probability assigned to parameter value h_j . Here and subsequently, the phrase "for all j " and summation over j are to be read as including all and only those values of j such that $h_j \in H^*$.

We require that

$$\begin{aligned} \sum_j q_j &= 1 &&) \\ j &&&) \\ &&&) \\ q_j &> 0 \text{ for all } j &&) \end{aligned} \quad \dots 3.2$$

The usual Bayes criterion of choice among discrete, mutually exclusive hypotheses outlined above then leads to the following condition on the set C of index sets:

$i \in c_j$ implies

$$p(x_i | h_j)q_j \geq p(x_i | h_{j'})q_{j'}, \text{ for all } i \text{ and } j' \neq j \quad \dots 3.3$$

Except perhaps for some observations where (3.3) gives exact equality, (3.3) is sufficient to define the index set c_j for all j , and hence to determine the mapping $m(x)$. Thus, (3.3) essentially allows us to solve the approximate discrete estimation problem. We have now to consider how the original problem can best be approximated by a discrete problem, that is, we have to decide how to select the discrete subset H^* , and how to assign prior probabilities to its members.

3.2 Basis of choice of the approximate problem

In replacing the given problem involving the continuum H and the density $f(h)$ by an approximate problem involving a discrete subset H^* of possible parameter values and their prior probabilities q , we must of course ensure that the discrete approximation to H is in some sense a close approximation. The "goodness" of this approximation can best be assessed by comparing the observation prior probabilities $\{r_i | i=1,2 \dots\}$ implied by the original problem with the observation prior probabilities which would obtain were the discrete model true, viz., the quantities $\{r_i^* | i=1,2 \dots\}$ where

$$r_i^* = \sum_j p(x_i | h_j)q_j = \sum_j b(h_j, x_i) \quad \dots 3.4$$

However, we shall not attempt to choose the discrete model on this basis. Instead, we shall choose the model to yield the highest possible values of the joint probability $b(h,x)$, following the general aim of Bayesean choice. Later, we shall return to the question of whether the resulting model is a reasonably good approximation to the given problem, on the basis of comparing r_i^* and r_i for all i .

It is clearly not meaningful to choose the model so as to maximise $b(h,x)$ for any particular observation x . Instead, we must choose the model to maximise some sort of average value of $b(h,x)$ over all x .

3.3 Expected log joint probability

In this section we shall argue that the discrete model should be chosen to maximise the expected value of the logarithm of the joint probability of the estimate and observation obtained within the context of the model. The argument can be expressed most simply in a context of assumptions which allows a frequency interpretation of the prior probability density. These assumptions follow.

We suppose that the subject of the experiment is a member of an infinite population whose members are each characterised

by a different unknown, value of the parameter h , and that the subject of the experiment may be regarded as being randomly selected from this population. We suppose that the selection process can be repeated indefinitely, each time yielding a different member of the population with, in general, a different value of h , and that the population and selection process are such that the probability on each selection of selecting a member with a value of h lying in some subset A of H is given by

$$\int_A f(h) dh \quad \dots 3.5$$

Note that we assume the prior density $f(h)$ to be already known exactly so that even if many members were selected and their parameters estimated, the probability that a further selection would give a member having $h \in A$ remains as given by (3.5).

In introducing the above assumptions we do not intend to suggest that the concept of prior probability density is only meaningful in such a context, or that the present estimation method can only be applied in this context. However, they permit a simple argument to be given for the desirability of maximising the expected log joint probability.

Suppose that a large number N of different subjects is selected from the population, and that the experiment specified in the statement of the point estimation problem is performed on each of these subjects in turn. Let the ordered set of observations so obtained be

$$Y = \{y_1, y_2, y_3, \dots, y_n, \dots, y_N\}$$

Each observation is of course some member of X , but we use the symbol y to avoid confusion with the use of x subscripted to label and distinguish the members of X . Thus we might in some particular sequence of experiments have $y_1 = x_{77}$, $y_2 = x_{31}$, $y_3 = x_{31}$, $y_4 = x_{99}$ and so on.

Let the ordered set of parameter values possessed by the N different subjects be

$$G = \{g_1, g_2, \dots, g_n, \dots, g_N\}$$

where $g_n \in H$ for all n .

Now consider the problem of forming estimates for each of these values. This may be considered as a single, combined, estimation problem, the result of which is an ordered set of N estimated values

$$K = \{k_1, k_2, \dots, k_n, \dots, k_N\}$$

where $k_n \in H^*$ for all n .

That is, the combined estimate is sought within the context of the approximate, discrete, model.

The joint probability of the observation sequence Y and parameter value sequence G may be written as

$$J(G,Y) = \prod_{n=1}^N b(g_n, y_n)$$

since each subject selection and experiment process is independent of all others in the sequence.

Our aim will be to choose the discrete model of the problem and the estimate sequence K so as to give the highest possible joint probability. That is, we will attempt to choose the model and K to maximise the function

$$J(K,Y) = \prod_{n=1}^N b(k_n, y_n) \quad \dots 3.6$$

For any given discrete model, we can immediately observe that each member of the estimate sequence K affects only one factor in the product (3.6), and that all factors have the same functional form. Hence the maximum possible value of (3.6) can be achieved by an estimate sequence satisfying

$$y_n = y_{n'} \rightarrow k_n = k_{n'}, \text{ for all } n, n'. \quad \dots 3.7$$

We will confine our search for an optimum K to such sequences. For any sequence satisfying (3.7), we can write

$$k_n = m(y_n) \text{ for all } n$$

where $m(\cdot)$ is some as yet undetermined function. Then we can write

$$J(K,Y) = \prod_{n=1}^N b(m(y_n), y_n)$$

The sequence Y is a sequence of members of X. Define D_i as the number of times observation value x_i occurs in Y, for all i. Then

$$J(K,Y) = \prod_i \{b(m(x_i), x_i)\}^{D_i} \quad \dots 3.8$$

We aim to choose the discrete model and function $m(\cdot)$ to maximise $J(K,Y)$. Direct maximisation of (3.8) is very difficult and requires knowledge of the hypothetical sequence of observations Y. However, if the sequence of experiments is very long, i.e. if N is large, then the number of occurrences of observation value x_i in the sequence Y will be approximately Nr_i . Hence we may approximate (3.8) by

$$J(K,Y) \approx \prod_i \{b(m(x_i), x_i)\}^{Nr_i} \quad \dots 3.9$$

and choose the discrete model and $m(\cdot)$ to maximise (3.9), or, equivalently, to maximise its logarithm divided by N , viz.,

$$\begin{aligned} \frac{1}{N} \sum_i N r_i \ln b(m(x_i), x_i) &= \sum_i r_i \ln b(m(x_i), x_i) \\ &= B \text{ (say)} \end{aligned} \quad \dots 3.10$$

Expression (3.10) can be recognised as the expected value of the logarithm of the joint probability of estimate and observation in a single experiment on a single, randomly chosen subject. The value of B depends upon the choice of the discrete subset H^* of parameter values and their prior probabilities q which characterise the discrete model, and upon the function $m(\cdot)$ which maps X onto H^* . All of these will be chosen to maximise B . The result will be to give a model and estimation rule which, for a long sequence of experiments, maximises the joint probability of the sequence of estimated parameters and the sequence of observations. We then argue that, since each subject and experiment of the sequence is independent of the others, and since knowledge of observations and estimates early in the sequence in no way improves or modifies our expectations about later members of the sequence, an estimation process which is optimal for the sequence is simply the repeated application, to each observation in turn, of a process which is optimal for a single experiment. That is, the model found by maximising B is also the optimal model and estimation function to use for a single experiment on a single, randomly chosen, subject.

The above justification for choosing the solution which maximises B is valid only for the frequentist context outlined at the beginning of this section. However, we believe that arguments can be advanced for adopting this solution in other contexts. These arguments will not be advanced in this paper, as they are rather lengthy, and also because the frequentist context assumed here models reasonably well the contexts of a significant proportion of practical estimation problems.

3.4 Maximization of B

The set of index sets C as defined by (2.4) is a partition of the set $\{i | x_i \in X\}$. Hence in (3.10), summation over i may be replaced by the nested summation

$$B = \sum_j \sum_{i \in C_j} r_i \ln b(m(x_i), x_i) \quad \dots 3.11$$

Using (2.3) and (3.1)

$$\begin{aligned} B &= \sum_j \sum_{i \in C_j} r_i \{ \ln p(x_i | h_j) + \ln q_j \} \\ &= \sum_j \left(\sum_{i \in C_j} r_i \right) \ln q_j + \sum_j \sum_{i \in C_j} r_i \ln p(x_i | h_j) \quad \dots 3.12 \end{aligned}$$

The above expression for B is a function of the original problem definition $\{H, X, f(\cdot), p(\cdot|\cdot)\}$, the discrete model described by H^* and $\{q_j | j = 1, 2, \dots\}$, and the mapping $m(x)$ or, equivalently C. To compute B, first the prior observation probabilities $\{r_i | x_i \in X\}$ are computed from (1.4). (These do not involve the discrete model problem or its solution.) The discrete subset H^* is then indexed in some convenient way, and the set C of index sets constructed by inspection of $m(x)$. Computation of B is then straightforward.

Unfortunately, it is not easy to choose H^* , the prior probabilities q , and $m(x)$ to maximise B. It can be shown that the optimum choice satisfies three conditions, which are of some help in finding it. These are:

(a) For each j , holding c_j fixed,

$$h_j \text{ maximises } \sum_{i \in c_j} r_i \ln p(x_i | h_j) \quad \dots 3.13$$

Thus h_j is in effect a compromise maximum likelihood estimate in that it maximises the expected logarithm of the conditional probabilities of observations resulting in that estimate. This condition follows by noting that, if c_j is fixed, variation of h_j affects only one term in the second sum of (3.12), namely, the term written in (3.13) above.

(b) For all j , holding H^* and C fixed,

$$q_j = \sum_{i \in c_j} r_i \quad \dots 3.14$$

This condition follows by noting that q_j appears only in the first sum of (3.12). To find the optimum value of q_j subject to (3.2), we use a Lagrange multiplier and differentiate, obtaining

$$\frac{\partial}{\partial q_j} \sum_{j'} \left(\sum_{i \in c_{j'}} r_i \right) \ln q_{j'} - \lambda \sum_{j'} q_{j'} = 0$$

$$\sum_{i \in c_j} r_i = \lambda q_j \quad \dots 3.15$$

$$\text{But, } \sum_j \sum_{i \in c_j} r_i = \sum_i r_i = 1$$

Therefore $\lambda = 1$, whence (3.14)

(c) For all i and j , and for all $j' \neq j$

$$i \in c_j \text{ implies: } p(x_i | h_j) q_j > p(x_i | h_{j'}) q_{j'}, \quad \dots 3.16$$

This condition is virtually a restatement of (3.3), which was obtained directly from the aim of maximising joint probabilities. However, it is worth noting that (3.3) is implied by maximization of B , and slightly strengthened in that the weak inequality of (3.3) is replaced by strict inequality in (3.16).

To show that (3.16) is implied by maximization of B , suppose the contrary, i.e., that there exists some observation x_i such that $m(x_i) = h_j$, and there exists another estimate $h_{j'}$, such that $p(x_i | h_j) q_j \leq p(x_i | h_{j'}) q_{j'}$. Now consider the change in the value of B which will occur if x_i is reassigned to estimate $h_{j'}$, keeping H^* and all q values constant. The value of B will change by the amount

$$\ln(p(x_i | h_{j'}) q_{j'}) - \ln(p(x_i | h_j) q_j)$$

which is not negative. Hence B is not decreased by the change. However, the change in assignment will make the old values of q_j and $q_{j'}$ no longer conform to (3.14), and q_j must be decreased by r_i , $q_{j'}$ increased by r_i . This change is easily shown to increase B . Moreover, the reassignment may leave h_j and $h_{j'}$ no longer satisfying (3.13), and any consequent change of them must further increase B . Hence unless (3.16) holds, a change to $m(x)$ can be made which will increase B .

Relations (a), (b) and (c) can be applied iteratively to improve a trial solution. Holding H^* and the q values constant, relation (c) is used to reassign observations to estimates. Then relations (a) and (b) are used to recompute the estimate values and prior probabilities. This cycle can be repeated until no change occurs, and is guaranteed to converge in a finite number of cycles. Unfortunately, many problems have large numbers of near-optimal solutions satisfying all three conditions, and hence not improved by the iteration.

4. SOME PROPERTIES OF THE SOLUTION

In section 3 we have shown that, given a Bayes estimation problem $\{H, X, f(\cdot), p(\cdot | \cdot)\}$, a discrete model of the problem may be constructed and solved, giving an estimation function $m(x)$ which maps X into H^* , a countable subset of the parameter space H . We have argued that the discrete model and its solution should be chosen so as to maximize the expected logarithm of the joint probability of the observation and the estimate.

Although the resulting solution is a conventional Bayesian

way of choosing one member of the discrete set of possible parameter values in the model, the model is only an approximation to the given problem, in that the continuum H and prior probability density $f(h)$ are replaced respectively by H^* (a discrete subset of H) and a discrete set of finite prior probabilities $\{q_j | h_j \in H^*\}$.

However, we believe that the solution to the model problem merits serious consideration as a solution to the original, continuous problem.

The value of any proposed solution to the point estimation problem is ultimately determined by the properties of the resulting estimates. Thus, the wide acceptance of maximum likelihood estimation is based on its possession of desirable properties such as invariance, dependence on sufficient statistics where these exist, etc.

In this section we discuss some of the properties of the estimation method derived in section 3.

4.1 Invariance

In the introduction we argued the desirability of finding a solution which is invariant under changes of description of the parameter space H . The solution obtained in section 3 has this property. The expression (3.12) for B depends on the prior probability density $f(h)$ only via the prior observation probabilities given by (1.4). Hence the value of B (and therefore the optimum $m(x)$) is unchanged by any transformation of H which leaves these unchanged. We can expect this to be the case for any useful description of the parameter space, as the prior observation probabilities have a significance independent of any description of H , and, in the frequential context of section 3.3, are in principle measurable.

4.2 Generality

The fact that $f(h)$ appears only in the integration of (1.4) means that H need not be a Euclidean space. We require only that H and $f(h)$ be such that integration of $p(x_i | h)$ with respect to the measure $f(h)dh$ be possible for all i . This generality has allowed us to use a computationally simplified approximation to the method in problems in numerical taxonomy where H is the union of a large number of continua of differing dimensionality (Wallace and Boulton, 1968). These problems cannot satisfactorily be solved by the method of Maximum Likelihood, as the likelihood function has no meaningful maximum.

More generally, the method is applicable to the estimation of the parameters of the components of a population which is a mixture of several simply distributed components. When the number of components is known the mixture problem is amenable to Maximum Likelihood estimation (Wolfe, 1970). However, the present method may be applied to populations which are mixtures

of an unknown number of components, and, as we hope to show in a later paper, has certain advantages over Maximum Likelihood estimation even when the number of components is known.

4.3 Continuous Observation Sets

In the derivation of the method, use was made of the fact that the set of observations which can arise from a given experiment is necessarily discrete. However, this fact is not essential to the method, and in many cases where the accuracy of measurement of observations is high, it would be more convenient to treat X as a region of a continuum. The set H^* of possible estimates remains discrete, but treating the observation x as a continuously-variable quantity in X , we replace

$p(x_i|h)$ by the conditional probability density $\pi(x|h)$;

$r(x_i)$ by the prior probability density $\rho(x)$

$$= \int_H \pi(x|h) f(h) dh;$$

and for each j such $h_j \in H^*$, we replace the index set c_j by a region v_j of X such that $x \in v_j$ implies $m(x) = h_j$. The set $V = \{v_j | h_j \in H^*\}$ forms a partition of X .

We can no longer write an expression analogous to (3.12) for the expected log joint probability, as no x value has a finite probability. However, in the discrete- X case, maximization of (3.12) is equivalent to maximization of

$$D = \sum_j \left(\sum_{i \in c_j} r_i \right) \ln q_j + \sum_j \sum_{i \in c_j} r_i \ln p(x_i|h_j) - \sum_j \sum_{i \in c_j} r_i \ln r_i \quad \dots 4.1$$

which differs from (3.12) only by the constant term

$$\sum_j \sum_{i \in c_j} r_i \ln r_i = \sum_i r_i \ln r_i$$

Writing

$$D = \sum_j \left(\sum_{i \in c_j} r_i \right) \ln q_j + \sum_j \sum_{i \in c_j} r_i \ln \{p(x_i|h_j)/r_i\} \quad \dots 4.2$$

we see that an analogous expression for the continuous- X case can be written as

$$D_c = \sum_j \left(\int_{v_j} \rho(x) dx \right) \ln q_j + \sum_j \int_{v_j} \rho(x) \ln \{\pi(x|h_j)/\rho(x)\} dx \quad \dots 4.3$$

In fact, if a continuous observation space is approximated by successively finer discretizations, the maximum value of D for

the discrete approximations to X approaches the maximum value of D_c as a limit under appropriate conditions, and the resulting mappings also approach the mapping obtained by maximization of D_c .

4.4 Sufficient Statistics

For some estimation problems, there exists a (possibly vector-valued) function $z(x)$ of the observation such that for all x and h , $p(x|h)$ may be expressed as $p(x|h) = p_1(z(x)|h) p_2(x)$...4.4

where the function p_2 does not involve h . Then $z(x)$ is a sufficient statistic for h in the Neymann-Pearson sense, and the value of $z(x)$ for any observation contains all the information about h contained by the observation.

Condition (3.16) on the optimum mapping function can then be rewritten as:

For all i and j , and for all $j' \neq j$,

$i \in c_j$ implies

$$p_1(z(x_i)|h_j) q_j p_2(x_i) > p_1(z(x_i)|h_{j'}) q_{j'} p_2(x_i)$$

i.e. $p_1(z(x_i)|h_j) q_j > p_1(z(x_i)|h_{j'}) q_{j'}$...4.5

Hence for any i, i'

$$z(x_i) = z(x_{i'}) \text{ implies } m(x_i) = m(x_{i'}) \quad \dots 4.6$$

That is, in our method, the estimate resulting from any observation depends only on the value of the sufficient statistic derived therefrom. It follows that in any estimation problem, the set X of possible observations may be replaced throughout the calculation by the set of distinct possible sufficient statistic values. The resulting solution to the problem has the same value of D (4.1) and gives the same estimate for every observation as the solution derived directly from X .

4.5 Closeness of model

In choosing the discrete model to maximise B , no attention was given to whether or not the resulting model was a good approximation to the given problem. In section 3.2 we suggested that the best criterion of the goodness of the approximation which could be considered within the very loose assumptions we have made about H is a comparison of the observation prior probabilities r implied by the given problem via (1.4) and the approximations r^* given by the model via (3.4).

Consider the function

$$\sum_i r_i \ln w_i \quad \dots 4.7$$

where the variables $\{w_i | x_i \in X\}$ are free variables constrained only by

$$\begin{aligned} \sum_i w_i &= 1 &&) \\ w_i &> 0 \text{ all } i &&) \end{aligned} \dots 4.8$$

It is easy to show that (4.7) is maximised with respect to the w's when

$$w_i = r_i \text{ for all } i$$

and that this maximum is unique. Hence we may obtain a measure of the degree to which the r^* approximate the r by computing the difference

$$\sum_i r_i \ln r_i^* - \sum_i r_i \ln r_i \dots 4.9$$

Since the quantities r^* satisfy (4.8), the value of (4.9) cannot be positive and can be zero only if $r_i^* = r_i$ for all i . Its value is a measure of the extent to which the model departs from reality.

Now for all i

$$\begin{aligned} r_i^* &= \sum_j b(h_j, x_i) \\ &> b(m(x_i), x_i) \end{aligned}$$

Hence

$$\begin{aligned} \sum_i r_i \ln r_i^* &\geq \sum_i r_i \ln b(m(x_i), x_i) \\ \sum_i r_i \ln r_i^* - \sum_i r_i \ln r_i &\geq \sum_i r_i \ln \{b(m(x_i), x_i)/r_i\} \dots 4.10 \end{aligned}$$

The right hand side of (4.10) can be recognised as the quantity D defined by (4.2), which differs from B only by a constant. Hence, in choosing the model to maximise B , we are maximising a lower bound to (4.9). To this extent at least, the method can be expected to construct discrete models which are reasonably close to the given problem.

The problem D is the expected logarithm of

$$b(h, x)/r$$

where h is the parameter value estimated from observation x , and r is the marginal probability of observation x .

From (3.1) we can write

$$b(h, x)/r = P(x|h)P(h)/P(x) \dots 4.11$$

Thus D resembles in form the expected logarithm of the posterior probability of the estimated parameter value. However, in (4.11), the value of $P(h)$ is the prior probability q associated

with the estimate in the discrete model problem, whereas $P(x) = r$ is the probability of observation x in the given problem. Thus D cannot be interpreted strictly as the expected log posterior probability. However, its value can, perhaps, be taken as giving some indication of the quality of the estimates produced by the method. Its continuous- X analog D_c has a similar significance for continuous- X problems.

4.6 Empirical Bayes Estimation

The method is well adapted to problems of empirical Bayes estimation, that is, to problems where the prior probability density $f(h)$ is unknown, but estimates are available of the marginal observation probabilities r . The usual approach to such problems (Maritz, 1970) is to assume some parameterised form $f(h, \theta)$ for $f(h)$, and then estimate the parameters θ of the assumed form by fitting

$$\int_H p(x_i | h) f(h, \theta)$$

to the known or estimated value r_i for all i . The present method allows the recovery of $f(h)$ from the marginal observation probabilities to be bypassed, since the method depends on $f(h)$ only via the values of r . When these are known, the method may proceed directly from them without any assumptions about or estimation of the prior probability density.

5. A NUMERICAL EXAMPLE

The problem considered is estimating the probability of success in a Bernoulli trial, given the results of M trials. Let h be the probability of success and let $f(h)$ ($0 \leq h \leq 1$) be the known prior probability density of h . By virtue of section 4.4, we can regard the observation as being the number of successes in M trials, since this is known to be a sufficient statistic for h . The set X is the set of integers from 0 to M inclusive.

Solutions to the problem were obtained for two sample sizes $M = 20$ and $M = 100$, and for three prior density functions

$$\begin{aligned} f(h) &= 1 \\ f(h) &= 2(1-h) \\ f(h) &= 3(1-h)^2 \end{aligned}$$

The above density functions, which are of Beta form, were chosen simply because they simplified the calculation and maximization of B . The fact that they happen to lead to posterior probability densities which also are of Beta form is irrelevant to the present method.

Tables 1 and 2 show the results for $M = 20$ and $M = 100$ respectively. Each table contains results for the three Bayes estimation problems resulting from the three different prior

density functions. For each estimation problem the set of estimates H^* is listed in the column headed Mh_j , where each estimate value is multiplied by the sample size M to allow a direct comparison with the success counts mapping into it. The range of x values mapping into each estimate is shown under heading c_j .

The value of B , the expected log joint probability, is shown for each problem. The value of D is also shown. As discussed in section 4.5, D resembles formally the expected log posterior probability, and also gives by its magnitude an upper bound on the departure of the model from the given problem.

Also, to show that the quality of the estimate is not markedly dependent on the particular evidence obtained, the table presents for each estimate, minus the value of $\ln(b(h,x)/r)$ averaged over all observations mapped into that estimate, i.e.

$$D_j = - \frac{\sum_{i \in c_j} r_i \{ \ln p(x_i | h_j) q_j / r_i \}}{\sum_{i \in c_j} r_i}$$

The results, together with others not detailed here, suggest the following general observations.

(a) The maximum of B is rather broad. Many non-optimum mappings were found with B values only 10^{-3} or 10^{-2} less than the optimum.

(b) In every case the width of the range of success counts mapping into an estimate h_j is approximately

$$3.5 \sqrt{Mh_j(1-h_j)}.$$

Thus the range of observations associated with a particular estimated mean is of the order of the spread of success counts expected of a process having that mean.

(c) The number of "significantly different" estimates produced by the method is approximately \sqrt{M} .

(d) The shape of the prior probability density has little effect on the solution. Its main effect is to shift the optimum estimate for each c_j towards the values of higher prior probability. The less uniform the prior density, the more it assists the estimation process, resulting in a higher value for B .

(e) The values of D are quite small, and the value of D is for no estimation problem less than -0.168 , corresponding to a "posterior probability" value of about 0.8 .

(f) For the prior distribution and $M = 100$, the problem is symmetrical about $h = 0.5$ and $x = 50$. However, there is an odd number of observation values in X , but an even number of estimate values (10) in H^* . Hence the solution is not exactly symmetrical. In this problem, there exist two equally good solutions, one as shown and the other its "mirror image".

| M = 20 | Flat $f(h) = 1$ | | | linear $f(h) = 2(1-h)$ | | | Quadratic $f(h) = 3(1-h)^2$ | | |
|--------|-----------------|--------|-------|------------------------|--------|-------|-----------------------------|--------|-------|
| | c_j | Mh_j | D_j | c_j | Mh_j | D_j | c_j | Mh_j | D_j |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1-6 | 3.5 | .153 | 1-6 | 3.3 | .144 | 1-6 | 3.2 | .139 |
| 3 | 7-14 | 10.5 | .164 | 7-14 | 10 | .164 | 7-14 | 9.6 | .180 |
| 4 | 15-19 | 17 | .162 | 15-20 | 16.7 | .196 | 15-20 | 16.3 | .284 |
| 5 | 20 | 20 | 0 | - | - | - | - | - | - |
| B | - 3.04452 | | | - 3.01584 | | | - 2.79433 | | |
| D | - 0.14437 | | | - 0.14285 | | | - 0.13697 | | |

Table 1: Results for three binomial Bayes Estimation Problems of sample size 20 and different prior probability densities.

| M = 100 | Flat $f(h) = 1$ | | | linear $f(h) = 2(1-h)$ | | | Quadratic $f(h) = 3(1-h)^2$ | | |
|---------|-----------------|--------|-------|------------------------|--------|-------|-----------------------------|--------|-------|
| | c_j | $M(h)$ | D_j | c_j | Mh_j | D_j | c_j | Mh_j | D_j |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1-6 | 3.5 | .159 | 1-6 | 3.5 | .157 | 1-6 | 3.4 | .155 |
| 3 | 7-17 | 12 | .171 | 7-17 | 11.9 | .169 | 7-17 | 11.8 | .168 |
| 4 | 18-32 | 25 | .173 | 18-32 | 24.8 | .172 | 18-32 | 24.5 | .172 |
| 5 | 33-49 | 41 | .174 | 33-49 | 40.6 | .173 | 33-49 | 40.2 | .175 |
| 6 | 50-66 | 58 | .174 | 50-66 | 57.4 | .175 | 50-66 | 56.9 | .181 |
| 7 | 67-81 | 74 | .173 | 67-81 | 73.3 | .178 | 67-82 | 73.0 | .191 |
| 8 | 82-93 | 87.5 | .173 | 82-93 | 86.6 | .182 | 83-94 | 86.8 | .218 |
| 9 | 94-99 | 96.5 | .159 | 94-100 | 96.8 | .209 | 95-100 | 96.3 | .328 |
| 10 | 100 | 100 | 0 | - | - | - | - | - | - |
| B | - 4.78312 | | | - 4.59468 | | | - 4.35839 | | |
| D | - 0.16800 | | | - 0.16780 | | | - 0.16544 | | |

Table 2: Results for three binomial Bayes Estimation Problems of sample size 100 and different prior probability densities.

Table 3 gives more detailed information about the performance of the method in one particular case, viz, sample size 20 and linear prior probability density. For each observation value, i.e. for each possible success count x_i , it lists the true marginal probability r_i (1.4), the marginal probability r_i^* resulting from the discrete model (3.4), the joint probability of the observation and the parameter value estimated from it $b(m(x_i), x_i)$, and the quantity $b(m(x_i), x_i)/r_i$ which formally resembles the posterior probability of the estimate $m(x_i)$.

It can be seen that the discrete model leads to values of r^* differing considerably from those of r , the ratio between r and r^* ranging between about 0.6 and 1.6. The discrete model is therefore a fairly rough model of the given problem.

The values of the "posterior probability" $b(m(x_i), x_i)/r_i$ also cover a rather wide range, and in some cases exceed unity. It is clear, therefore, that while D , the average logarithm of this quantity, may be useful as a measure of the overall performance of the estimation function, individual values of the quantity have little significance.

The horizontal lines divide the observation values into groups, such that values within the same group map into the same estimate. As might be expected, the values of r^* , joint probability and "posterior probability" are highest for those success counts near the value of Mh_j , and fall off towards the edges of the group.

It might be thought on the basis of these results that the discrete model is excessively coarse, and that it is unreasonable to map such a wide range of success counts into a single estimated parameter value. However, no observation value is significantly inconsistent with the parameter value estimated from it.

Consider the log - likelihood ratio

$$2 \ln \left[\frac{p(x_i | z_i)}{p(x_i | m(x_i))} \right]$$

where z_i is a conventional maximum - likelihood estimate derived from x_i . An approximate analysis of the method, not limited to the binomial example considered here, shows that provided $f(h)$ is slowly varying, and $p(x_i | h)$ considered as a function of h has an approximately quadratic behaviour about its maximum at $h=z_i$, then the maximum value achieved by the above log - likelihood ratio for any x is approximately equal to $(n+2)$, where n is the number of dimensions in H , i.e. the number of scalar parameters being estimated.

| x_i | r_i | r_i^* | $b(m(x_i), x_i)$ | $b(m(x_i), x_i)/r_i$ | Mh_j |
|-------|-------|---------|------------------|----------------------|--------|
| 0 | .0909 | .1028 | .0909 | 1.00 | 1.00 |
| 1 | .0866 | .0474 | .0474 | .55 | 3.3 |
| 2 | .0823 | .0902 | .0901 | 1.09 | |
| 3 | .0779 | .1085 | .1081 | 1.39 | |
| 4 | .0736 | .0936 | .0919 | 1.25 | |
| 5 | .0693 | .0642 | .0588 | .85 | |
| 6 | .0699 | .0429 | .0294 | .45 | |
| 7 | .0606 | .0387 | .0299 | .44 | |
| 8 | .0563 | .0475 | .0437 | .78 | |
| 9 | .0520 | .0593 | .0582 | 1.12 | |
| 10 | .0476 | .0643 | .0641 | 1.35 | |
| 11 | .0433 | .0585 | .0583 | 1.35 | |
| 12 | .0390 | .0445 | .0437 | 1.12 | |
| 13 | .0346 | .0292 | .0269 | .78 | |
| 14 | .0303 | .0193 | .0134 | .44 | |
| 15 | .0260 | .0171 | .0118 | .45 | 16.7 |
| 16 | .0217 | .0201 | .0184 | .85 | |
| 17 | .0173 | .0220 | .0216 | 1.25 | |
| 18 | .0130 | .0181 | .0180 | 1.39 | |
| 19 | .0087 | .0095 | .0095 | 1.09 | |
| 20 | .0043 | .0024 | .0024 | .55 | |

Table 3: Detailed results for sample size 20, $f(h) = 2(1-h)$

Under similar assumptions, it is known that the log likelihood ratio between the maximum likelihood and the likelihood of the true parameter value is approximately distributed as χ^2 with n degrees of freedom. Hence the estimate $m(x_i)$ could be regarded as inconsistent with observation x_i only if a value of $(n+2)$ could be rejected as a reasonable value for χ_n^2 . However, the probability that χ_n^2 should exceed $(n+2)$ is never small, and increases with increasing n . Thus, although the estimates produced by the present method are apparently coarse, they are always consistent with the observations when the approximations above are valid, at least according to the likelihood ratio test.

6. A FURTHER EXAMPLE

The problem considered is estimating the mean of a uniform distribution of known range B , given the values of M variates selected independently and randomly from the distribution. We assume that each variate value is measured to such high resolution that we can ignore the discretization of observations and treat the set of possible observations as a continuum, following section 4.3. We further assume that the known prior probability density $f(h)$ is constant over a range of h values much greater than R , and is zero outside this range. That is, H is some finite interval on the real line of length much greater than R .

There is no single scalar-valued sufficient statistic for h , but the highest and lowest of the M variate values are together jointly sufficient, or, alternatively, the sample range z and midrange y are together jointly sufficient. Following section 3.4, we will take the set X to be the set of possible (y,z) pairs, i.e., a two dimensional continuum. We write $x = (y,z)$. (The symbols y and z here have no connection to any previous use.)

$$z = (\text{highest variate} - \text{lowest variate})$$

$$y = (\text{highest variate} + \text{lowest variate})/2$$

$$\text{Clearly, } 0 \leq z < R \tag{...6.1}$$

Also, for all h , the conditional probability density

$$\pi(x|h) = \pi(y,z|h) = \begin{cases} M(M-1)z^{M-2}(1/R^M) & \text{if } |h-y| < (R-z)/2 \\ 0 & \text{otherwise.} \end{cases} \tag{...6.2}$$

Note that if $\pi(x|h) \neq 0$, it is independent of y and h .

This problem is interesting because for those observations for which z approaches R , the range of h values for which

$\pi(x|h)$ is non-zero becomes vanishingly small. Since any feasible solution must map every observation into an estimate h having non-zero $\pi(h|x)$, the set H^* of possible estimates must be everywhere dense in H . However, we will see that H^* remains countable, being a subset of the rational numbers.

We now derive the optimum solution. Recall from section 4.3 that for continuous x , we must maximize

$$D_c = \sum_j q_j \ln q_j + \sum_j \int_{v_j} \rho(x) \ln \{\pi(x|h_j)/\rho(x)\} dx \quad \dots 6.3$$

where for every j , v_j is the region of X such that $x \in v_j$ if $m(x) = h_j$. The set of regions $V = \{v_j | h_j \in H^*\}$ forms a partition of X .

The continuous- X analog of the optimising condition (3.16) is: For all $x \in X$ and all j and $j' \neq j$

$$x \in v_j \text{ implies } \pi(x|h_j)q_j \geq \pi(x|h_{j'})q_{j'} \quad \dots 6.4$$

To see how 6.4 applies in the present problem, define for every $h \in H^*$ a "possible region" of X bounded by the three lines

$$\begin{aligned} z &= 0 \\ z &= R - 2(y - h) \\ z &= R + 2(y - h) \end{aligned} \quad \dots 6.5$$

The possible region of h so defined includes all and only the observation values such that $\pi(x|h) \neq 0$.

Now, if some observation x lies in the possible regions of two estimates h_j and $h_{j'}$, then by 6.2,

$$\pi(x|h_j) = \pi(x|h_{j'})$$

and 6.4 reduces to

$$x \in v_j \text{ implies } q_j \geq q_{j'}$$

We can therefore express the effect of 6.4 thus:

If h_j and $h_{j'}$ are two estimates in H^* such that v_j and $v_{j'}$ have a boundary in common, and $q_j > q_{j'}$, then this boundary will follow the boundary of the "possible region" of h_j as defined by 6.5.

These considerations, together with the fact that $f(h)$ is independent of h for $h \in H$, and hence that $\rho(x)$ is independent of y except for y values near the limits of the possible range of y , lead to an optimum solution described by the V -partition of X shown in Figure 1. The vertical axis shows the sample range z , and the horizontal axis shows the sample mid-range y . Estimate values are also shown along the horizontal direction.

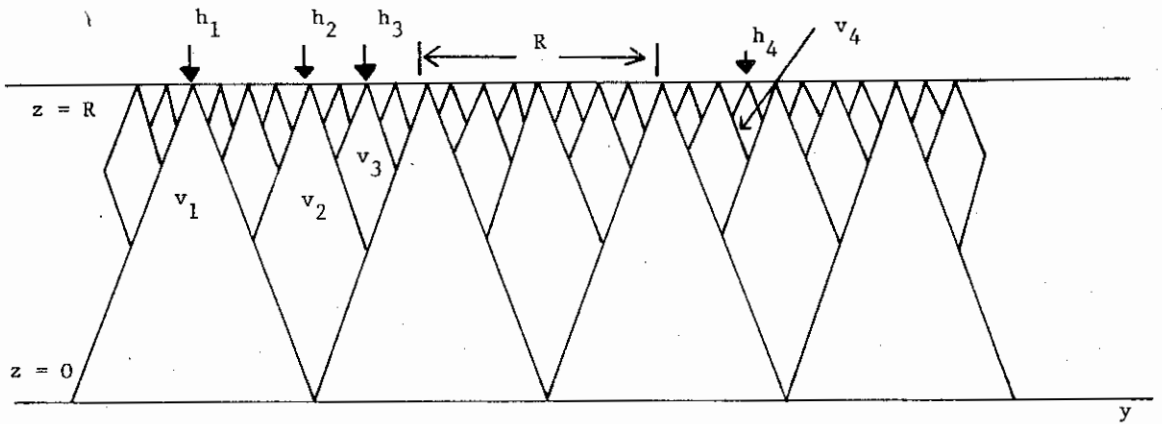


Figure 1: The partition of X into regions mapping into estimates.

For some estimates, such as h_1 in the figure, the associated region of X (e.g. v_1) includes the entire "possible region" of the estimate. Such estimate values are spaced at intervals R and have the highest prior probabilities. For all other estimates, the associated region is a lozenge having one corner on the line $z = R$ at a y coordinate equal to the estimate value. For any such region, say v_3 , the upper sides of the lozenge follow the boundaries of the possible region of the associated estimate, and the lower sides follow the boundaries of the possible regions of estimates having larger v regions, and hence larger prior probability. There is an infinite sequence of smaller regions approaching the line $z = R$.

The effect of the partition may be loosely described by saying that, if the estimated mean is expressed as a binary number in units of R , then the estimate resulting from some observation x is the binary value having the smallest number of digits following the binary point needed to make $\pi(x|h)$ non-zero.

The figure depicts a portion of the partition far from the extreme possible values of y , and the positioning of the pattern along the y axis would be determined by "end effects" near the extreme value of y . However, since it is assumed that the length of H is much greater than R , modifications to the pattern near the ends have negligible effect on the value of D_c .

The partition yields an infinite set of possible estimate values. Taking h_1 as given, any other estimate in H^* has a value of the form

$$h_1 + (n/2^m)R$$

where n and m are integers. Hence the estimates in H^* stand in

one-to-one correspondence with rational numbers of the form $(n/2^m)$, and so H^* is countable.

The prior probability of each estimate is given by the integral of $\rho(y,z)$ over its associated region, and is finite and non-zero.

The value of D_c is, neglecting end effects near the limits of H , independent of R and the length of H . It can be expressed as the sum of an infinite series having one term for each size of v region. The series is rapidly convergent and has been evaluated numerically for several values of the sample size M . The results are given in Table 4. It will be noted that although H^* contains an infinite number of estimates of very low prior probability, the value of D_c is reasonably high. If D_c is loosely interpreted, following section 4.5, as the expected log posterior probability, the logarithmically averaged probability exceeds 0.5.

| Sample size M | D_c |
|-----------------|---------|
| 2 | - 0.463 |
| 5 | - 0.579 |
| 10 | - 0.617 |
| 20 | - 0.636 |

Table 4: Expected log of $b(h,x)/r$ for estimates of the mean of a uniform distribution.

7. CONCLUSIONS AND GENERAL REMARKS

We have shown that it is possible to construct solutions to Bayesian point estimation problems in a way which is invariant under changes in the description of the parameter space H , and which makes no appeal to a cost function. The method may be applied to either discrete or continuous observation sets X and in the latter case is invariant under changes in the description of X . The parameter space need not have a simple structure, as we require only that it support integration over the whole space with respect to the prior probability measure $f(h)dh$. The method is well suited for empirical Bayes estimation as it can work directly from the marginal observation probabilities, if these are known rather than the prior probability density.

The method yields a countable set H^* of possible estimated parameter values, each having a finite non-zero prior probability, and a mapping of X into H^* . The crucial steps in the method are

- (a) replacing the given problem by an approximate model in which H is replaced by a discrete subset H^* of parameter values, each having a finite prior probability, and

- (b) choosing the model and estimation function to maximise the expected log joint probability of parameter and observation.

The argument presented here in support of step (b) is valid only for problems admitting a frequency interpretation of prior probability, but we believe the step can be justified in other contexts.

The two numerical examples illustrate the general tendency of the method to produce a rather coarse set of possible estimate values, with the result that estimates are yielded with a resolution or "accuracy" no higher than is warranted by the observation. Both examples show a tendency for D , which may loosely be interpreted as the expected log posterior probability, to fall with increasing sample size. However, an approximate analysis not given here shows that as the sample size increases, D approaches a limiting value.

The computations involved in maximising B are too difficult for the method to be attractive for general use. However, we have developed close approximations to the method which are computationally feasible and retain the main virtues of the method, viz., invariance of estimates under transformations of H , no dependence on a cost function, and ability to yield meaningful estimates in problems where H has a complex structure and where the likelihood function has no useful maximum. These approximate methods will be described in a later paper.

In this paper we have not touched on the question of the bias or variance of estimates produced by the method. These questions are meaningful only in relation to a specific coordinate system for H , and the present method is independent of the properties, or indeed the existence, of any such coordinate system.

REFERENCES

- Maritz, J.S. (1970). *Empirical Bayes Methods*. Methuen, London.
Wallace, C.S. and Boulton, D.M. (1968). An information measure for classification. *The Computer Journal*, 11, 185-194.
Wolfe, J.H. (1970). Pattern clustering by multivariate mixture analyses. *Multivariate Behavioral Research*, 5, 329-350.