SEQUENTIAL METHODS IN STATISTICS
BANACH CENTER PUBLICATIONS, VOLUME 16
PWN–POLISH SCIENTIFIC PUBLISHERS
WARSAW 1985

A SEQUENTIAL BAYESIAN APPROACH TO ESTIMATING THE DIMENSION OF A MULTINOMIAL DISTRIBUTION

C. G. E. BOENDER

Department of Mathematics, Erasmus University, Rotterdam, The Netherlands

and

RYSZARD ZIELIŃSKI

Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland

Introduction

Let $S_K = \{s_1, \ldots, s_K\}$ be a set and let $X_1, X_2, \ldots$ be independent identically distributed random variables with values in S_K, taking on the value s_j with probability $\theta_j > 0$, $\sum_{j=1}^{K} \theta_j = 1$. Neither K nor (s_j, θ_j), $j = 1, \ldots, K$, are known and, sampling $X_1, X_2, \ldots$ one after another, we are interested in discovering all "classes" s_j, $j = 1, \ldots, K$.

The following two examples reveal the origin of the problem and give some intuitions.

EXAMPLE 1. Suppose we are interested in estimating the number K of fish species living in a newly discovered and completely unknown lake. It is not known in advance what species $s_1, s_2, \ldots, s_K$ live in the lake. The random variable X is now interpreted as the type of a randomly captured fish and θ_j is a number proportional to the species s_j.

EXAMPLE 2. Let f be a continuous function on the unit interval and suppose we are interested in discovering all local minima of f. Now S_K is the unknown list of local minima. The procedure consists in making a random choice of point $x \in (0, 1)$ and then applying an iterative algorithm A for seeking local minima. Write $A(x) = s_i$ if, whenever we start from the point x, the algorthm leads to the local minimum s_i. The random variable X is now interpreted as the local minimum discovered when we start from a random

point, and θ_j is the probability that the starting point lies in the set $\{x: A(x) = s_j\}$ (the set of attraction of the jth local minimum).

The problem is rather old and goes back as far as Goodman's (1949) paper, which is the first paper in the statistical literature on the subject, as we can conclude from the lack of references in it. To the best of our knowledge there exists no general solution to the problem and our approach presents a new endeavour to attack it from the Bayesian standpoint. The problem has recently revived in the context of Example 2 above and our solution is in the spirit of the papers presented in Dixon and Szegö (1975, 1978), Archetti and Cugiani (1980), and a paper by the second of the present authors (Zieliński (1981)).

Optimal stopping problem

Let $w_n = w_n(X_1, X_2, \ldots, X_n)$ be the number od different X_i's in the sequence $X_1, X_2, \ldots, X_n$. Consider the following loss function:

$$L(w_n, K) = \begin{cases} Bn & \text{if } w_n = K, \\ A + Bn & \text{if } w_n \neq K, \end{cases}$$

where A and B are positive constants. The number A is interpreted as the loss connected with not discovering all classes and B is interpreted as the cost of one observation.

Let α_k, $k = 1, 2, \ldots$, be a prior distribution of K and let $\mu_k(d\theta)$ be a conditional prior distribution of $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ given $K = k$. We are looking for n minimizing the Bayes risk. The standard procedure (cf. DeGroot (1970), Ferguson (1967)) consists in evaluating, for any given $n = 1, 2, \ldots$, the posterior expected loss and optimal stopping of that stochastic sequence. In our case we have

$$\text{expected posterior loss} = Aq_n + Bn = A(q_n + cn),$$

where q_n is the posterior probability that all elements of S_K have not been discovered, and $c = B/A$. Now the problem reduces to that of optimal stopping for the stochastic sequence $Y_n = q_n + cn$, $n = 1, 2, \ldots$

Solution for a specific choice of the prior distribution

Given n, $X_1, X_2, \ldots, X_n$, and $w_n = w$, suppose that the discovered elements are $s_{i_1}, s_{i_2}, \ldots s_{i_w}$. Let N_j, $j = 1, 2, \ldots, w$, denote the number of X_j's which are equal to s_{i_j}. Let $[n_1, n_2, \ldots, n_w]$ denote the set of all permutations of numbers $n_1, n_2, \ldots, n_w$ such that $n_j > 0$ for all $j = 1, 2, \ldots, w$ and $n_1 + n_2 +$

$\ldots + n_w = n$. Denote by h_m the number of n_j's which are equal to m. Then

$$P\{w_n = w, (N_1, N_2, \ldots, N_w) \in [n_1, n_2, \ldots, n_w] \mid k, (\theta_1, \theta_2, \ldots, \theta_k)\}$$
$$= \sum_{[i_1, \ldots, i_w] \subset \{1, \ldots, k\}} \frac{1}{h_1! \ldots h_m!} \frac{n!}{n_1! \ldots n_w!} \theta_{i_1}^{n_1} \theta_{i_2}^{n_2} \ldots \theta_{i_w}^{n_w},$$

the summation being extended over all permutations of w different elements of the set $\{1, \ldots, k\}$. E.g., for $k = 3$ and $w = 2$ this means the summation with respect to (i_1, i_2) over the set $\{(1, 2), (2, 1), (1, 3), (3, 1), (2, 3), (3, 2)\}$. A detailed discussion of the above formula as well as of the posterior probabilities given below is presented in Boender and Rinnooy Kan (1982).

Take as the prior distributions $\alpha_k = \text{const}$, $k = 1, 2, \ldots$, (improper distribution on the set of all positive integers), and $\mu_k(d\theta) = \Gamma(k)\, d\theta_1\, d\theta_2 \ldots d\theta_k$ on the set $\{(\theta_1, \theta_2, \ldots, \theta_k)\colon\ 0 < \theta_j < 1, \sum \theta_j = 1\}$, $\mu_k(d\theta) = 0$ otherwise (uniform distribution on the unit simplex in R^k).Then, after some computations, we obtain the density of the posterior distribution

$$P\{K = k, (\theta_1, \ldots, \theta_k) \mid w_n = w, (N_1, \ldots, N_w) \in [n_1, \ldots, n_w]\}$$
$$= \frac{\Gamma(k)\Gamma(n)\Gamma(n-1) \sum_{[i_1, \ldots, i_w] \subset \{1, \ldots, k\}} \theta_{i_1}^{n_1} \theta_{i_2}^{n_2} \ldots \theta_{i_w}^{n_w}}{\Gamma(w)\Gamma(w+1)\Gamma(n-w-1) n_1!\, n_2! \ldots n_w!},$$

and integrating with respect to $(\theta_1, \ldots, \theta_k)$ gives us, for $1 \leqslant w \leqslant n-2$,

$$P\{K = k \mid w_n = w, (N_1, \ldots, N_w) \in [n_1, \ldots, n_w]\}$$
$$= \frac{\Gamma(k+1)\Gamma(k)\Gamma(n)\Gamma(n-1)}{\Gamma(k-w+1)\Gamma(n+k)\Gamma(w+1)\Gamma(w)\Gamma(n-w-1)},$$

which does not depend on $n_1, n_2, \ldots, n_w$ and will be denoted shortly by $P\{K = k \mid w_n = w\}$. If $w_n = n-1$ or $w_n = n$, the posterior distribution of K is again an improper one. Due to this fact we shall confine ourselves to the case where $n \geqslant 3$, which is nonrestrictive from the practical point of view. The above formula gives us

$$q_n = 1 - P\{K = w \mid w_n = w\} = 1 - \frac{\Gamma(n)\Gamma(n-1)}{\Gamma(n+w)\Gamma(n-1-w)}, \qquad 1 \leqslant w \leqslant n-2.$$

By similar arguments we obtain the posterior transition probabilities

$$P\{w_{n+1} = w+1 \mid w_n = w\} = \frac{w(w+1)}{n(n-1)},$$

$$P\{w_{n+1} = w \mid w_n = w\} = 1 - \frac{w(w+1)}{n(n-1)},$$

which describes the stochastic sequence (w_n) except for $w = n-1$ and $w = n$. Now the stochastic sequence (Y_n) is defined as

$$Y_n = 1 - \frac{\Gamma(n)\Gamma(n-1)}{\Gamma(n+w_n)\Gamma(n-1-w_n)} + cn, \qquad 1 \leqslant w_n \leqslant n-2,\ n = 3, 4, \ldots,$$

and the stopping rule N which minimizes the expected loss EY_N is that which maximizes EZ_N with

$$Z_n = Z_n(w_n) = \frac{\Gamma(n)\Gamma(n-1)}{\Gamma(n+w_n)\Gamma(n-1-w_n)} - cn, \qquad 1 < w_n < n-2,\ n = 3, 4, \ldots$$

Observe that

$$E(Z_{n+1} \mid Z_n) = Z_n + g(n, w_n) - c$$

where

$$g(n, w) = \frac{w(w+1)}{(n+w)(n+w+1)} \cdot \frac{\Gamma(n)\Gamma(n-1)}{\Gamma(n+w)\Gamma(n-1-w)}.$$

By the formula

$$g(n, w) = \frac{w+1}{w-1} \cdot \frac{n-w-1}{n+w+1} \cdot g(n, w-1)$$

we conclude that

$$\max_{1 \leqslant w \leqslant n-2} g(n, w) = g(n, v_n)$$

where v_n is the (unique) positive solution of the equation $(w+1)(n-w-1)/(w-1)(n+w+1) = 1$, i.e., $v_n = (\sqrt{1+4n}-1)/2$. It is easy to observe that

$$g(n, v_n) \leqslant \frac{1}{n+1+\sqrt{4n+1}},$$

and hence $g(n, w_n) - c < 0$ for all n greater than

$$N(c) = \left[\frac{1}{c} + 1 - \sqrt{\frac{4}{c}+1}\right] + 1$$

where $[x]$ denotes the integer part of x.

It follows that $(Z_n, n \geqslant N(c))$ is a supermartingale, so that the optimal stopping rule N for the sequence $\{Z_n, n \geqslant 3\}$ satisfies

$$P\{N \leqslant \max\{3, N(c)\}\} = 1.$$

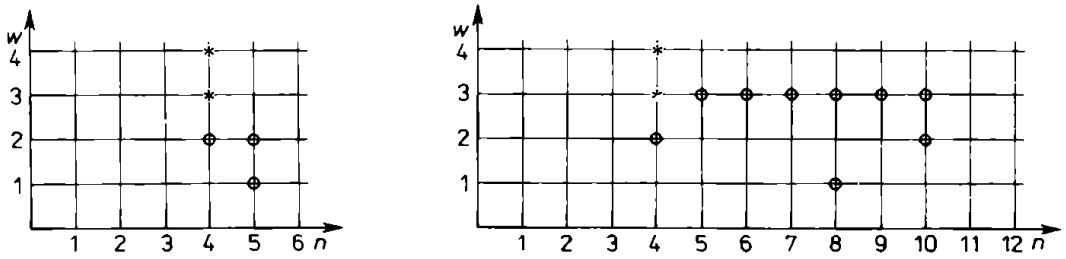
Given $N(c)$, the construction of the optimal stopping rule N may be performed by the backward induction technique (see, e.g, DeGroot (1970) or Ferguson (1968)).

The case $\{w_n = n-1 \text{ or } w_n = n\}$ needs some comments. If $w_n = n-1$ or $w_n = n$ for all $n = 3, 4, \ldots$, then $K = \infty$ and there is no reason to continue observations because of the loss increasing to infinity. If K is finite, then for some n we will pass from the situation $\{w_n = n-1 \text{ or } w_n = w\}$ to the situation $\{w_n = n-2\}$ and then we should proceed according to the general stopping rule constructed as above.

Observe that $g(n, n-2) = (n-2)\Gamma^2(n)/\Gamma(2n)$ decreases in n so that, given c, $g(n, n-2)-c < 0$ for all $n \geqslant N_2(c)$ with an appropriate positive integer $N_2(c)$. It follows that it is reasonable to stop the process Z_n with $w_n = n-1$ or $w_n = n$ not later than at the moment $N_2(c)$.

Numerical examples

Consider $P_n = (n, w_n)$, $n = 1, 2, \ldots$, as a random walk on the plane: the point p_3 emerges at (3,1), (3,2) or (3,3) "in a mysterious way" and afterwards passes "east" or "north-east", i.e., if $p_n = (n, w)$, then $p_{n+1} = (n+1, w)$ or $p_{n+1} = (n+1, w+1)$. The stopping rule may simply be presented by a list of the absorbing points of the random walk. For $c = 0.03$ and $c = 0.02$ the solutions are as follows (encircled are the stopping points with $w \leqslant n-2$ and stars represent p_n such that $n = N_2(c)$):



The above stopping rules have been computed by the backward induction assuming that, if p_n reaches one of the points $(N(c), w)$, $1 \leqslant w \leqslant N(c)$, the process will be stopped immediately.

References

[1] F. Archetti and C. Cugiani (ed.) (1980), *Numerical Techniques for Stochastic Systems*, North-Holland, Amsterdam.

[2] C. G. E. Boender and A. H. G. Rinnooy Kan (1983), *A Bayesian analysis of the number of cells of a multinomial distribution*, The Statistician 32, 240–248.

[3] M. H. DeGroot (1970), *Optimal Statistical Decisions*, McGraw-Hill, New York.

[4] L. C. W. Dixon and G. P. Szegö (ed.) (1975), *Towards Global Optimisation*, North-Holland, Amsterdam.

[5] –, – (1978), *Towards Global Optimisation 2*, North-Holland, Amsterdam.

[6] T. S. Ferguson (1967), *Mathematical Statistics: A Decision – Theoretic Approach*, Academic Press, New York.

[7] L. A. Goodman (1949), *On the estimation of the number of classes in a population*, Ann. Math. Statist. **20**, 572–579.

[8] R. Zieliński (1981), *A statistical estimate of the structure of multiextremal problems*, Math. Programming **21**, 348–356.

Presented to the semester
Sequential Methods in Statistics
September 7–December 11, 1981