The Annals of Statistics
1975, Vol. 3, No. 2, 512-517

APPROXIMATE BAYES SOLUTIONS TO SOME NONPARAMETRIC PROBLEMS

BY M. GOLDSTEIN

University of Oxford

The problem of making inferences about real functions of a probability distribution of unknown form is examined in a Bayesian nonparametric framework. With respect to a general quadratic loss function, Bayes estimates within the class of linear combinations of a given set of functions on the sample space are obtained for general functions on the distribution space. The result is then used to derive Bayes polynomial estimates of the moments of the distribution.

1. Introduction. Ferguson (1973) has written that "The Bayesian approach to statistical problems, though fruitful in many ways, has been rather unsuccessful in treating nonparametric problems. This is due primarily to the difficulty in finding workable prior distributions on the parameter space which in nonparametric problems is taken to be a set of probability distributions on a given sample space." As a result, the Bayesian approach to nonparametric problems seems, in practice, to proceed by approximating the given problem by a parametric problem which, hopefully, has many of the same features as the original problem.

This paper presents an alternative approach, approximating the solution of the original nonparametric problem, assuming a general prior distribution over the distribution space of the sample space. With respect to a general quadratic loss function, we look for the Bayes rule, for a general real-valued function on the distribution space, when the decision space is restricted to the set of linear combinations of some given set of real functions. The result is then used to derive Bayes polynomial estimates, of any given order, for the ith moment of the unknown distribution about the origin.

2. General formulation. A sample s is drawn from a probability distribution F, which is defined over a sample space S. The form of F is unknown, and we wish to make inferences about a real-valued quantity $g(F)$, where g is a measurable function from $\mathscr{F}$ (the space of all probability distributions defined over S) to the real line. The loss function for the problem is taken to be of the form

$$L(F, d) = w(F)(g(F) - d)^2 , \tag{2.1}$$

where $w(F) \geqq 0$ for all $F \in \mathscr{F}$.

We assign a prior measure $P(\cdot)$ over a σ-algebra of subsets of $\mathscr{F}$, with respect

Received September 1973, revised April 1974.

AMS 1970 *subject classifications.* Primary 62C10; Secondary 62G05.

Key words and phrases. Bayes nonparametric estimation, linear approximation, polynomial estimators for population moments.

to which $g(F)$, $w(F)$ are measurable, and, using the Bayes theorem, we evaluate $P_s(\cdot)$, the posterior measure over $\mathscr{F}$ given s. The Bayes rule $d(s)$ for $g(F)$, with respect to the loss function (2.1), is given by $\bar{g}(s)$, where

$$\bar{g}(s) = \int w(F)g(F)\, dP_s(F) / \int w(F)\, dP_s(F)\,, \tag{2.2}$$

(provided the posterior risk of $\bar{g}$ is finite for each s; Girshick and Savage (1951)).

Typically, $\bar{g}(s)$ will be difficult to evaluate explicitly, and we therefore seek to approximate it by a linear combination of m real-valued functions, $h_1(\cdot), \cdots, h_m(\cdot)$, defined on S; i.e. we seek the Bayes decision for $g(F)$ within the class of decision rules of the form $\sum_{i=1}^m \lambda_i h_i(s)$, where the λ_i are real.

We wish to find $\lambda = (\lambda_1, \cdots, \lambda_m)'$, to minimize

$$\int\int (g(F) - \textstyle\sum_{i=1}^m \lambda_i h_i(s))^2 w(F)\, dF(s)\, dP(F)\,. \tag{2.3}$$

From the general theory of Least Squares (see, for example, Rao (1973)), the value of λ minimizing (2.3), and the value, R, of (2.3) for this value of λ, are obtained by calculating $\mathbf{D}$, the dispersion matrix of $h_1 w^{\frac{1}{2}}, \cdots, h_m w^{\frac{1}{2}}$, and $\mathbf{b}$, the covariance vector of $gw^{\frac{1}{2}}$ with $h_1 w^{\frac{1}{2}}, \cdots, h_m w^{\frac{1}{2}}$.

We note that, defining

$$\begin{aligned} e &= \int g^2(F)w(F)\, dP(F)\,, \\ \bar{h}_i(F) &= \int h_i(s)\, dF(s)\,, \\ \bar{h}_{ij}(F) &= \int h_i(s)h_j(s)\, dF(s)\,, \end{aligned} \tag{2.4}$$

we have provided $\bar{h}_i(\cdot)$, $\bar{h}_{ij}(\cdot)$ are measurable functions of F,

$$\int\int g(F)h_i(s)w(F)\, dF(s)\, dP(F) = \int g(F)\bar{h}_i(F)w(F)\, dP(F) = b_i\,, \tag{2.5}$$

$$\int\int h_i(s)h_j(s)w(F)\, dF(s)\, dP(F) = \int \bar{h}_{ij}(F)w(F)\, dP(F) = d_{ij}\,. \tag{2.6}$$

Evaluation of (2.5) and (2.6) requires only a partial specification of P and will typically be more straightforward than evaluating (2.2). We have

THEOREM 2.1. *If $h_1(\cdot), \cdots, h_m(\cdot)$ are m real-valued functions on S, for which the values b_i, d_{ij} (defined in* (2.5), (2.6)) *exist for each i, j, then with respect to loss function* (2.1), *we obtain the following*:

If the matrix $\mathbf{D} = (d_{ij})$, $i, j, = 1, \cdots, m$, is invertible, and e (defined in (2.4)) *exists, then*:

(i) *the Bayes estimate for $g(F)$, in the class of linear combinations of the functions $h_1, \cdots, h_m$, is given by*

$$\lambda_1 h_1 + \cdots + \lambda_m h_m\,, \tag{2.7}$$

where, if $\lambda = (\lambda_1, \cdots, \lambda_m)'$, $\mathbf{b} = (b_1, \cdots, b_m)'$,

$$\lambda = \mathbf{D}^{-1}\mathbf{b}\,; \tag{2.8}$$

(ii) *the Bayes risk of estimate* (2.7) *is given by R, defined by*

$$R = \left|\begin{matrix} \mathbf{D} & \mathbf{b} \\ \mathbf{b}' & e \end{matrix}\right| / |\mathbf{D}|\,. \tag{2.9}$$

3. Bayes polynomial estimates for moment problems. An observation x comes from a probability distribution $F(x)$ over the real line. The form of $F(\cdot)$ is unknown, and we wish to estimate the function $\mu_i(F)$ defined by

$$\mu_i(F) = \int x^i \, dF(x) , \tag{3.1}$$

where i is some given positive integer. (The case of greatest practical interest is $i = 1$.) We assign a prior probability measure $P(\cdot)$ over $\mathscr{F}$, the space of all probability distributions over the real line, and we work with the quadratic loss function $L_i(F, d)$ defined by

$$L_i(F, d) = (\mu_i(F) - d)^2 . \tag{3.2}$$

With respect to this loss function, we seek the Bayes rule for $\mu_i(F)$ given x within the class of real-valued polynomials in x of order n. We denote this class by A_n, and we define $t_i(x) = x^i$, $i = 0, 1, \cdots, n$. Further, we define

$$\bar{\mu}_r = \int \mu_r(F) \, dP(F) , \tag{3.3}$$

$$\bar{\mu}_{r,s} = \int \mu_r(F)\mu_s(F) \, dP(F) , \tag{3.4}$$

$$\bar{t}_i(F) = \int t_i(x) \, dF(x) , \tag{3.5}$$

$$\bar{t}_{ij}(F) = \int t_i(x)t_j(x) \, dF(x) . \tag{3.6}$$

It follows that

$$\bar{t}_i(F) = \mu_i(F) , \qquad \bar{t}_{ij}(F) = \mu_{i+j}(F) , \tag{3.7}$$

so we have

$$\int \bar{t}_{ij}(F) \, dP(F) = \bar{\mu}_{i+j} , \qquad \int \mu_i(F)\bar{t}_j(F) \, dP(F) = \bar{\mu}_{i,j} . \tag{3.8}$$

But if $G(x)$ is our prior distribution for x, i.e.

$$G(x) = \int F(x) \, dP(F) , \tag{3.9}$$

then, assuming that the required expectations exist, we may apply Fubini's theorem to obtain

$$\begin{aligned} \int x^j \, dG(x) &= \int [\int x^j \, dF(x)] \, dP(F) \\ &= \bar{\mu}_j , \qquad j = 0, 1, 2, \cdots . \end{aligned} \tag{3.10}$$

The set $\{\bar{\mu}_0, \bar{\mu}_1, \bar{\mu}_2, \cdots\}$ is therefore the moment set about the origin of the distribution $G(\cdot)$. Thus, denoting by $\mathbf{K}_n$ the $(n + 1) \times (n + 1)$ matrix with entries K_{ij} defined by

$$K_{ij} = \bar{\mu}_{i+j} ; \qquad i, j = 0, 1, \cdots, n , \tag{3.11}$$

we see that $|\mathbf{K}_n|$ is the nth Hankel determinant of the distribution G, and hence, providing G has more than n points of increase, $|\mathbf{K}_n| > 0$, i.e. $\mathbf{K}_n$ is invertible, (see, for example, von Mises (1964)).

It follows that we can apply Theorem 2.1 directly. Here the sample space S is the real line, the function to be approximated is $\mu_i(F)$, $w(F) = 1$ for all F,

and the approximating functions are $t_0, t_1, \cdots, t_n$. Comparing (2.4), (2.6) with (3.6), (3.8) it is immediate that the matrix $\mathbf{D}$ defined in Theorem 2.1 is, in this instance, $\mathbf{K}_n$ (as defined in (3.11)). Provided G has more than n points of increase, $\mathbf{K}_n$ is invertible and the requisite condition of Theorem 2.1 is satisfied. Further, comparing (2.4), (2.5) with (3.5), (3.8), the vector $\mathbf{b}$ defined in Theorem 2.1 is, in this instance, the vector

$$\mathbf{b}_n{}^i = (\bar{\mu}_{i,0}, \bar{\mu}_{i,1}, \cdots, \bar{\mu}_{i,n}) . \tag{3.12}$$

Substituting the relevant quantities into Theorem 2.1, we thus obtain

THEOREM 3.1. (i) *Given the observation x, the Bayes estimate for $\mu_i(F)$, with respect to loss function* (3.2), *in the class A_n is*

$$\mathbf{n}(x)' \cdot \mathbf{a}_n{}^i \tag{3.13}$$

where $\mathbf{n}(x) = (1, x, x^2, \cdots, x^n)'$ *and*

$$\mathbf{a}_n{}^i = \mathbf{K}_n{}^{-1}\mathbf{b}_n{}^i . \tag{3.14}$$

This holds for any integer n such that the prior distribution for x has more than n points of increase, provided that the expectations detailed in (3.10) exist.

(ii) The Bayes risk incurred by using the estimate $\mathbf{n}(x)' \cdot \mathbf{a}_n{}^i$, is given by

$$|{}^{\mathbf{K}_n}_{\mathbf{b}_n{}^{i'}} \; {}^{\mathbf{b}_n{}^i}_{\bar{\mu}_{i,i}}|/|\mathbf{K}_n| . \tag{3.15}$$

As a simple application of the above theorem, suppose we are in a parametric framework where we know that the form of F is $N(\mu, \sigma^2)$, with σ^2 known. If the prior distribution for μ is $N(0, \tau^2)$ then, with respect to quadratic loss, the Bayes estimate for μ, in the class A_m for each $m \geqq 1$, is given by

$$\tau^2 x/(\tau^2 + \sigma^2) . \tag{3.16}$$

For a more complicated example, suppose that, as before, our prior distribution for μ is $N(0, \tau^2)$, but that the form of F is known to be $N(\mu, \sigma^2 + k\mu^2)$, $k > 0$. In the special case where $\sigma^2 = \tau^2$, the Bayes estimate for μ, in the class A_4, is given by

$$x/(2 + k) . \tag{3.17}$$

More generally, the Bayes estimate in A_4 is

$$x(c_1 + x^2 c_3)/c_2 , \tag{3.18}$$

where, defining

$$b_1 = (1 + k) , \qquad b_2 = (3k^2 + 6k + 1) , \qquad b_3 = (15k^3 + 45k^2 + 15k + 1) ,$$
$$a_1 = \sigma^2 + b_1\tau^2 , \qquad a_2 = 3(b_2\tau^4 + 2b_1\sigma^2\tau^2 + \sigma^4) ,$$
$$a_3 = 15(\tau^6 b_3 + 3\tau^4\sigma^2 b_2 + 3\tau^2\sigma^4 b_1 + \sigma^6) , \qquad a_4 = \tau^2 ,$$
$$a_5 = 3(\sigma^2\tau^2 + \tau^4(b_1 + 2k)) ,$$

then

$$c_2 = a_1a_3 - a_2^2\,, \qquad c_1 = a_3a_4 - a_2a_5\,, \qquad c_3 = a_1a_5 - a_2a_4\,; \tag{3.19}$$

so, for example,

$$c_3 = 6k\tau^4(\sigma^2 - \tau^2)\,. \tag{3.20}$$

For the majority of problems, of course, a complete specification of the prior measure over $\mathscr{F}$ will be impractical and the prior values $\bar{\mu}_k$, $\bar{\mu}_{k,i}$ will be estimated directly. The extension to the case where an independent sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is drawn from F is immediate. We define

$$\bar{\mu}_{j(1),\cdots,j(n)} = \int \prod_{i=1}^n \mu_{j(i)}(F)\, dP(F)\,. \tag{3.21}$$

We wish to approximate the Bayes rule for $g(F)$ by polynomials of order m in $\mathbf{x}$; i.e. by the set of functions

$$t_{j(i),\cdots,j(n)}(\mathbf{x}) = \prod_{i=1}^n x_i^{j(i)}\,, \qquad \sum_{i=1}^n j(i) \leqq m\,, \tag{3.22}$$

($j(i)$ is a nonnegative integer for each i).

Instead of (3.3) to (3.8), we have

$$\begin{aligned} &\int \bar{t}_{j(1),\cdots,j(n)}(F)\, dP(F) = \bar{\mu}_{j(1),\cdots,j(n)}\,, \\ &\int \bar{t}_{(j(1),\cdots,j(n))(k(1),\cdots,k(n))}(F)\, dP(F) = \bar{\mu}_{j(1)+k(1),\cdots,j(n)+k(n)}\,, \\ &\int \bar{t}_{j(1),\cdots,j(n)}(F)\mu_i(F)\, dP(F) = \bar{\mu}_{i,j(1),\cdots,j(n)}\,, \end{aligned} \tag{3.23}$$

and (3.10) becomes

$$\int \prod_{i=1}^n x_i^{j(i)}\, dG(x) = \bar{\mu}_{j(1),\cdots,j(n)}\,. \tag{3.24}$$

So, in the statement of Theorem 3.1, we therefore replace $\mathbf{b}_m{}^i$ by the vector $\mathbf{b}^i_{m,n}$ whose $(k(1), \cdots, k(n))$ entry is $\bar{\mu}_{i,k(1),\cdots,k(n)}$, $\mathbf{D}_m$ by the matrix $\mathbf{D}_{m,n}$ whose $(k(1), \cdots, k(n))$ row, $(j(1), \cdots, j(n)$ column entry is $\bar{\mu}_{j(1)+k(1),\cdots,j(n)+k(n)}$, and $\mathbf{m}(x)$ is replaced by $\mathbf{m}(\mathbf{x})$, whose $(j(1), \cdots, j(n))$, entry is $\prod_{i=1}^n x_i^{j(i)}$ where we have assigned some fixed ordering to the set of vectors $(a(1), \cdots, a(n))$, $(a(1), \cdots, a(n)$ are nonnegative integers), $\sum_{t=1}^n a(t) \leqq m$. For example, if $n = 2$, $m = 2$, and our ordering is (0, 0), (0, 1), (1, 0), (1, 1), (0,2), (2,0), then

$$\mathbf{b}^i_{m,n} = (\bar{\mu}_i, \bar{\mu}_{i,1}, \bar{\mu}_{i,1}, \bar{\mu}_{i,1,1}, \bar{\mu}_{i,2}, \bar{\mu}_{i,2})' \qquad \text{and}$$

$$\mathbf{m}(\mathbf{x}) = (1, x_2, x_1, x_1x_2, x_2^2, x_1^2)'\,.$$

Acknowledgments. I wish to thank my supervisor, Dr. A. F. M. Smith, for many helpful discussions, and the Science Research Council for their financial support. A referee's comments on an earlier version of the paper were useful.

REFERENCES

[1] FERGUSON, THOMAS S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

[2] GIRSHICK, M. A. and SAVAGE, L. J. (1951). Bayes and minimax estimates for quadratic loss functions. *Proc. Second Berkeley Symp. Math. Statist. Prob.* 53–73.

[3] RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, (2nd ed.). Wiley, New York.

[4] VON MISES, R. (1964). *Mathematical Theory of Probability and Statistics*. Academic Press, New York.

MATHEMATICAL INSTITUTE
UNIVERSITY OF OXFORD
ST. GILES
OXFORD, ENGLAND