DECISION THEORY FOR SOME NONPARAMETRIC MODELS

KJELL A. DOKSUM
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction and summary

The problem considered in this paper is that of obtaining optimal decision rules when a parametric form of the distribution of the observations is not known exactly. Thus we assume that the underlying distribution function F of the X_i in the random sample $X = (X_1, \cdots, X_n)$ is in a class Ω of distribution functions, and Ω is not indexed in a natural way by a parameter θ in m dimensional Euclidean space R^m. Let $R(F, d)$ denote the risk of the decision rule $d = d(X)$ when F is the true distribution. Minimax procedures that minimize the maximum risk sup $\{R(F, d); F \in \Omega\}$ have been obtained in special cases by Hoeffding [8], Ruist [14], Huber [9], [10], [11], and Doksum [3]. In particular, Huber was able to show that if Ω is the class of all distributions in a neighborhood of a normal distribution, then the minimax procedures are based on statistics that are, approximately, trimmed means. Most stringent procedures that minimize the maximum shortcoming $\sup_F \{R(F, d) - \inf_d R(F, d)\}$ have been considered by Schaafsma [15].

Another approach would be to define a probability (weight function) P on Ω and then minimize the average (Bayes) risk $\int_\Omega R(F, d)\, P(dF)$, thereby obtaining what is called the Bayes solution. This approach has been taken by Kraft and van Eeden [13], Ferguson [6], and Antoniak [1], who were able to obtain explicit Bayes solutions for some probabilities P. Their work is closely related to the work of Fabius [5], who considered properties of posterior distributions for a class of probability measures P that essentially contains those of Kraft and van Eeden and of Ferguson. Fabius' work in turn is related to that of Freedman [7], who considered properties of Bayes procedures in the case where the X_i are discrete random variables. The relationship between these papers will be discussed further in Section 5.

In this paper, we introduce a criterion which involves minimizing a quantity between the maximum risk and the average risk. This criterion is appropriate when the probability P on Ω is not fully specified, but only the distribution of $F(t_1), \cdots, F(t_k)$ is known for some $t_1 < \cdots < t_k$. Thus past records may

Prepared with the partial support of the National Science Foundation under Grant GP-15283, and with the partial support of the Office of Naval Research under Contract N00014--69-A-0200-1036. Completed while the author was at the University of Oslo.

be available for $F(1)$, $F(2)$, and so on, but not for $F(e)$, $F(\pi)$ and so on. The criterion is to minimize the average maximum risk, where the average is computed with respect to the distribution λ of $F(t_1), \cdots, F(t_k)$. More precisely, let $t_1 < \cdots < t_k$ be fixed and let $\Omega(q, k)$ be the class of distribution functions in Ω that pass through $(t_1, q_1), (t_2, q_2), \cdots, (t_k, q_k)$, $0 \leqq q_1 \leqq \cdots \leqq q_k \leqq 1$. We define the average maximum risk (or mixed risk) as $\int_{R^k} [\sup_{F \in \Omega(q,k)} R(F,d)] \lambda(dq)$. The decision rule that minimizes this risk is called the mixed Bayes-minimax rule, or mixed rule. It will be shown in Sections 2 and 3 that, under certain conditions on the risk function, the mixed rule can be obtained by computing the posterior distribution of a multinomial parameter $p = (p_1, \cdots, p_{k-1})$ having as prior distribution the distribution μ of $F(t_2) - F(t_1), F(t_3) - F(t_2), \cdots, F(t_k) - F(t_{k-1})$.

The mixed Bayes-minimax rule d_k can be thought of as an approximation to the Bayes rule. In Section 4, Prohorov's theorem is used to show that if Ω is contained in $C[0, 1]$ or $D[0, 1]$, and if a Bayes solution d exists, then d_k converges to d in the sense of convergence of Bayes risks. Thus in situations where P is known, but the Bayes rule is hard to compute, one can use the mixed rule as an approximation. If the limit $\lim_{k \to \infty} d_k$ can be computed, then it gives a method of obtaining the Bayes solution. Note that $k = 0$ corresponds to the minimax problem.

2. The mixed Bayes-minimax problem

Let $X_1, \cdots, X_n$ be independent, identically distributed random variables with distribution function F, where F belongs to some specified class Ω of distribution functions. It will be convenient to assume that Ω is a measurable subset of some larger class Γ of functions with a σ-field $\mathscr{S}$. We further assume that there is a probability P on $(\Gamma, \mathscr{S})$, with $P(\Omega) = 1$.

Let $L(F, d)$ be a real valued function that denotes the loss of the real valued decision rule $d = d(X_1, \cdots, X_n)$ when $F \in \Omega$ is the true distribution. Let $X = (X_1, \cdots, X_n)$. Then the risk of d is

$$R(F, d) = E_F L(F, d(X)). \tag{2.1}$$

If there is a rule (procedure) that minimizes the maximum risk

$$R(d) = \sup_{F \in \Omega} R(F, d), \tag{2.2}$$

then it is called a minimax procedure. Similarly, if there is a d that minimizes the average risk

$$r(P, d) = \int_\Omega R(F, d)\, P(dF), \tag{2.3}$$

then it is a Bayes procedure. Here F is thought of as a random distribution function with distribution P; that is, F is a stochastic process with sample paths $F(t)$, $t \in R$. Computing Bayes procedures will involve computing the posterior distribution of F given X.

The mixed Bayes-minimax procedure minimizes a function that is between R(d) and $r(P, d)$. This function is the average maximum risk when a finite dimensional distribution corresponding to P is known, and the average is taken with respect to this distribution. We now proceed with the definition. The *carrier* of a given distribution is in general the smallest compact set whose probability under the given distribution is one. Let $C(F)$ denote the carrier of $F \in \Omega$; we define the *support* of Ω to be $S(\Omega) = \cup_{F \in \Omega} C(F)$. Let $t_1 < \cdots < t_k$ be k points in $S(\Omega)$. The distribution of $F(t_1), \cdots, F(t_k)$ under P will be denoted by λ, or $\lambda(\cdot\,; P, k)$. Thus, if we write $q = (q_1, \cdots, q_k)$, $0 \leqq q_1 \leqq \cdots \leqq q_k \leqq 1$, then

$$\lambda(q; P, k) = P(F\colon F(t_1) \leqq q_1, \cdots, F(t_k) \leqq q_k). \tag{2.4}$$

The class of distributions in Ω whose value at t_i is q_i, $i = 1, \cdots, k$, will be denoted by $\Omega(q, k)$, that is,

$$\Omega(q, k) = \{F \in \Omega\colon F(t_i) = q_i, i = 1, \cdots, k\}. \tag{2.5}$$

The *average maximum risk* of a decision rule d is now defined as

$$r_k(P, d) = \int_{R^k} [\sup_{F \in \Omega(q, k)} R(F, d)] \lambda(dq; P, k). \tag{2.6}$$

If there is a d_0 that minimizes $r_k(P, d)$, d_0 is called a mixed Bayes-minimax (or mixed) procedure.

As we are dealing with functions of $X \in R^n$ and $F \in \Omega$, we need a joint distribution for X and a random element F of Ω with distribution P. If $\mathscr{B}^n$ denotes the σ-field of Borel sets in R^n, we define the probability $\tilde{P}$ on $(R^n \times \Gamma, \mathscr{B}^n \times \mathscr{S})$ by

$$\tilde{P}(B \times S) = \int_S \mu_F(B)\, P(dF), \qquad B \in \mathscr{B}^n, \qquad S \in \mathscr{S}, \tag{2.7}$$

where μ_F is the probability in R^n corresponding to the distribution of X, and it is assumed that μ_F is $\mathscr{S}$ measurable. We also assume that a conditional distribution of X given F exists and satisfies

$$\tilde{P}(X_1 \leqq x_1, \cdots, X_n \leqq x_n | F) = \textstyle\prod_1^n F(x_i). \tag{2.8}$$

Furthermore, we assume that F has a conditional distribution given $F(t_i) = q_i$, $i = 1, \cdots, k$; we denote this conditional distribution by P^q. For a further discussion of these definitions, see Fabius [5]. The assumptions involved in the definitions are satisfied for the complete, separable metric spaces considered in Section 4.

The following inequalities follow at once from the definitions. Note that when more than one set of $t_1, \cdots, t_k$ is considered, we will use double subscripts and write $t_{m,1}, \cdots, t_{m,k_m}$.

LEMMA 2.1. *The risks defined above satisfy the following relations:*

(i) $R(d) = r_0(P, d) \geqq r_k(P, d) \geqq r(P, d)$, $\quad k \geqq 1$;

(ii) *if* $\{\prod_m\colon t_{m,1} < \cdots < t_{m,k_m}\}$, $m = 1, 2, \cdots$, *is a sequence of partitions such that each partition is a refinement of the previous one, then*

$$r_{k_m}(P, d) \geqq r_{k_\ell}(P, d) \qquad \textit{for } m < \ell. \tag{2.9}$$

We next give a parametric example in which the mixed risk r_k equals the Bayes risk r.

EXAMPLE 2.1. Let $\Gamma = \Omega$ be the class of normal distribution functions F_θ with mean θ and variance unity. Suppose that P is the measure for which θ has a normal distribution with mean ξ and variance unity. Let $\mathscr{S}$ be the class of sets of the form $\{F_\theta : \theta \in B\}$, where B is a Borel subset of the reals. All the quantities of this section are now defined. Moreover, since $F(t_1)$ determines θ and θ determines $F(t_1)$, then $r_k(P, d) = r(P, d)$, $k \geqq 1$.

Next we consider a "discrete" example in which the mixed risk eventually equals the Bayes risk.

EXAMPLE 2.2. Let $\Omega = \Gamma$ be a countable class $\{F_1, F_2, \cdots\}$ and let $\mathscr{S}$ be the collection of all subsets of Ω. If Ω is discrete (that is, Ω has no limit points for the sup norm), and if $\{t_{m,j}\}$ of Lemma 2.1 (ii) becomes dense in $S(\Omega)$ as $m \to \infty$, then there exists $m_1 \geqq 1$ such that $r_{k_m}(P, d) = r(P, d)$ for all $m \geqq m_1$. To see this, note that, by our assumptions, there exists m_1 such that

$$(2.10) \qquad \big(F_i(t_{m,1}), \cdots, F_i(t_{m,k_m})\big) \neq \big(F_j(t_{m,1}), \cdots, F_j(t_{m,k_m})\big)$$

for $m \geqq m_1$ and all $i \neq j$.

We now define a decision rule d_k that, in some cases, will be the mixed Bayes-minimax procedure. Let $t_1 = \inf\{t : t \in S(\Omega)\}$ and $t_k = \sup\{t : t \in S(\Omega)\}$. It follows that $F(t_k) = 1$. We will assume that $F(t_1) = 0$ a.s. (P), $-\infty < t_1 < t_k < \infty$, and that $k \geqq 3$. Let $q = \{q_1, \cdots, q_k\}$ with $0 = q_1 \leqq \cdots \leqq q_k = 1$. Let $F_{q,k}$ be the polygonal distribution function that equals q_i at t_i, $i = 1, \cdots, k$, and is linear over each interval $[t_i, t_{i+1}]$, $i = 1, \cdots, k-1$. Let F_k denote the random distribution function obtained by letting q in $F_{q,k}$ have distribution $\lambda = \lambda(\cdot\,; P, k)$. We assume that F_k is a measurable function on some measure space to $(\Gamma, \mathscr{S})$. Let P_k denote the distribution of F_k. Finally, d_k will denote the Bayes solution for $F_{q,k}$ when q has prior λ, that is, d_k minimizes

$$(2.11) \qquad r(P_k, d) = \int_{R_k} R(F_{q,k}, d)\lambda(dq).$$

THEOREM 2.1. *If $F_{q,k} \in \Omega$ for almost all q in the carrier C_λ of λ, if d_k minimizing* (2.11) *exists, and if*

$$(2.12) \qquad L(F, d_k) = L(G, d_k)$$

for all $F, G \in \Omega(q, k)$ and almost all q in C_λ then d_k is the mixed Bayes-minimax procedure.

PROOF. Let N_i denote the number of X in $(t_i, t_{i+1}]$, $i = 1, \cdots, k-1$. Then $\mathbf{N} = (N_1, \cdots, N_{k-1})$ is sufficient for $F_{q,k}$, and d_k depends on the X only through $\mathbf{N}$. Thus the distribution of d_k as a function of the X is the same for all F in $\Omega(q, k)$. This and (2.12) imply that $R(F, d_k) = R(G, d_k)$ for all $F, G \in \Omega(q, k)$. Thus $r_k(P, d_k) = r(P_k, d_k)$, and since d_k is optimal for P_k, then $r(P_k, d_k) \leqq r(P_k, d)$ for all other rules d. Finally, since $F_{q,k} \in \Omega$ for almost all q, $r(P_k, d) \leqq r_k(P, d)$ and the results follow.

REMARK 2.1. From the above proof, it is clear that (2.12) can be replaced by the condition

$$r_k(P, d_k) = r(P_k, d_k) \tag{2.13}$$

and that (2.13) is weaker than (2.12).

EXAMPLE 2.3. If $0 \in \{t_1, \cdots, t_k\}$ and $L(F, d) = [d - F(0)]^2$, then (2.12) is satisfied. Generalizing this, we have that (2.12) is satisfied for any loss function depending on F only through $F(t_1), \cdots, F(t_k)$; that is, the loss is defined through those points where we have information about F. Such a loss function corresponding to squared error when estimating the mean of F would be $[d - \mu(F, k)]^2$, where

$$\mu(F, k) = \tfrac{1}{2} \sum_{i=1}^{k-1} (t_{i+1} + t_i)[F(t_{i+1}) - F(t_i)] \tag{2.14}$$

In the next section, we consider a testing problem for which (2.12) is satisfied.

3. A testing problem

Suppose $\Omega = \Omega_0 \cup \Omega_1$ with $\Omega_0 \cap \Omega_1$ empty, $\Omega_0, \Omega_1 \in \mathscr{S}$, and we want to test $H_0 : F \in \Omega_0$ against $\mathrm{H}_1 : F \in \Omega_1$. $\varphi = \varphi(X)$ will denote a test function, and $L(F, \varphi)$ will be the usual loss for the testing problem; that is, $L(F, \varphi) = L_i$, a positive constant, if H_i is falsely rejected, $i = 0, 1$, and $L(F, \varphi) = 0$ otherwise. The Bayes risk is

$$r(P, \varphi) = L_0 \int_{\Omega_0} E_F(\varphi)\, P(dF) + L_1 \int_{\Omega_1} [1 - E_F(\varphi)]\, P(dF) \tag{3.1}$$

and the average maximum risk is

$$\begin{aligned} r_k(P, \varphi) = L_0 \int_{Q_0} [\sup_{F \in \Omega(q,k)} E_F(\varphi)] \lambda(dq : P, k) \\ + L_1 \int_{Q_1} [\sup_{F \in \Omega(q,k)} [1 - E_F(\varphi)] \lambda(dq ; P, k), \end{aligned} \tag{3.2}$$

where $Q_i = \{(F(t_1), \cdots, F(t_k)) : F \in \Omega_i\}$, $i = 1, 2$; that is, Q_0 and Q_1 are the sets in R^k corresponding to Ω_0 and Ω_1. We assume that $Q_0 \cap Q_1$ has λ probability zero. This assumption is needed to obtain (3.2) above. For this loss function it is clear that (2.12) of Theorem 2.1 is satisfied; if in addition $F_{q,k} \in \Omega$, then the result can be applied. If $F_{q,k} \notin \Omega$, then $L(F_{q,k}, \varphi)$ is not defined. However, there is a natural way of defining $L(F_{q,k}, \varphi)$ and making $F_{q,k}$ a member of $\Omega(q, k)$: for a given q not in $Q_0 \cap Q_1$, $L(F, \varphi)$ has the same value for each $F \in \Omega(q, k)$; define $L(F_{q,k}, \varphi)$ to be this value.

In what follows, it will be assumed that either $F_{q,k} \in \Omega$ or $L(F_{q,k}, \varphi)$ has been defined as above. Let $f(x|q)$ denote the density of $F_{q,k}$. Then from (3.2) and Theorem 2.1 we can conclude that the mixed Bayes-minimax procedure φ_k rejects H_0 when

$$L_1 \int_{Q_1} \prod_{i=1}^{k-1} (q_{i+1} - q_i)^{N_i} \lambda(dq) \geqq L_0 \int_{Q_0} \prod_{i=1}^{k-1} (q_{i+1} - q_i)^{N_i} \lambda(dq),\ i = 1, 2. \tag{3.3}$$

Let $p_i = q_{i+1} - q_i$, $i = 1, \cdots, k-1$ and

$$A_i = \{(F(t_2) - F(t_1)), \cdots, (F(t_k) - F(t_{k-1})) : F \in \Omega_i\}. \tag{3.4}$$

Then (3.3) becomes

$$L_1 \int_{A_1} \prod_{i=1}^{k-1} p_i^{N_i} \pi(dp) \geqq L_0 \int_{A_0} \prod_{i=1}^{k-1} p_i^{N_i} \pi(dp), \tag{3.5}$$

where π is the distribution of $(F(t_2) - F(t_1), \cdots, F(t_k) - F(t_{k-1}))$ when F has distribution P. Note that (3.5) is the solution to a Bayesian multinomial testing problem, in which $(N_1, \cdots, N_{k-1} | P)$ is a multinomial variable with parameters n and p, and we are testing $H_0 : p \in A_0$ versus $H_1 : p \in A_1$. The solution is the test φ_k that rejects H_0 when the ratio of the posterior probability of H_1 to the posterior probability of H_0 exceeds the ratio of the losses L_0/L_1; that is, if p has the posterior density $g(p|\mathbf{N})$, $\mathbf{N} = (N_1, \cdots, N_{k-1})$, then the test φ_k rejects H_0 when

$$\frac{\int_{A_1} g(p|\mathbf{N})\, dp}{\int_{A_0} g(p|\mathbf{N})\, dp} \geqq \frac{L_0}{L_1}. \tag{3.6}$$

EXAMPLE 3.1. Consider the goodness of fit problem where Ω_0 contains only the uniform distribution F_0 on (0, 1) and

$$\Omega_1 = \{F : C_F \subset [0, 1], F \leqq F_0\} - \Omega_0. \tag{3.7}$$

If P assigns probability $\frac{1}{2}$ to Ω_0, and if $t_1 = 0$, $t_2 = \frac{1}{2}$, $t_3 = 1$, then from (3.5)

$$\varphi_k = \begin{cases} 1 & \text{if } 2^{-n} \int_0^{1/2-0} p_1^{N_1}(1-p_1)^{n-N_1} \pi_1(dp_1) \geqq \dfrac{L_1}{L_0} \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

where $\pi_1(p_1) = P(F(\frac{1}{2}) \leqq p_1)$, so that $\pi_1(\frac{1}{2} - 0) = \frac{1}{2}$ and $\pi_1(\frac{1}{2}) = 1$. Thus φ_k is based on a decreasing function of N_1 = number of X less than or equal to $\frac{1}{2}$. When $2\pi_1$ is the beta distribution, then φ_k can be obtained from the tables of the incomplete beta function.

REMARK 3.1. The loss function of this section is not the only one for which optimal mixed tests can be obtained. For instance, if Ω_0 and Ω_1 are as in Example 3.1, then the loss for deciding H_0 when $F \in \Omega_1$ could be defined to be $L_1(\frac{1}{2} - F(\frac{1}{2}))$. The conditions of Theorem 2.1 would then still be satisfied and the optimal test can be obtained from the corresponding multinomial problem.

4. Convergence of the mixed solutions to the Bayes solution

In this section, we will consider classes of distribution functions on [0, 1], that is, Ω is such that its support $S(\Omega)$ is contained in [0, 1]. Then Ω is a subset of the class $D[0, 1]$ of right continuous functions on [0, 1] with limits from the left. On $D[0, 1]$ we use the Skorohod topology with the modified Skorohod metric (see for example [2], p. 113) so that $D[0, 1]$ is a complete separable metric space. In the notation of Section 2, $\Gamma = D[0, 1]$ and $\mathscr{S}$ is the σ-field generated by the open sets, F is a random function (a measurable function on some measure space to $(\Gamma, \mathscr{S})$) with distribution P. We assume that the probability P on $(\Gamma, \mathscr{S})$ satisfies $P(F \in \Omega) = 1$. If $\{F_k\}$ is a sequence of random functions with distributions $\{P_k\}$, then F_k is said to *converge in distribution* to F if

$$\int_\Gamma h(F) P_k(dF) \to \int_\Gamma h(F) P(dF) \tag{4.1}$$

for each continuous, bounded, real valued function h on Γ. If (4.1) holds, then P_k is also said to *converge weakly* to P.

If F_k is the polygonal random distribution function of Section 2, then F_k does not necessarily converge in distribution to F. However, we will show that the Bayes risk of the Bayes solution for the prior P_k converges to the Bayes risk of the Bayes solution for the prior P. To do this, we will make use of the random distribution function G_k that is constant over each interval $[t_i, t_{i+1})$, and for which the joint distribution of $G_k(t_1), \cdots, G_k(t_k)$ equals that of $F(t_1), \cdots, F(t_k)$. The symbol Q_k will denote the distribution of G_k. We will need double subscripts on the t and again write $t_{m,1}, \cdots, t_{m,k_m}$ instead of $t_1, \cdots, t_k$. We assume that $\{\Pi_m\colon t_{m,1} < \cdots < t_{m,k_m}\}$, $m = 1, \cdots, k_m$ is a sequence of partitions of $[0, 1]$ such that Π_{m+1} is a refinement of Π_m and $\max_i |t_{m,(i+1)} - t_{m,i}| \to 0$ as $m \to \infty$.

It is now easy to show using Prohorov's theorem that:

LEMMA 4.1. *If $\Omega \subset D[0, 1]$, then G_k converges in distribution to F.*

PROOF. For each $F_0 \in \Omega$ and $\delta \in (0, 1)$ define

$$v(F_0, \delta) = \sup \min [F_0(t) - F_0(s), F_0(u) - F_0(t)], \tag{4.2}$$

where the sup is over $s \leqq t \leqq u$, $u - s = \delta$. Similarly, define

$$w_0(F_0, \delta) = \sup [F_0(u) - F_0(s)], \tag{4.3}$$

where the sup is over $1 - \delta \leqq s < u < 1$. By Prohorov's theorem applied to $D[0, 1]$ it is enough to show that (i) the finite dimensional distributions of Q_k converge weakly to the finite dimensional distributions of P at points $s_1, \cdots, s_r$ in the set $\{t\colon P[F(t) \neq F(t^-)] = 0\} \cup \{0, 1\}$, and that (ii) for each $\varepsilon, \eta > 0$, there exists $\delta \in (0, 1)$ and an integer k_0 such that

$$Q_k(v(G_k, \delta) < \varepsilon) > 1 - \eta \qquad \text{for} \quad k \geqq k_0, \tag{4.4}$$

and

$$Q_k(w_0(G_k, \delta) < \varepsilon) > 1 - \eta \qquad \text{for} \quad k \geqq k_0. \tag{4.5}$$

(See for example [2], pp. 125–126.)

The convergence of the indicated finite dimensional distributions follows from the definition of G_k. The inequalities (4.4) and (4.5) are easy to establish using the tightness of P. We omit the first subscript on the t to simplify the notation. Let k_0 be such that $t_{(i+1)} - t_i < \delta$, $i = 1, \cdots, k$, for all $k \geqq k_0$. Then since G_k is constant between points t_i,

$$v(G_k, \delta) \leqq \max \min \left[F(t_i) - F(t_j), F(t_j) - F(t_\ell)\right], \tag{4.6}$$

where the max is over $t_\ell \leqq t_j \leqq t_i$, $t_i - t_\ell \leqq 2\delta$. The right side is bounded above by $v(F, 2\delta)$. Thus

$$v(G_k, \delta) \leqq v(F, 2\delta) \qquad \text{for} \quad k \geqq k_0. \tag{4.7}$$

Similarly, $w_0(G_k, \delta) \leqq w_0(F, 2\delta)$. Since P is tight, we can choose δ so that $P(v(F, 2\delta) < \varepsilon) > 1 - \eta$ and $P(w_0(F, 2\delta) < \varepsilon) > 1 - \eta$. This implies the result.

Let $G_{q,k}$ denote the distribution function that is constant on $[t_i, t_{i+1})$ and whose value at t_i is q_i, where $0 = q_1 \leqq \cdots \leqq q_k = 1$. Thus Q_k is the distribution of $G_{q,k}$ when q has distribution λ, and G_k is the random distribution function obtained by letting q in $G_{q,k}$ have distribution λ.

Recall that d_k is the decision rule that minimizes the Bayes risk $r(P_k, d)$.

Theorem 4.1. *If $G_{q,k}$ is an element of Ω for almost all q in C_λ, if the conditions of Theorem 2.1 are satisfied, if a Bayes solution d exists, and if d has a continuous (in F) bounded risk $R(F, d)$, then for $\Omega \subset D[0, 1]$, d_k converges to d in the sense that the Bayes risk $r(P, d_k)$ of d_k converges to the Bayes risk $r(P, d)$ of d.*

Proof. Consider the Bayes solution $\hat{d}_k$ for the prior Q_k. Since $G_{q,k}$ is constant between points t_i, $\hat{d}_k$ will depend on the X only through $\mathbf{S} = (S_1, \cdots, S_{k-1})$, where S_i = number of X equal to t_{i+1}, $i = 1, \cdots, k - 1$. (Recall that $F(t_1) = F(0) = 0$ a.s. (P) by assumption.) Note that $\mathbf{S}$ and $\mathbf{N}$ have the same distribution under $G_{q,k}$ and that this is the distribution $\mathbf{N}$ has under $F_{q,k}$. This implies that d_k is also a Bayes solution for the prior Q_k. We now have

$$r_k(P, d_k) = r(P_k, d_k) = r(Q_k, d_k) = r(Q_k, \hat{d}_k) \leqq r(Q_k, d). \tag{4.8}$$

By Lemma 4.1,

$$\lim_{k \to \infty} r(Q_k, d) = r(P, d). \tag{4.9}$$

Equations (4.8) and (4.9) yield

$$\limsup_{k \to \infty} r_k(P, d_k) \leqq r(P, d). \tag{4.10}$$

On the other hand, by Lemma 2.1,

$$r_k(P, d_k) \geqq \mathrm{r}(P, d_k). \tag{4.11}$$

Since d is the Bayes solution for P,

$$r(P, d_k) \geqq r(P, d). \tag{4.12}$$

Putting these inequalities together, we get

$$\lim_{k\to\infty} r_k(P, d_k) = r(P, d), \qquad \lim_{k\to\infty} r(P, d_k) = r(P, d). \tag{4.13}$$

If Ω is a class of continuous distribution functions on $[0, 1]$, then it is a subset of the class $C[0, 1]$ of continuous functions on $[0, 1]$. On $C[0, 1]$ we use the sup norm and the σ-algebra generated by the open sets.

LEMMA 4.2. *If $\Omega \subset C[0, 1]$, then F_k converges in distribution to F.*

PROOF. The convergence of finite dimensional distributions follows from the definition of F_k. For each of $F_0 \in \Omega$ and $\delta \in (0, 1)$, let

$$w(F_0, \delta) = \sup_{t-s=\delta} [F_0(t) - F_0(s)]. \tag{4.14}$$

Let k_0 be such that $t_{i+1} - t_i < \delta$, $i = 1, \cdots, k$, for all $k \geqq k_0$. Now the tightness of $\{P_k)\}$ follows from the inequality $w(F_k, \delta) \leqq w(F, 3\delta)$ and the tightness of P. Thus the result follows from Prohorov's theorem applied to $C[0, 1]$.

We can now prove that d_k converges to d under fewer conditions than when $\Omega \subset D[0, 1]$.

THEOREM 4.2. *If $F_{q,k} \in \Omega$ for almost all q in C_λ, if d_k is a mixed Bayes-minimax solution, if a Bayes solution d exists, and if d has a continuous bounded risk $R(F, d)$, then for $\Omega \subset C[0, 1]$, $\lim_{k\to\infty} r(P, d_k) = r(P, d)$.*

PROOF. Since d_k is Bayes for P_k, then

$$r_k(P, d_k) = r(P_k, d_k) \leqq r(P_k, d). \tag{4.15}$$

By Lemma 4.2,

$$\lim_{k\to\infty} r(P_k, d) = r(P, d). \tag{4.16}$$

The rest of the proof now follows on the lines of the proof of Theorem 4.1.

5. Examples of random distribution functions

In order to obtain the mixed Bayes-minimax solutions, we have to specify a distribution for the random distribution function F_k which is linear between the points $(t_1, q_1), \cdots, (t_k, q_k)$, $0 = q_1 \leqq \cdots \leqq q_k = 1$. Equivalently, we have to define a probability λ on

$$A_k = \{q \in R^k : 0 = q_1 \leqq \cdots \leqq q_k = 1\}, \tag{5.1}$$

or a probability π on

$$B_k = \left\{p \in R^{k-1} : 0 \leqq p_i \leqq 1, \sum_{i=1}^{k-1} p_i = 1\right\}. \tag{5.2}$$

Here q_i is thought of as $F(t_i)$ and p_i as $F(t_{i+1}) - F(t_i)$. We say that $p = (p_1, \cdots, p_{k-1})$ has distribution π.

One way to obtain a class of distributions π of p (Freedman [7], Fabius [5], Connor and Mosimann [16]), is to let p have the same distribution as the vector whose ith coordinate is

$$Z_i \prod_{j=1}^{i-1} (1 - Z_j), \qquad i = 1, \cdots, k-1, \tag{5.3}$$

where $Z_1, \cdots, Z_{k-1}$ are independent random variables satisfying

$$0 < Z_i \leqq 1, \qquad Z_{k-1} = 1. \tag{5.4}$$

For each choice of distributions $H_1, \cdots, H_{k-2}$ of the Z, we obtain a probability π on B_k. For this class of probabilities, it is easy ([7], p. 1401 and [5], p. 848) to compute posterior probabilities $\pi(p|\mathbf{N})$ of p given $\mathbf{N} = (N_1, \cdots, N_{k-1})$. Such probabilities π are called *tailfree* by Freedman [7] and Fabius [5] and *neutral* by Connor and Mosimann [16]. If we let each Z_i have a beta distribution $B(r_i, s_i)$ with parameters r_i and s_i, then π is called the *generalized Dirichlet* distribution [16]. If in addition,

$$s_i = \sum_{j=i+1}^{k-1} r_j, \qquad i = 1, \cdots, k-2, \quad s_{k-1} = 0, \tag{5.5}$$

then π is called the *Dirichlet* distribution with parameters $r_1, \cdots, r_{k-1}$.

Extensions of the definition of π on B_k to

$$B_\infty = \left\{ p \in R^\infty : 0 \leqq p_i \leqq 1, \sum_{i=1}^{\infty} p_i = 1 \right\} \tag{5.6}$$

are obtained by replacing (5.4) by

$$0 \leqq Z_i \leqq 1, \qquad \lim_{r \to \infty} \prod_{i=1}^{r} (1 - Z_i) = 0 \qquad \text{a.s.} \tag{5.7}$$

If the Z have beta distributions, then the resulting π on B_∞ is called the *infinite dimensional generalized Dirichlet* distribution (Freedman [7]).

Note that, if p has a Dirichlet distribution, then Cov $(p_i, p_j) < 0$. However, for the generalized Dirichlet distribution, it is possible to have Cov $(p_i, p_j), > 0$ (see [16], p. 198).

EXAMPLE 5.1. Consider the problem of estimating the mean $\mu(F) = E_F(X_1)$ when $p = [F(t_2) - F(t_1), \cdots, F(t_k) - F(t_{k-1})]$ has a generalized Dirichlet distribution π with parameters $(r_1, s_1), \cdots, (r_{k-1}, s_{k-1})$. If we consider the squared error loss function, then the Bayes estimate of

$$\mu(F_k) = \tfrac{1}{2} \sum_{i=1}^{k-1} [F(t_{i+1}) - F(t_i)](t_{i+1} + t_i) \tag{5.8}$$

is

$$\hat{\mu}_k = E_\pi(\mu(F_k)|X) = \tfrac{1}{2} \left(\sum_{i=1}^{k-1} [t_{i+1} + t_i] \, E_\pi(p_i|\mathbf{N}) \right), \tag{5.9}$$

where (see [5])

$$E_\pi(p_i|\mathbf{N}) = \frac{r_i + N_i}{r_1 + s_1 + n} \prod_{\ell=1}^{i-1} \frac{s_\ell + n - \sum_{j=1}^{\ell} N_j}{r_{\ell+1} + s_{\ell+1} + n - \sum_{j=1}^{\ell} N_j}. \tag{5.10}$$

Following Ferguson [6], let α be a finite, finitely additive measure on R and let

$$r_i = \alpha(t_i, t_{i+1}], \qquad s_i = \sum_{j=i+1}^{k-1} r_j, \qquad s_{k-1} = 0. \tag{5.11}$$

Assume that α assigns measure zero to the region outside $(t_1, t_k]$. Then

$$\begin{aligned} E_\pi(\mu(F_k)|X) &= (\alpha(R) + n)^{-1} \sum_{i=1}^{k-1} \tfrac{1}{2}(t_{i+1} + t_i)(\alpha(t_i, t_{i+1}] + N_i) \\ &= \alpha_n E(\mu(F_k)) + (1 - \alpha_n)\bar{X}', \end{aligned} \tag{5.12}$$

where $\alpha_n = \alpha(R)[\alpha(R) + n]^{-1}$ and $\bar{X}'$ is the average of the random variables X' obtained by replacing each X in the interval $(t_i, t_{i+1}]$ by the midpoint $\frac{1}{2}(t_{i+1} + t_i)$. Note that if the t become dense in $(t_1, t_k]$ as in Section 4, then from (5.12)

$$\lim_{k\to\infty} E_\pi(\mu(F_k)|X) = \alpha_n\mu_0 + (1 - \alpha_n)\bar{X}, \tag{5.13}$$

where $\mu_0 = \alpha(R)^{-1}\int x\alpha(dx)$. This is the estimate obtained by Ferguson [6]. Note that in addition to being the Bayes estimate of $\mu(F_k)$, the estimate (5.9) is the mixed Bayes-minimax estimate of $\mu(F)$ for the loss function $[d - \mu(F, k)]^2$ of Example 2.3.

EXAMPLE 5.2. Suppose again that p has a generalized Dirichlet distribution π. If $s \in \{t_1, \cdots, t_k\}$, the mixed Bayes-minimax estimate of $F(s)$ using squared error loss $[d - F(s)]^2$ is

$$\hat{F}(s) = \sum_{j=1}^{i-1} E_\pi(p_j|\mathbf{N}), \tag{5.14}$$

where $s = t_i$ and $E_\pi(p_j|\mathbf{N})$ is given by (5.10). If in addition (5.11) is satisfied, then (5.14) becomes

$$\alpha_n\alpha_0(s) + (1 - \alpha_n)F_n(s), \tag{5.15}$$

where $\alpha_0(s) = \alpha(-\infty, s]/\alpha(R)$ and $\mathrm{F}_n(s)$ is the empirical distribution function of the sample. This is exactly the estimate obtained by Ferguson [6].

Next we consider the problem of defining a probability P on a set Ω of distribution functions F in such a way that it is possible to compute the posterior of F given X under the prior P. Ferguson [6] shows that for each finite, finitely additive measure α on R, it is possible to define a *Dirichlet process* F in such a way that $P_F(A_1), \cdots, P_F(A_m)$ has a Dirichlet distribution with parameters $\alpha_1, \cdots, \alpha_m$, where $d_i = \alpha(A_i)$, and $A_1, \cdots, A_m$ is a measurable partition of R.

He shows that the posterior of F given X is again a Dirichlet process with α replaced by $\alpha + \Sigma_{i=1}^{n} \delta(x_i)$, where $\delta(x)$ is the measure giving mass one to the point x. This makes it possible to compute the Bayes procedure for this prior. The estimate (5.15) in Example 5.2 above is both the Bayes and the mixed Bayes-minimax estimate for this prior.

Fabius ([5], p. 853) gives a general construction of probabilities on the set Ω of all distribution functions on [0, 1] that include the Dirichlet process on [0, 1], the processes of Kraft [12], Kraft and van Eeden [13], and those special cases of the processes of Dubins and Freedman that are contained in [13]. Kraft and van Eeden [13] compute the Bayes estimate for one of these processes for a problem in bioassay. Ferguson shows that if F is the Dirichlet process and Ω_1 is the class of discontinuous distribution functions, then $P(F \in \Omega_1) = 1$. Kraft [12] shows that it is possible to use the construction of Fabius to obtain a process F such that $P(F \in \Omega^*) = 1$, where Ω^* is the class of absolutely continuous distribution functions.

Using definitions (2.2) and (2.3) of Fabius [5], it is possible to check that the Dirichlet process is tailfree for all trees of partitions. Thus one can use expression (2.4) of [5] for the posterior distribution of a tailfree process to conclude that the posterior of a Dirichlet process is again Dirichlet.

◇ ◇ ◇ ◇ ◇

I am indebted to Lucien LeCam, Thomas Ferguson, Jaap Fabius, and others for helpful discussions, and to Michael Stuart for a careful reading of the manuscript that led to many improvements.

REFERENCES

[1] C. Antoniak, "Mixtures of Dirichlet processes with application to some nonparametric problems," Ph.D. thesis, Mathematics Department, University of California, Los Angeles, 1969.

[2] P. Billingsley, *Convergence of Probability Measures*, New York, Wiley, 1968.

[3] K. A. Doksum, "Minimax results for IFRA scale alternatives, *Ann. Math. Statist.*, Vol. 40 (1969), pp. 1778–1783.

[4] L. E. Dubins and D. A. Freedman, "Random distribution functions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1966, Vol. 2, pp. 183–214.

[5] J. Fabius, "Asymptotic behavior of Bayes' estimates," *Ann. Math. Statist.*, Vol. 35 (1964), pp. 846–856.

[6] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," 1970, submitted for publication.

[7] D. A. Freedman, "On the asymptotic behavior of Bayes' estimates in the discrete case," *Ann. Math. Statist.*, Vol. 34 (1963), pp. 1386–1403.

[8] W. Hoeffding, "'Optimum' nonparametric tests," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1951, pp. 83–92.

[9] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, Vol. 36 (1965), pp. 1753–1758.

[10] ———, "Robust confidence limits," *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, Vol. 10 (1968), pp. 269–278.

[11] ———, *Théorie de l'Inférence Statistique Robuste*, Montreal, Les Presses de l'Université de Montréal, 1969.

[12] C. H. Kraft, "A class of distribution function processes which have derivatives," *J. Appl. Probability*, Vol. 1 (1964), pp. 385–388.

[13] C. H. Kraft and C. van Eeden, "Bayesian bio-assay," *Ann. Math. Statist.*, Vol. 35 (1964), pp. 886–890.

[14] E. Ruist, "Comparison of tests for non-parametric hypotheses," *Arkiv Math.*, Vol. 3 (1954), pp. 133–163.

[15] W. Schaafsma, *Hypothesis Testing Problems with the Alternative Restricted by a Number of Inequalities*, Groningen, Noordhoff, 1966.

[16] R. J. Connor and J. E. Mosimann, "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *J. Amer. Statist. Assoc.*, Vol. 64 (1969), pp. 194–206.