ADAPTIVE CLASSIFICATION PROCEDURES*

by

Andrew L. Rukhin
Purdue University

Technical Report #82-34

Department of Statistics
Purdue University

September 1982
(Revised April 1983)

*This work was supported by National Science Foundation Grant MCS-8101670.

ABSTRACT

An explicitly computable necessary and sufficient condition for the existence of an adaptive classification procedure is obtained. By definition, an adaptive procedure, which classifies a sample as coming from one of alternative distributions known only up to a finite-valued nuisance parameter, is required to have the same asymptotic behavior of error probability for these families as asymptotically optimal rules for each of the families. We investigate the conditions under which the overall maximum likelihood procedure is adaptive and derive a rule which is adaptive if any procedure is. The consistency of these procedures is studied. Several exponential-family examples illustrate their form, and small-sample study of error probabilities is performed.

Key words: Classification procedures, information divergence, adaptive procedures, exponential family, maximum likelihood procedures.

1. INTRODUCTION

Let us consider a problem of classifying a random sample $\underline{x} = (x_1,\ldots,x_n)$ as coming from one of m possible probability distributions $P_1,\ldots,P_m$ given over a probability space $\mathcal{X}$. Denote by $\pi_1,\ldots,\pi_m$ prior probabilities of these distributions and by $p_i(\underline{x}) = \prod_1^n f_i(x_j)$, $i = 1,\ldots,m$ probability density functions of $\underline{x}$ for given value of the finite parameter i.

It is well known that the decision rule $\tilde{\delta}$ which minimizes the probability of error,

$$P_e = \sum_{i=1}^{m} P_i(\tilde{\delta}(\underline{x}) \neq i)\pi_i,$$

is the maximum likelihood procedure, i.e. $\tilde{\delta}(\underline{x}) = i$ if $p_i(\underline{x})\pi_i = \max_k p_k(\underline{x})\pi_k$.

It is also known (see Renyi 1969, Krafft and Puri 1974) that P_e tends to zero exponentially fast as the sample size n increases. More precisely

$$\lim_{n\to\infty} P_e^{1/n} = \max_{i\neq k} \inf_{0 \leq s \leq 1} \int f_i^s(x) f_k^{1-s}(x)d\mu(x)$$

$$= \max_{i\neq k} \rho(P_i, P_k). \tag{1.1}$$

Several useful bounds for P_e have been developed by Kailath (1967), Hellman and Raviv (1970) and Ben-Bassat and Raviv (1978).

In this paper we assume that the distributions $P_1,\ldots,P_m$ are not known exactly, but only up to a finite nuisance (shape) parameter α, $\alpha = 1,\ldots,A$. For instance, there might be A experiment types and for each fixed (but unknown to the observer) type α the measurements have one of m alternative distributions. Another example is the transmission of a message in one of A possible languages which use the same alphabet. The message is sent n times over a discrete memoryless channel and the choice between m possible messages (or between m probability distributions which correspond to them) has to be made. Many more examples of this situation, which are important in statistical application in the case A = 2 and for continuous structural parameter,have been considered by Cox (1961, 1962). In the former paper (p. 122) the treatment of a "discontinuous" parameter is suggested as an open problem.

Thus for each value of α a family of distributions $P_1^\alpha, \ldots, P_m^\alpha$ is given, and one can construct a decision rule (which typically depends on α) such that (1.1) holds. If such a rule can be chosen independently of α it is called adaptive. In other terms, a decision rule δ_a is adaptive if for any α

$$\lim_{n\to\infty} [P_e^\alpha]^{1/n} = \lim_{n\to\infty} [\sum_{i=1}^{m} \pi_i \; P_i^\alpha \; (\delta_a(\underline{x}) \neq i)]^{1/n}$$

$$= \max_{i\neq k} \inf_{0 \leq s \leq 1} \int f_i^s \; (x,\alpha) \; f_k^{1-s} \; (x,\alpha) \; d\,\mu(x) = \rho_\alpha .$$

Here f_i $(\cdot, \alpha)$ denotes the density of P_i^α.

Thus an adaptive procedure is asymptotically optimal for any value of the nuisance parameter and does not depend on it.

This formulation of adaptation for a finite structural parameter appears to be the natural analogue of adaptation in the case of a continuous parameter as initiated by Stein (1956). Bickel (1982) gives a general definition in the latter case and derives necessary condition and sufficient conditions for adaptation, which are used to construct adaptive procedures in situations studied earlier by Beran 1974, Kraft and van Eeden 1970, Policello and Hettmansperger 1976, Sacks 1975, Stone 1975, and van Eeden 1970. A review of adaptive robust procedures is provided in the paper of Hogg (1974).

One of the methods of estimation in the presence of a nuisance parameter suggested by Cox (1961) and Hogg, Uthoff, Randles and Davenport (1972) consists of the following. Estimate (or eliminate) the nuisance parameter by the maximum likelihood method, i.e. define α by the formula

$$\max_{\theta} \pi_\theta \, p_\theta \, (\underline{x}, \alpha) = \max_{\beta} \max_{\theta} \pi_\theta \, p_\theta \, (\underline{x}, \beta),$$

where $$p_\theta \, (\underline{x}, \alpha) = \prod_1^n f_\theta \, (x_j, \alpha).$$

For this value of α use the maximum likelihood procedure $\tilde{\delta}$. The combination of these two methods leads to the overall maximum likelihood classification rule $\hat{\delta}$

$$\{\hat{\delta} = i\} = \{\max_{\beta} \pi_i \, p_i \, (\underline{x}, \beta) = \max_k \max_{\beta} \pi_k \, p_k \, (\underline{x}, \beta)\}.$$

A closely related procedure δ^* corresponds to the Bayes method of eliminating the nuisance parameter α by means of prior weights w_α, i.e.

$$\{\delta^* = i\} = \{\pi_i \sum_\alpha w_\alpha \, p_i \, (\underline{x}, \alpha) = \max_k \pi_k \sum_\alpha w_\alpha \, p_k \, (\underline{x}, \alpha)\}.$$

In this paper we investigate the conditions under which procedures $\hat{\delta}$ and δ^* are adaptive. The existence of adaptive procedures has been studied by the author (Rukhin 1982), who obtained a necessary and a sufficient condition for the existence of such rules. In this paper we derive one very simple necessary and sufficient condition for adaptation. It turns out that an adaptive procedure exists if and only if the classification problem for any fixed value of the nuisance parameter is "at least as difficult" as the classification problem for distributions belonging to different values of this parameter. The "difficulty" of the problem is described here by an information divergence function ρ introduced in (1.1). We construct procedure $\hat{\delta}_a$, which is adaptive if any adaptive procedure exist, and its consistency property when the model is uncorrect is studied. We also consider several examples of adaptive procedures for exponential families and investigate small sample behavior of $\hat{\delta}$ and $\hat{\delta}_a$. The necessary mathematical results and some notation are gathered in the Appendix.

2. CONDITIONS FOR THE EXISTENCE OF ADAPTIVE PROCEDURES

For any two probability distributions P and Q we define

$$\rho(P, Q) = \inf_{0 \leq s \leq 1} \int p^s (x)\, q^{1-s} (x)\, d\, \mu(x) = \rho(Q, P).$$

Here p and q are densities of P and Q with respect to some measure μ. Clearly $0 \leq \rho(P, Q) \leq 1$, and if P and Q are different, then $\rho(P, Q) < 1$, and if P and Q are not mutually singular, then $\rho(P, Q) > 0$. Thus the quantity $\rho(P, Q)$ characterizes the divergence or discrimination between P and Q. In particular, as follows from (1.1), the larger is $\max_{i \neq k} \rho(P_i, P_k)$ the "more difficult" is the classification problem in the sense of the rate of convergence to zero of error probability. (See Vajda 1970 for further properties of the function ρ).

We shall need the discrimination $\rho(F, G)$ between any two mutually absolutely continuous positive (non-normed) measures F and G which is defined by the formula

$$\rho(F, G) = \inf_{s>0} \int [dG/dF]^s\, dF.$$

Theorem 1. An adaptive procedure exists if and only if

$$\max_{\alpha\neq\beta}\max_{i\neq k} \rho\ (F_i^\alpha,\ F_k^\beta) \leq 1. \tag{2.1}$$

Here $F_i^\alpha = e^{c_\alpha} P_i^\alpha$, $c_\alpha = -\log \rho_\alpha$, $i = 1,\ldots,m, \alpha = 1,\ldots,A$.

Proof. If an adaptive procedure exists then Lemma 1 with $a = c_\alpha$, $b = c_\beta$ implies

$$1 = \max\ [e^{c_\alpha} \rho_\alpha,\ e^{c_\beta} \rho_\beta]$$

$$\geq \max\ [\ \rho\ (F_i^\alpha,\ F_k^\beta),\ \rho\ (F_k^\beta,\ F_i^\alpha)].$$

Since α, β, i, k are arbitrary, (2.1) holds.

Because of (5.1),

$$\max_\alpha e^{c_\alpha} \rho_\alpha\ (c_1,\ldots,c_A) \leq \max_{\alpha,\beta}\max_{i\neq k}\ \rho(F_i^\alpha,\ F_k^\beta).$$

But for any α

$$\max_{i\neq k} \rho\ (F_i^\alpha,\ F_k^\alpha) = 1, \tag{2.2}$$

so that (2.1) implies that, for any α,

$$\max_\alpha e^{c_\alpha} \rho_\alpha\ (c_1,\ldots,c_A) \leq 1$$

or

$$\rho_\alpha\ (c_1,\dots,c_A) \le e^{-c_\alpha} = \rho_\alpha.$$

The existence of an adaptive procedure follows now from Theorem A.

According to (2.1) if an adaptive procedure exists then for all $\alpha \neq \beta$, $i \neq k$,

$$\rho(F_i^\alpha,\ F_k^\beta) \le 1.$$

Since $F_i^\alpha\ (\mathcal{X}) = e^{c_\alpha} > 1$,

$$\min\ [F_i^\alpha(\mathcal{X}),\ F_k^\alpha(\mathcal{X})] > \ \rho(F_i^\alpha,\ F_k^\beta) = e^{c_\alpha} \inf_{0<s<1}\ E_i^\alpha\ [\tilde{f}_k(X,\beta)/\tilde{f}_i\ (X,\alpha)]^s$$

$$= \rho(F_k^\beta,\ F_i^\alpha),$$

where $\tilde{f}_i\ (X,\alpha) = f_i\ (X,\alpha)\ e^{c_\alpha}$, $i = 1,\dots,m$, $\alpha = 1,\dots,A$. Since always

$$\rho(F_i^\alpha,\ F_k^\beta) \le e^{c_\alpha} \inf_{0<s<1}\ E_i^\alpha\ [\tilde{f}_k(X,\beta)/\tilde{f}_i\ (X,\alpha)]^s,$$

we can reformulate Theorem 1 in the following way.

Theorem 1^1. An adaptive procedure exists if and only if for all $\alpha \neq \beta$

$$\max_{i \neq k} \inf_{0<s<1} E_i^{\alpha} [\tilde{f}_k (X,\beta)/\tilde{f}_i (X,\alpha)]^s \leq \rho_\alpha .$$

Theorem 2. If condition (2.1) is met then any of the following procedures based on the modified likelihood functions $\prod_1^n \tilde{f}_i (x_j,s)$ is adaptive:

(i) modified overall maximum likelihood estimator $\hat{\delta}_a$

$$\{\hat{\delta}_a = i\} = \{\max_\alpha \pi_i \prod_1^n \tilde{f}_i(x_j,\alpha) = \max_k \max_\alpha \pi_k \prod_1^n \tilde{f}_k(x_j, \alpha)\};$$

(ii) modified Bayes-maximum likelihood rule δ_a^*

$$\{\delta_a^* = i\} = \{\pi_i \sum_\alpha w_\alpha \prod_1^n \tilde{f}_i(x_j, \alpha) = \max_k \pi_k \sum_\alpha w_\alpha \prod_1^n \tilde{f}(x_j,\alpha)\},$$

where $w_1,\dots,w_A$ are arbitrary but fixed positive weights.

The proof of Theorem 2 follows from Lemma 2.

Our next result provides a sufficient condition for the adaptation of the modified overall maximum likelihood procedure $\hat{\delta}_b$ based on $p_i(\underline{x}, \alpha)$. In particular when $b_1 = \dots = b_A = 0$ Theorem 3 gives a condition for the adaptation of $\hat{\delta}$.

Theorem 3. For fixed real constants $b_1, \dots, b_A$ define procedure $\hat{\delta}_b$ by the formula

$$\{\hat{\delta}_b = i\} = \{\pi_i \max_\alpha p_i(\underline{x}, \alpha)e^{nb_\alpha} = \max_k \pi_k \max_\alpha p_k(\underline{x}, \alpha)e^{nb_\alpha}\} .$$

If for all $\alpha \neq \beta$

$$\max_{i\neq k} \min_{\gamma} \inf_{s>0} \exp\{s(b_\beta - b_\gamma)\} E_i^\alpha[f_k(X,\beta)/f_i(x,\gamma)]^s \leq \rho_\alpha$$

then $\hat{\delta}_b$ is adaptive.

Proof. Because of Corrollary 1 to Lemma 3

$$\rho_\alpha \leq \rho_\alpha\ (b_1, \ldots, b_A)$$

$$= \max_{\beta} \max_{i\neq k} \inf_{s_1,\ldots,s_A\geq 0} \exp\{\sum_r s_r(b_\beta - b_r)\} E_i^\alpha \prod_{r=1}^{A} [f_k(X,\beta)/f_i(X,r)]^{s_r}$$

$$\leq \max_{\beta} \max_{i\neq k} \min_{\gamma} \inf_{s>0} \exp\{s(b_\beta - b_\gamma)\} E_i^\alpha(f_k(X,\beta)/f_i(X,\gamma))^s \ .$$

Thus the condition of Theorem 3 implies that

$$\rho_\alpha = \rho_\alpha\ (b_1, \ldots, b_A),$$

and because of Theorem A $\hat{\delta}_b$ is adaptive.

Corollary. If for all $\alpha \neq \beta$

$$\max_{i\neq k} \rho(F_i^\alpha, F_k^\beta) \leq \min\ [1, \exp(c_\beta - c_\alpha - b_\beta + b_\alpha)],$$

then $\hat{\delta}_b$ is an adaptive procedure.

Proof of Corollary. The condition of this Corollary implies the existence of an adaptive procedure. If $c_\beta - b_\beta \geq c_\alpha - b_\alpha$ then

$$\max_{i\neq k} \inf_{s>0} \exp\{s(b_\beta - b_\alpha)\}\ E_i^\alpha[f_k(X,\beta)/f_i(X,\alpha)]^s$$

$$\leq e^{-c_\alpha} \max_{i\neq k} \rho(F_i^\alpha, F_k^\beta) \leq \rho_\alpha .$$

If $c_\beta - c_\alpha < b_\beta - b_\alpha$ then according to Theorem 1[1]

$$\exp\,(c_\beta - c_\alpha - b_\beta + b_\alpha) \geq \max_{i \neq k}\ (F_i^\alpha,\ F_k^\beta)$$

$$= \max_{i \neq k}\ e^{c_\alpha} \inf_{0<s<1} e^{s(c_\beta - c_\alpha)}\ E_i^\alpha[f_k(X,\beta)/f_i(X,\alpha)]^s$$

$$\geq \exp\,(c_\alpha + c_\beta - c_\alpha - b_\beta + b_\alpha)\ \max_{i \neq k}\ \inf_{0<s<1} e^{s(b_\beta - b_\alpha)} E_i^\alpha[f_k(X,\beta)/f_i(X,\alpha)]^s,$$

so that the condition of Theorem 3 holds true in both cases.

Our last result in this Section gives a condition for the consistency of the procedure $\hat{\delta}_b$ for arbitrary family of distributions $\{Q_i^\alpha,\ i=1,\dots,\ m;\ \alpha=1,\dots,A\}$.

Theorem 4. Procedure $\hat{\delta}_b$ is consistent for any family $\{Q_i^\alpha\}$ if for any $\alpha \neq \beta$, $i \neq k$ there exists γ such that

$$b_\gamma - K(Q_i^\alpha, Q_i^\gamma) > b_\beta - K(Q_i^\alpha, Q_k^\beta).$$

Here $K(Q,P) = E^Q \log(dQ/dP)$ is the information number.

Proof. According to Lemma 3 $\hat{\delta}_b$ is consistent if

$$\rho_\alpha(b_1,\ \dots,\ b_A) = \max_i\ \ell_i^\alpha(b_1,\ \dots,\ b_A) < 1\ .$$

If the condition of Theorem 4 is met then with $g_k(X,\beta)$ denoting the density of Q_k^β

$$\rho_\alpha(b_1,\ \dots,\ b_A) \leq \max_\beta\ \max_{i \neq k}\ \inf_{s>0}\ e^{s(b_\beta - b_\gamma)} E_i^\alpha[g_k(X,\beta)/g_i(X,\gamma)]^s$$

$$< 1,$$

since the derivative of the latter function at $s = 0$ is negative and its value at $s = 0$ is 1.

Corollary. Procedure $\hat{\delta}_b$ is consistent if for any $\alpha \neq \beta$

$$b_\beta - b_\alpha < \min_{i \neq k} K(Q_i^\alpha, Q_k^\rho).$$

Procedure $\hat{\delta}$ is consistent for any family $\{Q_i^\alpha\}$ such that $Q_i^\alpha \neq Q_k^\beta$ for $i \neq k$, $\alpha \neq \beta$.

To prove this Corollary put $\gamma = \alpha$ in Theorem 4.

Notice that Theorems 3 and 4 hold true also for procedures δ_b^* of the form

$$\{\delta_b^* = i\} = \{\pi_i \sum_\alpha w_\alpha e^{nb_\alpha} p_i(\underline{x}, \alpha) = \max_k \pi_k \sum_\alpha w_\alpha e^{nb_\alpha} p_k(\underline{x}, \alpha)\}$$

where $w_1, \ldots, w_A$ are fixed positive numbers.

3. DISCUSSION

Because of (2.2) Theorem 1 has a very clear-cut meaning: An adaptive classification procedure exists if and only if the information divergence between members of one family is not smaller than the divergence between distributions in any two different families. In other words, such a rule exists if and only if the classification problem for any fixed value of the nuisance parameter is at least as difficult as the classification problems formed by the distributions corresponding to different values of this parameter.

Condition (2.1) is explicitly computable and if it is met i.e., if an adaptive procedure exists, then the modified overall maximum likelihood procedure $\hat{\delta}_a$ is adaptive. In Section 4 we perform a small sample study of the procedures $\hat{\delta}_a$ and $\hat{\delta}$ in several cases when (2.1) is satisfied. According to these results, the procedure $\hat{\delta}_a$ exhibits very reasonable behavior compared to the optimal (Bayes) procedure which uses the knowledge about the nuisance parameter α. Therefore it can be recommended in application if the family $\{P_i^\alpha\}$ satisfies (2.1). In Section 4 we show that in classification problem of normal population $\hat{\delta}_a$ and $\hat{\delta}$ have the same parametric region of adaptation. However in classification problem of exponential population $\hat{\delta}_a$ may be adaptive when $\hat{\delta}$ is not. Moreover, as we shall see, $\hat{\delta}_a$ may have smaller error probability than $\hat{\delta}$ for all sample sizes.

The drawback of procedure $\hat{\delta}_a$ is that when (2.1) is violated it may not be consistent. Because of Corollary to Theorem 4 $\hat{\delta}$ is always consistent. Therefore intermediate weights b_α, $0 \leq b_\alpha \leq c_\alpha$, may be used in constructing modified maximum likelihood procedures $\hat{\delta}_b$. Indeed assume that the parametric region, where the adaptation is desired, is described by the inequalities

$$\max_{i \neq k} \rho(F_i^\alpha, F_k^\beta) \leq \min\,[1, \exp(c_\beta - c_\alpha - b_\beta + b_\alpha)]$$

for some constants $b_1, \ldots, b_A$. Then the estimator $\hat{\delta}_b$ will be consistent for any family $\{Q_i^\alpha\}$ such that the condition of Theorem 4 is satisfied.

4. EXAMPLES

Let the distributions P_k^α be members of an exponential family over Euclidean space, i.e. the densities $f_k\,(x, \alpha)$ have the form

$$f_k\,(x, \alpha) = \exp\,\{\theta_\alpha'\,(k)x - \chi\,(\theta_\alpha\,(k))\},$$

$\alpha = 1,\ldots,A$, $k = 1,\ldots,m$. Here $\theta_\alpha\,(k)$ and x are vectors, and ' denotes the transposition. Since the distributions P_k^α are supposed to be different, the common support of these distributions contains at least two points, and the function χ is strictly convex.

If

$$F = e^{b_\alpha} P_i^\alpha, \; G = e^{b_\beta} P_k^\beta,$$

then

$$\log \rho(F, G) = \inf_{s>0} [\chi((1-s)\theta + s\xi) - s\chi(\xi) - (1-s)\chi(\theta) + sb_\beta + (1-s)b_\alpha],$$

where $\theta = \theta_\alpha (i)$, $\xi = \theta_\beta (k)$.

In particular

$$c_\alpha = - \log \rho_\alpha$$

$$= - \max_{i\neq k} \inf_{s>0} [\chi((1-s)\theta_\alpha(i) + s\,\theta_\alpha (k)) - s\,\chi(\theta_\alpha(k)) - (1-s)\,\chi(\theta_\alpha(i))]. \quad (4.1)$$

According to Theorem 1 an adaptive procedure exists if and only if

$$\max_{\alpha\neq\beta} \max_{i\neq k} \inf_{s>0} [\chi((1-s)\theta_\alpha(i) + s\,\theta_\beta(k)) - s\chi(\theta_\beta(k))$$

$$- (1-s)\chi(\theta_\alpha(i)) + s\,c_\beta + (1-s)\,c_\alpha] \leq 0, \quad (4.2)$$

where c_α are defined by (4.1).

Let us consider some specific situations.

1. P_i^α is multivariate normal distribution with mean $\eta_\alpha(i)$ and nonsingular covariance matrix $\sum$. Then

$$\chi(\theta) = \theta' \sum \theta/2 = ||\theta||^2/2, \ \theta_\alpha(i) = \sum^{-1} \eta_\alpha(i)$$

and

$$c_\alpha = \min_{i \neq k} ||\theta_\alpha(i) - \theta_\alpha(k)\ ||^2/8.$$

Because of (4.2), a necessary and sufficient condition for adaptation is that for all $\alpha \neq \beta$, $i \neq k$

$$(2c_\alpha)^{1/2}||\theta_\beta(k) - \theta_\alpha(i)|| \leq ||\theta_\beta(k) - \theta_\alpha(i)||^2/2 + c_\alpha - c_\beta.$$

For instance, when m = 2 this condition means that

$$2\ ||\xi_\beta - \theta_\alpha|| \geq ||\xi_\alpha - \theta_\alpha|| + ||\xi_\beta - \theta_\beta||.$$

Here $\xi_\beta = \theta_\beta(2)$, $\theta_\alpha = \theta_\alpha(1)$.

In this case the condition of Theorem 3 with $b_1 = \ldots = b_A = 0$ is met for all parametric points in the region defined by (2.1), and the procedures $\hat{\delta}_a$ and $\hat{\delta}$ are adaptive simultaneously.

In Table 1 we reported the results of a small sample study of the efficiencies

$$e_n = -n^{-1} \log [\pi_1\ P_1^\alpha(\hat{\delta}_a \neq 1) + \pi_2\ P_2^\alpha(\hat{\delta}_a \neq 2)]$$

and

$$d_n = -n^{-1} \log [\pi_1 P_1^\alpha(\hat{\delta} \neq 1) + \pi_2 P_2^\alpha(\hat{\delta} \neq 2)],$$

for $\pi_1 = \pi_2 = .5$, $-\theta_1 = \xi_1 = 1$, $\theta_2 = 3$, $\xi_2 = 4$ and different sample sizes n. In this case

$$\{\hat{\delta}_a = 1\} = \{\bar{x} < 0\} \cup \{35/16 < \bar{x} < 7/2\},$$

$$\{\hat{\delta} = 1\} = \{\bar{x} < 0\} \cup \{2 < \bar{x} < 7/2\}.$$

Along with e_n and d_n we also tabulate, for given α, the efficiencies f_n of the Bayes procedure δ_0, which in this case is just maximum likelihood rule. For $\alpha = 1$

$$\{\delta_0 = 1\} = \{\bar{x} < 0\}$$

and for $\alpha = 2$

$$\{\delta_0 = 2\} = \{\bar{x} > 7/2\}.$$

For given α the quantities e_n, d_n, f_n converge to their common limiting value c_α rather slowly (in this and the next examples the differences $e_n - c_\alpha$, $d_n - c_\alpha$, $f_n - c_\alpha$ are of order $\log (Cn)/n$ for some constants C). (Similar phenomenon has been reported by Goeoneboom and Oosterhoff, 1980). However it. follows from Table 1 that for $n \geq 25$, the procedure $\hat{\delta}_a$ is practically as good as δ_0. A well known approximation to the standard normal distribution function (see Feller 1968, p. 193) was used to evaluate the probabilities of large deviations which arose in e_n, d_n, f_n.

2. P_i^α is a distribution over the real line with a density of the form

$$f_k(y, \alpha) = C \exp\{-\theta_\alpha(k)|y|^{a-1} + a \log \theta_\alpha(k)\},$$

or of the form

$$f_k(y, \alpha) = C \exp\{-\theta_\alpha(k)y^{a-1} + a \log \theta_\alpha(k)\}, \ y > 0.$$

These families include normal, exponential and Weibull distributions with unknown scale parameter. In this case $\chi(\theta) = - a \log \theta$, $\theta > 0$, $a > 0$, and

$$c_\alpha = a \min_{i \neq k} [r_\alpha(i,k) - \log r_\alpha(i,k) - 1],$$

where $r_\alpha(i,k) = \log(\theta_\alpha(i)/\theta_\alpha(k))/(\theta_\alpha(i)/\theta_\alpha(k) - 1)$.

Condition (4.2) means that

$$\min_{\alpha \neq \beta} \min_{i \neq k} [q_{\alpha\beta}(i,k) - \log q_{\alpha\beta}(i,k) - 1 - c_\alpha] \geq 0,$$

where

$$q_{\alpha\beta}(i,k) = (\log(\theta_\alpha(i)/\theta_\beta(k)) + c_\beta - c_\alpha)/(\theta_\alpha(i)/\theta_\beta(k) - 1).$$

We have also evaluated the efficiencies e_n, d_n, f_n for $\hat{\delta}_a$, $\hat{\delta}$ and δ_0, respectively, when $\theta_2 = 2.5$, $\xi_2 = 3.5$ (Table 2). In this case $\hat{\delta}_a$ is adaptive, but, for $\alpha = 1$, $\hat{\delta}$ is not adaptive. To evaluate numerically the quantities $\Gamma(n, nt)$ which enter in the formulae for e_n, d_n, f_n, an algorithm suggested by Gautschi (1979) was used.

3. P_i is the binomial distribution with parameters N and $p_\alpha(i)$. An easy calculation shows that

$$c_\alpha = N \min_{i\neq k}[r_\alpha(i,k)\log(r_\alpha(i,k)/p_\alpha(i)) + (1 - r_\alpha(i,k))\log((1-r_\alpha(i,k))/(1-p_\alpha(i)))],$$

where

$$r_\alpha(i,k) = \log((1-p_\alpha(i))/(1-p_\alpha(k)))/\log(p_\alpha(k)(1-p_\alpha(i))/p_\alpha(i)/(1-p_\alpha(k))).$$

An adaptive procedure exists if and only if

$$\min_{\alpha\neq\beta}\min_{i\neq k}[q_{\alpha\beta}(i,k)\log(q_{\alpha\beta}(i,k)/p_\alpha(i)) + (1-q_{\alpha\beta}(i,k))\log((1-q_{\alpha\beta}(i,k))/(1-p_\alpha(i))$$

$$- c_\alpha] \geq 0,$$

where

$$q_{\alpha\beta}(i,k) = \log((1-p_\alpha(i))/(1-p_\beta(k))/\log(p_\beta(k)(1-p_\alpha(i))/p_\alpha(i)/(1-p_\beta(k))).$$

In Table 3 the efficiencies e_n, d_n, f_n are tabulated for $p_2 = .7$, $q_2 = .65$. In this situation $\hat{\delta}_a$ is adaptive, but $\hat{\delta}$ is not. For $\alpha = 1$

$$d_n \to q \log q/p_1 + (1-q) \log ((1-q)/(1-p_1)) < c_1,$$

$$q = q_{1,2}\ (1,2).$$

In fact, the procedure $\hat{\delta}_a$ is preferable to $\hat{\delta}$ for all sample sizes.

To evaluate the probabilities of large deviations for the binomial distribution which entered into the quantities e_n, d_n, f_n, we used an approximation for the latter distribution due to Bahadur (1960).

More examples of adaptive procedures in the estimation problem of a shift parameter on a cyclic group are given in Rukhin (1983b).

Acknowledgement. Thanks are due to Professors Walter Gautschi and Herman Rubin of Purdue University for consultation in the calculations of Example 2. The author is also grateful to the referee for many penetrating comments.

APPENDIX

In this Appendix we give three Lemmas and a Theorem needed to prove Theorems 1-4 of Section 2.

Let α, β be two different values of the nuisance parameter and let i, k be two different values of the structural parameter. Define positive measures F and G in the following way: $F = e^a P_i^\alpha$, $G = e^b P_k^\beta$, where a, b are fixed numbers.

Lemma 1. For any procedure δ,

$$\max \{e^a \liminf_{n \to \infty} [P_i^\alpha (\delta(\underline{x}) \neq i)]^{1/n}, e^b \liminf_{n \to \infty} [P_k^\beta (\delta(\underline{x}) \neq k)]^{1/n}\}$$

$$\geq \max [\rho (F, G), \rho (G, F)].$$

Proof. Let δ_1 be a rule which takes only two values, i and k, and is of the form

$$\{\delta_1 = i\} = \{p_i (\underline{x}, \alpha) e^{na} > p_k (\underline{x}, \beta) e^{nb}\}$$

and

$$\{\delta_1 = k\} = \{p_k (\underline{x}, \beta) e^{nb} > p_i (\underline{x}, \alpha) e^{na}\}.$$

The definition of δ_1 in the case of the tie, $p_i(\underline{x}, \alpha) e^{na} = p_k (\underline{x}, \beta) e^{nb}$, is immaterial. Clearly δ_1 is a Bayes rule against the uniform prior distribution over {i, k} for the loss function

$$L\ (\alpha,\delta) = \begin{cases} 0 & \text{if } \alpha = \delta \\ e^{na} & \text{if } \alpha = i,\ \ \delta = k \\ e^{nb} & \text{if } \alpha = k,\ \ \delta = i\ . \end{cases}$$

Therefore for any procedure δ

$$e^{na}\ P_i^\alpha\ (\delta(\underline{x}) \neq i) + e^{nb}\ P_k^\beta\ (\delta(\underline{x}) \neq k)$$

$$\geq e^{na}\ P_i^\alpha\ (\delta_1\ (\underline{x}) \neq i) + e^{nb}\ P_k^\beta\ \ (\delta_1\ (\underline{x}) \neq k).$$

But

$$P_i^\alpha\ (\delta_1\ (\underline{x}) \neq i) = P_i^\alpha\ (\delta_1\ (\underline{x}) = k)$$

$$= P_i^\alpha\ (\sum_1^n\ \ (\log f_k\ (x_j,\ \beta) - \log f_i\ (x_j,\ \alpha) + b - a) > 0).$$

Because of Chernoff's Theorem (Chernoff 1952) we have

$$\lim_{n\to\infty} [P_i^\alpha(\delta_1(\underline{x}) \neq i)]^{1/n} = \inf_{s>0} e^{s(b-a)}\ E_i^\alpha\ [f_k(X,\beta)/f_i(X,\alpha)]^s.$$

$$= e^{-a}\ \rho(F,\ G).$$

Analogously

$$\lim_{n\to\infty} [P_k^\beta(\delta_1(\underline{x}) \neq k)]^{1/n} = \inf_{s>0} e^{s(a-b)}\ E_k^\beta\ [f_i(X,\alpha)/f_k(X,\beta)]^s$$

$$= e^{-b}\ \rho(G,\ F),$$

and Lemma 1 is proven.

The following Lemma is proved with the help of multivariate version of Chernoff's Theorem (Groeneboom, Oosterhoff, Ruymgaart 1979). Details of the proof can be reproduced from Rukhin (1982).

Lemma 2. Let c_n, $n = 1, 2, \ldots,$ be a sequence of positive numbers such that $n^{-1} \log c_n$ converges to a finite limit L. Also let a_i, b_i, $i = 1,\ldots,A$, be real constants and let p_i, q_i, $i = 1,\ldots,A,$ be strictly positive measurable functions. If, for all positive probabilities v_i, $i = 1,\ldots,A,$ and all k,

$$\Pr \{\sum_1^A v_i \, [\log(p_k(X)/q_i(X)) + a_k - b_i] > L\} > 0$$

then for any positive weights, $w_1,\ldots,w_A$,

$$\lim_{n\to\infty} [\Pr \{\sum_k w_k \prod_1^n (p_k(x_j)e^{a_k}) \geq c_n \sum_k w_k \prod_1^n (q_k(x_j)e^{b_k})\}]^{1/n}$$

$$= \lim_{n\to\infty} [\Pr \{\max_k w_k \left[\prod_1^n p_k(x_j)e^{a_k}\right] \geq c_n \max_k w_k \prod_1^n \left[q_k(x_j) \; e^{b_k}\right]\}]^{1/n}$$

$$= \max_{1\leq k\leq A} \quad \inf_{s_1,\ldots,s_A \geq 0} \quad \exp \{-\sum_i s_i(a_k - b_i)\} \, E \prod_{i=1}^A (p_k(X)/q_i(X))^{s_i}.$$

Lemma 2 motivates the following notation for a collection of mutually absolutely continuous probability measures P_i^α, $i = 1, \ldots, m$, $\alpha = 1, \ldots, A$

$$\ell_i^\alpha(b_1,\ldots,b_A)$$

$$= \max_{k:k\neq i} \; \max_\beta \; \inf_{s_1,\ldots,s_A \geq 0} \; \exp\{\sum_r s_r(b_\beta - b_r)\} E_i^\alpha \prod_{r=1}^A [f_k(X,\beta)/f_i(X,r)]^{s_r},$$

$$\rho_\alpha(b_1,\ldots,b_A) = \max_i \; \ell_i^\alpha(b_1,\ldots,b_A).$$

Here $b_1,\dots,b_A$ are real numbers. Notice that

$$e^{b_\alpha}\rho_\alpha(b_1,\dots,b_A)$$

$$\leq \max_\beta \max_{i\neq k} \inf_{s>0} \exp\,(b_\alpha + s(b_\beta - b_\alpha))\; E_i^\alpha[f_k(X,\beta)/f_i(X,\alpha)]^s$$

$$= \max_\beta \max_{i\neq k} \rho(F_i^\alpha,\, F_k^\beta), \tag{5.1}$$

where $F_i^\alpha = e^{b_\alpha}\, P_i^\alpha$, $i = 1,\dots,m$, $\alpha = 1,\dots,A$.

Lemma 3. Let $\hat{\delta}_b$ be a maximum likelihood procedure based on the likelihood function $\max_\alpha p_i\,(\underline{x},\,\alpha)e^{nb_\alpha}$, i.e.

$$\{\hat{\delta}_b = i\} = \{\pi_i \max_\alpha p_i\,(\underline{x},\,\alpha)e^{nb_\alpha} = \max_k \pi_k \max_\alpha p_k\,(\underline{x},\,\alpha)e^{nb_\alpha}\}.$$

Then

$$\lim_{n\to\infty} [P_i^\alpha\,(\hat{\delta}_b(\underline{x}) \neq i)]^{1/n} = \ell_i^\alpha\,(b_1.,,,.b_A)$$

and, for any procedure δ,

$$\max_\alpha \{e^{b_\alpha} \liminf_{n\to\infty} \max_i\, [P_i^\alpha\,(\delta(\underline{x}) \neq i)]^{1/n}\}$$

$$\geq \max_\alpha \{e^{b_\alpha}\rho_\alpha\,(b_1,\dots,b_A)\}.$$

Proof. The first statement of Lemma 3 follows from Lemma 2 applied to functions $e^{b_\alpha} f_k(\underline{x},\alpha)$. We derive the second from the inequalities

$$A \sum_i \pi_i \max_\alpha [P_i^\alpha (\delta(\underline{x}) \neq i) e^{nb_\alpha}]$$

$$\geq \sum_i \pi_i \int_{\{\delta \neq i\}} \cdots \int \max_\alpha [p_i (\underline{x}, \alpha) e^{nb_\alpha}] \, d\mu(\underline{x})$$

$$\geq \sum_i \pi_i \int_{\{\hat{\delta}_b \neq i\}} \cdots \int \max_\alpha [p_i (\underline{x}, \alpha) e^{nb_\alpha}] \, d\mu(\underline{x})$$

$$\geq \sum_i \pi_i \max_\alpha [P_i^\alpha (\hat{\delta}_b (\underline{x}) \neq i) e^{nb_\alpha}].$$

The penultimate inequality here follows from the fact that $\hat{\delta}_b$ is the Bayes rule with respect to zero-one loss and the density proportional to $\max_\alpha [p_i(\underline{x},\alpha) e^{nb_\alpha}]$.

Corollary 1. For any α

$$\rho_\alpha (b_1.,,,.b_A) \geq \rho_\alpha .$$

This corollary follows from Lemma 3 and (1.1).

The asymptotical minimaxness of procedures similar to $\hat{\delta}_b$ in classification problems without nuisance parameter has been studied in Rukhin (1983a).

Theorem A. If an adaptive procedure exists then for all real $b_1,\dots,b_A$

$$\max_\alpha e^{b_\alpha} \rho_\alpha \geq \max_\alpha e^{b_\alpha} \rho_\alpha (b_1,\dots,b_A).$$

If for all α and some real $b_1,\dots,b_A$

$$\rho_\alpha \geq \rho_\alpha (b_1,\dots,b_A),$$

then an adaptive procedure exists.

Proof. Assume that an adaptive rule δ_a exists. Then because of Lemma 3

$$\max_\alpha e^{b_\alpha} \rho_\alpha = \max_\alpha e^{b_\alpha}[\lim \max_i [P_i^\alpha (\delta_a(\underline{x}) \neq i)]^{1/n}$$

$$\geq \max_\alpha e^{b_\alpha} \rho_\alpha (b_1,\dots,b_A).$$

Because of the Corollary 1, the inequalities

$$\rho_\alpha \geq \rho_\alpha (b_1,\dots,b_A) \quad \text{for } \alpha = 1, \dots, A$$

imply the equality of these quantities, which establishes the adaptiveness of procedure $\hat{\delta}_b$ of Lemma 3.

Corollary 2. The existence of an adaptive procedure implies that for all real $b_1,\dots,b_A$

$$\max_\alpha e^{b_\alpha} \rho_\alpha (b_1,\dots,b_A) = \max e^{b_\alpha} \rho_\alpha.$$

Indeed one has

$$\max_\alpha e^{b_\alpha} \rho_\alpha \leq \max_\alpha e^{b_\alpha} \rho_\alpha (b_1,\ldots,b_A),$$

so that Corollary 2 follows from the first part of Theorem A.

Corollary 3. If for some $i \neq k$ and $\alpha \neq \beta$, $P_i^\alpha = P_k^\beta$, then there is no adaptive procedure.

Indeed in this case

$$\rho_\alpha (0,\ldots,0) = 1,$$

and

$$\max_\alpha \rho_\alpha < \max_\alpha \rho_\alpha (0,\ldots,0) = 1,$$

so that an adaptive procedure cannot exist.

REFERENCES

BAHADUR, R.R. (1960), "Some Approximations to the Binomial Distribution Function", Annals of Mathematical Statistics, 31, 43-54.

BEN-BASSAT, MOSHE, and RAVIV, JOSEPH (1978), "Renyi's Entropy and the Probability of Error," IEEE Transactions on Information Theory, IT-24, 324-331.

BERAN, R. (1974), "Asymptotically Efficient Adaptive Rank Estimates in Location Models", Annals of Statistics, 2, 63-74.

BICKEL, P. J. (1982), "On Adaptive Estimation," Annals of Statistics, 10, 647-671.

CHERNOFF, H. (1952), "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," Annals of Mathematical Statistics, 23, 493-507.

COX, D. R. (1961), "Tests of Separate Families of Hypotheses," Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 105-123.

__________.(1962), "Further Results on Tests of Separate FAmilies of Hypotheses," Journal of the Royal Statistical Society, Ser. B, 24, 406-424.

FELLER, WILLIAM (1968), An Introduction to Probability Theory and Its Applications, I (3rd ed.), New York: John Wiley & Sons.

GAUTSCHI, WALTER (1979), "Algorithm 542 - Incomplete Gamma Function," ACM Transactions on Mathematical Software, 5, 466-481.

GROENEBOOM, P., OOSTERHOFF, J., and RUYMGAART, F. H. (1979), "Large Deviation Theorems for Empirical Probability Measures," Annals of Probability, 7, 553-586.

GROENEBOOM, P., and OOSTERHOFF, J. (1980), Bahadur Efficiency and Small-Sample Efficiency: A Numerical Study, Afdeling Mathematische Statistiek, 68, Mathematisch Centrum, Amsterdam.

HELLMAN, MARTIN E., and RAVIV, JOSEPH (1970), "Probability of Error, Equivocation, and the Chernoff Bound," *IEEE Transactions on Information Theory*, IT-16, 368-372.

HOGG, ROBERT V. (1974), "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory," *Journal of the American Statistical Association*, 69, 909-923.

HOGG, ROBERT Y., UTHOFF, V.A., RANDLES, R. H., and DAVENPORT, A. S. (1972), "On the Selection of the Underlying Distribution and Adaptive Estimation," *Journal of the American Statistical Association*, 67, 597-600.

KAILATH, T. (1967), "The Divergence and Bhattacharya Distance in Signal Selection," *IEEE Transactions in Communications Technology*, COM-15, 52-60.

KRAFFT, OLAF and PURI, MADAN L. (1974), "The Asymptotic Behavior of the Minimax Risk for Multiple Decision Problems," *Sankhyā*, 36, 1-12.

KRAFT, CHARLES, and VAN EEDEN, CONSTANCE (1970), "Efficient Linearized Estimates Based on Ranks," in *Nonparametric Techniques in Statistical Inference*, ed. M. L. Puri, London: Cambridge University Press.

POLICELLO, GEORGE E. II, and HETTMANSPERGER, THOMAS P. (1976), "Adaptive Robust Procedures for the One-Sample Location Problem," *Journal of the American* Statistical Association, 71, 624-633.

RENYI, A. (1969), "On Some Problems of Statistics From the Point of View of Information Theory," in *Proceedings of the Colloquium on Information Theory, Debrecen*, 343-357, also in *Selected Papers of Alfred Renyi*, Academiai Kiado, Budapest, 1976, Vol. 3, 560-576.

RUKHIN, ANDREW L. (1982), "Adaptive Procedures in Multiple Decision Problems and Hypothesis Testing," *Annals of Statistics*, 10, 1148-1162.

________________________________ (1983a), "Convergence Rates of Estimators of a Finite Parameter: How Small Can Error Probabilities Be?", Annals of Statistics, 11, 202-207.

________________________________ (1983b), "Inference About Permutation Parameter in Large Samples," Journal of Statistical Planning and Inference, 7, 000-000.

SACKS, JEROME (1975), "An Asymptotically Efficient Sequence of Estimators of a Location Parameter," Annals of Statistics, 3, 285-298.

STONE, C. J. (1975), "Adaptive Maximum Likelihood Estimators of a Location Parameter," Annals of Statistics, 3, 267-284.

STEIN, CHARLES (1956), "Efficient Nonparametric Testing and Estimation," Proceedings of Third Berkeley Symposium on Mathematical Statistics and Probability, 1, 187-196.

VAN EEDEN, CONSTANCE (1970), "Efficiency Robust Estimation of Location," Annals of Mathematical Statistics, 41, 172-181.

VAJDA, IGOR (1970), "On the Amount of Information Contained in a Sequence of Independent Observations," Kybernetika, 6, 306-323.

TABLE TITLES

Table 1

Efficiencies e_n, d_n and f_n of Procedures $\hat{\delta}_a$, $\hat{\delta}$ and δ_0 for Different Sample Sizes (Normal Case)

	$\alpha = 1$			$\alpha = 2$		
n	e_n	d_n	f_n	e_n	d_n	f_n
2	.796	.722	.925	.256	.293	.367
4	.728	.671	.772	.249	.269	.286
6	.688	.640	.708	.235	.246	.252
8	.661	.620	.671	.223	.229	.231
10	.641	.605	.646	.213	.216	.217
25	.575	.559	.575	.175	.175	.175
50	.544	.535	.544	.156	.156	.156
100	.529	.527	.532	.143	.144	.150
500	.507	.507	.508	.130	.130	.132
1000	.504	.504	.504	.128	.128	.129
∞	.500	.500	.500	.125	.125	.125

Table 2

Efficiencies e_n, d_n and f_n of Procedures $\hat{\delta}_a$, $\hat{\delta}$ and δ_0 for Different Sample Sizes (Exponential Case)

		$\alpha = 1$			$\alpha = 2$	
n	e_n	d_n	f_n	e_n	d_n	f_n
2	.5079	.4947	.5703	.3770	.3833	.4461
4	.2874	.2730	.3473	.1933	.2009	.2479
6	.2153	.2012	.2666	.1359	.1439	.1787
8	.1807	.1671	.2239	.1083	.1163	.1428
10	.1604	.1473	.1971	.0920	.0998	.1207
25	.1113	.1031	.1262	.0526	.0584	.0642
50	.0918	.0804	.0982	.0379	.0414	.0428
100	.0795	.0673	.0818	.0289	.0305	.0307
500	.0653	.0500	.0656	.0185	.0186	.0186
1000	.0629	.0470	.0629	.0167	.0167	.0167
∞	.0597	.0433	.0597	.0141	.0141	.0141

Table 3

Efficiencies e_n, d_n and f_n of Procedures $\hat{\delta}_a$, $\hat{\delta}$ and δ_0 for Different Sample Sizes (Binomial Case)

	$\alpha = 1$			$\alpha = 2$		
n	e_n	d_n	f_n	e_n	d_n	f_n
2	.4547	.4547	.4547	.3815	.3754	.3815
4	.2554	.2554	.2554	.1965	.1923	.1965
6	.1495	.1345	.1839	.1072	.0929	.1333
8	.1189	.1152	.1460	.0834	.0785	.1032
10	.0997	.1020	.1221	.0681	.0679	.0839
25	.0521	.0477	.0654	.0301	.0273	.0372
50	.0366	.0279	.0435	.0169	.0139	.0208
100	.0285	.0170	.0313	.0111	.0082	.0122
500	.0191	.0058	.0191	.0043	.0035	.0043
1000	.0171	.0038	.0171	.0031	.0027	.0031
∞	.0145	.0016	.0145	.0014	.0014	.0014