The Annals of Statistics
1978, Vol. 6, No. 3, 515-533

ASYMPTOTICALLY EFFICIENT SOLUTIONS TO THE CLASSIFICATION PROBLEM

BY LOUIS GORDON[1] AND RICHARD A. OLSHEN[2]

Alza Corporation and University of California, San Diego

We study a class of decision rules based on an adaptive partitioning of an Euclidean observation space. The class of partitions has a computationally attractive form, and the related decision rule is invariant under strictly monotone transformations of coordinate axes. We provide sufficient conditions that a sequence of decision rules be asymptotically Bayes risk efficient as sample size increases. The sufficient conditions involve no regularity assumptions on the underlying parent distributions.

1. Introduction and summary. In this paper we present asymptotic results for the nonparametric, multivariate classification problem. The rules we study are asymptotically Bayes risk efficient with no assumptions whatever regarding the underlying distributions. Our results apply to show asymptotic efficiency of variations of many successive partitioning solutions to the classification problem, for example those of Anderson (1966) and of Morgan and Sonquist (described in Sonquist (1970)). No result in such generality for this type of rule has been available before. Recently Stone (1977) also has presented completely general asymptotically efficient solutions to the classification problem. The rules to which his and our results apply are different, and naturally the proofs are quite different also.

Our work flows from ideas of Friedman (1977). He offers convincing evidence that the recursive partitioning algorithms he studies are computationally attractive for large data sets and are well suited for dealing with highly nonlinear discrimination problems. Our interest in the present paper grew out of our attempts to demonstrate the asymptotic efficiency of Friedman's algorithms. First we review some related previous work of others.

Fix and Hodges (1951) showed in effect that when both parent distributions are Lebesgue absolutely continuous with almost everywhere continuous densities, then kth nearest neighbor classification rules are asymptotically Bayes risk efficient as k and the sample size increase indefinitely in a prescribed way. The Fix–Hodges result was the first in a long line of results which in effect show that when the parent distributions are Lebesgue absolutely continuous, then

Received November 1975; revised October 1976.

[1] Now with the Energy Information Administration, U.S. Department of Energy.

[2] This research was supported in part by NSF Grants GP-37769 to the University of Michigan and MCS 76-08314 to the University of California, San Diego, and in part by U.S. ERDA at the Stanford Linear Accelerator Center. Most of this work was completed while Olshen was on leave from the University of Michigan at Stanford University and the Stanford Linear Accelerator Center.

AMS 1970 *subject classifications.* Primary 62H30; Secondary 62G20, 62P99.

Key words and phrases. Nonparametric discrimination, nonparametric classification.

consistent estimates of densities provide asymptotically efficient solutions to the classification problem. For example, Van Ryzin (1966) showed that under various regularity conditions both kernel and orthogonal expansion estimates of probability densities also give rise to asymptotically Bayes risk efficient classification procedures; moreover, he obtained rates of convergence. Pelto (1969) studied adaptive asymptotically efficient kth nearest neighbor rules from a decision theoretic viewpoint.

In work which has become famous in pattern recognition, Cover and Hart (1967) showed that (under regularity conditions) simple nearest neighbor classification asymptotically achieves at most twice the Bayes risk of the Bayes classification rule for 0-1 loss. Stone's cited work contains far reaching extensions of the Fix–Hodges, Pelto, and Cover–Hart results. His work explicitly and our work implicitly also contain substantial general results on nonlinear regression. Stone's paper includes an extensive and useful bibliography on related matters.

The rules discussed in the last paragraph lack an invariance enjoyed by the classification problem: invariance under all strictly monotone transformations of the coordinate axes. The maximal invariants are vectors of coordinate-wise ordered population labels of the training sets. This invariance was discussed by Anderson (1966, pages 22 to 24) in his work on statistically equivalent blocks. Anderson's rules are invariant and do partition the "feature space" by hyperplanes as do ours, but their partitioning is not based on the data. Our theorems show that Anderson's rules can be universally asymptotically efficient.

Morgan and Sonquist have developed algorithms for nonlinear regression which specialize to the classification problem when the dependent variable assumes only two values (see Sonquist (1970)). They recursively partition boxes (rectangular parallelopipeds with sides parallel to the coordinate axes) so as to effect the greatest reduction in the variance of the dependent indicator variable. Their rules for classification are invariant in the sense described. From our theorems it follows that suitable extensions of their procedures are asymptotically Bayes risk efficient.

Our work can be viewed as a continuation of the approach to classification through consistent estimation of a pair of densities, but our densities are with respect to a perfectly general dominating measure which need not be known to the experimenter. In order that our mathematical results be pertinent to data from the medical and social sciences, they must apply to distributions which are partly discrete and partly continuous. In fact, they apply to data sampled from arbitrary mixtures of purely discrete, absolutely continuous, and continous singular distributions. Some of the technical difficulties encountered below are the price paid for the full generality of the main result.

It is useful here to introduce some notation. Throughout the paper we shall use F and G to denote the distribution functions of our two populations. We assume that in classifying a "test point" we suffer losses l_F and l_G for misclassifying when, respectively, F and G apply. We lose nothing for a correct classification. Moreover, by π_F and π_G we mean the respective prior probabilities that a test point is from F and from G. Throughout the body of the paper we suppose that $\pi_F l_F = \pi_G l_G$. (Extensions to the general case are easy and are mentioned briefly in Section 5.) Statements of both Bayes classification rules and our proofs in subsequent sections are facilitated by the introduction of $H = \pi_F F + \pi_G G$, the unconditional distribution function of the data.

(1.01) A Bayes rule for classifying an observation x assigns x to F whenever $(dF/dH)(x)$ exceeds 1, and to G otherwise, where dF/dH is some version of the derivative. We say a procedure is *asymptotically Bayes risk efficient* if, as the sizes of both training samples become arbitrarily large, the Bayes risk in classifying one additional observation approaches the Bayes risk of a Bayes procedure based on complete knowledge of the underlying distributions F and G.

We study classification rules associated with partitions of a feature space into boxes. Each partition of interest is a result of successive refinements of ancestor partitions. The classification rule is by relative majority vote within each box of the final partition. In essence, this relative majority vote provides an estimate of the ratio of densities within the box in question. We introduce the notion of a p-quantile cut with reference to the successive refinement of partitions to enable us to obtain our principal contribution: if increasingly many p-quantile cuts are performed relative to each coordinate axis as the training sample sizes grow large, then the decision rule obtained from the final partition is asymptotically Bayes risk efficient, regardless of the parent distributions or intervening cuts. Hence, in proving asymptotic efficiency, we may force p-quantile cuts, or we may restrict our attention to parent distributions which guarantee that such cuts will be made. Application of the result to Anderson's and Morgan and Sonquist's classification schemes illustrates the former alternative.

The classification rules of this paper all involve a preprocessing of training samples prior to any classification of data of unknown origin. A preprocessing approach to a classification rule is foreshadowed, though not fully exploited, in the work of Belson (1959) and Hills (1967). Both authors emphasize the classification problem based on data with dichotomous features, and so typically they do not have to deal with the question of how to cut again on a previously sectioned axis. Belson remarks that he has implemented his procedure by means of a card sorter.

The rules we introduce informally in (1.02) and study in detail in Section 4 are all asymptotically Bayes risk efficient for any pair of parent distributions. In Section 2 we give precise conditions that any successive partitioning rule be efficient in the described sense. The proofs there use the (martingale) argument that the conditional expectations of an integrable function relative to an increasing sequence of sigma-fields tend almost surely to the conditional expectation relative to the limiting sigma-field. Our sigma-fields are not nested, however, and we obtain only convergence in probability at the crucial juncture. From one point of view, the arguments of Section 2 resemble those of, for example, Jessen, Marcinkiewicz and Zygmund (1935) on the almost-everywhere differentiability of multiple integrals where, in effect, the sigma-fields need not be nested nor, for bounded functions, need the diameters of the boxes of the partitions be controlled. However, the cited results pertain only to differentiation with respect to Lebesgue measure. Recent related work on strong maximal theorems (for example, Cordoba and Fefferman (1975)) do not seem to help either.

In Section 3 we apply the conditions of Section 2 to a set of criteria which may be enforced upon any successive partitioning rule. Our basic mathematical tools here are two-fold: first, a large deviation result of Kiefer (1961) concerning deviations of multivariate empirical distribution functions and their expectations and, second, an optional stopping inequality on the expected marginal probability content of certain projections determined by the boxes of a partition. (The Kiefer result is also used in Section 2.) The connection between the mathematics outlined so far and the actual Bayes risk of a rule is made by an argument in Section 2. There only the dominated convergence theorem and the splitting of an integral are needed.

Section 4 contains a theorem whose implications are discussed informally in (1.02) and also elaborations on the algorithms of (1.02). Section 5 includes extensions of the results of the first four sections.

(1.02) *Nonatomic marginal distributions.* In the remainder of this section we present applications of our theorem in a highly specialized context. Suppose that we are given $X_1, \cdots, X_{\sharp_F}$ and $Y_1, \cdots, Y_{\sharp_G}$, two training samples consisting of independent observations from respective multivariate cumulative distribution functions F and G. We assume here that the one-dimensional marginal cumulative distribution functions for each coordinate axis are nonatomic and that $\sharp_F = \sharp_G = n$. The first assumption obviates the need for the technical considerations which preoccupy us in the remaining sections, and the second makes for simpler prose.

We defer formal definition of box, and refer to any rectangular parallelopiped with sides parallel to the coordinate axis as a box. Our decision rules may be ambiguously defined on the boundary of a finite number of boxes, but this is necessarily a set of $F + G$ measure 0.

We associate with any partition $Q^{(n)}$ of feature space the decision rule d_n: classify any new observation having value x as drawn from F or G depending on the sign of $\hat{F}_n(B) - \hat{G}_n(B)$. The box in $Q^{(n)}$ containing x is denoted by B, while $\hat{F}_n$ and $\hat{G}_n$ are the empirical probability measures associated with the X and Y training samples. This rule corresponds to the choice of prior probabilities $\pi_F = \pi_G = \frac{1}{2}$.

The algorithm (1.03) introduces the successive partitioning rules to which our results apply, though as it stands we cannot prove that the associated decision rule is asymptotically Bayes risk efficient.

(1.03) We simplify Friedman's (1977) algorithm to the following operations:

(a) Given a box check whether it contains fewer than $n^{\frac{5}{8}}$ points of the combined training sample. If it does, refine this box no further. Classify any new observation lying in this box as F or G by the cited criterion. If the box has more than $n^{\frac{5}{8}}$ points, continue.

(b) Compute the d coordinatewise conditional empirical marginal cdf's of the training samples restricted to the box in question. Evaluate the Kolmogorov–Smirnov distance between each of the d pairs of marginal distributions, and cut the original box into two boxes at a sample point maximizing the Kolmogorov–Smirnov distance with a cut perpendicular to that axis yielding maximum separation over all axes. Repeat (a) for each of the two resulting boxes.

(By Kolmogorov–Smirnov distance, we mean the distance between two cumulative distribution functions with regard to the uniform metric, i.e., the two-sided Kolmogorov–Smirnov statistic.)

Motivation for maximizing the Kolmogorov–Smirnov criterion when $\pi_F l_F = \pi_G l_G$ flows from an observation of Stoller (1954). For suppose that F and G are *known* univariate distributions, and that we are to cut the real line at a point, assigning the left region to one distribution and the right region to the other so as to minimize the Bayes risk. Stoller observed that a solution consists of choosing a point which maximizes the Kolmogorov–Smirnov distance between the two distributions and making the obvious assignment.

Note that the algorithm (1.03) gives rise to a sequence of successively refined partitions, and that, in the course of refinement, a parent box is cut into at most two daughter boxes at each implementation of step b.

The following example illustrates a potential problem with successive partitioning algorithms (such as the one given) which are completely determined by the coordinatewise marginal distributions of the box which next is to be cut.

(1.04) Suppose the "feature space" is the unit cube U in R^3. Let F and G have densities f and g, where

$$\begin{aligned} f(x) &= 1 \\ g(x) &= (x_3 + \tfrac{1}{2})(\cos[2\pi(x_1 + x_2)] + 1)\,. \end{aligned}$$

Both distributions have uniform marginal distributions on the first two coordinates; for both, the third coordinate is independent of the first two. Therefore, part (a) of the Friedman algorithm applied to the true cumulative distribution functions would always cut orthogonal to the x_3 axis (in fact at the dyadic rationals) and never on the x_1 or x_2 axes. The resulting decision rule would not ultimately arbitrarily well approximate the Bayes rule. Indeed, the conditional marginal cdf's would become increasingly similar as the number of cuts increased. With training samples one might hope that the algorithm would homogenize the sample variation in the x_3 dimension within the given box to such a degree that random fluctuations alone would eventually prevail to force a cut in one of the other directions. Once one such cut were made, the apparent uniformity of marginals within the cut box would disappear, and further cuts in the x_1 and x_2 directions would be forced. Rather than rely on such thin hopes, we here introduce an elementary notion of p-quantile cuts in the context of parent distributions with nonatomic marginals. A substantial generalization is necessitated in Section 3 by consideration of general parent distributions.

We say an ith p-quantile cut has been achieved if a box B is refined by a cut perpendicular to coordinate axis i so that at most p of the original contents of B lie in either of the two daughter boxes obtained. Note that of necessity p is at least $\frac{1}{2}$. In practice, p is close to 1. Our theorem (4.01) below shows that if, in arriving at each box belonging to the final partition on which the decision rule is based, an algorithm ultimately performs arbitrarily many quantile cuts relative to each coordinate axis, then that rule is asymptotically Bayes risk efficient.

In the following, we reformulate the algorithm (1.03) so that our theorem (4.01) applies to it. The modifications force every coordinate to be used occasionally and preclude refining cuts from being made too near the face of a box.

(1.05) The following algorithm produces a decision rule which is asymptotically Bayes risk efficient for parent distributions with nonatomic coordinatewise marginals and prior probabilities $\pi_F = \pi_G = \frac{1}{2}$.

(a) Given a box, check whether it contains fewer than $n^{\frac{5}{8}}$ points in the combined training sample. If it does, refine this box no further. Classify any new observation lying in this box as an F or G by the cited criterion. If the box has more than $n^{\frac{5}{8}}$ points, continue.

(b) The user chooses a large integer M. If any axis i has not been cut in the preceding M refinements of boxes ancestor to the box in question, cut that box at the median ith coordinate of the combined X's and Y's lying in that box, and return to step (a). Otherwise, employ the Kolmogorov–Smirnov criterion as in algorithm (1.03b) to determine on which axis i the refining cut is to be made. Cut perpendicular to axis i at the middle one of the following three values: the $(1 - p)$th quantile of the i-coordinates of the points belonging to the box, the pth quantile of the i-coordinates of the points belonging to the box, a point at which the Kolmogorov–Smirnov statistic is attained.

(1.06) The Morgan Sonquist AID algorithms (with sorting of classes suppressed) described in Sonquist (1971, especially pages 221–230) may be used to implement rules for the classification problem it the dependent variable is treated as a 0-1 indicator. The partition ultimately produced is used in the usual classification scheme by relative majority rule.

Since the dependent variable is two-valued, the Morgan–Sonquist criterion for the decision to cut box B into boxes B_1 and B_2 perpendicular to axis i reduces to a quantity proportional to the uncorrected chi-square statistic for a 2×2 contingency table. It is then straightforward to show that both daughter boxes must have at least $v/2$ of the contents of the parent box whenever a Morgan–Sonquist cut is mandated. Here v, a parameter specified by the user of the AID algorithm, represents a minimal fraction of total variance explained. Therefore all Morgan–Sonquist cuts are of necessity $(1 - v/2)$-quantile cuts. Note that since each Morgan–Sonquist cut is predicated on an absolute reduction of the within-partition sum of squares, no more than $1/v$ Morgan–Sonquist cuts can be made in toto on any given axis.

(1.07) A prescription for minor modifications to the AID rules which render them asymptotically Bayes risk efficient for parent distributions with nonatomic marginals now follows:

(a) Given a box, check whether it contains fewer than $n^{\frac{5}{8}}$ points of the combined training sample. If it does, refine this box no further. Classify any new observation lying in this box as F or G by the cited criterion. If the box has more than $n^{\frac{5}{8}}$ points, continue.

(b) Choose $v(n)$ such that $v(n) \to 0$ and $1/v(n) = o[\log(n)]$. Refine partitions according to the AID criterion whenever possible. When no Morgan–Sonquist cut can be made, cut the least recently cut axis at the median value of the coordinates of training sample points in the box.

Notice that if the majority votes in all daughter boxes of a given box produce the same winner, then, with regard to the classification rule, the partitioning of the given box was gratuitous. Thus, we believe that the Morgan–Sonquist algorithm alone suffices to produce asymptotically Bayes risk efficient procedures for many problems.

2. Consistent estimation of densities. We here begin to face the technical problems which result from allowing the parent distributions F and G to be arbitrary. For example, suppose the data are two-dimensional, and that we are required to perform a p-quantile cut on a box, perpendicular to the horizontal axis. It may happen that all the data of the box lie on a vertical line running through the box. We could cut the box in the vertical direction by partitioning the line on which the data lie so that part of the line is assigned the left daughter box, and the remainder is assigned to the right daughter box. The reader should see from this example that care is required in order to define exactly the objects upon which our algorithms operate. It is to that end that the following definitions are made.

(2.01) A *basic box* in $\mathbb{R}^d$ is a triple (a, b, r) of vectors with a, b in $\mathbb{R}^d$ and

$r_i \in \{0, 1, 2, 3\}$, for all $i \leqq d$. We identify the basic box with the subset:

(2.02) $$B = \bigcap_{\{r_i=0\}} \{x \in \mathbb{R}^d \mid a_i < x_i < b_i\} \cap \bigcap_{\{r_i=1\}} \{x \in \mathbb{R}^d \mid a_i \leqq x_i < b_i\}$$ $$\cap \bigcap_{\{r_i=2\}} \{x \in \mathbb{R}^d \mid a_i < x_i \leqq b_i\} \cap \bigcap_{\{r_i=3\}} \{x \mid \mathbb{R}^d \;\; a_i \leqq x_i \leqq b_i\} .$$

(2.03) A *vertex* of (a, b, r) is a vector v in $\mathbb{R}^d$ with $v_i = a_i$ or $v_i = b_i$ for all dimensions i. The vertex b is called the *upper vertex* and a is the *lower vertex* of the basic box (a, b, r).

(2.04) The *dimension* of (a, b, r) is the cardinality of $\{i \mid a_i < b_i\}$.

(2.05) A *subside* of a basic box $B = (a, b, r)$ is a basic box $B' = (a', b', r')$ for which:

(i) there exists a dimension i_0 for which $a_{i_0} < b_{i_0}$ and
$$a'_{i_0} = b'_{i_0} = a_{i_0} \qquad \text{or} \qquad a'_{i_0} = b'_{i_0} = b_{i_0} ;$$

(ii) $a_i \leqq a_i' \leqq b_i' \leqq b_i$ for all dimensions i;

(iii) at least one vertex of B' is a vertex of B.

(2.06) A box is a union of a basic box and a set of subsides such that, for each dimension i, for at most one subside is $a_i' = b_i' = b_i$ and for at most one subside is $a_i' = b_i' = a_i$. Note that by definition, any box may be considered a union of at most $2d + 1$ basic boxes, and that all boxes are convex.

(2.07) The closure of any box is a basic box. Given any box, we therefore refer to the upper and lower vertices of its closure as the upper and lower vertices of the original box. In (1.02) and in Sections 4 and 5 we explicitly describe algorithms that provide classification rules invariant under the class of coordinate by coordinate strictly monotone transformation of observations (training samples). Therefore, without loss of generality, we assume that $F(U) = G(U) = 1$, where U is the unit cube $[0, 1]^d$.

(2.08) For the sake of a uniform notation, we reserve Q as a generic symbol for a partition of U, all of whose component subsets are boxes B. For $x \in U$, we denote by $B(x)$ the unique box in Q containing x. The upper and lower vertices of B are denoted $b(B)$ and $a(B)$. We occasionally suppress the explicitly stated dependence of a and b on B. If a sequence of partitions is discussed, the index is superscripted, and the same indexing is carried to boxes. For example, $Q^{(n)}$ denotes an element in a sequence of partitions and $B^{(n)}(x)$ is that box in $Q^{(n)}$ containing x.

(2.09) Suppose that we are confronted with a sequence of similar classification problems and training samples on U, and that we apply a successive partitioning algorithm to each. We then obtain a triangular array $Q^{(n,j)}$ of partitions of U such that: (i) for the nth problem, $Q^{(n,j+1)}$ is a refinement of $Q^{(n,j)}$, and (ii) each partition is composed of a set of boxes. There is, however, no refinement relation between $Q^{(n,j)}$ and $Q^{(n',j')}$ for $n \neq n'$. Proposition (2.10) is addressed to that issue.

(2.10) PROPOSITION. *Let F, G be two arbitrary multivariate cumulative distribution functions on U the unit cube in $\mathbb{R}^d$. Assume $H = \phi F + (1 - \phi)G$ for some ϕ in* (0, 1). *Let $Q^{(n)}$ be a sequence of partitions of U satisfying*:

(2.11) $$\textit{for each} \quad n\,, \quad Q^{(n)} \quad \textit{is a finite set of boxes};$$

(2.12) *there exists K with $H(K) = 1$ such that for each $x \in K$ and each dimension i,*

$$\sup\{y_i \mid y \in B^{(n)}(x) \cap K\} - \inf\{y_i \mid y \in B^{(n)}(x) \cap K\} \to 0\,;$$

(2.13) $$H\{x \mid H(B^{(n)}(x)) > 0\} \to 1\,.$$

It follows that for any $\varepsilon > 0$,

(2.14) $$H\{x \mid |F(B^{(n)}(x))/H(B^{(n)}(x)) - dF/dH(x)| > \varepsilon\} \to 0\,.$$

PROOF. We show that any sequence satisfying (2.11), (2.12), (2.13) has a subsequence satisfying (2.14). Let $\varepsilon > 0$ be arbitrary. H has only countably many atoms, so we may proceed by the Cantor diagonal method to obtain a subsequence $\hat{Q}^{(n)}$ such that for each direction i and atom x, eventually either

(2.15) $$x_i = \hat{b}_i(\hat{B}^{(n)}(x)) \qquad \text{for all large} \quad n > N(x)$$

or

(2.16) $$x_i = \hat{a}_i(\hat{B}^{(n)}(x)) \qquad \text{for all large} \quad n > N(x)$$

or

(2.17) $$\hat{a}_i(\hat{B}^{(n)}(x)) < x_i < \hat{b}_i(\hat{B}^{(n)}(x)) \qquad \text{for all} \quad n > N(x)\,.$$

We now construct a monotone sequence of subsets $A_1 \supset A_2 \supset \cdots$ and an increasing sequence of integers $n_1 < n_2 < \cdots$ for which

(2.18) $$H(A_j) > 1 - \varepsilon$$

(2.19) $$x \in A_{j+1} \qquad \text{implies} \quad \hat{B}^{(n_j)}(x) \supset \hat{B}^{(n_{j+1})}(x)$$

(2.20) $$x \in A_j \qquad \text{implies} \quad K \cap \hat{B}^{(n_j)}(x) \subset A_j\,.$$

Choose n_1 such that (2.13) is true for all $n > n_1$ on a set of mass at least $1 - (\varepsilon/2)$. Let $A_1 = K$. Given n_j and A_j, use (2.15–2.17) on the atoms and dominated convergence elsewhere to obtain an n_{j+1} with $H\{x \mid K \cap \hat{B}^{(n_{j+1})}(x) \subset K \cap \hat{B}^{(n_j)}(x)\} > 1 - \varepsilon$. Now take $A_{j+1} = \{x \in K \mid K \cap \hat{B}^{(n_{j+1})}(x) \subset K \cap \hat{B}^{(n_j)}(x)\}$.

Now denote by $\tilde{S}_j$ the σ-algebra generated by $\{\hat{B}^{(n_k)}(x) \mid x \in K, k \leqq j\}$. Let $\tilde{S}$ be the σ-algebra generated by the collection of $\tilde{S}_j$. The assumption (2.12) implies that on K, $\tilde{S}$ coincides with the Borel σ-algebra restricted to K. By a well known corollary to the martingale convergence theorem,

(2.21) $$dF/dH \mid \tilde{S}_j \to dF/dH \mid \tilde{S} = dF/dH \quad (H\text{-almost surely})\,.$$

Further, (2.13) implies that on $\cap A_j$, $dF/dH \mid \tilde{S}_j$ eventually H-almost surely equals $F(\hat{B}^{(n_j)}(x))/H(\hat{B}^{(n_j)}(x))$, and so $F(B^{(n)})/H(B^{(n)}) - dF/dH$ tends to 0 H-almost everywhere for the subsequence n_j.

(2.22) Denote by $X_1, \cdots$ and $Y_1, \cdots$ two independent sequences of independent random vectors having respective distribution functions F, G with support in U. Let $H = \phi F + (1 - \phi)G$ where ϕ is a constant in $(0, 1)$. Let $\#_F(n)$ and $\#_G(n)$ be sequences of integers with $\#_F(n) + \#_G(n) = n$. We interpret $\#_F$ and $\#_G$ as numbers of observations available at time n. Denote by $\hat{F}_n$ the empirical cumulative distribution function of $X_1, \cdots, X_{\#_F}$, and similarly for $\hat{G}_n$. Let $\hat{H}_n = \phi\hat{F}_n + (1 - \phi)\hat{G}_n$. Given a set A in U denote by $\#_n(A)$ the number of observations of either type in A:

$$\#_n(A) = \#_F(n)\hat{F}_n(A) + \#_G(n)\hat{G}_n(A)\,. \tag{2.23}$$

(2.24) PROPOSITION. *Let $Q^{(n)}$ be as in* (2.10). *If in addition, eventually*

$$\#_F(n)/n \in (\theta, 1 - \theta) \qquad \textit{for some fixed, positive} \quad \theta\,; \tag{2.25}$$

$$H\{x \mid H(B^{(n)}(x)) > 0\} \to 1\,; \tag{2.26}$$

$$\textit{there exists} \quad k(n) \quad \textit{with} \quad k(n)/n^{\frac{1}{2}} \to \infty \quad \textit{and} \tag{2.27}$$
$$H\{x \mid \#_n(B^{(n)}(x)) > k(n)\} \to 1 \quad \textit{in probability;} \qquad \textit{then}$$

$$H\{x \mid |\hat{F}_n(B^{(n)}(x))/\hat{H}_n(B^{(n)}(x)) - F(B^{(n)}(x))/H(B^{(n)}(x))| > \varepsilon\} \to 0 \tag{2.28}$$

in probability for all positive ε.

PROOF. According to a theorem of Kiefer (1961, Theorem 1-m), with arbitrarily high probability

$$|\hat{F}_n(B^{(n)}) - F(B^{(n)})| < h_F(n) \tag{2.29}$$

$$|G_n(B^{(n)}) - G(B^{(n)})| < h_G(n) \tag{2.30}$$

simultaneously for all $B^{(n)} \in Q^{(n)}$ where $n(h_F(n) + h_G(n))/k(n) \to 0$. It therefore suffices to observe that

$$\begin{aligned} &|\hat{F}_n(B^{(n)}(x))/\hat{H}(B^{(n)}(x)) - F(B^{(n)}(x))/H(B^{(n)}(x))| \\ &\qquad < K(\theta)(nh_F(n)/k(n) + [F(B^{(n)}(x))/H(B^{(n)}(x))][nh_G(n)/k(n)]) \end{aligned} \tag{2.31}$$

eventually with arbitrarily high probability for sufficiently large n on a subset of U having arbitrarily large H-measure. This completes the proof of (2.24).

Combination of the preceding two results proves

(2.32) PROPOSITION. *If $Q^{(n)}$ is a sequence of partitions satisfying* (2.11), (2.12), (2.13) *of Proposition* (2.10) *and also satisfying* (2.25), (2.26), (2.27) *of Proposition* (2.24), *then*

$$H\{x \mid |\hat{F}_n(B^{(n)}(x))/\hat{H}_n(B^{(n)}(x)) - dF/dH(x)| > \varepsilon\} \to 0 \tag{2.33}$$

in probability for all positive ε.

The proceeding result allows us to estimate consistently the Bayes rule (1.01).

(2.34) *Notation.* We identify each classification rule with its associated indicator function $T(x)$, where $T(x) = 1$ if the decision rule assigns an observation having value x to population F, and where $T(x) = 0$ if the observation is assigned to G. We denote the Bayes risk of procedure T by $R(t)$ where

$$R(t) = E \int \pi_F l_F (1 - T) \frac{dF}{dH} + \pi_G l_G T \frac{dG}{dH} \, dH \, . \tag{2.35}$$

The expectation reflects the data dependent nature of the decision rule T.

(2.36) THEOREM. *Let $Q^{(n)}$ be a sequence of partitions derived from training samples $X_1, \cdots, X_{\sharp_F(n)}$ and $Y_1, \cdots, Y_{\sharp_G(n)}$ satisfying* (2.11), (2.12), (2.13), (2.25), (2.26), *and* (2.27). *Let $T(x) = I_{\{dF/dH(x) > 1\}}$ be the Bayes classification rule* (1.01). *Let*

$$T^{(n)}(x) = I_{\{\hat{F}^{(n)}(B^{(n)}(x))/\hat{H}^{(n)}(B^{(n)}(x)) > 1\}} \tag{2.37}$$

be the approximating rule based on $Q^{(n)}$ suggested by (2.33). *Conclude that $R(T^{(n)})$ tends to $R(T)$ as n tends to ∞.*

PROOF. Recall $\pi_F l_F = \pi_G l_G$ by assumption. Let

$$R^{(n)} = \pi_F l_F + \int \pi_F l_F \left[\frac{dG}{dH} - \frac{dF}{dH} \right] T^{(n)} \, dH \, . \tag{2.38}$$

Note that $R^{(n)}$ is bounded. Since

$$\left| \frac{dF}{dH} - 1 \right| < \varepsilon \quad \text{implies} \quad \left| \frac{dG}{dH} - 1 \right| < \frac{\pi_F}{\pi_G} \varepsilon \, , \tag{2.39}$$

(2.32) implies that $R^{(n)}$ tends in probability to $\pi_F l_F + \int \pi_F l_F (dG/dH - dF/dH) \, T \, dH$. This last assertion is made clear by breaking (2.38) into three integrals: one over $\{dF/dH > 1 + \varepsilon\}$, one over $\{dF/dH < 1 - \varepsilon\}$, and one over $\{|dF/dH - 1| \leq \varepsilon\}$. The dominated convergence theorem is applicable to the first two integrals, while (2.39) is applicable to the third.

We next provide a criterion alternate to (2.12). This new criterion is easier to verify.

3. Quantile cuts. The notion of p-quantile cut was motivated and introduced in (1.02). When marginal distributions are continuous, it is intuitively clear what such a partition of a box ought to be. However, the example discussed at the outset of Section 2 suggests that more care is required when the distributions F and G are arbitrary. In this section we give a careful definition of p-quantile cut of a box in a specified direction, and explain how to perform one. We expect that typically in practice the algorithms will automatically perform these cuts with minimal external intervention. Also, in Section 2, the condition (2.12) of (2.10) is not stated in a manner that appears to depend only on the order statistics of the coordinates of the combined F and G observations in the training sample. In what follows we define the norm of a partition and show that convergence of that norm to 0 over a sequence of partitions ensures that the condition (2.12) of the basic Proposition (2.32) holds.

(3.01) Denote by H a Borel probability measure on U. Let $H_i(x) = H\{y \mid y_i \leqq x_i\}$. H_i is the ith marginal of H, save that its argument has been altered for notational convenience.

(3.02) Let B_1, B_2, B_3 be boxes contained in U. Let $p \in [\frac{1}{2}, 1)$ be a constant, and i index a coordinate in $\mathbb{R}^d$. We say that B_2 and B_3 comprise an *ith p-quantile cut of B_1 relative to H* if any of (3.03), (3.04) or (3.05) obtain.

(3.03) (a) through (e) hold.

(a) $B_1 = B_2 \cup B_3$, a disjoint union;
(b) $a_j(B_1) = a_j(B_2) = a_j(B_3)$ for all coordinate indices $j \neq i$;
(c) $b_j(B_1) = b_j(B_2) = b_j(B_3)$ for all coordinate indices $j \neq i$;
(d) $r_j(B_1) = r_j(B_2) = r_j(B_3)$ for all coordinate indices $j \neq i$;
(e) $H(B_j) \leqq pH(B_1)$ for $j = 2, 3$.

(3.04) (a) through (d) of (3.03) hold, and

(e′) $H(B_3) \leqq H(B_2)$ and $H_i(b(B_3)) - H_i(a(B_3)) \geqq H_i(b(B_2)) - H_i(a(B_2))$.

(3.05) (a) $B_1 = B_2 \cup B_3$, either a disjoint union or $B_2 = B_3$;
(b) $a_i(B_1) = b_i(B_1)$.

Note that (3.03) implies we may renumber B_2 and B_3 to obtain $a_i(B_2) \leqq b_i(B_2) = a_i(B_3) \leqq b_i(B_3)$.

Note also that one can always perform an ith p-quantile cut on a given box B. If one can satisfy neither (3.03) nor (3.05), then $B = B_1 \cup B_2 \cup B_3$ a disjoint union of boxes which satisfy (3.03b—d). Further, we may take $b_i(B_2) = a_i(B_2)$ a constant. If $H_i(b(B_1)) - H_i(a(B_1)) \leqq H_i(b(B_3)) - H_i(a(B_3))$ then $B_1 \cup B_2$ and B_3 comprise a p-quantile cut of B satisfying (3.04). Otherwise, B_1 and $B_2 \cup B_3$ comprise a p-quantile cut satisfying (3.04).

(3.06) We next define the ith norm relative to H of a partition Q composed of boxes:

$$\|Q\|_i^H = \sum \{[H_i(b(B)) - H_i(a(B))]H(B) \mid B \in Q\},$$

where i is a specified dimension in $\mathbb{R}^d$. Note that $\|Q\|_i^H$ is the expected value of $H_i(b(X)) - H_i(a(X))$ where X is distributed as H. As a result, if $Q^{(2)}$ refines $Q^{(1)}$, then $\|Q^{(2)}\|_i^H \leqq \|Q^{(1)}\|_i^H$. The impact of the next proposition is to provide a condition implying (2.12) which can be verified for the class of algorithms presented in (1.02) and in Sections 4 and 5. The first hypothesis assures us that the algorithm generating the partition frequently cuts boxes at about their centers on each of the axes. The proof of (3.07) is immediate from the combination of Lemmas (3.09), (3.11), and (3.13), whose proofs complete this section.

(3.07) PROPOSITION. *Let $Q^{(n)}$ be a sequence of partitions of U composed of boxes. Let p in $[\frac{1}{2}, 1)$ be given. Let θ in $(0, 1)$, F, G, H, $X_1, \cdots, X_{\sharp F(n)}$ and $Y_1, \cdots, Y_{\sharp G(n)}$ be given as in Section 2. In addition, assume*:

(a) *for each $Q^{(n)}$ there exists $Q^{(n,1)}, \cdots, Q^{(n,n)} = Q^{(n)}$ a finite sequence of partitions of U composed of boxes such that $Q^{(n,j+1)}$ is a refinement of $Q^{(n,j)}$*;

(b) *for each direction i, and fixed integer m, $\hat{H}_n\{x \mid B^{(n,j+1)}(a(B^{(n,j)}(x)))$ and $B^{(n,j+1)}(b(B^{(n,j)}(x)))$ comprise an* ith *p-quantile cut of $B^{(n,j)}(x)$ relative to $\hat{H}_n$ for at least m distinct j's$\} \to 1$*;

(c) (*without loss of generality for our applications*) *that each H_i is of the form $H_i(t) = P(D_i(z) \leqq t)$, where z is distributed as D_i, an arbitrary univariate cumulative distribution function*;

(d) *$k(n)$ has properties* (2.27).

Conclude that

(3.08) $$H\{x \mid b_i(B^{(n)}(x)) - a_i(B^{(n)}(x)) > \varepsilon\} \to 0$$

in probability for each positive ε.

The following three lemmas all subsume the hypotheses and notation of the previous proposition. The proofs of all three follow their statement. The lemma (3.09) is the critical mathematical observation in the entire paper. The reader is encouraged to intuit its implication for Lebesgue absolutely continuous data, where (3.03e) is the important assumption.

(3.09) LEMMA. *Let $M_n = \{x \mid B^{(n,j+1)}(b(B^{(n,j)}(x)))$ and $B^{(n,j+1)}(a(B^{(n,j)}(x)))$ are* ith *p-quantile cuts of $B^{(n,j)}(x)$ with respect to $\hat{H}_n$ for at least m distinct values $j\}$. Conclude that for all n sufficiently large,*

(3.10) $$\|Q^{(n)}\|_i^H < p^m H(M_n) + (1-p)^{-1} H(M_n^c) + o_p(1) ,$$

where $o_p(1) = (n/k(n))n^{-\frac{1}{2}} + s_n + O_p(n^{-\frac{1}{2}})$, and s_n is a remainder term contributed by the aggregate of those boxes containing fewer than $k(n)$ observations.

(3.11) LEMMA. *Let Z be the univariate random variable with cumulative distribution function D. If $\alpha < \beta$ are elements of* $[0, 1]$, *β in the support of the measure induced by $D(Z)$, and $P\{D(Z) \in (\alpha, \beta)\} = 0$, then*

(3.12) $$P\{D(Z) = \beta\} \geqq \beta - \alpha .$$

(3.13) LEMMA. *If $\|Q^{(n)}\|_i^H \to 0$ then there exists K with $H(K) = 1$ for which*

(3.14) $$H\{x \mid \text{diameter } (K \cap B^{(n)}(x)) \to 0\} \to 1$$

and

(3.15) *if x_i is an atom of the* ith *marginal of H, then*

$$I_{\{a_i(B^{(n)}(x))=x_i\}} \to 1 .$$

PROOF OF (3.09). Eventually an arbitrarily large proportion of observations are in boxes of $\hat{H}_n$ content greater than $\theta k(n)/n$. Hence, there are at most $n/\theta k(n)$ such boxes in $Q^{(n)}$, and, from Kiefer (1961) the difference between the $\hat{H}_n$ content and the H content of each is $O_p(n^{-\frac{1}{2}})$, so that

(3.16) $$\|Q^{(n)}\|_i^{\hat{H}_n} - \|Q^{(n)}\|_i^H = o_p(1) .$$

Now define iteratively the "stopping times"

$$L_0(x) = 0 \tag{3.17}$$

$$L_j(x) = \min\,(n, \min\{k > L_{j-1}(x) \mid B^{(k+1)}(b(B^{(k)}(x))) \text{ and } B^{(k+1)}(a(b^{(k)}(x))) \textit{ are } i\text{th } \textit{p-quantile cuts of } B^{(k)}(x) \textit{ with respect to } \hat{H}\})\,. \tag{3.18}$$

Note that $\min\,(L_j(x), k)$ is constant on the sets in $Q^{(k)}$. Define

$$\hat{Q}^{(n,2j)} = \{B^{(\min(n, L_j(x)+1))}(x) \mid x \in U\} \tag{3.19}$$

$$\hat{Q}^{(n,2j-1)} = \{B^{L_j(x)}(x) \mid x \in U\}\,. \tag{3.20}$$

The $\hat{Q}(n, l)$ are partitions composed of boxes. L_j is measurable with respect to the σ-algebra generated by $\hat{Q}(n, 2j - 1)$. For $B \in \hat{Q}^{(n,2j-1)}$ and $B \subset \{L_{2j-1} < n\}$, define

$$V(B) = \hat{H}(B_2)/\hat{H}(B) \tag{3.21}$$

and

$$W(B) = [\hat{H}_i(b(B_2)) - \hat{H}_i(a(B_2))]/[\hat{H}_i(b(B)) - \hat{H}_i(a(B))] \tag{3.22}$$

where B_2 is the box in $\hat{Q}^{(n,2j)}$ containing $a(B)$. In case of division by 0, take V or W to be 0. By backwards induction

$$\|\hat{Q}^{(n,2j)}\|_i^{\hat{H}} \leqq \sum \{[\hat{H}_i(b(B)) - \hat{H}_i(a(B))]\hat{H}(B) \mid B \in \hat{Q}^{(n,2j)} \text{ and } L_{2j-1}(x) < n \text{ for } x \in B\} + \hat{H}\{x \mid L_{2m-1}(x) = n\} \tag{3.23}$$

$$\leqq \sum \{[\hat{H}_i(b(B)) - \hat{H}_i(a(B))]\hat{H}(B) \times [V(B)W(B) + (1 - V(B))(1 - W(B))] \mid B \in \hat{Q}^{(n,2j-1)} \text{ and } L_{2j-1}(x) < n \text{ for } x \in B\} + \hat{H}\{x \mid L_{2m-1}(x) = n\} \tag{3.24}$$

where, from the remark preceding (3.06), $V(B)$ and $W(B)$ are in $[0, 1]$, since an ith p-quantile cut was done on B in going from $\hat{Q}^{(n,L_{2j-1})}$ to $\hat{Q}^{(n,L_{2j})}$. From (3.02), (3.21) and (3.22)

$$\|\hat{Q}^{(n,2j)}\|_i^{\hat{H}} \leqq \sum \{p[\hat{H}_i(b(B)) - \hat{H}_i(a(B))] \mid B \in \hat{Q}^{(n,2j-1)} \text{ and } L_{2j-1}(x) < n \text{ for } x \in B\} + \hat{H}\{x \mid L_{2m-1}(x) = n\}\,, \tag{3.25}$$

and so, from (3.16)

$$\|\hat{Q}^{(n,2j)}\|_i^{H} \leqq p[\|\hat{Q}^{(n,2j-2)}\|_i^{H}] + H\{L_{2m-1} = n\} + o_p(1)\,. \tag{3.26}$$

PROOF OF (3.11). It is well known that

$$P\{D(Z) \leqq D(t)\} \geqq D(t) \qquad \text{for any scalar } t \tag{3.27}$$

and that

$$P\{D(Z) < \gamma\} \leqq \gamma \qquad \text{for any } \gamma \text{ in } [0, 1]. \tag{3.28}$$

From the hypotheses, $D(D^{-1}[\beta, \beta + \delta]) > 0$ for all scalar $\delta > 0$, so there exists

t with $D(t) = \beta$. Hence $P\{D(Z) \in [\alpha + \delta, \beta]\} \geqq \beta - (\gamma + \delta)$ for all positive δ sufficiently small.

PROOF OF (3.13). The Markov inequality implies that $H\{x \mid H_i(b(B^{(n)}(x))) - H_i(a(B^{(n)}(x))) > \varepsilon\}$ goes to 0 for any positive ε. Choose K with $H(K) = 1$ such that x in K implies that for each direction i, either $H_i\{x\}$ is positive or that for each d positive, both $H_i(x + de_i) - H_i(x)$ and $H_i(x) - H_i(x - de_i)$ are positive, where e_i is the unit vector having 1 in the ith coordinate and 0 elsewhere.

If $x \in K$ and $H_i\{x\} = 0$, then $H_i(b(B^{(n)}(x))) - H_i(a(B^{(n)}(x)))$ goes to 0, only if $b_i(B^{(n)}(x)) - a_i(B^{(n)}(x))$ goes to 0, by the construction of K.

If instead $H_i\{x\} > 0$, then by (3.11), $a_i(B^{(n)}(x)) = x_i$ for all sufficiently large n, establishing (3.15). Another argument by contradiction using (3.11) establishes the convergence of $b_i(B^{(n)}(x))$ to $a_i(B^{(n)}(x))$.

4. Implementation of efficient rules.

We continue by restating the results of the preceding two sections in a usable form. We then sketch four sets of rules. The first two extend (1.05) and are similar to the variable metric classification algorithms of Friedman. The next is from rules studied by Anderson. The last extends (1.07) and is motivated by the Morgan–Sonquist algorithms. All the rules studied here are asymptotically Bayes risk efficient for all parent distributions F and G.

Combination of (2.32) and (3.07) yields:

(4.01) PROPOSITION. *Let* $X_1, \cdots, X_{\sharp_F(n)}$ *and* $Y_1, \cdots, Y_{\sharp_G(n)}$ *be as in* (2.32). *Assume that an algorithm applied to the observations satisfies* (a) *through* (d) *and either* (e) *or* (e′):

(a) *for each* n, *the algorithm generates a sequence of successively refined partitions* $Q^{(n,1)}, \cdots, Q^{(n,n)} = Q^{(n)}$ *of d-dimensional space consisting of a finite set of boxes*;

(b) *there exists* $k(n)$ *such that* $k(n)/n^{\frac{1}{2}} \to \infty$, $k(n)/n \to 0$;

(c) $\hat{H}_n\{x \mid \sharp_n(B(x)) > k(n)$ *for* $B \in Q^{(n)}\} \to 1$ *in probability*;

(d) *the algorithm produces partitions* $Q^{(n)}$ *invariant under strictly monotone transformations of coordinates of the observations*;

(e) *there exist monotone nondecreasing sequences* $m_n \to \infty$ *and* $p_n \in [\frac{1}{2}, 1)$ *for which* $(p_n)^{m_n} \to 0$ *and* $\hat{H}_n\{x \mid$ *for at least* m_n *indices* j, $B^{(n,j+1)}(a)$ *and* $B^{(n,j+1)}(b)$ *comprise an* ith p_n*-quantile cut of* $B^{(n,j)}(x)$ *relative to* $\hat{H}_n$, *where a and b are the upper and lower vertices of* $B^{(n,j)}(x)\} \to 1$ *in probability.*

(e′) $\|Q^{(n)}\|_i^{\hat{H}_n} \to 0$ *in probability for all i.*

Then the decision rule (4.02) *below applied to the partition* $Q^{(n)}$ *is asymptotically Bayes risk efficient, for the prior placing mass* π_F *on* F *and* π_G *on* G.

(4.02) Classify a new observation Z to population F if

$$\hat{F}_n(B^{(n)}(Z))/\hat{H}_n(B^{(n)}(Z)) > 1,$$

where $\hat{H}_n = \pi_F \hat{F}_n + \pi_G \hat{G}_n$. Compare with (1.01).

The first two rules discussed below have similar structures. We first describe the simpler rule, indicate its limitations, and then describe a possible modification.

(4.03) RULE 1. Choose $k(n)$ satisfying (4.01b) and p in $[\frac{1}{2}, 1)$. Define a subroutine that occasionally but arbitrarily often ensures an ith p-quantile cut relative to $\hat{H}_n$.

Define iteratively partitions $Q^{(n,1)}, \cdots, Q^{(n,n)}$ according to the following scheme:

(a) Set $Q^{(n,1)} = \{\mathbb{R}^d\}$.

(b) For each box B in $Q^{(n,j)}$, if $\#_n(B) < k(n)$, then copy B into $Q^{(n,j+1)}$. Otherwise go to (c).

(c) Decide whether to demand an ith p-quantile cut, and, if so, for which of the d directions. This decision may include the need for completing the partitioning started in step (e). Also, note that the decision is sometimes made while performing step (e). If a quantile cut is necessary go to (d); otherwise, go to (e).

(d) One can always perform an ith p-quantile cut satisfying (3.03), (3.04) or (3.05). If possible select a cut with at least $k(n)/2$ observations in each of the two (possibly identical) boxes comprising the cut.

If such a cut cannot be made, then at least $\#_n(B) - k(n)$ observations in B share the identical ith coordinate. Isolate all those points sharing that coordinate in a box unto itself. Relegate the other observations in B to at most two additional boxes. These operations may be performed in two consecutive steps, consisting of two ith p-quantile cuts, and result in three boxes in $Q^{(n,j+2)}$. Note that (3.05) is trivially satisfied on the axis cut at each subsequent partition for that box containing the $\#_n(B) - k(n)$ of the observations originally in B.

(e) Compute the Kolmogorov–Smirnov distance between the marginal conditional empirical cumulative distribution functions $\hat{F}_i^{(n)}(\cdot)/\hat{F}_i^{(n)}(B)$ and $\hat{G}_i^{(n)}(\cdot)/\hat{G}_i^{(n)}(B)$. Select a direction i_0 at which the maximum Kolmogorov–Smirnov separation is attained. Attempt to cut B perpendicular to the direction i_0. Actually perform the cut only if both boxes resulting from the cut would contain at least $k(n)$ observations; otherwise, go to (c), and there perform an i_0-quantile cut relative to $\hat{H}_n$.

We explicitly verify condition (3.07d). Let $k'(n) = n^{\frac{1}{2}}k(n)^{\frac{1}{2}}$. Recall that a box with fewer than $k(n)$ points can be created only in (4.01d). Further, when such a box is created from box B, we have at worst $B = B_1 \cup B_2 \cup B_3$, a disjoint union, where $\#_n(B_1 \cup B_3) < k(n)$, $a_i(B_2) = b_i(B_2)$. One of these 3 boxes must have no fewer than $k(n)/3$ points, and additional cuts on the other axes will be made only if $\#_n(B_2) \geqq k(n)$. No further cuts can be made along coordinate i. Hence the terminal boxes can be matched so that each box with fewer than $k(n)/3$ points is associated with a box having more than $k(n)/3$ points. In addition, no terminal box with more than $k(n)/3$ points is associated with more than $2d$ terminal boxes having fewer. Let m_2 be the number of terminal boxes with fewer than $k'(n)$ points. Let m_1 be the number of terminal boxes having more than $k(n)/3$ points. Then $m_2 \leqq 2dm_1$, and $m_1k(n)/3 \leqq n$. Hence the empirical measure of all boxes having more than $k'(n)$ points is at least $1 - (6dk'(n)/k(n))$ which converges to 1 as n increases indefinitely. Hence $k'(n)$ satisfies (2.27). Note that the partitions at each stage are composed solely of simple boxes. While this algorithm is asymptotically Bayes risk efficient, it is not local in its selection of cut points, since the quantile cut criterion (3.02) depends on the marginals $\hat{H}_i$ and $\hat{F}_i$, for $i \leqq d$.

(4.04) RULE 2. It is with a view toward proposing a remedy to the nonlocality of the preceding example that the propositions were proved for boxes, rather than simple boxes. We sketch in the following paragraphs an alternative way to perform an ith p-quantile cut that always satisfies (3.03) or (3.05).

(4.05) LEMMA. *Let $A_1, \cdots, A_n$ be events in the same sample space. Suppose that $P(A_i) \geqq (1 - 2\alpha)$ and that $\alpha \leqq 1/4n$. Then $P(A_1 \cap \cdots \cap A_n) \geqq \frac{1}{2}$.*

PROOF. Boole's inequality implies that $P(A_1{}^c \cup \cdots \cup A_n{}^c) \leqq \sum_1^n 2 \cdot 1/4n = \frac{1}{2}$.

A box B and axis on which a quantile cut is to be made are assumed given. For convenience, and without loss, assume the axis to be the dth. For present purposes, by a "quantile cut" we mean a cut at, say, $x_d{}^*$ with this property. If $H^*_{n,B}$ is the conditional empirical measure in the box induced by the first n observations, then $H^*_{n,B}(B \cap \{x_d < x_d{}^*\}) \geqq p$ and $H^*_{n,B}(B \cap \{x_d > x_d{}^*\}) \geqq p$. It may happen that there is no such $x_d{}^*$. Instead there is a set $\{x_d = x_d{}^{**}\}$ which has $H^*_{n,B}$ probability at least $1 - 2p$. Order the coordinates 1 through $d - 1$ according to their importance: the number of times they have been cut previously in arriving at B. (The ordering may not be unique.) Cycle through the axes and pick the first axis which permits a cut of the given box intersected with $\{x_d = x_d{}^{**}\}$ yielding two subboxes of the original box with the stated "quantile cut property." Were there not to exist such an axis, there would, according to (4.05) and the implicit assumption that B has at least k observations, be a point z of $B \cap \{x_d = x_d{}^{**}\}$ with at least $\frac{1}{2}k$ observations. To complete our prescription, suppose now that the cited z exists. Separate z as a box unto itself in the following manner. Cycle through the axes, cutting by hyperplanes parallel to the axes, where possible reducing the dimension of the box containing z by 1 with each hyperplane. Make no cuts on any box not containing z. Thereby, at most $2d + 1$ new boxes are created. The box containing z has dimension 0. The discussion following (4.03e) shows that (2.27) is satisfied for the now completed rule.

(4.07) RULE 3. A particularly simple application of Theorem (4.01) is to Anderson's (1966, page 26) class of rules employing statistically equivalent blocks. These rules are completely specified for parent distributions F and G having nonatomic marginals. Conditions (b), (c), and (e') can be verified directly for the sequence of rules in question, since for the rules based on these

$Q^{(n)}$, $\hat{H}\{x \mid \sharp_n(B(x)) > k(n)\}$ and $\|Q^{(n)}\|_i^{\hat{H}_n}$ are strictly deterministic functions. Hence a sequence of Anderson's rules may be verified to be asymptotically Bayes risk efficient for any pair of parent distributions having nonatomic marginals. The latter restriction is convenient because Anderson's rules are not well defined as stated for parent distributions whose marginals do possess atoms.

(4.08) RULE 4. The Morgan–Sonquist AID algorithms (with sorting of classes suppressed) have been described in (1.06). The general Morgan–Sonquist algorithm uses (3.03)—(3.05), or (4.05), in place of cuts at the medians of (b) of (1.07); (4.01) applies in a straightforward manner.

5. Extensions.

Results of the previous sections have been contingent upon the assumption that $l_F \pi_F = l_G \pi_G$, where the l's and π's are losses and prior probabilities introduced in Section 1. When the cited equality does not hold, easy modifications of the previous algorithms are appropriate. Thus, in place of the rule (1.01), a more general Bayes rule assigns x to F whenever $\pi_F l_F(dF/dH)(x)$ exceeds $\pi_G l_G(dG/dH)(x)$. In place of (4.02), therefore, a generalized rule would classify Z to population F if

$$(5.01) \qquad \pi_F l_F \hat{F}_n(B^{(n)}(Z)) > \pi_G l_G \hat{G}_n(B^{(n)}(Z)) .$$

Recall that Rules 1 and 2 of the previous section both make repeated use of the Kolmogorov–Smirnov distance between within-box empirical marginal distributions and that motivation for maximizing the Kolmogorov–Smirnov criterion when $l_F \pi_F = l_G \pi_G$ flows from an observation of Stoller (1954). In case $l_F \pi_F \neq l_G \pi_G$, the solution to Stoller's problem, if one exists, consists of picking x_0 which minimizes

$$(5.02) \qquad \min\{\min_x (l_F \pi_F F(x) + l_G \pi_G(1 - G(x))),\ \min_x (l_G \pi_G G(x) + l_F \pi_F(1 - F(x)))\}$$

and again making the obvious assignment. When F and G are replaced by $\hat{F}$ and $\hat{G}$, solutions do exist, and we therefore recommend that the Kolmogorov–Smirnov criterion for partitioning be replaced by the generalization which follows from (5.02): partition on that axis and at that point for which (5.02) is minimized for $\hat{F}_i^{(n)}$ and $\hat{G}_i^{(n)}$.

Our theorems show that when the algorithms are so modified—but quantile cuts are kept—then asymptotic Bayes risk efficiency obtains as before. Moreover, it obtains also when the Kolmogorov–Smirnov criterion is retained but the criterion (5.01) is employed.

Our theorems are stated for $\mathbb{R}^d$ (or without loss $[0, 1]^d$) valued data. Because no assumptions on F or G are required for the results, they apply as well when some coordinates are manufactured from others given by the "raw data." For example, the sum or difference of two coordinates, or for that matter some nonlinear function of a given set of coordinates, may be of scientific interest a priori. These "fictitious coordinates" can be incorporated into the present rules. Both F and G will necessarily be singular with respect to Lebesgue measure, but our conclusions are unaffected. So long as the input for the algorithms is understood to consist of both "raw" and "fictitious" coordinates, the rules retain their invariance.

Acknowledgments. We thank Jerome Friedman and Charles Stone for many conversations which were invaluable in the preparation of the paper. Also, several conversations with Rick Greer, John McGuire, and Richard Savage were helpful.

REFERENCES

ANDERSON, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. *In Multivariate Analysis* (P. R. Krishnaiah, ed.) 5–27. Academic Press, New York.

BELSON, W. A. (1959). Matching and prediction on the principle of biological classification. *Appl. Statist.* **8** 65–75.

CORDOBA, A. and FEFFERMAN, R. (1975). A geometric proof of the strong maximal theorem. *Bull. Amer. Math. Soc.* **81** 941.

COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Information Theory* **IT-13** 21–27.

FIX, E. and HODGES, J. L. (1951). Discriminatory analysis, nonparametric classifications. *USAF Sch. Aviat. Med.*, Randolph Field, Texas, Project 21-49-004, Report 4, Contract AF41 (128)-31.

FRIEDMAN, J. H. (1977). A variable metric decision rule for nonparametric classification. *IEEE Trans. Computers* **C-26** 404–408.

HILLS, M. (1967). Discrimination and allocation with discrete data. *Appl. Statist.* **16** 237–250.

JESSEN, B., MARCINKIEWICZ, J. and ZYGMUND, A. (1935). Note on the differentiability of multiple integrals. *Fund. Math.* **25** 217–234.

KIEFER, J. (1961). On large deviations of the empiric d.f. of vector chance variables and a law of iterated logarithm. *Pacific J. Math.* **11** 649–660.

PELTO, C. R. (1969). Adaptive nonparametric classification. *Technometrics* **11** 775–792.

SONQUIST, J. (1970). *Multivariate Model Building: The Validation of a Search Strategy.* Institute for Social Research, Univ. of Michigan, Ann Arbor.

STOLLER, D. C. (1954). Univariate two-population distribution-free discrimination. *J. Amer. Statist. Assoc.* **49** 770–775.

STONE, C. J. (1977). Nonparametric regression and its applications (with discussion). *Ann. Statist.* **5** 595–645.

VAN RYZIN, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhyā Ser. A* **28** 261–270.

ALZA RESEARCH
950 PAGE MILL ROAD
PALO ALTO, CALIFORNIA 94304

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093