SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS

J. MacQUEEN
UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

The main purpose of this paper is to describe a process for partitioning an N-dimensional population into k sets on the basis of a sample. The process, which is called 'k-means,' appears to give partitions which are reasonably efficient in the sense of within-class variance. That is, if p is the probability mass function for the population, $S = \{S_1, S_2, \cdots, S_k\}$ is a partition of E_N, and u_i, $i = 1, 2, \cdots, k$, is the conditional mean of p over the set S_i, then $w^2(S) = \sum_{i=1}^{k} \int_{S_i} |z - u_i|^2 \, dp(z)$ tends to be low for the partitions S generated by the method. We say 'tends to be low,' primarily because of intuitive considerations, corroborated to some extent by mathematical analysis and practical computational experience. Also, the k-means procedure is easily programmed and is computationally economical, so that it is feasible to process very large samples on a digital computer. Possible applications include methods for similarity grouping, nonlinear prediction, approximating multivariate distributions, and nonparametric tests for independence among several variables.

In addition to suggesting practical classification methods, the study of k-means has proved to be theoretically interesting. The k-means concept represents a generalization of the ordinary sample mean, and one is naturally led to study the pertinent asymptotic behavior, the object being to establish some sort of law of large numbers for the k-means. This problem is sufficiently interesting, in fact, for us to devote a good portion of this paper to it. The k-means are defined in section 2.1, and the main results which have been obtained on the asymptotic behavior are given there. The rest of section 2 is devoted to the proofs of these results. Section 3 describes several specific possible applications, and reports some preliminary results from computer experiments conducted to explore the possibilities inherent in the k-means idea. The extension to general metric spaces is indicated briefly in section 4.

The original point of departure for the work described here was a series of problems in optimal classification (MacQueen [9]) which represented special

This work was supported by the Western Management Science Institute under a grant from the Ford Foundation, and by the Office of Naval Research under Contract No. 233(75), Task No. 047-041.

cases of the problem of optimal information structures as formulated by Marschak [11], [12]. (For an interesting treatment of a closely related problem, see Blackwell [1].) In one instance the problem of finding optimal information structures reduces to finding a partition $S = \{S_1, S_2, \cdots, S_k\}$ of E_N which will minimize $w^2(S)$ as defined above. In this special model, individual A observes a random point $z \in E_N$, which has a known distribution p, and communicates to individual B what he has seen by transmitting one of k messages. Individual B interprets the message by acting as if the observed point z is equal to a certain point $\hat{z}$ to be chosen according to the message received. There is a loss proportional to the squared error $|z - \hat{z}|^2$ resulting from this choice. The object is to minimize expected loss. The expected loss becomes $w^2(S)$, where the i-th message is transmitted if $z \in S_i$, since the best way for B to interpret the information is to choose the conditional mean of p on the set associated with the message received. The mean, of course, minimizes the squared error. Thus the problem is to locate a partition minimizing $w^2(S)$. This problem was also studied by Fisher [5], who gives references to earlier related works.

The k-means process was originally devised in an attempt to find a feasible method of computing such an optimal partition. In general, the k-means procedure will not converge to an optimal partition, although there are special cases where it will. Examples of both situations are given in section 2.3. So far as the author knows, there is no feasible, general method which always yields an optimal partition. Cox [2] has solved the problem explicitly for the normal distribution in one dimension, with $k = 2, 3, \cdots, 6$, and a computational method for finite samples in one dimension has been proposed by Fisher [5]. A closely related method for obtaining reasonably efficient 'similarity groups' has been described by Ward [15]. Also, a simple and elegant method which would appear to yield partitions with low within-class variance, was noticed by Edward Forgy [7] and Robert Jennrich, independently of one another, and communicated to the writer sometime in 1963. This procedure does not appear to be known to workers in taxonomy and grouping, and is therefore described in section 3. For a thorough consideration of the biological taxonomy problem and a discussion of a variety of related classification methods, the reader is referred to the interesting book by Sokal and Sneath [14]. (See *Note added in proof* of this paper.)

Sebestyen [13] has described a procedure called "adaptive sample set construction," which involves the use of what amounts to the k-means process. This is the earliest explicit use of the process with which the author is familiar. Although arrived at in ignorance of Sebestyen's work, the suggestions we make in sections 3.1, 3.2, and 3.3, are anticipated in Sebestyen's monograph.

2. K-means; asymptotic behavior

2.1. *Preliminaries.* Let $z_1, z_2, \cdots$ be a random sequence of points (vectors) in E_N, each point being selected independently of the preceding ones using a fixed probability measure p. Thus $P[z_1 \in A] = p(A)$ and $P[z_{n+1} \in A | z_1, z_2, \cdots, z_n] =$

$p(A)$, $n = 1, 2, \cdots$, for A any measurable set in E_N. Relative to a given k-tuple $x = (x_1, x_2, \cdots, x_k)$, $x_i \in E_N$, $i = 1, 2, \cdots, k$, we define a *minimum distance partition* $S(x) = \{S_1(x), S_2(x), \cdots, S_k(x)\}$ of E_N, by

$$S_1(x) = T_1(x), S_2(x) = T_2(x)S_1'(x), \cdots,$$
$$S_k(x) = T_k(x)S_1'(x)S_2'(x) \cdots S_{k-1}'(x), \tag{2.1}$$

where

$$T_i(x) = \{\xi: \xi \in E_N, |\xi - x_i| \leq |\xi - x_j|, j = 1, 2, \cdots, k\}. \tag{2.2}$$

The set $S_i(x)$ contains the points in E_N nearest to x_i, with tied points being assigned arbitrarily to the set of lower index. Note that with this convention concerning tied points, if $x_i = x_j$ and $i < j$ then $S_j(x) = \varnothing$. Sample k-means $x^n = (x_1^n, x_2^n, \cdots, x_k^n)$, $x_i^n \in E_N$, $i = 1, \cdots, k$, with associated integer weights $(w_1^n, w_2^n, \cdots, w_k^n)$, are now defined as follows: $x_i^1 = z_i$, $w_i^1 = 1$, $i = 1, 2, \cdots, k$, and for $n = 1, 2, \cdots$, if $z_{k+n} \in S_i^n$, $x_i^{n+1} = (x_i^n w_i^n + z_{n+k})/(w_i^n + 1)$, $w_i^{n+1} = w_i^n + 1$, and $x_j^{n+1} = x_j^n$, $w_j^{n+1} = w_j^n$ for $j \neq i$, where $S^n = \{S_1^n, S_2^n, \cdots, S_k^n\}$ is the minimum distance partition relative to x^n.

Stated informally, the k-means procedure consists of simply starting with k groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After a point is added to a group, the mean of that group is adjusted in order to take account of the new point. Thus at each stage the k-means are, in fact, the means of the groups they represent (hence the term k-means).

In studying the asymptotic behavior of the k-means, we make the convenient assumptions, (i) p is absolutely continuous with respect to Lebesgue measure on E_N, and (ii) $p(R) = 1$ for a closed and bounded convex set $R \subset E_N$, and $p(A) > 0$ for every open set $A \subset R$. For a given k-tuple $x = (x_1, x_2, \cdots, x_k)$—such an entity being referred to hereafter as a k-point—let

$$W(x) = \sum_{i=1}^{k} \int_{S_i} |z - x_i|^2 \, dp(z),$$
$$V(x) = \sum_{i=1}^{k} \int_{S_i} |z - u_i(x)|^2 \, dp(z), \tag{2.3}$$

where $S = \{S_1, S_2, \cdots, S_k\}$ is the minimum distance partition relative to x, and $u_i(x) = \int_{S_i} z \, dp(z)/p(S_i)$ or $u_i(x) = x_i$, according to whether $p(S_i) > 0$ or $p(S_i) = 0$. If $x_i = u_i(x)$, $i = 1, 2, \cdots, k$ we say the k-point x is *unbiased*.

The principal result is as follows.

THEOREM 1. *The sequence of random variables $W(x^1)$, $W(x^2)$, $\cdots$ converges a.s. and $W_\infty = \lim_{n\to\infty} W(x^n)$ is a.s. equal to $V(x)$ for some x in the class of k-points $x = (x_1, x_2, \cdots, x_k)$ which are unbiased, and have the property that $x_i \neq x_j$ if $i \neq j$.*

In lieu of a satisfactory strong law of large numbers for k-means, we obtain the following theorem.

THEOREM 2. *Let $u_i^n = u_i(x^n)$ and $p_i^n = p(S_i(x^n))$; then*

$$\sum_{n=1}^{m} \left(\sum_{i=1}^{k} p_i^n |x_i^n - u_i^n| \right) \Big/ m \underset{\text{a.s.}}{\to} 0 \quad \text{as} \quad m \to \infty. \tag{2.4}$$

2.2. *Proofs.* The system of k-points forms a complete metric space if the distance $\rho(x, y)$ between the k-points $x = (x_1, x_2, \cdots, x_k)$ and $y = (y_1, y_2, \cdots, y_k)$, is defined by $\rho(x, y) = \sum_{i=1}^{k} |x_i - y_i|$. We designate this space by M and interpret continuity, limits, convergence, neighborhoods, and so on, in the usual way with respect to the metric topology of M. Of course, every bounded sequence of k-points contains a convergent subsequence.

Certain difficulties encountered in the proof of theorem 1 are caused by the possibility of the limit of a convergent sequence of k-points having some of its constituent points equal to each other. With the end in view of circumventing these difficulties, suppose that for a given k-point $x = (x_1, x_2, \cdots, x_k)$, $x_i \in R$, $i = 1, 2, \cdots, k$, we have $x_i = x_j$ for a certain pair i, j, $i < j$, and $x_i = x_j \neq x_m$ for $m \neq i, j$. The points x_i and x_j being distinct in this way, and considering assumption (ii), we necessarily have $p(S_i(x)) > 0$, for $S_i(x)$ certainly contains an open subset of R. The convention concerning tied points means $p(S_j(x)) = 0$. Now if $\{y^n\} = \{(y_1^n, y_2^n, \cdots, y_k^n)\}$ is a sequence of k-points satisfying $y_i^n \in R$, and $y_i^n \neq y_j^n$ if $i \neq j$, $n = 1, 2, \cdots$, and the sequence y^n approached x, then y_i^n and y_j^n approach $x_i = x_j$, and hence each other; they also approach the boundaries of $S_i(y^n)$ and $S_j(y^n)$ in the vicinity of x_i. The conditional means $u_i(y^n)$ and $u_j(y^n)$, however, must remain in the interior of the sets $S_i(y^n)$ and $S_j(y^n)$ respectively, and thus tend to become separated from the corresponding points y_i^n and y_j^n. In fact, for each sufficiently large n, the distance of $u_i(y^n)$ from the boundary of $S_i(y^n)$ *or* the distance of $u_j(y^n)$ from the boundary of $S_j(y^n)$, will exceed a certain positive number. For as n tends to infinity, $p(S_i(y^n)) + p(S_j(y^n))$ will approach $p(S_i(x)) > 0$—a simple continuity argument based on the absolute continuity of p will establish this—and for each sufficiently large n, at least one of the probabilities $p(S_i(y^n))$ or $p(S_j(y^n))$ will be positive by a definite amount, say δ. But in view of the boundedness of R, a convex set of p measure at least $\delta > 0$ cannot have its conditional mean arbitrarily near its boundary. This line of reasoning, which extends immediately to the case where some three or more members of $(x_1, x_2, \cdots, x_k)$ are equal, gives us the following lemma.

LEMMA 1. *Let $x = (x_1, x_2, \cdots, x_k)$ be the limit of a convergent sequence of k-points $\{y^n\} = \{(y_1^n, y_2^n, \cdots, y_k^n)\}$ satisfying $y_i^n \in R$, $y_i^n \neq y_j^n$ if $i \neq j$, $n = 1, 2, \cdots$. If $x_i = x_j$ for some $i \neq j$, then $\liminf_n \sum_{i=1}^{k} p(S_i(y^n))|y_i^n - u_i(y^n)| > 0$. Hence, if $\lim_{n\to\infty} \sum_{i=1}^{k} p(S_i(y^n))|y_i^n - u_i(y^n)| = 0$, each member of the k-tuple $(x_1, x_2, \cdots, x_k)$ is distinct from the others.*

We remark that if each member of the k-tuple $x = (x_1, x_2, \cdots, x_k)$ is distinct from the others, then $\pi(y) = (p(S_1(y)), p(S_2(y)), \cdots, p(S_k(y)))$, regarded as a mapping of M onto E_k, is continuous at x—this follows directly from the absolute continuity of p. Similarly, $u(y) = (u_1(y), u_2(y), \cdots, u_k(y))$ regarded as a mapping from M onto M is continuous at x—because of the absolute continuity of p and the boundness of R (finiteness of $\int z\, dp(z)$ would do). Putting this remark together with lemma 1, we get lemma 2.

LEMMA 2. *Let $x = (x_1, x_2, \cdots, x_k)$ be the limit of a convergent sequence of k-points $\{y^n\} = \{(y_1^n, y_2^n, \cdots, y_k^n)\}$ satisfying $y_i^n \in R$, $y_i^n \neq y_j^n$ if $i \neq j$, $n = 1, 2,$*

$\cdots$. *If* $\lim_{n\to\infty} \sum_{i=1}^{k} p(S_i(y^n))|y_i^n - u_i(y^n)| = 0$, *then* $\sum_{i=1}^{k} p(S_i(x))|x_i - u_i(x^n)| = 0$ *and each point* x_i *in the* k*-tuple* $(x_1, x_2, \cdots, x_k)$ *is distinct from the others.*

Lemmas 1 and 2 above are primarily technical in nature. The heart of the proofs of theorems 1 and 2 is the following application of martingale theory.

LEMMA 3. *Let* $t_1, t_2, \cdots$, *and* $\xi_1, \xi_2, \cdots$, *be given sequences of random variables, and for each* $n = 1, 2, \cdots$, *let* t_n *and* ξ_n *be measurable with respect to* β_n *where* $\beta_1 \subset \beta_2 \subset \cdots$ *is a monotone increasing sequence of* σ*-fields (belonging to the underlying probability space). Suppose each of the following conditions holds a.s.:* (i) $|t_n| \lesssim K < \infty$, (ii) $\xi_n \geq 0$, $\sum \xi_n < \infty$, (iii) $E(t_{n+1}|\beta_n) \leq t_n + \xi_n$. *Then the sequences of random variables* $t_1, t_2, \cdots$ *and* $s_0, s_1, s_2, \cdots$, *where* $s_0 = 0$ *and* $s_n = \sum_{i=1}^{n} (t_i - E(t_{i+1}|\beta_i))$, $n = 1, 2, \cdots$, *both converge a.s.*

PROOF. Let $y_n = t_n + s_{n-1}$ so that the y_n form a martingale sequence. Let c be a positive number and consider the sequence $\{\tilde{y}_n\}$ obtained by stopping y_n (see Doob [3], p. 300) at the first n for which $y_n \leq -c$. From (iii) we see that $y_n \geq -\sum_{i=1}^{n-1} \xi_i - K$, and since $y_n - y_{n-1} \geq 2K$, we have $\tilde{y}_n \geq \max(-\sum_{i=1}^{n-1} \xi_i - K, -(c + 2K))$. The sequence $\{\tilde{y}\}$ is a martingale, so that $E\tilde{y}_n = E\tilde{y}_1$, $n = 1, 2, \cdots$, and being bounded from below with $E|\tilde{y}_1| \leq K$, certainly $\sup_n E|\tilde{y}_n| < \infty$. The martingale theorem ([3], p. 319) shows $\tilde{y}_n$ converges a.s. But $y_n = \tilde{y}_n$ on the set A_c where $-\sum_{i=1}^{\infty} \xi_i > -c - K$, $i = 1, 2, \cdots$, and (ii) implies $P[A_c] \to 1$ as $c \to \infty$. Thus $\{y_n\}$ converge a.s. This means $s_n = y_{n+1} - t_{n+1}$ is a.s. bounded. Using (iii) we can write $-s_n = \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \Delta_i$ where $\Delta_i \geq 0$. But since s_n and $\sum_1^n \xi_i$ are a.s. bounded, $\sum \Delta_i$ converges a.s., s_n converges a.s., and finally, so does t_n. This completes the proof.

Turning now to the proof of theorem 1, let ω_n stand for the sequence $z_1, z_2, \cdots, z_{n-1+k}$, and let A_1^n be the event $[z_{n+k} \in S_1^n]$. Since S^{n+1} is the minimum distance partition relative to x^{n+1}, we have

$$
\begin{aligned}
(2.5) \qquad E[W(x^{n+1})|\omega_n] &= E\left[\sum_{i=1}^{k} \int_{S_i^{n+1}} |z - x_i^{n+1}|^2 \, dp(z)|\omega_n\right] \\
&\leq E\left[\sum_{i=1}^{k} \int_{S_i^n} |z - x_i^{n+1}|^2 \, dp(z)|\omega_n\right] \\
&= \sum_{j=1}^{k} E\left[\sum_{i=1}^{k} \int_{S_i^n} |z - x_i^{n+1}|^2 \, dp(z)|A_j^n, \omega_n\right] p_j^n.
\end{aligned}
$$

If $z_{n+k} \in S_j^n$, $x_i^{n+1} = x_i^n$ for $i \neq j$. Thus we obtain

$$
\begin{aligned}
(2.6) \qquad E[W(x^{n+1})|\omega_n] &\leq W(x^n) - \sum_{j=1}^{k} \left(\int_{S_j^n} |z - x_j^n|^2 \, dp(z)\right) p_j^n \\
&\quad + \sum_{j=1}^{k} E\left[\int_{S_j^n} |z - x_j^{n+1}|^2 \, dp(z)|A_j^n, \omega_n\right] p_j^n.
\end{aligned}
$$

Several applications of the relation $\int_A |z - x|^2 \, dp(z) = \int_A |z - u|^2 \, dp(z) + p(A)|x - u|^2$, where $\int_A (u - z) \, dp(z) = 0$, enables us to write the last term in (2.6) as

$$(2.7)\qquad \sum_{j=1}^{k}\Big[\int_{S_j^n}|z-x_j^n|^2\,dp(z)\,p_j^n-(p_j^n)^2|x_j^n-u_j^n|^2 + (p_j^n)^2|x_j^u-u_j^n|^2(w_j^n/(w_j^n+1))^2+\int_{S_j^n}|z-u_j^n|^2\,dp(z)\,p_j^n/(w_j^n+1)_2\Big].$$

Combining this with (2.6), we get

$$(2.8)\qquad E(W(x^{n+1})|\omega_n]\le W(x^n)-\sum_{j=1}^{k}|x_j^n-u_j^n|^2(p_j^n)^2(2w_j^n+1)/(w_j^n+1)^2 + \sum_{j=1}^{k}\sigma_{n,j}^2(p_j^n)^2/(w_j^n+1)^2,$$

where $\sigma_{n,j}^2=\int_{S_j^n}|z-u_j^n|^2\,dp(z)/p_j^n$.

Since we are assuming $p(R)=1$, certainly $W(x^n)$ is a.s. bounded, as is $\sigma_{n,j}^2$. We now show that

$$(2.9)\qquad \sum_n (p_j^n)^2/(w_j^n+1)^2$$

converges a.s. for each $j=1,2,\cdots,k$, thereby showing that

$$(2.10)\qquad \sum_n\Big(\sum_{j=1}^{k}[\sigma_{n,j}^2(p_j^n)^2/(w_j^n+1)^2]\Big)$$

converges a.s. Then lemma 3 can be applied with $t_n=W(x^n)$ and $\xi_n=\sum_{j=1}^{k}\sigma_{n,j}^2(p_j^n)^2/(w_j^n+1)^2$.

It suffices to consider the convergence of

$$(2.11)\qquad \sum_{n\ge 2}(p_j^n)^2/[(\beta+1+w_j^n)(\beta+1+w_j^{n+1})]$$

with $\beta>0$, since this implies convergence of (2.9). Also, this is convenient, for $E(I_j^n|\omega_n)=p_j^n$ where I_j^n is the characteristic function of the event $[z_{n+k}\in S_j^n]$, and on noting that $w_j^{n+1}=1+\sum_{i=1}^{n}I_j^i$, an application of theorem 1 in [4], p. 274, says that for any positive numbers α and β,

$$(2.12)\qquad P\Big[\beta+1+w_j^{n+1}\ge 1+\sum_{i=1}^{n}p_j^i-\alpha\sum_{i=1}^{n}v_j^i \text{ for all } n=1,2,\cdots\Big] > 1-(1+\alpha\beta)^{-1},$$

where $v_j^i=p_j^i-(p_j^i)^2$ is the conditional variance of I_j^i given ω_i. We take $\alpha=1$, and thus with probability at least $1-(1+\beta)^{-1}$ the series (2.11) is dominated by

$$(2.13)\qquad \sum_{n\ge 2}(p_j^n)^2\Big/\Big[\Big(1+\sum_{i=1}^{n-1}(p_j^i)^2\Big)\Big(1+\sum_{i=1}^{n}(p_j^i)^2\Big)\Big] = \sum_{n\ge 2}\Big[1\Big/\Big(1+\sum_{i=1}^{n-1}(p_j^i)^2\Big)-1\Big/\Big(1+\sum_{i=1}^{n}(p_j^i)^2\Big)\Big],$$

which clearly converges.

The choice of β being arbitrary, we have shown that (2.9) converges a.s. Application of lemma 3 as indicated above proves $W(x^n)$ converges a.s.

To identify the limit W_∞, note that with t_n and ξ_n taken as above, lemma 3

entails a.s. convergence of $\sum_n [W(x^n) - E[W(x^{n+1})|\omega_n]]$, and hence (2.8) implies a.s. convergence of

$$\sum_n \left(\sum_{j=1}^k |x^n - u_j^n|^2 (p_j^n)^2 (2w_j^n + 1)/(w_j^n + 1)^2 \right). \tag{2.14}$$

Since (2.14) dominates $\sum_n (\sum_{j=1}^k p_j^n |x_j^n - u_j^n|)/kn$, the latter converges a.s., and a little consideration makes it clear that

$$\sum_{j=1}^k p_j^n |x_j^n - u_j^n| = \sum_{j=1}^k p(S_j(x^n)) |x_j^n - u_j(x^n)| \tag{2.15}$$

converges to zero on a subsequence $\{x^{n_t}\}$ and that this subsequence has itself a convergent subsequence, say $\{x^{n_t}\}$. Let $x = (x_1, x_2, \cdots, x_k) = \lim_{t\to\infty} x^{n_t}$. Since $W(x) = V(x) + \sum_{j=1}^k p(S_j(x))|x_j - u(x)|^2$ and in particular,

$$W(x^n) = V(x^n) + \sum_{j=1}^k p(S_j(x^n)) |x_j^n - u(x_j^n)|^2, \tag{2.16}$$

we have only to show

(a) $\lim_{t\to\infty} W(x^{n_t}) = W_\infty = W(x)$, and

(b) $\lim_{t\to\infty} \sum_{j=1}^k p(S_j(x^{n_t})) |x_j^{n_t} - u(x_j^{n_t})|^2 = 0 = \sum_{j=1}^k p(S_j(x)) |x_j - u_j(x)|^2$.

Then $W(x) = V(x)$ and x is a.s. unbiased. (Obviously, $\sum_{i=1}^k p_i |a_i| = 0$ if and only if $\sum_{i=1}^k p_i |a_i|^2 = 0$, where $p_i \geq 0$.)

We show that (a) is true by establishing the continuity of $W(x)$. We have

$$\begin{aligned} W(x) &\leq \sum_{j=1}^k \int_{S_j(y)} |z - x_j|^2 \, dp(z) \\ &\leq \sum_{j=1}^k \int_{S_j(y)} |z - y_j|^2 + \sum_{j=1}^k [p(S_j(y)) |x_j - y_j|^2 \\ &\quad + 2|x_j - y_j| \int_{S_j(y)} |z - x_j| \, dp(z)], \end{aligned} \tag{2.17}$$

with the last inequality following easily from the triangle inequality. Thus $W(x) \leq W(y) + o(\rho(x, y))$, and similarly, $W(y) \leq W(x) + o(\rho(x, y))$.

To establish (b), lemma 2 can be applied with $\{y^n\}$ and $\{x^{n_t}\}$ identified, for a.s. $x_i^n \neq x_j^n$ for $i \neq j$, $n = 1, 2, \cdots$. It remains to remark that lemma 2 also implies a.s. $x_i \neq x_j$ for $i \neq j$. The proof of theorem 1 is complete.

Theorem 2 follows from the a.s. convergence of $\sum_n (\sum_{i=1}^k p_i^n |x_i^n - u_i^n|)/nk$ upon applying an elementary result (c.f. Halmos [8], theorem C, p. 203), which says that if $\sum a_n/n$ converges, $\sum_{i=1}^n a_i/n \to 0$.

2.3. *Remarks.* In a number of cases covered by theorem 1, all the unbiased k-points have the same value of W. In this situation, theorem 1 implies $\sum_{i=1}^k p_i^n |x_i^n - u_i^n|$ converges a.s. to zero. An example is provided by the uniform distribution over a disk in E_2. If $k = 2$, the unbiased k-point (x_1, x_2) with $x_1 \neq x_2$ consist of the family of points x_1 and x_2 opposite one another on a diameter, and at a certain fixed distance from the center of the disk. (There is one unbiased k-point with $x_1 = x_2$, both x_1 and x_2 being at the center of the disk in this case.)

The k-means thus converge to some such relative position, but theorem 1 does not quite permit us to eliminate the interesting possibility that the two means oscillate slowly but indefinitely around the center.

Theorem 1 provides for a.s. convergence of $\sum_{i=1}^{k} p_i^n|x_i^n - u_i^n|$ to zero in a slightly broader class of situations. This is where the unbiased k-points $x = (x_1, x_2, \cdots, x_k)$ with $x_i \neq x_j$ for $i \neq j$, are all *stable* in the sense that for each such x, $W(y) \geq W(x)$ (and hence $V(y) \geq V(x)$) for all y in a neighborhood of x. In this case, each such x falls in one of finitely many equivalence classes such that W is constant on each class. This is illustrated by the above example, where there is only a single equivalence class. If each of the equivalence classes contains only a single point, theorem 1 implies a.s. convergence of x^n to one of those points.

There are unbiased k-points which are not stable. Take a distribution on E_2 which has sharp peaks of probability at each corner of a square, and is symmetric about both diagonals. With $k = 2$, the two constituent points can by symmetrically located on a diagonal so that the boundary of the associated minimum distance partition coincides with the other diagonal. With some adjustment, such a k-point can be made to be unbiased, and if the probability is sufficiently concentrated at the corners of the square, any small movement of the two points off the diagonal in opposite directions, results in a decrease in $W(x)$. It seems likely that the k-means *cannot* converge to such a configuration.

For an example where the k-means converge with positive probability to a point x for which $V(x)$ is not a minimum, take equal probabilities at the corner points of a rectangle which is just slightly longer on one side than the other. Number with 1 the corner points, and 2 at the end points of one of the short edges, and 3 and 4, at the end points of the other short edge, with 1 opposite 3 on the long edge. Take $k = 2$. If the first four points fall at the corner points 1, 2, 3, 4 in that order, the two means at this stage are directly opposite one another at the middle of the long edges. New points falling at 1 and 3 will always be nearer the first mean, and points falling at 2 and 4 will always be nearer the second mean, unless one of the means has an excursion too near one of the corner points. By the strong law of large numbers there is positive probability this will *not* happen, and hence with positive probability the two means will converge to the midpoints of the long edges. The corresponding partition clearly does not have minimum within-class variance.

3. Applications

3.1. *Similarity grouping: coarsening and refining.* Perhaps the most obvious application of the k-means process is to the problem of "similarity grouping" or "clustering." The point of view taken in this application is *not* to find some unique, definitive grouping, but rather to simply aid the investigator in obtaining qualitative and quantitative understanding of large amounts of N-dimensional data by providing him with reasonably good similarity groups. The method should be used in close interaction with theory and intuition. Consequently, the

computer program actually prepared for this purpose involved several modifications of the k-means process, modifications which appear to be helpful in this sense.

First, the program involves two parameters: C for 'coarsening,' and R for 'refinement.' The program starts with a user specified value of k, and takes the first k points in the sample as initial means. The k-means process is started, each subsequent sample point being assigned to the nearest mean, the new mean computed, and so on, except that after each new point is added, and for the initial means as well, the program determines the pair of means which are nearest to each other among all pairs. If the distance between the members of this pair is less than C, they are averaged together, using their respective weights, to form a single mean. The nearest pair is again determined, their separation compared with C, and so on, until all the means are separated by an amount of C or more. Thus k is reduced and the partition defined by the means is coarsened. In addition, as each new point is processed and its distance from the nearest of the current means determined, this distance is compared with R. If the new point is found to be further than R from the nearest mean, it is left by itself as the seed point for a new mean. Thus k is increased and the partition is refined. Ordinarily we take $C \leq R$. After the entire sample is processed in this way, the program goes back and reclassifies all the points on the basis of nearness to the final means. The points thus associated with each mean constitutes the final grouping. The program prints out the points in each group along with as many as 18 characters of identifying information which may be supplied with each point. The distance of each point from its nearest mean, the distances between the means, the average for each group, of the squared distance of the points in each group from their respective defining means, and the grand average of these quantities over groups, are all printed out. The latter quantity, which is not quite the within-group variance, is called the within-class variation for purposes of the discussion below. If requested, the program determines frequencies of occurrence within each group of the values of discrete variables associated with each point. Up to twelve variables, with ten values for each variable, can be supplied. This makes it convenient to determine whether or not the groups finally obtained are related to other attributes of interest. (Copies of this experimental program are available from the author on request.)

The program has been applied with some success to several samples of real data, including a sample of five dimensional observations on the students' environment in 70 U.S. colleges, a sample of twenty semantic differential measurements on each of 360 common words, a sample of fifteen dimensional observations on 760 documents, and a sample of fifteen physiological observations on each of 560 human subjects. While analysis of this data is still continuing, and will be reported in detail elsewhere, the meaningfulness of the groups obtained is suggested by their obvious pertinence to other identifiable properties of the objects classified. This was apparent on inspection. For example, one group of colleges contained Reed, Swarthmore, Antioch, Oberlin, and Bryn

Mawr. Another group contained the Universities of Michigan, Minnesota, Arkansas, and Illinois, Cornell, Georgia Tech, and Purdue. Selecting at random a half-dozen words from several groups obtained from the semantic differential data, we find in one group the words calm, dusky, lake, peace, sleep, and white; in another group the words beggar, deformed, frigid, lagging, low; and in another group the words statue, sunlight, time, trees, truthful, wise.

When the sample points are rearranged in a new random order, there is some variation in the grouping which is obtained. However, this has not appeared to be a serious concern. In fact, when there are well separated clusters, as determined by inspection of the between-mean distances in relation to the within-class variation, repeated runs give virtually identical groupings. Minor shifts are due to the unavoidable difficulty that some points are located between clusters.

A degree of stability with respect to the random order in which the points are processed is also indicated by a tendency for the within-class variation to be similar in repeated runs. Thus when a sample of 250 points in five dimensions with $k = 18$, was run three times, each time with the points in a different random order, the within-class variation (see above) changed over the three runs by at most 7%. A certain amount of stability is to be expected simply because the within-class variation is the mean of k dependent random variables having the property that when one goes up the others generally go down. We can reasonably expect the within-class stability to generally increase with k and the sample size. Actually, it will usually be desirable to make several runs, with different values of C and R, and possibly adding, deleting, or rescaling variables, and so on, in an effort to understand the basic structure of the data. Thus any instabilities due to random ordering of the sample will be quickly noted. Being able to make numerous classifications cheaply and thereby look at the data from a variety of different perspectives is an important advantage.

Another general feature of the k-means procedure which is to be expected on intuitive grounds, and has been noted in practice, is a tendency for the means and the associated partition to avoid having the extreme of only one or two points in a set. In fact, there is an appreciable tendency for the frequency to be evenly split over groups. If there are a few relatively large groups, these tend to have relatively low within-class variation, as would be expected from a tendency for the procedure to approximate minimum variance partitions.

Running times of the above program on the IBM 7094 vary with C, R, the number of dimensions, and the number of points. A conservative estimate for 20-dimensional data, with C and R set so that k stays in the vicinity of 20, is one minute for two hundred sample points. Most of this computation time results from the coarsening and refining procedure and the auxiliary features. A limited amount of experience indicates the undecorated k-means procedure with $k = 20$ will process five hundred points in 20 dimensions in something like 10 seconds.

3.2. *Relevant classifications.* Suppose it is desired to develop a classification scheme on the basis of a sample, so that knowing the classification of a new point, it will be possible to predict a given dependent variable. The values of the de-

pendent variable are known for the sample. One way to do this, closely related to a procedure proposed by Fix and Hodges [6], is illustrated by the following computer experiment. A sample of 250 four-dimensional random vectors was prepared, with the values on each dimension being independently and uniformly distributed on the integers 1 through 10. Two of the dimensions were then arbitrarily selected, and if with respect to these two dimensions a point was either 'high' (above 5) on both *or* 'low' (5 or less) on both, it was called an A; otherwise, it was called a B. This gave 121 A's and 129 B's which were related to the selected dimensions in a strongly interactive fashion. The k-means with $k = 8$ were then obtained for the A's and B's separately. Finally, using the resulting 16 (four-dimensional) means, a prediction, A or B, was made for each of a *new* sample of 250 points on the basis of whether or not each point was nearest to an A mean or a B mean. These predictions turned out to be 87% correct.

As this example shows, the method is potentially capable of taking advantage of a highly nonlinear relationship. Also, the method has something to recommend it from the point of view of simplicity, and can easily be applied in many dimensions and to more than two-valued dependent variables.

3.3. *Approximating a general distribution.* Suppose it is desired to approximate a distribution on the basis of a sample of points. First the sample points are processed using the k-means concept or some other method which gives a minimum distance partition of the sample points. The approximation, involving a familiar technique, consists of simply fitting a joint normal distribution to the points in each group, and taking as the approximation the probability combination of these distributions, with the probabilities proportional to the number of points in each group.

Having fitted a mixture of normals in this way, it is computationally easy (on a computer) to do two types of analysis. One is predicting unknown coordinates of a new point given the remaining coordinates. This may be done by using the regression function determined on the assumption that the fitted mixture is the true distribution. Another possible application is a kind of nonlinear discriminant analysis. A mixture of k normals is fitted in the above fashion to two samples representing two given different populations; one can then easily compute the appropriate likelihood ratios for deciding to which population a new point belongs. This method avoids certain difficulties encountered in ordinary discriminant analysis, such as when the two populations are each composed of several distinct subgroups, but with some of the subgroups from one population actually between the subgroups of the other. Typically in this situation, one or several of the k-means will be centered in each of the subgroups—provided k is large enough—and the fitted normals then provide a reasonable approximation to the mixture.

To illustrate the application of the regression technique, consider the artificial sample of four-dimensional A's and B's described in the preceding section. On a fifth dimension, the A's were arbitrarily given a value of 10, and the B's a value of 0. The k-means procedure with $k = 16$ was used to partition the combined

sample of 250 five-dimensional points. Then the mixture of 16 normal distributions was determined as described above for this sample. The second sample of 250 points was prepared similarly, and predictions were made for the fifth dimension on the basis of the original four. The standard error of estimate on the new sample was 2.8. If, in terms of the original A-B classification, we had called a point on A if the predicted value exceeded 5, and a B otherwise, 96% of the designations would have been correct on the new sample. The mean of the predictions for the A's was 10.3, and for B's, 1.3.

Considering the rather complex and highly nonlinear relationship involved in the above sample, it is doubtful that any conventional technique would do as well. In the few instances which were tested, the method performed nearly as well as linear regression on normally distributed samples, provided k was not too large. This is not surprising inasmuch as with $k = 1$ the method *is* linear regression. In determining the choice of k, one procedure is to increase k as long as the error of estimate drops. Since this will probably result in "over fitting" the sample, a cross validation group is essential.

3.4. *A scrambled dimension test for independence among several variables.* As a general test for relationship among variables in a sample of N-dimensional observations, we propose proceeding as follows. First, the sample points are grouped into a minimum distance partition using k-means, and the within-class variance is determined. Then the relation among the variables is destroyed by randomly associating the values in each dimension; that is, a sample is prepared in which the variables are unrelated, but which has exactly the same marginal distributions as the original sample. A minimum distance partition and the associated within-class variance is now determined for this sample. Intuition and inspection of a few obvious examples suggest that on the average this "scrambling" will tend to *increase* the within-class variance, more or less regardless of whatever type of relation might have existed among the variables, and thus comparison of the two variances would reveal whether or not any such relation existed.

To illustrate this method, a sample of 150 points was prepared in which points were distributed uniformly outside a square 60 units on a side, but inside a surrounding square 100 units on a side. This gave a sample which involves essentially a zero correlation coefficient, and yet a substantial degree of relationship which could not be detected by any conventional quantitative technique known to the author (although it could be detected immediately by visual inspection). The above procedure was carried out using k-means with $k = 12$. As was expected, the variance after scrambling was increased by a factor of 1.6. The within-class variances were not only larger in the scrambled data, but were apparently more variable. This procedure was also applied to the five-dimensional sample described in the preceding section. Using $k = 6$, 12, and 18, the within-class variance increased after scrambling by the factors 1.40, 1.55, and 1.39, respectively.

A statistical test for nonindependence can be constructed by simply repeating the scrambling and partitioning a number of times, thus obtaining empirically a

sample from the conditional distribution of the within-class variance under the hypothesis that the variables are unrelated *and* given the marginal values of the sample. Under the hypothesis of independence, the unscrambled variance should have the same (conditional) distribution as the scrambled variance. In fact, the rank of the unscrambled variance in this empirical distribution should be equally likely to take on any of the possible values 1, 2, $\cdots$, $n + 1$, where n is the number of scrambled samples taken, regardless of the marginal distributions in the underlying population. Thus the rank can be used in a nonparametric test of the hypothesis of independence. For example, if the unscrambled variance is the lowest in 19 values of the scrambled variance, we can reject the hypothesis of independence with a Type I error of .05.

A computer program was not available to do the scrambling, and its being inconvenient to set up large numbers of scrambled samples using punched cards, further testing of this method was not undertaken. It is estimated, however, that an efficient computer program would easily permit this test to be applied at, say, the .01 level, on large samples in many dimensions.

The power of this procedure remains to be seen. On the encouraging side is the related conjecture, that for fixed marginal distributions, the within-class variance for the optimal partition as defined in section 1 is maximal when the joint distribution is actually the product of the marginals. If this is true (and it seems likely that it is, at least for a large class of reasonable distributions), then we reason that since the k-means process tends to give a good partition, this difference will be preserved in the scrambled and unscrambled variances, particularly for large samples. Variation in the within-class variance due to the random order in which the points are processed, can be reduced by taking several random orders, and averaging their result. If this is done for the scrambled runs as well, the Type I error is preserved, while the power is increased somewhat.

3.5. *Distance-based classification trees.* The k-means concept provides a number of simple procedures for developing lexigraphic classification systems (filing systems, index systems, and so on) for a large sample of points. To illustrate, we describe briefly a procedure which results in the within-group variance of each of the groups at the most refined level of classification being no more than a specified number, say R. The sample k-means are first determined with a selected value of k, for example, $k = 2$. If the variance of any of the groups of points nearest to these means is less than R, these groups are not subclassified further. The remaining groups are each processed in the same way, that is, k-means are determined for each of them, and then for the points nearest each of these, and so on. This is continued until only groups with within-group variance less than R remain. Thus for each mean at the first level, there is associated several means at the second level, and so on. Once the means at each level are determined from the sample in this fashion, the classification of a new point is defined by the rule: first, see which one of the first level k-means the point is nearest; then see which one of the second-level k-means associated with that mean the point is nearest,

and so on; finally the point is assigned to a group which in the determining sample has variance no more than R.

This procedure has some promising features. First, the amount of computation required to determine the index is approximately linear in the sample size and the number of levels. The procedure can be implemented easily on the computer. At each stage during the construction of the classification tree, we are employing a powerful heuristic, which consists simply of putting points which are near to each other in the same group. Each of the means at each level is a fair representation of its group, and can be used for certain other purposes, for instance, to compare other properties of the points as a function of their classification.

3.6. *A two-step improvement procedure.* The method of obtaining partitions with low within-class variance which was suggested by Forgy and Jennrich (see section 1.1) works as follows. Starting with an arbitrary partition into k sets, the means of the points in each set are first computed. Then a new partition of the points is formed by the rule of putting the points into groups of the basis of nearness to the first set of means. The average squared distance of the points in the new partition from the first set of means (that is, from their nearest means) is obviously less than the within-class variance of the first partition. But the average within-class *variance* of the new partition is even lower, for the variance of the squared distance of the points in each group from their respective means, and the mean, of course, is that point which minimizes the average squared distance from itself. Thus the new partition has lower variance. Computationally, the two steps of the method are (1) compute the means of the points in each set in the initial partition and (2) reclassify the points on the basis of nearness to these means, thus forming a new partition. This can be iterated and the series of the partitions thus produced have decreasing within-class variances and will converge in a finite number of steps.

For a given sample, one cycle of this method requires about as much computation as the k-means. The final partition obtained will depend on the initial partition, much as the partition produced by k-means will depend on random variation in the order in which the points are processed. Nevertheless, the procedure has much to recommend it. By making repeated runs with different initial starting points, it would seem likely that one would actually obtain the sample partition with minimum within-class variance.

4. General metric spaces

It may be something more than a mere mathematical exercise to attempt to extend the idea of k-means to general metric spaces. Metric spaces other than Euclidian ones do occur in practice. One prominent example is the space of binary sequences of fixed length under Hamming distance.

An immediate difficulty in making such an extension is the notion of mean itself. The arithmetic operations defining the mean in Euclidian space may not be available. However, with the communication problem of section 1 in mind,

one thinks of the problem of representing a population by a point, the goal being to have low average error in some sense. Thus we are led to proceed rather naturally as follows.

Let M be a compact metric space with distance ρ, let $\mathfrak{F}$ be the σ-algebra of subsets of M, and let p be a probability measure on $\mathfrak{F}$. For the measure p, a centroid of order $r \geq 0$ is any point in the set $\mathcal{C}^r$ of points x^* such that $\int \rho^r(x^*, z)\, dp(z) = \inf_x \int \rho^r(x, z)\, dp(z)$. The quantity $\int \rho^r(x^*, z)\, dp(z)$ is the r-th moment of p. The compactness and the continuity of ρ guarantee that $\mathcal{C}^r$ is nonempty. For finite samples, sample centroids are defined analogously, each point in the sample being treated as having measure $1/n$ where n is the sample size; namely, for a sample of size n, the sample centroid is defined up to an equivalence class $\mathcal{C}_n^r$ which consists of all those points $\hat{x}_n$ such that $\sum_{i=1}^n \rho^r(\hat{x}_n, z_i) = \inf_x \sum_{i=1}^n \rho^r(x, z_i)$, where $z_1, z_2, \cdots, z_n$ is the sample.

Note that with M the real line, and ρ ordinary distance, $r = 2$ yields the ordinary mean, and $r = 1$ yields the family of medians. As r tends to ∞, the elements of $\mathcal{C}_r$ will tend to have (in a manner which can easily be made precise) the property that they are centers for a spherical covering of the space with minimal radius. In particular, on the line, the centroid will tend to the mid-range. As r tends to zero, one obtains what may with some justification be called a mode, for on a compact set, $\rho^r(x, y)$ is approximately 1 for small r, except where x and y are very near, so that minimizing $\int \rho^r(x, y)\, dp(y)$ with respect to x, involves attempting to locate x so that there is a large amount of probability in its immediate vicinity. (This relationship can also be made precise.)

We note that the optimum communication problem mentioned in section 1.1 now takes the following general form. Find a partition $S = \{S_1, S_2, \cdots, S_k\}$ which minimizes $w = \sum_{i=1}^k \int_{S_i} \rho^r(x_i^*, y)\, dp(y)$, where x_i^* is the centroid of order r with respect to the (conditional) distribution on S_i. If there is any mass in a set S_i nearer to x_j than to x_i, $j \neq i$, then w can be reduced by modifying S_i and S_j so as to reassign this mass to S_j. It follows that in minimizing w we can restrict attention to partitions which are minimum distance partitions, analogous to those defined in section 2, that is, partitions of the form $S(x) = \{S_1(x), S_2(x), \cdots, S_k(x)\}$ where $x = (x_1, x_2, \cdots, x_k)$ is a k-tuple of points in M, and $S_i(x)$ is a set of points at least as near x_i (in terms of ρ) as to x_j if $j \neq i$. In keeping with the terminology of section 2, we may say that a k-tuple, or "k-point," $x = (x_1, x_2, \cdots, x_k)$ is unbiased if x_i, $i = 1, 2, \cdots, k$, belongs to the class of points which are centroids within $S_i(x)$.

It is now clear how to extend the concept of k-means to metric spaces; the notion of centroid replaces the more special concept of mean. The first 'k-centroid' $(x_1^1, x_2^1, \cdots, x_k^1)$ consists of the first k points in the sample, and thereafter as each new point is considered, the nearest of the centroids is determined. The new point is assigned to the corresponding group and the centroid of that group modified accordingly, and so on.

It would seem reasonable to suppose that the obvious extension of theorem 1 would hold. That is, under independent sampling, $\sum_{i=1}^k \int_{S_i(x^n)} \rho^r(z, x_i^n)\, dp(z)$ will

converge a.s., and the convergent subsequences of the sequence of sample k-centroids will have their limits in the class of unbiased k-points. This is true, at any rate, for $k = 1$ and $r = 1$, for if $z_1, z_2, \cdots, z_n$ are independent, $\sum_{i=1}^{n} \rho(z_i, y)/n$ is the mean of independent, identically distributed random variables, which because M is compact, are uniformly bounded in y. It follows (cf. Parzen [13]) that $\sum_{i=1}^{n} \rho(z_i, y)/n$ converges a.s. to $\int \rho(z, y)\, dp(z)$ uniformly in y. By definition of the sample centroid, we have $\sum_{i=1}^{n} \rho(z_i, x^*)/n \geq \sum_{i=1}^{n} \rho(z_i, \hat{x}_n)/n$; hence, $\int \rho(z, x^*)\, dp(z) \geq \limsup \sum_{i=1}^{n} \rho(z_i, \hat{x}_n)/n$ with probability 1. On the other hand, from the triangle inequality, $\sum_{i=1}^{n} \rho(z_i, y)/n \leq \sum_{i=1}^{n} \rho(z_i, \hat{x}_n)/n + \rho(\hat{x}_n, y)$. Using this inequality on a convergent subsequence $\hat{x}_{n1}, \hat{x}_{n2}, \cdots$, chosen so that

$$\lim_{t\to\infty} \sum_{i=1}^{n_t} \rho(z_i, \hat{x}_{n_t})/n_t = \liminf \sum_{i=1}^{n} \rho(z_i, \hat{x}_n)/n, \tag{4.1}$$

we see that with probability 1,

$$\int \rho(z, x^*)\, dp(z) \leq \int \rho(z, y)\, dp(z) \leq \liminf \sum_{i=1}^{n} \rho(z_i, \hat{x}_n)/n, \tag{4.2}$$

where $y = \lim_{t\to\infty} \hat{x}_{n_t}$.

Provided the necessary computations can be accomplished, the methods suggested in sections 3.1, 3.2, 3.4, 3.5, and 3.6 can all be extended to general metric spaces in a quite straightforward fashion.

ACKNOWLEDGMENTS

The author is especially indebted to Tom Ferguson, Edward Forgy, and Robert Jennrich, for many valuable discussions of the problems to which the above results pertain. Richard Tenney and Sonya Baumstein provided the essential programming support, for which the author is very grateful. Computing facilities were provided by the Western Data Processing Center.

Note added in proof. The author recently learned that C. S. Wallace of the University of Sidney and G. H. Ball of the Stanford Research Institute have independently used this method as a part of a more complex procedure. Ball has described his method, and reviewed earlier literature, in the interesting paper "Data analysis in the social sciences: What about the details?", *Proceedings of the Fall Joint Computer Conference*, Washington, D.C., Spartan Books, 1965.

REFERENCES

[1] DAVID BLACKWELL, "Comparison of experiments," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1951, pp. 93–102.

[2] D. R. Cox, "Note on grouping," *J. Amer. Statist. Assoc.*, Vol. 52 (1957), pp. 543–547.
[3] J. L. Doob, *Stochastic Processes*, New York, Wiley, 1953.
[4] L. E. Dubins and L. J. Savage, "A Tchebycheff-like inequality for stochastic processes," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 53 (1965), pp. 274–275.
[5] W. D. Fisher, "On grouping for maximum homogeneity," *J. Amer. Statist. Assoc.*, Vol. 53 (1958), pp. 789–798.
[6] Evelyn Fix and J. L. Hodges, Jr., "Discriminatory Analysis," USAF Project Report, School of Aviation Medicine, Project Number 21-49-004, No. 4 (1951).
[7] Edward Forgy, "Cluster analysis of multivariate data: efficiency vs. interpretability of classifications," abstract, *Biometrics*, Vol. 21 (1965), p. 768.
[8] Paul R. Halmos, *Measure Theory*, New York, Van Nostrand, 1950.
[9] J. MacQueen, "The classification problem," Western Management Science Institute Working Paper No. 5, 1962.
[10] ———, "On convergence of k-means and partitions with minimum average variance," abstract, *Ann. Math. Statist.*, Vol. 36 (1965), p. 1084.
[11] Jacob Marschak, "Towards an economic theory of organization and information," *Decision Processes*, edited by R. M. Thrall, C. H. Coombs, and R. C. Davis, New York, Wiley, 1954.
[12] ———, "Remarks on the economics of information," Proceedings of the scientific program following the dedication of the Western Data Processing Center, University of California, Los Angeles, January 29–30, 1959.
[13] Emanuel Parzen, "On uniform convergence of families of sequences of random variables," *Univ. California Publ. Statist.*, Vol. 2, No. 2 (1954), pp. 23–54.
[14] George S. Sebestyen, *Decision Making Process in Pattern Recognition*, New York, Macmillan, 1962.
[15] Robert R. Sokal and Peter H. Sneath, *Principles of Numerical Taxonomy*, San Francisco, Freeman, 1963.
[16] Joe Ward, "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, Vol. 58 (1963), pp. 236–244.