ASYMPTOTIC PROPERTIES OF UNIVARIATE SAMPLE K-MEANS CLUSTERS

M. Anthony Wong
Massachusetts Institute of Technology

Working Paper #1341-82

ASYMPTOTIC PROPERTIES OF UNIVARIATE SAMPLE K-MEANS CLUSTERS

M. Anthony Wong
Massachusetts Institute of Technology

Working Paper #1341-82

Key words and phrases: k-means clusters, within-clusters sum of squares, cluster lenghths, non-standard asymptotics.

AMS 1980 subject classification: Primary 62H30.

ABSTRACT

A random sample of size N is divided into k clusters that minimize the within clusters sum of squares locally. Some large sample properties of this k-means clustering method (as k approaches ∞ with N) are obtained. In one dimension, it is established that the sample k-means clusters are such that the within-cluster sums of squares are asymptotically equal, and that the sizes of the cluster intervals are inversely proportional to the one-third power of the underlying density at the midpoints of the intervals. The difficulty involved in generalizing the results to the multivariate case is mentioned.

1. INTRODUCTION

Let the univariate observations $x_1, x_2, \ldots, x_N$ be sampled from a distribution F with density function f. Suppose that these observations are partitioned into k groups with means $\bar{z}_1, \ldots, \bar{z}_k$ such that no movement of an observation from one group to another will reduce the within groups sum of squares

$$WSS_N = \sum_{i=1}^{N} \min_{1 \le j \le k} || x_i - \bar{z}_j ||^2.$$

This method for division of a sample into k groups to minimize the within groups sum of squares locally is known in the clustering literature as k-means. In one dimension, the partition will be specified by k-1 cutpoints; the observations lying between common cutpoints are in the same group. See Hartigan (1975) for a detailed description of the k-means method, and see Hartigan and Wong (1979) for an efficient computational algorithm. This method has been widely used in various clustering applications (see Blashfield and Aldenderfer, 1978). Its asymptotic properties (when $N \to \infty$) for fixed k have been studied by MacQueen (1967), Hartigan (1978), and Pollard (1979). Here, the sampling properties of k-means clusters when k approaches ∞ with N are presented.

The properties of the univariate population k-means clusters when $k \to \infty$ are given in Wong (1982a). It is shown that for large k, the optimal population partition is such that the within-cluster sums of squares are equal, and that the sizes of the cluster intervals are inversely proportional to the one-third power of the underlying density at the midpoints of the intervals. In this paper, non-standard asymptotics are used to obtain the asymptotic properties (when $k \to \infty$ with N) of the locally optimal k-means clusters for samples from a general population F on $[0,1]$; in particular, it is shown that the locally optimal partition approaches the population optimal partition under certain regularity conditions. The special difficulties in showing this result are: (1) the number of clusters k approaches ∞ with N (such that the length of each cluster interval approaches zero while the number of observations in each cluster approaches infinity), and (2) the locations and sizes of the cluster intervals are determined by an optimization procedure (so all results concerning the clusters need to hold uniformly for all clusters).

Theorem 1 and Theorem 2, respectively, gives the asymptotic expression for the lengths and the within-cluster sums of squares of the locally optimal k-means clusters. The result of Theorem 1 is obtained in Section 2, and Theorem 2 is proved in Section 3. Some concluding remarks are given in Section 4, in which the difficulties involved in generalizing these univariate results to many dimensions are also mentioned.

2. ASYMPTOTIC LENGTHS OF THE LOCALLY OPTIMAL SAMPLE K-MEANS CLUSTERS

Let $x_1, x_2, \ldots, x_N$ be a random sample from a density f which is positive and has four bounded derivatives in [0,1]. Denote the ith derivative of f at x by $f^{(i)}(x)$, and let $B = \sup_{0\le x\le 1} f(x)$ and $b = \inf_{0\le x\le 1} f(x)$. Suppose that the N observations are grouped into k_N clusters with means $\bar{z}_1<\bar{z}_2<\ldots<\bar{z}_{k_N}$ so that the within clusters um of squares of this locally optimal k_N-partition cannot be decreased by moving any single observation from its present cluster to any other cluster. Denote the pth order statistic by $x_{(p)}$, and let n_j be the number of observations in the jth cluster. Then $x_{(\sum_{t=0}^{j-1} n_t+1)} < \ldots <$ $x_{(\sum_{t=0}^{j} n_t)}$ are the observations in the jth cluster, where $n_0=0$. And the length e_j and the midpoint m_j of the jth cluster are defined to be $[x_{(\sum_{t=o}^{j} n_t+1)} - x_{(\sum_{t=0}^{j-1} n_t)}]$ and $1/2\,[x_{(\sum_{t=0}^{j} n_t+1)} + x_{(\sum_{t=0}^{j-1} n_t)}]$ respectively, where $x_{(0)} = 0$ and $x_{(N+1)} = 1$.

There may be many locally optimal k-means partitions. Theorem 1 states that for any such partition, if $k_N = o([N/\log N]^{1/3})$, then

$$\max_{1\le j\le k_N} |k_N e_j f(m_j)^{1/3} - \int_0^1 f(x)^{1/3}dx| = o_p(1).$$

To show the result of Theorem 1, we need a few lemmas. Lemma 1 is a direct consequence of a theorem of large deviations given in Feller (1971). It is useful in proving Lemma 2 which states that if the n_j's are large enough, the cluster means are suitably close to the cluster midpoints. This result is then used in the proof of Theorem 1 to determine the relationship between the lengths of neighboring clusters.

Lemma 1:

Let $x_1, x_2, \ldots, x_N$ be a random sample from a distribution F with density f which is positive and has four bounded derivatives in $[0,1]$. Put $B = \sup_{0 \le x \le 1} f(x)$ and $b = \inf_{0 \le x \le 1} f(x)$. Denote the n observations contained in an open subinterval I of $[0,1]$ by $z_1, z_2, \ldots, z_n$, and let $E[z_i - u_I] = 0$ and $E[(z_i - u_I)^2] = \sigma_I^2$.

Then there exist constants C, D, and N_0 not depending on u_I and σ_I (or I) such that if $N \ge N_0$ and $n \ge N(\log N/N)^{1/3} b/16$,

$$P_r\{\sigma_I^{-1} n^{1/2} \left|\frac{z_1 + \ldots + z_n}{n} - u_I\right| \ge C (\log N)^{1/2}\} \le DN^{-2} (\log N)^{-1/2}.$$

Proof:

From the theorem of large deviations given in Feller (1971, p. 549), since $(4\log N)^{1/2}\, n^{-1/6} \to 0$ as $n \to \infty$, we have

$$P_r\{\sigma_I^{-1} n^{1/2} \left|\frac{z_1 + \ldots + z_n}{n} - u_I\right| \ge (4\log N)^{1/2}\}/(2\pi)^{-1/2}(4\log N)^{-1/2} N^{-2} \to 1$$

as $n \to \infty$.

Now since f has bounded derivatives and $\sigma_I^{-1} n^{1/2} \left|\frac{z_1 + \ldots + z_n}{n} - u_I\right|$ has a distribution not depending on u_I and σ_I (or I), the lemma follows. (In the proof of Theorem 1, it will be shown that when N is large enough and if $k_N = o([N/\log N]^{1/3})$, then each of the k_N clusters contains at least $N(\log N/N)^{1/3} b/16$ observations, and hence the result of this lemma can be applied.)

In the application of Lemma 1, n will be the number of observations in an open subinterval I of $[0,1]$. Therefore, n is approximately $NF(I)$, where $F(I) = \int_I dF$. More precisely, using Donsker's theorem for empirical processes (see Billingsley 1968, p. 141), we have

$$\sup_I | n_I - NF(I) | = 0_p(N^{1/2}) \tag{2.1}$$

where n_I is the number of observations in I, and the sup is taken over all open subintevals of I. Both Lemma 1 and Equation (2.1) will be used in the proof of Lemma 2 to give a uniform estimate of the deviations between cluster means and midpoints for those clusters that are large enough.

Lemma 2:

Let $x_1, x_2, \ldots, x_N$ be a random sample from F whose density f is positive and has four bounded derivatives in $[0,1]$. Put $B = \sup_{o \leq x \leq 1} f(x)$, $b = \inf_{0 \leq x \leq 1} f(x)$, and $F(I) = \int_I dF$. Let $\bar{z}_I$ be the mean of the observations in an open interval I, $u_I = \int_I xdF/F(I)$ be the conditional mean of F on I, and s_I be the size of the interval I. Then there exists a constant C_0 such that $P_r\{\sup_I s_I^{-1/2}|\bar{z}_I - u_I| \leq C_0 \,(\log N/N)^{1/2}\} = 1 - o(1)$, where the sup is taken over all open intervals I (whose boundary points are order statistics) containing at least $N(\log N/N)^{1/3}b/16$ observations.

Proof:

For any $N \geq N_0$, consider an interval I of the form

$(x_{(p)}, x_{(p+n_I+1)})$, where $n_I \geq N\,(\log N/N)^{1/3} b/16$.

Using Lemma 1 (first conditioning on the two order statistics and then integrating out), we obtain

$$P_r\{\sigma_I^{-1}\, n_I^{1/2} |\, \bar{z}_I - u_I\,| \geq C(\log N)^{1/2}\} \leq DN^{-2}\,(\log N)^{-1/2}, \qquad (2.2)$$

where $\sigma_I^2 = \int_I (x - u_I)^2\, dF/F(I)$ = conditional variance of F on I. Now, by the Taylor series expansion of f, $f(x) = f(m_I) + (x - m_I)\, f^{(1)}(m_I) + \frac{1}{2}(x-m_I)^2 f^{(2)}(q_x)$ for all x in I, where m_I is the midpoint of I and q_x is between x and m_I. Therefore,

$$F(I) = f(m_I)\, s_I\, [1+0(s_I^2)]$$

$$u_I = m_I + \frac{1}{12} \cdot \frac{f^{(1)}(m_I)}{f(m_I)} \cdot s_I^2 + 0(s_I^4), \text{ and}$$

$$\sigma_I^2 = \frac{1}{12}\, s_I^2\, [1 + 0(s_I^2)]\ .$$

(Note that the constant in the 0 term depends on the bound of the second derivative of f).

It follows that (2.2) can be written as:

$$P_r\{\sqrt{12}\ \ s_I^{-1}\, [1 + 0(s_I^2)]\ \ n_I^{1/2}\, |\, \bar{z}_I - u_I| \geq C(\log N)^{1/2}\}$$
$$\leq DN^{-2}\,(\log N)^{-1/2}$$

Since the number of possible intervals I is bounded by N^2, we have

$$P_r \{ \sup_I s_I^{-1} [1 + O(s_I^2)] n_I^{1/2} |\bar{z}_I - u_I| \leq \frac{1}{\sqrt{12}} C (\log N)^{1/2} \}$$

$$\geq 1 - D(\log N)^{-1/2} = 1 - o(1).$$

Now from (2.1), we have uniformly in I, $n_I \geq \frac{1}{2} NF(I) \geq \frac{1}{2} Nb\, s_I$ with probability tending to one.

Therefore,

$$P_r \{ \sup_I s_I^{-1/2} |\bar{z}_I - u_I| \leq C_0 (\log N/N)^{1/2} \} \to 1 \text{ as } N \to \infty,$$

where $C_0 = \frac{2}{\sqrt{6}} b^{-1/2} C$, and the sup is taken over all intervals of the form $(x_{(p)}, x_{(p+n_I+1)})$ with $n_I \geq N (\log N/N)^{1/3} b/16$.

The result of Lemma 2 shows that if the k-means clusters are large enough, the cluster means are suitably close to the cluster midpoints. When combined with some well-known properties of locally optimal k-means partitions, this result is useful in establishing the relationship between the lengths of neighboring clusters. The main difficulty to be overcome in the proof of Theorem 1 is showing that all the clusters are large enough.

Theorem 1:

Let $x_1, x_2, \ldots, x_N$ be a random sample from a distribution F whose density f is positive and has four bounded derivatives in [0,1]. Put $B = \sup_{0 \leq x \leq 1} f(x)$ and $b = \inf_{0 \leq x \leq 1} f(x)$, let e_j $(j = 1, 2, \ldots, k_N)$ be the length of the jth cluster of a locally optimal k_N-partition of the sample. Then, provided that $k_N = o([N/\log N]^{1/3})$, we have

$$\max_{1 \le j \le k_N} | k_N e_j f(m_j)^{1/3} - \int_0^1 f(x)^{1/3} dx | = o_p(1),$$

where m_j is the midpoint of the jth cluster.

Proof:

Consider a locally optimal k_N-partition with $k_N = o([N/\log N]^{1/3})$. Denote the open interval (whose boundary points are order statistics) containing the jth cluster by I_j. Then, as before, we have

$$u_j = \int_{I_j} x \, dF / F(I_j) = m_j + \frac{1}{12} \cdot \frac{f^{(1)}(m_j)}{f(m_j)} \cdot e_j^2 + O(e_j^4) \qquad (2.3)$$

The proof is in three parts. In part I, it is shown that if a cluster is of length bounded by $2(B/b)^{1/3}/k_N$ and $1/(2k_N)$, then both it and its neighboring clusters contain at least $N(\log N/N)^{1/3}$ b/16 observations. In part II, using the result of Lemma 2, the relationship between the lengths of neighboring clusters is established; a bound of the ratio $\frac{e_{j-1}}{e_j} \cdot (\frac{f(m_{j-1})}{f(m_j)})^{1/3}$ is given by $1 + k_N^{-1} o_p(1)$. Since at least one of the k_N clusters satisfies $2(B/b)^{1/3}/k_N \ge e_j \ge 1/(2k_N)$, applying parts I and II repeatedly gives the result of this theorem.

To avoid wordiness, statements are to be read as if they included the qualification: "with probability tending to one as N approaches infinity".

[I] Suppose that $2(B/b)^{1/3}/k_N \ge e_j \ge 1/(2k_N)$. Then $F(I_j) \ge b/(2k_N)$. By (2.1), the jth cluster contains at least $Nb/(2k_N) - O_p(N^{1/2})$ observations. Therefore, the number of observations in the jth cluster exceeds $Nb/4k_N$ eventually. Since $k_N = o([N/\log N]^{1/3})$, this number exceeds $N(\log N/N)^{1/3}$ b/4.

Applying Lemma 2 to the jth cluster, we have

$$|\bar{z}_j - u_j| \leq C_0 \ (\log N/N)^{1/2} \ e_j^{\ 1/2}, \tag{2.4}$$

where $\bar{z}_j$ is the mean of the observations in the jth cluster. Consider the (j-1)st cluster, a cluster adjacent to the jth cluster. The largest observation in the (j-1)st cluster is $x_{(\sum_{t=0}^{j-1} n_t)}$, and the smallest observation in the jth cluster is $x_{(\sum_{t=0}^{j-1} n_t + 1)}$. Then by local optimality, the midpoint M between $\bar{z}_{j-1}$ and $\bar{z}_j$ must lie between $x_{(\sum_{t=0}^{j-1} n_t)}$ and $x_{(\sum_{t=0}^{j-1} n_t + 1)}$. And

$$e_{j-1} \geq M - \bar{z}_{j-1} = \bar{z}_j - M = \bar{z}_j - x_{(\sum_{t=0}^{j-1} n_t)} + 0_p \ ([\log N/N]),$$

since the largest gap between successive order statistics is $0_p([\log N/N])$. It follows from (2.3) that

$$e_{j-1} \geq (\bar{z}_j - u_j) + \frac{1}{2}e_j + 0_p([\log N/N])$$

And from (2.4), we obtain

$$e_{j-1} \geq \frac{1}{2} \ e_j - C_0 \ [\log N/N]^{1/2} \ e_j^{\ 1/2} + 0_p \ ([\log N/N])$$

$$\geq \frac{1}{4} \ e_j \geq 1/(8k_N), \text{ since } \ e_j \geq 1/(2k_N).$$

Therefore, it follows from (2.1) that the (j-1)st interval contains at least $Nb/(8k_N) - 0_p(N^{1/2}) \geq N \ [\log N/N]^{1/3} \ b/16$ observations eventually.

[II] Now, applying Lemma 2 to the (j-1)st cluster, we have

$$|\bar{z}_{j-1} - u_{j-1}| \leq C_0 \ (\log N/N)^{1/2} \ e_{j-1}^{\ 1/2}. \tag{2.5}$$

Since $M - \bar{z}_{j-1} = \bar{z}_j - M$, we have

$$[m_{j-1} + \frac{1}{2}e_{j-1} + 0_p([\log N/N])] - \bar{z}_{j-1} = \bar{z}_j - [m_j - \frac{1}{2}e_j + 0_p([\log N/N])] \quad (2.6)$$

From (2.3), (2.6) can be written as

$$[u_{j-1} - \frac{1}{12} \cdot \frac{f^{(1)}(m_{j-1})}{f(m_{j-1})} \cdot e^2_{j-1} - 0\ (e^4_{j-1}) + \frac{1}{2}\ e_{j-1}] - \bar{z}_{j-1} = \bar{z}_j -$$

$$[u_j - \frac{1}{12} \cdot \frac{f^{(1)}(m_j)}{f(m_j)} \cdot e^2_j - 0(e^4_j) - \frac{1}{2}\ e_j] + 20_p([\log N/N]) \quad (2.7)$$

Let f^* be the density at the midpoint between m_j and m_{j-1}. Then

$$e_j[1 + \frac{1}{6} \cdot \frac{f^{(1)}(m_j)}{f(m_j)} \cdot e_j + 0(e^3_j)] = e_j\ [(1 - \frac{1}{2} \cdot \frac{f^{(1)}(m_j)}{f(m_j)} \cdot e_j)^{1/3} +$$

$$0(e^2_j)] = e_j\ [f^*/f(m_j)]^{-1/3}\ [1 + 0(e^2_j) + 0_p\ (\log N/N)],$$

since $f^* = f(m_j) - \frac{1}{2}\ f^{(1)}(m_j)\ e_j + 0(e^2_j) + 0_p\ ([\log N/N])$ by the expansion of f about m_j.

Similarly, by the expansion of f about m_{j-1}, we have

$$e_{j-1}[1 - \frac{1}{6} \cdot \frac{f^{(1)}(m_{j-1})}{f(m_{j-1})} + 0(e^2_{j-1})] = e_{j-1}[f^*/f(m_{j-1})]^{-1/3}\ 1 + 0(e^2_{j-1}) +$$

$$2\ 0_p(\log N/N)].$$

Therefore, (2.7) becomes

$$(u_{j-1} - \bar{z}_{j-1}) + \frac{1}{2}\ e_{j-1}\ [f^*/f(m_{j-1})]^{-1/3}[1 + 0(e^2_{j-1})] = (\bar{z}_j - u_j) +$$

$$\frac{1}{2}\ e_j\ [f^*/f(m_j)]^{-1/3}[1 + 0(e^2_j)] + 2\ 0_p\ (\log N/N). \quad (2.8)$$

Combining (2.4), (2.5), and (2.8), we can first show the ratio e_{j-1}/e_j is bounded, and then

$$\left|\frac{e_{j-1}}{e_j} \cdot [f(m_{j-1})/f(m_j)]^{1/3} - 1\right| \leq 4\, C_0\, [f^*/f(m_j)]^{1/3}(\log N/N)^{1/2} e_j^{-1/2} + 2e_j^{-1}\, 0_p(\log N/N) + 0\,(e_j^2) \qquad (2.9)$$

$$\leq k_N^{-1}\, [4\, \sqrt{2}\;\; C_0 (B/b)^{1/3} k_N^{3/2} (\log N/N)^{1/2} + 4k_N^2\, o_p(\log N/N) + 0(k_N^{-1})]$$

$$= k_N^{-1}\, o_p(1);$$

and this bound does not depend on the intervals involved.

[III] From the first inequality in (2.9), we can show by contradiction that at least one of the k_N cluster intervals satisfies $2(B/b)^{1/3}/k_N \geq e_j \geq 1/(2k_N)$. Then using the bound in (2.9) and carrying on at most k_N comparisons of adjacent clusters, we obtain

$$(e_i/e_j) \cdot [f(m_i)/f(m_j)]^{1/3} = [1 + k_N^{-1}\, o_p(1)]^{k_N} = 1 + o_p(1)$$

uniformly in $1 \leq i, j \leq k_N$.

Since $\sum_1^{k_N} e_j\; f(m_j)^{1/3} \to \int_0^1 f(x)^{1/3}\, dx$ as $N \to \infty$, the theorem follows.

3. ASYMPTOTIC WITHIN-CLUSTER SUMS OF SQUARES OF THE LOCALLY OPTIMAL SAMPLE K-MEANS CLUSTERS

As in Section 2, let $x_{(\sum_{t=0}^{j-1} n_t + 1)} < \ldots < x_{(\sum_{t=0}^{j} n_t)}$ denote the n_j observations in the jth cluster of a locally optimal k-means partition of a sample from f, and let $\bar{z}_j$ be the jth cluster mean. The within-cluster sum of squares of the jth cluster, WSS_j, is defined to be $\sum_{p=1}^{n_j} (x_{(\sum_{t=0}^{j-1} n_t + p)} - \bar{z}_j)^2$. In this section, we will show that the within-cluster sums of squares of the k_N clusters are asymptotically equal. First, a direct consequence of Feller's theorem on large deviations (Lemma 3) is used in the proof of Lemma 4 to obtain a uniform estimate of the within-cluster sum of squares, which is a function of the length of the cluster interval and the density at the midpoint of the interval. The result of Theorem 2 then follows from Theorem 1 and Lemma 4.

Lemma 3:

Let $x_1, x_2, \ldots, x_N$ be a random sample from a distribution F with density f which is positive and has four bounded derivatives in $[0,1]$. Put $B = \sup_{0 \le x \le 1} f(x)$ and $b = \inf_{0 \le x \le 1} f(x)$. Denote the n observations contained in an open interval I by $z_1, z_2, \ldots, z_n$, and let $E[z_i - u_I] = 0$, $E[(z_i - u_I)^2] = \sigma_I^2$, and $E[(z_i - u_I)^4] = \gamma_I < \infty$. Then there exist constants C^*, D^*, and N_0^*, not depending on I such that if $N \ge N_0^*$ and $n \ge N(\log N/N)^{1/3} b/16$,

$$P_r\{\gamma_I^{-1/2} n^{-1/2} \left| \sum_{i=1}^{n} (z_i - \bar{z})^2 - n\sigma_I^2 \right| \ge C^*(\log N)^{1/2}\} \le D^* N^{-2}(\log N)^{-1/2}.$$

(Since the proof is similar to that of Lemma 1, it will not be given here.)

Lemma 4:

Let $x_1, x_2, \ldots, x_N$ be a random sample from F whose density f is positive and has four bounded derivatives in [0,1]. Put $B = \sup_{0 \le x \le 1} f(x)$, $b = \inf_{0 \le x \le 1} f(x)$, and $F(I) = \int_I dF$. Let $\bar{z}_I$ be the mean of the observations in an open interval I, WSS_I be the within-cluster sum of squares of the observations in I, m_I be the midpoint of I, and s_I be the size of the interval I. Then there exists a constant C* such that

$$P_r \{\sup_I s_I^{-5/2} | WSS_I - \frac{1}{12} N f(m_I) s_I^3 [1 + O(s_I^2)]| \le C_0^* N(\log N/N)^{1/2}\}$$

$$= 1 - o(1),$$

where the sup is taken over all open intervals I (whose boundary points are order statistics) containing at least $N(\log N/N)^{1/3} b/16$ observations.

The proof of this lemma is similar to that of Lemma 2; therefore, only the differences between the two proofs are outlined here. In this proof, the Taylor series expansion is carried out to the fourth order terms. And after some series manipulations, we obtain

$$\sigma_I^2 = \int_I (x - u_I)^2 dF/F(I) = \frac{1}{12} s_I^2 [1 + O(s_I^2)], \tag{3.1}$$

and $$\gamma_I^2 = \int_I (x - u_I)^4 dF/F(I) = \frac{1}{180} s_I^4 [1 + O(s_I^2)]. \tag{3.2}$$

Applying (3.1) and (3.2) to the result of Lemma 3 will give this lemma because the number of possible intervals I is again bounded by N^2.

Theorem 2:

Let $x_1, x_2, \ldots, x_N$ be a random sample from a distribution F whose density f is positive and has four bounded derivatives in [0,1]. Put $B = \sup_{0 \le x \le 1} f(x)$ and $b = \inf_{0 \le x \le 1} f(x)$. Let WSS_j $(j = 1, 2, \ldots, k_N)$ be the within-cluster sum of squares of the jth cluster of a locally optimal k_N-partition of the sample. The provide that $k_N = o_p([N/\log N]^{1/3})$, we have

$$\max_{1 \le j \le k_N} \left| 12N^{-1}k_N{}^3 \, WSS_j - (\textstyle\int_0^1 f(x)^{1/3}dx)^3 \right| = o_p(1).$$

Proof:

Consider a locally optimal k_N-partition with $k_N = o([N/\log N]^{1/3})$. It is shown in Theorem 1 that for all N large enough, with probability tending to one, (i) the number of observations in each cluster exceeds $N(\log N/N)^{1/3}b/16$, and (ii) $e_j = k_N^{-1} f(m_j)^{-1/3} G[1+o_p(1)]$, where $G = \int_0^1 f(x)^{1/3} dx$. From (i), we can apply Lemma 4 to obtain $e_j{}^{-5/2} \left| WSS_j - \frac{1}{12} Nf(m_j) e_j^3 [1+O(e_j^2)] \right| \le C_0^* N(\log N/N)^{1/2}$, uniformly in $1 \le j \le k_N$. It follows from (ii) that, we have uniformly in $1 \le j \le k_N$,

$$|WSS_j - \frac{1}{12} Nk_N^{-3} G^3 [1+o_p(1)]| \le 2 C_0^* N(\log N/N)^{1/2} k_N{}^{-5/2} b^{-5/6} G^{5/2}.$$

Therefore,

$$|12N^{-1}k_N^3 WSS_j - G^3| \le o_p(1) + C^{**} k_N{}^{1/2} (\log N/N)^{1/2} = o_p(1),$$

where $C^{**} = 2C_0^* b^{-5/6} G^{5/2}$. And the theorem is proved.

Since the globally optimal k-means partition of the sample is necessarily locally optimal, the results of Theorem 1 and Theorem 2 also apply to the globally optimal k_N-partition. Moreover, the generalization of these results to densities with finite support [a,b] is immediate.

4. DISCUSSION

In this paper, the properties of univariate sample k-means clusters when k approaches infinity with the sample size N are presented. This study is of interest in its own right because we want to examine the sampling properties of this widely-used clustering method. The result in Section 2 also indicates that the k-means method would partition a sample from a distribution with density f on $[a,b]$ in such a way that the sizes of the cluster intervals are adaptive to f; the intervals are large when the density is low while the intervals are small where the density is high. Therefore, k-means can potentially be used as a method for constructing variable-cell histograms. Indeed, a histogram estimate of f based on the k-means method is proposed in Wong (1982b); and it is shown to be uniformly consistent in probability.

The multivariate case requires further investigation as the generalization of univariate results to many dimensions is not straightforward. An important first step is to determine the configuration of the optimal population k-means partition in several dimensions. There is some indication that the best partition in R^2 is into regular hexagons (Wong 1982c); but it is not clear that the best partition is into regular polytopes for higher dimensions.

REFERENCES

Billingsley, P. (1968). Convergence of Probability Measures. New York: John Wiley and Sons.

Blashfield, R.K., and Aldenderfer, M.S. (1978). The literature on cluster analysis. Multivariate Behavioral Research, 13, 271-95.

Feller, W. (1971). An Introduction to Probability Theory and Its Applications. Volume II, New York: John Wiley and Sons, 549-553.

Hartigan, J.A. (1975). Clustering Algorithms. New York: John Wiley and Sons.

_________(1978). Asymptotic distributions for clustering criteria. Annals of Statistics, 6, 117-131.

_________, and Wong, M.A. (1979). Algorithm AS136: A k-means clustering algorithm. Applied Statistics, 28, 100-108.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Probability and Statistics, 281-97.

Pollard, D. (1981). Strong consistent of k-means clustering. Annals of Statistics, 9, 135-140.

Wong, M.A. (1982a). Asymptotic properties of univariate population k-means clusters. Unpublished manuscript, Sloan School of Management, M.I.T.

_________(1982b). Using the k-means clustering method as a density estimation procedure. Unpublished manuscript, Sloan School of Management, M.I.T.

_________(1982c). Asymptotic properties of bivariate k-means clusters. Communications in Statistics, Vol. A11, No. 10 (to appear).

Sloan School of Management
Massachusetts Institute of Technology, E53-335
Cambridge, MA 02139
U.S.A.