WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

ASYMPTOTIC PROPERTIES OF UNIVARIATE POPULATION K-MEANS CLUSTERS

M. Anthony Wong
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

Working Paper #1339-82

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

ASYMPTOTIC PROPERTIES OF UNIVARIATE

POPULATION K-MEANS CLUSTERS

M. Anthony Wong
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

Working Paper #1339-82

Key Words and Phrases: population k-means clusters; within-cluster sums of squares; cluster lengths.

ABSTRACT

Let f be a density function defined on the closed interval [a, b]. The k-means partition of this interval is defined to be the k-partition with the smallest within cluster sum of squares. The properties of this k-means partition when k becomes large will be obtained in this paper. The results suggest that the k-means clustering procedure can be used to construct a variable-cell histogram estimate of f using a sample of observations taken from f.

1. INTRODUCTION

Let the univariate observations $X_1, X_2, \ldots, X_N$ be sampled from a distribution F with density function F. In cluster analysis, the k-means clustering method (see Hartigan (1975), Chapter 4) is often used to partition the sample of N observations into k clusters with means $U_1, \ldots, U_k$. The resultant clusters satisfy the property that no movement of an observation from one cluster to another reduces the sample within cluster sum of squares

$$WSS_N = \sum_{i=1}^{N} \min_{1 \le j \le k} || X_i - U_j ||^2.$$

For these sample k-means clusters, if I_j is used to denote the interval containing all points in R^1 closer to U_j than to any other cluster means, then $\{I_1, \ldots, I_k\}$ defines a k-partition of the sampled space. The corresponding k-means partition in the population F is defined by the k-population means $m_1, \ldots, m_k$, which are selected in such a way that the within cluster (or interval) sum of squares

$$WSS = \int \inf_{1 \le j \le k} || x - m_j ||^2 \, dF$$

is minimized.

The k-means method has been widely used in clustering applications (see Blashfield and Aldenderfer, 1978), and the efficient computational algorithm given in Hartigan and Wong (1979) has been included in the multivariate programs BMDPKM of the BMDP statistical package. The properties of <u>sample</u> k-means clusters have also been studied by several investigators. In Fisher (1958), and Fisher and Van Ness (1971), it is shown that k-means clusters are convex, i.e., if an observation is a weighted average of observations in a cluster, the observation is also in the cluster. And the asymptotic convergence (as $N \to \infty$) of the sample k-means clusters to the population k-means cluster for fixed number of clusters k has been studied by MacQueen (1967), Hartigan (1978), and Pollard (1981), in which conditions that ensure the almost sure convergence of the set of means of the k-means clusters can be found. However, little work have been done in examining the properties of population k-means clusters, especially when k becomes large. In Dalenius (1951), it is shown that the cut-point between neighboring population clusters is the average of the means in the clusters, and in Cox (1957), the cut-points for the k-means clusters in the standard normal distribution are given for $k = 1, 2, \ldots, 6$.

In this paper, the asymptotic properties (as k becomes large) of the population k-means clusters in one dimension are obtained. It is shown in Section 2 that the optimal population partition is such that the within cluster sums of squares of the k cluster intervals 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 implications of these results are discussed in Section 3.

2. ASYMPTOTIC PROPERTIES OF POPULATION K-MEANS CLUSTERS

Let $f(x)$ be a density function defined on the interval $[a,b]$, and denote the ith derivative of f at x by $f^{(i)}(x)$. Let the k-partition of $[a,b]$ specified by the k-1 cutpoints $a < y_1 < y_2 < \dots < y_{k-1} < b$ be the k-partition with the smallest within cluster sum of squares

$$WSS = \sum_{i=1}^{k} WSS_i = \sum_{i=1}^{k} \int_{y_{i-1}}^{y_i} (x - m_i)^2 f(x)\, dx,$$

where $a = y_o$, $b = y_k$, and

$$m_i = \int_{y_{i-1}}^{y_i} x\, f(x)\, dx \Big/ \int_{y_{i-1}}^{y_i} f(x) dx.$$

In this section, we will describe the properties of this k-means partition of a finite interval [a,b] as the number of cluster intervals (or cells) becomes large.

<u>Theorem</u>: Let $f(x)$ denote a density function on the interval [a,b]. And let $a = y_{0k} < y_{1k} < \dots < y_{(k-1)k} < y_{kk} = b$ be the cutpoints specifying the k-means partition of [a,b]. If f is positive and has four bounded derivatives in [a,b], then we have uniformly in $1 \leq i \leq k$,

$$k\, e_{ik}\, f_{ik}^{1/3} \to \int_a^b [f(x)]^{1/3} dx \tag{2.1}$$

$$k\, p_{ik}\, f_{ik}^{-2/3} \to \int_a^b [f(x)]^{1/3} dx \tag{2.2}$$

$$k^3 WSS_{ik} \to [\int_a^b [f(x)]^{1/3} dx]^3/12 \tag{2.3}$$

as $k \to \infty$,

where $e_{ik} = y_{ik} - y_{(i-1)k}$

$$f_{ik} = f\,(1/2\, y_{ik} + 1/2\, y_{(i-1)k})$$

$$p_{ik} = \int_{y_{(i-1)k}}^{y_{ik}} f(x)\, dx$$

and $WSS_{ik} = \int_{y_{(i-1)k}}^{y_{ik}} [x - \int_{y_{(i-1)k}}^{y_{ik}} x\, f(x) dx/p_{ik}]^2\, f(x) dx.$

(The theorem states that, for large k, the within cluster sums of squares of the k intervals are nearly equal; it follows that the length of the interval containing a point x of density $f(x)$ is proportional to $f(x)^{-1/3}$.)

Proof: The proof is in four parts.

(I) The k-partition of [a,b] consisting of k equal intervals has a within cluster sum of squares of order k^{-2}; the contribution from the ith interval to the optimal within cluster sum of squares is of order e_{ik}^3. Therefore, $\sum_{i=1}^{k} e_{ik}^3 = O(k^{-2})$, which implies that $\sup_i e_{ik} = O(k^{-2/3})$. To avoid complexity of notation, the k's indexing partition will be dropped.

(II) In this part of the proof, it will be shown that lengths of neighboring clusters are of the same order of magnitude. Let m_i be the mean of the ith interval. Then

$$m_i = \int_{y_{i-1}}^{y_i} x\, f(x)dx / \int_{y_{i-1}}^{y_i} f(x)dx.$$

Consider any two neighboring intervals e_j and e_{j+1}. By the optimality of the partition, as is shown in Dalenius (1951),

$$y_j - m_j = m_{j+1} - y_j.$$

Thus, $e_j \geq y_j - m_j$

$$= m_{j+1} - y_j$$

$$= \int_{y_j}^{y_{j+1}} x\, f(x)\, dx / \int_{y_j}^{y_{j+1}} f(x)\, dx - y_j$$

$$= \int_{o}^{e_{j+1}} x\, f(x + y_j)\, dx / \int_{o}^{e_{j+1}} f(x + y_j)\, dx$$

$$\geq \frac{1}{2} \cdot \frac{M_1}{M_u} \cdot e_{j+1}, \text{ where } M_1 = \inf_{a \leq x \leq b} f(x) \text{ and}$$

$$M_u = \sup_{a \leq x \leq b} f(x).$$

Similarly, $e_{j+1} \geq \frac{1}{2} \cdot \frac{M_1}{M_u} \cdot e_j.$

(III) We will now establish the asymptotic relationship between the lengths of neighboring intervals. Denote the center of the ith interval by C_i $(i = 1, \ldots, k)$. It follows that $C_i = y_{i-1} + \frac{1}{2} e_i$. Using the Taylor series expansion, we have, for any x in the ith interval, $f(x) = f(C_i) + (x-C_i) \cdot f^{(1)}(C_i) + \frac{1}{2}(x-C_i)^2 \cdot f^{(2)}(C_i) + \frac{1}{6}(x-C_i)^3 \cdot f^{(3)}(C_i) + \frac{1}{24}(x-C_i)^4 \cdot f^{(4)}(q_x)$, where q_x is between x and C_i. Since the first four derivatives are bounded on [a,b], it follows from the above series expansion that we have simultaneously for all $1 \le i \le k$,

$$p_i = \int_{y_{i-1}}^{y_i} f(x)dx = e_i[f(C_i) + \frac{1}{24} f^{(2)}(C_i)e_i^{\ 2} + O(e_i^{\ 4})], \qquad (2.4)$$

and

$$\int_{y_{i-1}}^{y_i} xf(x)\,dx = e_i[C_i f(C_i) + \frac{1}{12} f^{(1)}(C_i)e_i^{\ 2} + \frac{1}{24} C_i\, f^{(2)}(C_i)\cdot e_i^{\ 2} + O(e_i^{\ 4})] \qquad (2.5)$$

(Note that the universal bound contained in the O term depends only on the various bounds of the derivatives of f and is independent of i.)

Therefore,

$$m_i = \int_{y_{i-1}}^{y_i} x\, f(x)dx/p_i = C_i + \frac{1}{12} \cdot \frac{f^{(1)}(C_i)}{f(C_i)} e_i^{\ 2} + O(e_i^{\ 4}) \qquad (2.6)$$

Since the partition is optimal, we have simultaneously for all $1 \le i \le k$, $(C_i + \frac{1}{2} e_i) - m_i = m_{i+1} - (C_{i+1} - \frac{1}{2} e_{i+1})$, which when combined with (2.6) gives

$$e_i - \frac{1}{6} \cdot \frac{f^{(1)}(C_i)}{f(C_i)} e_i^{\ 2} + O(e_i^{\ 4}) = e_{i+1} + \frac{1}{6} \cdot \frac{f^{(1)}(C_{i+1})}{f(C_{i+1})} e_{i+1}^2 + O(e_{i+1}^4).$$

Since it has been shown in part [II] that e_i and e_{i+1} are of the same order of magnitude, we have for all $1 \le i \le k$,

$$e_{i+1} + \frac{1}{6} \cdot \frac{f^{(1)}(C_{i+1})}{f(C_{i+1})} e_{i+1}^2 = e_i - \frac{1}{6} \cdot \frac{f^{(1)}(C_i)}{f(C_i)} e_i^2 + O(e_i^4).$$

It follows that

$$e_{i+1} = e_i \{1 - \frac{1}{6} (\frac{f^{(1)}(C_i)}{f(C_i)} e_i + \frac{f^{(1)}(C_{i+1})}{f(C_{i+1})} \cdot \frac{e_{i+1}^2}{e_i} + O(e_i^3)\}.$$

After some Taylor series manipulation, we have

$$e_{i+1}/e_i = [f(C_{i+1})/f(C_i)]^{-1/3} \cdot [1 + O(e_i^2)]. \tag{2.7}$$

Moreover, since it can be shown from (2.4), (2.5), and (2.6) that

$$WSS_i = \int_{y_{i-1}}^{y_i} (x-m_i)^2 f(x)dx = \frac{1}{12} f(C_i)e_i^3[1 + O(e_i^3)],$$

and from (2.4), $p_i = f(C_i)\ e_i\ [1 + O(e_i^2)]$, we obtain from 2.7 that

$$WSS_{i+1}/WSS_i = 1 + O(e_i^2) \tag{2.8}$$

and $$p_{i+1}/p_i = [f(C_{i+1})/f(C_i)]^{2/3} [1 + O(e_i^2)]. \tag{2.9}$$

[IV] Finally, we will now establish the relationship between e_i and e_j for <u>any</u> $1 \le i < j \le k$. It follows from (2.7) that for any pair of values of $1 \le i < j \le k$,

$$e_i/e_j = [f(C_i)/f(C_j)]^{-1/3}\{[1+0(e_i^2)][1+0(e_{i+1}^2)]\cdots[1+0(e_j^2)]\}.$$

But it has been shown in part [I] that $\sup_i e_i = 0(k^{-2/3})$.

$$\text{Hence,} \quad e_i/e_j = [f(C_i)/f(C_j)]^{-1/3}\,[1+0(k^{-4/3})]^k$$

$$= [f(C_i)/f(C_j)]^{-1/3}\,[1+0(k^{-1/3})]$$

for all $1 \le i < j \le k$, which implies that we have uniformly in $1 \le i < j \le k$,

$$(e_i/e_j) \cdot [f(C_i)/f(C_j)]^{1/3} \to 1 \text{ as } k \to \infty.$$

Since $\sum_{i=1}^{k} e_i\, f(C_i)^{1/3} \to \int_a^b f(x)^{1/3}\,dx$, (2.1) follows.

Similarly, from (2.8) and (2.9), we have uniformly in $1 \le i < j \le k$.

$$WSS_i/WSS_j \to 1, \text{ and } (p_i/p_j)\cdot[f(C_i)/f(C_j)]^{-2/3} \to 1 \text{ as}$$

$k \to \infty$, which in turn gives (2.3) and (2.2) respectively. And the theorem is proved.

3. DISCUSSION

In this paper, our effort is directed towards obtaining the properties of univariate population k-means clusters when k becomes large. The properties given in Section 2 indicate that the lengths of the population k-means intervals (or cells) are adaptive to the underlying density function: the intervals are large when the density is low, while the intervals are small where the density is high. This result suggests that the k-means clustering procedure can be used to construct a variable-cell histogram estimate of an underlying density using a sample of observations taken from that density (see Wong, 1980). Such a density estimation method is of interest because it makes use of the computationally efficient k-means clustering procedure (Hartigan and Wong, 1979) which is also applicable to multivariate data.

REFERENCES

Blashfield, R.K., and Aldenderfer, M.S. (1978), "The Literature on Cluster Analysis," Multivariate Behavioral Research, 13, 271-295.

Cox, D.R. (1957), "Notes on grouping," Journal of the American Statistical Association, 52, 543-547.

Dalenius, T. (1951), "The problem of optimum stratification", Skandinavisk Aktuarietidskrift, 34, 133-148.

Fisher, W.D. (1958), "On grouping for maximum homogeneity," Journal of the American Statistical Association, 53, 789-798.

Fisher, L., and Van Ness, J.N. (1971), "Admissible clustering procedures," Biometrika, 58, 91-104.

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, 281-297.

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

Wong, M.A. (1980), "Asymptotic properties of k-means clustering algorithm as a density estimation procedures," Sloan School of Management Working Paper #2000-80, M.I.T., Cambridge, MA.