The Annals of Statistics
1981, Vol. 9, No. 1, 135–140

STRONG CONSISTENCY OF K-MEANS CLUSTERING[1]

BY DAVID POLLARD

Yale University

A random sample is divided into the k clusters that minimise the within cluster sum of squares. Conditions are found that ensure the almost sure convergence, as the sample size increases, of the set of means of the k clusters. The result is proved for a more general clustering criterion.

1. Introduction. The k-means clustering procedure prescribes a criterion for partitioning a set of points into k groups: to divide points $x_1, x_2, \ldots, x_n$ in $\mathbb{R}^s$ according to this criterion, first choose cluster centres $a_1, a_2, \ldots, a_k$ to minimise

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le j \le k} \| x_i - a_j \|^2,$$

where $\| \cdot \|$ denotes the usual Euclidean norm, then assign each x_i to its nearest cluster centre. In this way, each centre a_i acquires a subset C_i of the x's as its cluster. The mean of the points in C_i must equal a_i – otherwise W_n could be decreased by first replacing a_i by that cluster mean then, if necessary, reassigning some of the x's to new centres. The criterion is, therefore, equivalent to that of minimising the within cluster sum of squares.

In this paper $\{x_1, x_2, \ldots, x_n\}$ is assumed to be a sample of independent observations on some distribution P. Conditions are given that ensure the almost sure convergence of the cluster centres as the sample size increases. This generalises one of the results of Hartigan (1978), who gave a detailed analysis for the splitting of observations in one dimension into two clusters; he proved convergence in probability for the point defining the optimal split.

MacQueen (1967) obtained weaker consistency results for a k-means algorithm that distributes points $x_1, x_2, \ldots$ sequentially amongst k clusters. With this algorithm, the centres are not chosen to minimise W_n; instead, each x_n is assigned to the cluster with the nearest cluster centre, then that centre is shifted to the mean of the enlarged cluster. MacQueen proved that the corresponding W_n converges almost surely; he did not prove convergence of the cluster centres.

Because of the difficulties that can arise from ambiguities in the labelling of the points $x_1, \ldots, x_n$ and the centres $a_1, \ldots, a_k$, it is advantageous to regard W_n as a function of the set of cluster centres and of the empirical measure P_n obtained from the sample by placing mass n^{-1} at each of $x_1, x_2, \ldots, x_n$. That is, the problem is to minimise

$$W(A, P_n) := \int \min_{a \in A} \| x - a \|^2 \, P_n(dx)$$

over all possible choices of the set A containing k (or fewer) points. For each fixed A, a strong law of large numbers (SLLN) argument shows that

$$W(A, P_n) \to W(A, P) := \int \min_{a \in A} \| x - a \|^2 \, P(dx), \qquad \text{a.s.}$$

It might be expected therefore that A_n, the set of optimal cluster centres for the sample, should

Received July 1979; revised October 1979.

[1] This research was supported by the Air Force Office of Scientific Research, Contract No. F49620-79-C-0164.

AMS 1970 *subject classifications.* Primary 62H30; secondary 60F15.

Key words and phrases. Clustering criterion, minimising within cluster sum of squares, k-means, strong consistency, uniform strong law of large numbers.

lie close to $\bar{A}$, the set of centres that minimises $W(\cdot, P)$, provided that $\bar{A}$ is uniquely determined. With an appropriate definition of closeness, this is the result to be proved in the next section. It implies that there is a labelling $a_{n1}, a_{n2}, \ldots, a_{nk}$ of the points in A_n, and a labelling $\bar{a}_1, \ldots, \bar{a}_k$ of the points in $\bar{A}$, such that $a_{ni} \to \bar{a}_i$ a.s. This approach also avoids problems with possible coincidence of two of the cluster centres—a possibility that caused MacQueen (1967) so much trouble.

In practice, finding an A at which $W(\cdot, P_n)$ attains its global minimum involves a prohibitive amount of calculation, except in the one dimensional ($s = 1$) case (Fisher, 1958). This problem has been discussed by Hartigan (1975, Chapter 4; 1978). However, there do exist efficient algorithms (see Hartigan and Wong (1979) or Wong (1979), for example) for finding locally optimal partitions of the sample points into k clusters. I do not know whether the techniques to be developed in this paper can be applied to prove consistency results for these locally optimal partitions.

My method of proof is based on repeated application of the SLLN; the arguments apply to almost all sample points ω. First it is shown that the optimal sample cluster centres A_n eventually lie in some compact region of $\mathbb{R}^s$. The proof of this takes an inductive form, starting from the simple 1-mean case. Once the consistency result for $(k-1)$-means is established, the k-means case is treated by first showing that there exists a large compact ball $B(M)$ that contains at least one point of the optimal A_n, for all n large enough. If this were not so, $W(\cdot, P_n)$ could be decreased (when n is large enough) by moving all of the cluster centres to a single point. Then an even larger compact ball $B(5M)$ is shown to contain all of the points in A_n. Indeed, $B(5M)$ can be made so large that any points of A_n outside $B(5M)$ can be deleted without increasing $W(A_n, P_n)$ by very much. This would give a set of at most $k-1$ points for which, if n is large enough, $W(\cdot, P_n)$ is less than the minimum value for $(k-1)$-means: a contradiction.

The second stage in the proof involves showing that, almost surely, $W(A, P_n) - W(A, P)$ converges to zero uniformly over those subsets of $B(5M)$ containing k or fewer points. The proof of this uniform SLLN is deferred to Section 4. Minimising $W(\cdot, P_n)$ is therefore asymptotically equivalent to minimising $W(\cdot, P_n)$—the desired result.

The proof just outlined will apply to more general clustering criteria. For example, the cluster centres could be chosen to minimise a quantity based on absolute deviations

$$\int \min_{a \in A} \|x - a\| \, P_n(dx),$$

or even a criterion with a robustness appeal:

$$\int \min_{a \in A} \|x - a\| \wedge 1 \, P_n(dx).$$

The main theorem is proved in a generality that includes such possibilities: a general increasing function $\phi(\|x - a\|)$ of the deviations $\|x - a\|$ can be used in defining a within cluster sum of deviations.

2. The consistency theorem. Let $x_1, x_2, \ldots$ be independent $\mathbb{R}^s$-valued random variables with common distribution P. Write P_n for the corresponding empirical measure. The sample $\{x_1, x_2, \ldots, x_n\}$ is to be divided into k clusters by minimising a within clusters sum of deviations; a consistency result for the cluster centres is to be proved.

For each probability measure Q on $\mathbb{R}^s$ and each (finite) subset A of $\mathbb{R}^s$ define

$$\Phi(A, Q) := \int \min_{a \in A} \phi(\|x - a\|) \, Q(dx)$$

and

$$m_k(Q) := \inf\{\Phi(A, Q) : A \text{ contains } k \text{ or fewer points}\}.$$

For a given k, the set $A_n = A_n(k)$ of optimal sample cluster centres is to be chosen to satisfy $\Phi(A_n, P_n) = m_k(P_n)$; the optimal population cluster centres $\bar{A} = \bar{A}(k)$ satisfy $\Phi(\bar{A}, P) = m_k(P)$. The aim is to show that $A_n \to \bar{A}$, a.s.

The convergence of sets should be taken to mean convergence as determined by the Hausdorff metric $H(\cdot, \cdot)$, which is defined for compact subsets A, B of $\mathbb{R}^s$ by: $H(A, B) < \delta$ if and only if every point of A is within (Euclidean) distance δ of at least one point of B, and vice versa. Suppose that A contains exactly k distinct points, and that δ is chosen less than one half of the minimum distance between points of A. Then if B is any set of k or fewer points for which $H(A, B) < \delta$, it must contain exactly k distinct points, each of which lies within a distance δ of a uniquely determined point of A. Almost sure convergence of A_n in the Hausdorff sense could, therefore, be translated into almost sure convergence of individual cluster centres under a suitable labelling.

For the minimisation procedures just described to make sense, the function ϕ must satisfy some regularity conditions. We shall need ϕ continuous and nondecreasing, with $\phi(0) = 0$. In order to control the growth of ϕ in the tails, assume that there exists some constant λ such that $\phi(2r) \le \lambda\phi(r)$ for every $r > 0$. As long as $\int \phi(\|x\|)\, P(dx)$ is finite, this ensures that $\Phi(A, P)$ is finite for each A: for each $a \in \mathbb{R}^s$,

$$\begin{aligned}\int \phi(\|x - a\|)\, P(dx) &\le \int \phi(\|x\| + \|a\|)\, P(dx) \\ &\le \phi(2\|a\|) + \int_{\|x\| \ge \|a\|} \phi(2\|x\|)\, P(dx) \\ &\le \phi(2\|a\|) + \lambda \int \phi(\|x\|)\, P(dx).\end{aligned}$$

These assumptions on ϕ will remain in force throughout the paper.

THEOREM. *Suppose that $\int \phi(\|x\|)\, P(dx) < \infty$ and that for each $j = 1, 2, \ldots, k$ there is a unique set $\bar{A}(j)$ for which $\Phi(\bar{A}(j), P) = m_j(P)$. Then $A_n \to \bar{A}(k)$* a.s., and $\Phi(A_n, P_n) \to m_k(P)$ a.s.

The uniqueness condition on the $\bar{A}(j)$'s carries a lot of information: not only is it needed for the inductive argument outlined in Section 1, but also it implies that $m_1(P) > m_2(P) > \cdots m_k(P)$. For suppose that $m_{j-1}(P) = m_j(P)$ for some j. Then $\bar{A}(j-1)$ could be augmented by any arbitrary point to give a (nonunique) set A, of no more than j distinct points, for which $\Phi(A, P) = m_j(P)$. The condition similarly implies that $\bar{A}(j)$ contains exactly j distinct points.

Since the conclusions of the theorem are in terms of almost sure convergence, there might be aberrant null sets of ω's for which the convergence does not hold. Rather than add the qualifying "except for a possible null set of ω's" to each assertion, I leave the casting out of the null sets to the reader; the argument is written as if almost sure statements held everywhere.

3. Proof of the theorem. The first step consists of finding an M (not depending on ω) so large that, when n is large enough, at least one point of A_n is contained in the closed ball $B(M)$ centred at the origin and of radius M. It is convenient to assume that $\phi(r) \to \infty$ as $r \to \infty$; the proof for ϕ bounded is only slightly more complicated.

Find an r so that the ball K of radius r and centred at the origin has positive P measure. For the purposes of this first step it will suffice that M be large enough to make $\phi(M - r)\, P(K) > \int \phi(\|x\|)\, P(dx)$; for the second and third steps two further requirements will be placed on M.

By assumption $\Phi(A_n, P_n) \le \Phi(A_0, P_n)$ for any set A_0 containing at most k points. Choose A_0 to consist of a single point at the origin. Then

$$\Phi(A_0, P_n) = \int \phi(\|x\|)\, P_n(dx) \to \int \phi(\|x\|)\, P(dx) \quad \text{a.s.}$$

If, for infinitely many values of n, no point of A_n were contained in $B(M)$, then

$$\limsup_n \Phi(A_n, P_n) \geq \lim_n \phi(M - r)\, P_n(K) = \phi(M - r)\, P(K) \quad \text{a.s.}$$

This would make $\Phi(A_n, P_n) > \Phi(A_0, P_n)$ infinitely often: a contradiction. Without loss of generality we may therefore assume that A_n always contains at least one point of $B(M)$.

If $k = 1$ the next step in the proof can be skipped; if $k > 1$ then we have to show that, for n large enough, the closed ball $B(5M)$, of radius $5M$ and centred at the origin, contains all the points of A_n. For the purposes of an inductive argument, assume that the conclusions of the theorem are valid when applied to optimal allocation of 1, 2, $\cdots$ $k - 1$ cluster centres. If A_n were not eventually contained in $B(5M)$, the cluster centres outside $B(5M)$ could be deleted to obtain a set of $k - 1$ (or fewer) centres that would reduce $\Phi(\cdot, P_n)$ below its minimum over all sets of $k - 1$ points. To obtain such a contradiction, we need M large enough to ensure that

$$\lambda \int_{\|x\| \geq 2M} \phi(\|x\|)\, P(dx) < \epsilon,$$

where $\epsilon > 0$ is chosen to satisfy $\epsilon + m_k(P) < m_{k-1}(P)$. This is the second requirement placed on M.

Suppose that A_n contains at least one point outside $B(5M)$. What would be the effect on $\Phi(A_n, P_n)$ of deleting such points as cluster centres? At worst, the centre a_1 that is known to lie in $B(M)$ might have to accept into its own cluster all of the sample points presently assigned to cluster centres outside $B(5M)$. These sample points must have been a distance at least $2M$ from the origin, otherwise they would have been closer to the cluster centre a_1 than to any centre outside $B(5M)$. The extra contribution to $\Phi(\cdot, P_n)$ due to deleting centres outside $B(5M)$ would therefore be at most

$$\begin{aligned}
\int_{\|x\| \geq 2M} \phi(\|x - a_1\|)\, P_n(dx) &\leq \int_{\|x\| \geq 2M} \phi(\|x\| + \|a_1\|)\, P_n(dx) \\
&\leq \int_{\|x\| \geq 2M} \phi(2\|x\|)\, P_n(dx) \\
&\leq \lambda \int_{\|x\| \geq 2M} \phi(\|x\|)\, P_n(dx).
\end{aligned}$$

The set A_n^* obtained by deleting from A_n all centres outside $B(5M)$ is a candidate for minimising $\Phi(\cdot, P_n)$ over sets of $k - 1$ or fewer points; it is therefore beaten by the optimal set, B_n say, of $k - 1$ centres. Thus

$$\Phi(A_n^*, P_n) \geq \Phi(B_n, P_n),$$

which, by the inductive hypothesis, converges almost surely to $m_{k-1}(P)$. If $A_n \not\subseteq B(5M)$ along some subsequence $\{n_i\}$ of values of n, we therefore get

$$\begin{aligned}
m_{k-1}(P) &\leq \liminf_i \Phi(A_{n_i}^*, P_n) \text{ a.s.} \\
&\leq \limsup_n \left[\Phi(A_n, P_n) + \lambda \int_{\|x\| \geq 2M} \phi(\|x\|) P_n(dx)\right] \\
&\leq \limsup_n \Phi(A, P_n) + \lambda \int_{\|x\| \geq 2M} \Phi(\|x\|) P(dx) \quad \text{a.s.}
\end{aligned}$$

for any fixed A with k or fewer points. Choose $A = \bar{A}(k)$, the optimal set of k centres for $\Phi(\cdot, P)$. Then, because of the second requirement placed on M, this last bound is less than $\Phi(\bar{A}(k), P) + \epsilon = m_k(P) + \epsilon < m_{k-1}(P)$: a contradiction.

We now know that, for n large enough, it suffices to search for A_n amongst the class of sets $\mathscr{E}_k := \{A \subseteq B(5M)$: A contains k or fewer points$\}$. For the final requirement on M, we assume it is large enough to ensure that $\mathscr{E}_k$ contains $\bar{A}(k)$; the function $\Phi(\cdot, P)$ therefore achieves its unique minimum on $\mathscr{E}_k$ at $\bar{A}(k)$. Under the topology induced by the Hausdorff metric, $\mathscr{E}_k$ is

compact (this follows from the compactness of $B(5M)$) and, as is proved at the end of Section 4, the map $A \to \Phi(A, P)$ is continuous on $\mathscr{E}_k$. The function $\Phi(\cdot, P)$ therefore enjoys an even stronger minimisation property on $\mathscr{E}_k$: given any neighbourhood $\mathscr{N}$ of $\bar{A}(k)$ there exists an $\eta > 0$, depending on $\mathscr{N}$, such that

$$\Phi(A, P) \geq \Phi(\bar{A}(k), P) + \eta$$

for every A belonging to $\mathscr{E}_k \backslash \mathscr{N}$.

The proof can now be completed by an appeal to a uniform SLLN:

$$\sup_{A \in \mathscr{E}_k} |\Phi(A, P_n) - \Phi(A, P)| \to 0 \text{ a.s.}$$

This result is proved in Section 4. We need to show that A_n is eventually inside the neighbourhood $\mathscr{N}$. It is enough to check that $\Phi(A_n, P) < \Phi(\bar{A}(k), P) + \eta$ eventually. This follows from

$$\Phi(A_n, P_n) \leq \Phi(\bar{A}(k), P_n)$$

and

$$\Phi(A_n, P_n) - \Phi(A_n, P) \to 0 \text{ a.s.}$$

and

$$\Phi(\bar{A}(k), P_n) - \Phi(\bar{A}(k), P) \to 0 \text{ a.s.}$$

Similarly, for n large enough,

$$\Phi(A_n, P_n) = \inf\{\Phi(A, P_n) : A \in \mathscr{E}_k\} \to \inf\{\Phi(A, P) : A \in \mathscr{E}_k\} \text{ a.s.} = m_k(P). \qquad \square$$

4. Proofs of the uniform SLLN and the continuity of $\Phi(\bullet, P)$. Let $\mathscr{G}$ denote the family of P-integrable functions on $\mathbb{R}^s$ of the form $g_A(x) := \min_{a \in A} \phi(\|x - a\|)$, where A ranges over all subsets of $\mathscr{E}_k$ containing k or fewer points. We need to prove the uniform SLLN:

$$(*) \qquad \sup_{g \in \mathscr{G}} \left| \int g \, dP_n - \int g \, dP \right| \to 0 \text{ a.s.}$$

Many results like this are to be found in the literature—see Section 1.1 of the survey by Gaenssler and Stute (1979). A sufficient condition for $(*)$ to hold is: for each $\epsilon > 0$ there exists a finite class $\mathscr{G}_\epsilon$ of functions such that to each $g \in \mathscr{G}$ there are functions $\mathring{g}, \bar{g} \in \mathscr{G}_\epsilon$ with $\mathring{g} \leq g \leq \bar{g}$ and $\int (\bar{g} - \mathring{g}) dP < \epsilon$. (The proof uses the SLLN applied to each function in the countable class $\mathscr{G}_{1/2} \cup \mathscr{G}_{1/3} \cup \mathscr{G}_{1/4} \cup \ldots$, together with the bound $\int (\bar{g} - g_0) dP + \max\{|\int \bar{g} dP_n - \int \bar{g} dP|, |\int \mathring{g} dP_n - \int \mathring{g} dP|\}$ for $|\int g dP_n - \int g dP|$.)

To find a suitable $\mathscr{G}_\epsilon$, let D_δ be a finite subset of $B(5M)$ such that each point of $B(5M)$ is within a distance δ of at least one point of D_δ. The appropriate value of δ will be specified shortly. Write $\mathscr{E}_{k,\delta}$ for $\{A \in \mathscr{E}_k; A \subseteq D_\delta\}$. Take $\mathscr{G}_\epsilon$ as the class of functions of the form

$$\min_{a \in A'} \phi(\|x - a\| + \delta) \quad \text{or} \quad \min_{a \in A'} \phi(\|x - a\| - \delta),$$

where A' ranges over $\mathscr{E}_{k,\delta}$.

Given an $A = \{a_1, a_2, \ldots, a_k\}$ in $\mathscr{E}_k$, there exists an $A' = \{a_1', a_2', \ldots, a_k'\}$ in $\mathscr{E}_{k,\delta}$ with $H(A, A') < \delta$—just choose $a_i' \in D_\delta$ with $\|a_i - a_i'\| < \delta$, for each i. Corresponding to $g_A \in \mathscr{G}$ take

$$\bar{g}_A := \min_{a \in A'} \phi(\|x - a\| + \delta)$$

and

$$\mathring{g}_A := \min_{a \in A'} \phi(\|x - a\| - \delta),$$

where $\phi(r)$ is defined as zero if r is negative. Since ϕ is monotone and

$$\|x - a_i'\| - \delta \leq \|x - a_i\| \leq \|x - a_i'\| + \delta$$

for each i and each $x \in \mathbb{R}^s$, this choice ensures that $\mathring{g}_A \le g_A \le \bar{g}_A$. Also, if R is greater than $5M + \delta$,

$$\int \bar{g}_A(x) - \mathring{g}_A(x)\, P(dx)$$

$$\le \int \sum_{i=1}^{k} [\phi(\|x - a_i'\| + \delta) - \phi(\|x - a_i'\| - \delta)] P(dx)$$

$$\le k \sup_{\|x\| \le R} \sup_{a \in B(5M)} |\phi(\|x - a\| + \delta) - \phi(\|x - a\| - \delta)|$$
$$+ k \int_{\|x\| \ge R} 2\lambda\, \phi(\|x\|)\, P(dx).$$

The second term can be made less than $\epsilon/2$ by choosing R large enough, then the uniform continuity of ϕ on bounded sets can be used to find a $\delta > 0$ small enough to make the first term less than $\epsilon/2$. This completes the proof of (*).

A similar argument can be employed to prove continuity of the map $A \to \Phi(A, P)$ on $\mathscr{E}_k$. If $A, B \in \mathscr{E}_k$ and $H(A, B) < \delta$ then to each $b \in B$ there is a point $a(b) \in A$ such that $\|b - a(b)\| < \delta$. Then

$$\Phi(A, P) - \Phi(B, P) = \int \min_{a \in A} \phi(\|x - a\|) - \min_{b \in B} \phi(\|x - b\|)\, P(dx)$$
$$\le \int \max_{b \in B} [\phi(\|x - a(b)\|) - \phi(\|x - b\|)]\, P(dx)$$
$$\le \int \sum_{b \in B} [\phi(\|x - b\| + \delta) - \phi(\|x - b\|)]\, P(dx)$$

which is less than ϵ if δ is chosen as in the preceding paragraph. The other inequality needed to complete the continuity proof is obtained by interchanging the roles of A and B.

Notice that the arguments in this section, and those in Section 3, do not really depend on the underlying sample space being $\mathbb{R}^s$ equipped with its usual norm—any metric space for which all the closed balls are compact would do. For example, the theorem would still hold if the Euclidean norm on $\mathbb{R}^s$ were replaced by a norm such as $\|x\| := \max\{|x_i| = i = 1, \ldots, s\}$; and for compact metric spaces, we would obtain generalisations of results of Sverdrup-Thygeson (1981).

Acknowledgment. I am grateful to John Hartigan, who aroused my interest in this problem and made many perceptive comments on my attempts to solve it. My thanks go also to the referee who made several useful suggestions for improving the presentation.

REFERENCES

FISHER, W. D. (1958). On grouping for maximum homogeneity. *J. Amer. Statist. Assoc.* **53** 789–798.

GAENSSLER, P. and STUTE, W. (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *Ann. Probability* **7** 193–243.

HARTIGAN, J. A. (1975). *Clustering Algorithms.* Wiley, New York.

HARTIGAN, J. A. (1978). Asymptotic distributions for clustering criteria. *Ann. Statist.* **6** 117–131.

HARTIGAN, J. A. and WONG, M. A. (1979). Algorithm AS136: A *K*-means clustering algorithm. *Appl. Statist.* **28** 100–108.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 281–297.

SVERDRUP-THYGESON, H. (1981). Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces. *Ann. Statist.* **9** 141–145.

WONG, M. A. (1979). Hybrid clustering. Doctoral dissertation, Yale University.

DEPARTMENT OF STATISTICS
YALE UNIVERSITY
BOX 2179, YALE STATION
NEW HAVEN, CONNECTICUT 06520