The Annals of Probability
1982, Vol. 10, No. 4, 919–926

A CENTRAL LIMIT THEOREM FOR k-MEANS CLUSTERING[1]

BY DAVID POLLARD

Yale University

A set of n points in Euclidean space is partitioned into the k groups that minimize the within groups sum of squares. Under the assumption that the n points come from independent sampling on a fixed distribution, conditions are found to assure asymptotic normality of the vector of means of the k groups. The method of proof makes novel application of a functional central limit theorem for empirical processes—a generalization of Donsker's theorem due to Dudley.

1. Introduction. In this paper a central limit theorem is proved for the procedure known in the cluster analysis literature (see, for example, Hartigan, 1975) as the method of k-means. The theorem generalizes a result of Hartigan (1978) to a multidimensional setting.

Independent observations $x_1, x_2, \cdots, x_n$ are made on a probability measure P on $\mathbb{R}^d$. The k-means procedure prescribes a criterion for partitioning these observations into k groups, or clusters: minimize the within cluster sum of squares. Equivalently, a vector of optimal centers $\mathbf{b}_n = [b_{n1}, b_{n2}, \cdots, b_{nk}]$ can be chosen to minimize

$$W_n(\mathbf{a}) = n^{-1} \sum_{i=1}^n \min_{1 \le j \le k} \|x_i - a_j\|^2$$

as a function of the vector $\mathbf{a} = [a_1, \cdots, a_k]$. Associated with each center a_j is the convex polyhedron A_j of all points in $\mathbb{R}^d$ closer to a_j than to any other center. (The precise convention adopted for allocating points common to the boundaries of two or more of the A_j's is unimportant.) The polyhedral regions corresponding to the optimal centers $\mathbf{b}_n$ partition $\{x_1, x_2, \cdots, x_n\}$ into the optimal clusters, which minimize the within cluster sum of squares; each b_{nj} is the mean of those x_i's in its cluster. The main result of this paper (stated and proved in Section 3) gives conditions under which $\mathbf{b}_n$, suitably normalized, has an asymptotic normal distribution.

The difficulty in extending Hartigan's (1978) central limit theorem beyond one dimension is due to added complication in the way A_j responds to changes in the cluster centers. For $d = 1$, small changes in $\mathbf{a}$ shift the boundary points of A_j by only a small amount. For $d \ge 2$, a small change in $\mathbf{a}$ can augment or reduce A_j by an unbounded wedge-shaped region; the contributions to $W_n(\mathbf{a})$ from such unbounded wedges are hard to handle. This difficulty is overcome (Lemma B) by application of a generalized Donsker theorem for empirical measures, due to Dudley (1981).

The proof of the main theorem depends on a quadratic approximation

$$W_n(\mathbf{a}) \approx W_n(\boldsymbol{\mu}) - n^{-1/2} Z_n'(\mathbf{a} - \boldsymbol{\mu}) + \tfrac{1}{2}(\mathbf{a} - \boldsymbol{\mu})'\Gamma(\mathbf{a} - \boldsymbol{\mu}), \tag{1}$$

for $\mathbf{a}$ in a neighbourhood of a fixed vector $\boldsymbol{\mu}$, where Γ is a fixed positive definite matrix and $\mathbf{Z}_n$ is an asymptotically normally distributed random vector. The optimal $\mathbf{b}_n$ that minimizes $W_n(\cdot)$ lies close to the vector $\boldsymbol{\mu} + n^{-1/2}\Gamma^{-1}\mathbf{Z}_n$ that minimizes the righthand side of (1), in the sense that $n^{1/2}(\mathbf{b}_n - \boldsymbol{\mu}) - \Gamma^{-1}\mathbf{Z}_n$ converges to zero in probability. A precise formulation of approximation (1) appears in Lemma D of Section 2 as the culmination of three lemmas establishing differentiability properties of $W_n(\cdot)$.

Received February 4, 1981; revised April 2, 1982.

[1] This research was supported in part by the Air Force Office of Scientific Research, Contract No. F49620-79-C-0164, and in part by the National Science Foundation, Grant No. MCS-8102725.

Key words and phrases. k-means clustering, central limit theorem, minimized within cluster sum of squares, differentiability in quadratic mean, Donsker classes of functions, functional central limit theorem for empirical processes.

AMS 1980 *subject classification.* Primary, 62H30; secondary, 60F05; 60F17.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Probability.* www.jstor.org

For typographical convenience, all integrals in this paper are written in linear functional notation, and sets are identified with their indicator functions. Thus, instead of $\int_A f\,dP$, I write $P(fA)$. Convergence in distribution will be denoted by the symbol $\rightsquigarrow$. The stochastic order symbols $o_p(\cdot)$ and $O_p(\cdot)$ will be employed with random arguments—the definitions given by Chernoff (1956) carry over without change.

2. The quadratic approximation. Most of the regularity conditions needed for the proof of the main theorem are introduced in this section in the context of the three differentiability lemmas needed to formalize (1), which is justified by a Taylor expansion to quadratic terms of a deterministic component of $W_n(\cdot)$ together with a linear approximation to a stochastic component.

The decomposition of $W_n(\cdot)$ into these two components can be expressed most easily in terms of the empirical measure P_n, obtained by placing mass n^{-1} at each of $x_1, x_2, \cdots, x_n$, and the associated empirical process $X_n(\cdot) = n^{1/2}(P_n(\cdot) - P(\cdot))$. For every vector $\mathbf{a} = [a_1, a_2, \cdots, a_k]$ in $\mathbb{R}^{kd}$ and every x in $\mathbb{R}^d$ define

$$\phi(x, \mathbf{a}) = \min_{1\le j\le k}\|x - a_j\|^2.$$

Then

$$W_n(\mathbf{a}) = P_n\phi(\cdot, \mathbf{a}) = P\phi(\cdot, \mathbf{a}) + n^{-1/2}X_n\phi(\cdot, \mathbf{a}). \tag{2}$$

The deterministic component $P\phi(\cdot, \mathbf{a})$ is known as the population within cluster sum of squares; I shall denote it by $W(\mathbf{a})$.

LEMMA A. *Suppose $P\|x\|^2 < \infty$ and that P gives zero measure to every hyperplane in $\mathbb{R}^d$. Then the map $\mathbf{a} \to \phi(\cdot, \mathbf{a})$ from $\mathbb{R}^{kd}$ into $\mathscr{L}^2(P)$ is differentiable in quadratic mean. In consequence, the map $\mathbf{a} \to P\phi(\cdot, \mathbf{a})$ is differentiable.*

PROOF. If $x \in \operatorname{int} A_j$ and $\mathbf{h}$ is small enough,

$$\phi(x, \mathbf{a} + \mathbf{h}) = \|x - a_j - h_j\|^2 = \phi(x, \mathbf{a}) - 2h_j'(x - a_j) + \|h_j\|^2.$$

Define $\mathbf{\Delta}(x, \mathbf{a})$ to be the k vector of $\mathscr{L}^2(P)$ functions $-2A_j(x - a_j)$. Then, because the boundary of each A_j has zero P measure, the function $\phi(\cdot, \mathbf{a} + \mathbf{h})$ can be expanded for all $x \in \mathbb{R}^d$ into

$$\phi(x, \mathbf{a} + \mathbf{h}) = \phi(x, \mathbf{a}) + \mathbf{h}'\mathbf{\Delta}(x, \mathbf{a}) + \|\mathbf{h}\|R(x, \mathbf{a}, \mathbf{h}) \tag{3}$$

where

$$R(x, \mathbf{a}, \mathbf{h}) \to \mathbf{0} \quad \text{for } P \text{ almost all } x \text{ as } \mathbf{h} \to \mathbf{0}.$$

The first part of the lemma then follows from the domination

$$\begin{aligned} |R(x, \mathbf{a}, \mathbf{h})| &\le \|\mathbf{h}\|^{-1}(|\mathbf{h}'\mathbf{\Delta}(x, \mathbf{a})| + \max_j|\|x - a_j - h_j\|^2 - \|x - a_j\|^2|) \\ &\le \|\mathbf{\Delta}(x, \mathbf{a})\| + \|\mathbf{h}\|^{-1}\textstyle\sum_{j=1}^k|\|x - a_j - h_j\|^2 - \|x - a_j\|^2| \\ &\le C(1 + \|x\|) \quad \text{for } \mathbf{h} \text{ small enough and some constant } C \\ &\in \mathscr{L}^2(P). \end{aligned}$$

Convergence in $\mathscr{L}^2(P)$ implies convergence in $\mathscr{L}^1(P)$. From expansion (3) and what has just been proved, it therefore follows that $P\phi(\cdot, \mathbf{a})$ is differentiable with derivative

$$\gamma(\mathbf{a}) = P\mathbf{\Delta}(\cdot, \mathbf{a}). \qquad \square$$

To convert $\mathscr{L}^2(P)$ differentiability of $\phi(\cdot, \mathbf{a})$ into stochastic differentiability of $X_n\phi(\cdot, \mathbf{a})$, I need to use the properties of Donsker classes of functions. If $\mathscr{G} \subseteq \mathscr{L}^2(P)$, the empirical process X_n may be thought of as a stochastic process indexed by $\mathscr{G}$. Dudley (1981) called such a $\mathscr{G}$ a Donsker class for P if a functional central limit theorem holds for the sequence of processes $\{X_n(g) : g \in \mathscr{G}\}$. The key property of a Donsker class is that to every $\varepsilon > 0$

and $\eta > 0$ there exists a $\delta > 0$ and an n_0 such that, for all n greater than n_0,

$$\mathbb{P}\{\sup_{[\delta]} | X_n(g_1) - X_n(g_2) | > \eta\} < \varepsilon. \tag{4}$$

The $[\delta]$ indicates that the supremum here runs over all pairs of functions g_1, g_2 in $\mathcal{G}$ less than δ apart in $\mathcal{L}^2(P)$ norm. The class $\mathcal{G}$ in this paper will consist of all functions $R(\cdot, \boldsymbol{\mu}, \mathbf{a} - \boldsymbol{\mu})$ with $\mathbf{a}$ ranging over a neighbourhood of a fixed $\boldsymbol{\mu}$, chosen so that $| R(x, \boldsymbol{\mu}, \mathbf{a} - \boldsymbol{\mu}) | \le C(1 + \| x \|)$ for all $\mathbf{a}$ in the neighbourhood and all x, for some constant C. (The processes $\{X_n(g) : g \in \mathcal{G}\}$ are separable in the sense of Section 5 of Pollard (1981a) so there is no measurability problem with the supremum in (4).)

As before, write $A_1, \cdots A_k$ for the convex polyhedra associated with the centres $a_1, \cdots a_k$; write $M_1, \cdots M_k$ for the polyhedra associated with $\mu_1, \cdots \mu_k$. Then

$$\begin{aligned} R(x, \boldsymbol{\mu}, \mathbf{a} - \boldsymbol{\mu}) &= \| \mathbf{a} - \boldsymbol{\mu} \|^{-1} [\phi(x, \mathbf{a}) - \phi(x, \boldsymbol{\mu}) - (\mathbf{a} - \boldsymbol{\mu})' \Delta(x, \boldsymbol{\mu})] \\ &= \textstyle\sum_{i,j} M_i A_j \| \mathbf{a} - \boldsymbol{\mu} \|^{-1} [\| x - a_j \|^2 - \| x - \mu_i \|^2 + 2(a_i - \mu_i)'(x - \mu_i)] \\ &= \textstyle\sum_{i,j} M_i A_j \| \mathbf{a} - \boldsymbol{\mu} \|^{-1} [2(a_i - a_j)' x + \| \mu_i \|^2 - 2\mu_i' a_i + \| a_j \|^2]. \end{aligned}$$

Thus every function in $\mathcal{G}$ can be written as a sum of k^2 members of the class $\mathcal{F}$ of all functions f with these properties:

(i) $| f(x) | \le C(1 + \| x \|) = F(x)$ for all x;
(ii) $f = LQ$, where L is a linear function and Q is a convex region expressible as an intersection of at most $2k$ open or closed half spaces.

By Theorem 10(i) of Pollard (1982a), to prove that $\mathcal{G}$ is a Donsker class it suffices to check the entropy condition of Theorem 7 of that same paper for the class $\mathcal{F}$. To do this I borrow an idea of LeCam (1981).

Let S be any finite subset of $\mathbb{R}^k$. For any function h on $\mathbb{R}^k$ define

$$\| h \|_S = [\textstyle\sum_{x \in S} h(x)^2]^{1/2}.$$

Given $\delta > 0$, choose a maximal subclass $\{f_1, \cdots, f_m\}$ of $\mathcal{F}$ for which

$$\| f_i - f_j \|_S > \delta \| F \|_S \quad \text{for} \quad i \ne j.$$

To check the entropy condition it is enough to show that $m \le B\delta^{-W}$ for some constants B and W.

Represent each f_i by its graph

$$G_i = \{(x, t) \in \mathbb{R}^{k+1} : 0 < t \le f_i(x) \quad \text{or} \quad 0 > t \ge f_i(x)\}.$$

By the defining property (ii) of $\mathcal{F}$, each G_i is a union of at most two convex regions in $\mathbb{R}^{k+1}$, each expressible as an intersection of $2k + 2$ open or closed half spaces. The collection of all such sets has the Vapnik-Červonenkis property (Dudley, 1978, Proposition 7.12).

Define a probability measure Q on $\mathbb{R}^{k+1}$ in two steps. First select a point in S according to the distribution that places mass $F(x)^2 / \| F \|_S^2$ on x. Given x, choose t according to the uniform distribution on $[-F(x), F(x)]$. The maximality property of the subclass implies that for $i \ne j$,

$$Q(G_i \Delta G_j) = \textstyle\sum_{x \in S} \dfrac{F(x)^2}{\| F \|_S^2} \dfrac{| f_i(x) - f_j(x) |}{2F(x)} \ge \sum_{x \in S} \dfrac{F(x)^2}{\| F \|_S^2} \dfrac{[f_i(x) - f_j(x)]^2}{4F(x)^2} \ge \delta^2/4.$$

The rest of the argument follows Theorem 9 of Pollard (1982a) almost verbatim. Thus $\mathcal{G}$ is a Donsker class.

LEMMA B. *Let $\{\mathbf{a}_n\}$ be a sequence of random vectors with $\| \mathbf{a}_n - \boldsymbol{\mu} \| = o_p(1)$, for some fixed vector $\boldsymbol{\mu}$. Then, under the conditions of Lemma A,*

$$X_n \phi(\cdot, \mathbf{a}_n) = X_n \phi(\cdot, \boldsymbol{\mu}) + (\mathbf{a}_n - \boldsymbol{\mu})' X_n \boldsymbol{\Delta}(\cdot, \boldsymbol{\mu}) + o_p(r_n) \tag{5}$$

where $r_n = \| \mathbf{a}_n - \boldsymbol{\mu} \|$.

PROOF. From the expansion (3), the remainder term in (5) can be written as $\| \mathbf{a}_n - \boldsymbol{\mu} \| X_n R(\cdot, \boldsymbol{\mu}, \mathbf{a}_n - \boldsymbol{\mu})$. Remember from Lemma A that $R(\cdot, \boldsymbol{\mu}, \mathbf{a} - \boldsymbol{\mu})$ converges in $\mathscr{L}^2(P)$ norm to zero as $\mathbf{a} \to \boldsymbol{\mu}$. If $\mathbf{a}$ is close enough to $\boldsymbol{\mu}$, the pair $g_1 = R(\cdot, \boldsymbol{\mu}, \mathbf{a} - \boldsymbol{\mu})$ and $g_2 = 0$ therefore falls into the class $[\delta]$ appearing in inequality (4). With high probability, this reasoning also applies to $\mathbf{a}_n$, if n is large enough. □

Second derivatives of $P\phi(\cdot, \mathbf{a})$ involve integrals over faces of the A_j's with respect to $(d-1)$ dimensional Lebesgue measure $\sigma(\cdot)$. In the next lemma, F_{ij} denotes the face (possibly empty) common to A_i and A_j, and I_d denotes the $d \times d$ identity matrix.

LEMMA C. *Suppose $P \| x \|^2 < \infty$ and that $P(\cdot)$ has a continuous density $f(\cdot)$ with respect to d dimensional Lebesgue measure $\lambda(\cdot)$. Assume that the integral $\sigma[F_{ij} f(x)(x-m)(x-m)']$ exists and depends continuously on the location of the centres, for each i and j and for each fixed $m \in \mathbb{R}^d$. Then, if the centres a_i are all distinct, the map $\mathbf{a} \to P\phi(\cdot, \mathbf{a})$ has a second derivative Γ made up of $d \times d$ blocks*

$$\Gamma_{ij} = \begin{cases} 2PA_i I_d - 2 \sum_{\alpha \neq i} r_{i\alpha}^{-1} \sigma[F_{i\alpha} f(x)(x - a_i)(x - a_i)'] & \text{for } j = i \\ -2 r_{ij}^{-1} \sigma[F_{ij} f(x)(x - a_i)(x - a_j)'] & \text{for } i \neq j \end{cases} \tag{6}$$

where $r_{ij} = \| a_i - a_j \|$.

PROOF. From Lemma A, $P\phi(\cdot, \mathbf{a})$ has derivative $\gamma(\mathbf{a})$ with components $-2PA_i(x - a_i)$ for $i = 1, 2, \cdots, k$. I shall prove differentiability of $\gamma(\cdot)$ at any fixed $\boldsymbol{\mu}$ by demonstrating differentiability of the map $\mathbf{a} \to PA_i(x - \mu_i)$. This suffices because $(\partial/\partial a_i) PA_i(x - a_i) = -PA_i Id$, which depends continuously on $\mathbf{a}$.

Write $PA_i(x - \mu_i)$ as $\lambda[A_i f(x)(x - \mu_i)]$. An application of Stokes's theorem (justified by the continuity assumptions on the surface integrals), as in Theorem 1 of Baddeley (1977), establishes differentiability of this integral as a function of $\mathbf{a}$. The derivative is given in differential form at $\mathbf{a} = \boldsymbol{\mu}$ by

$$dPA_i(x - \mu_i) = \sigma[\partial M_i f(x)(x - \mu_i) \mathbf{v}_i'] \, d\mathbf{a} \tag{7}$$

where the kd vector $\mathbf{v}_i$ denotes the velocity vector for motion of A_i orthogonal to its boundary ∂A_i, evaluated at $\mathbf{a} = \boldsymbol{\mu}$. The form of this velocity vector depends on the boundary face F_{ij}.

Write $n_{ij} = r_{ij}^{-1}(\mu_j - \mu_i)$ for the unit normal pointing outward from the face F_{ij}. Elementary calculations show that on F_{ij},

$$\mathbf{v}_i' \, d\mathbf{a} = -(x - \tfrac{1}{2}(\mu_i + \mu_j))' \, dn_{ij} + \tfrac{1}{2} n_{ij}' (da_i + da_j),$$

the first term coming from rotation of the face and the second from translation. With the substitution

$$dn_{ij} = r_{ij}^{-1} \Pi_{ij} (da_j - da_i),$$

where Π_{ij} denotes the matrix for projection onto the affine hull of F_{ij}, this last expression reduces to

$$\begin{aligned} \mathbf{v}_i' \, d\mathbf{a} &= -r_{ij}^{-1}(x - \tfrac{1}{2}(\mu_i + \mu_j))'(da_j - da_i) + (2r_{ij})^{-1}(\mu_j - \mu_i)'(da_i + da_j) \\ &= r_{ij}^{-1}(x - \mu_i')\, da_i - r_{ij}^{-1}(x - \mu_j)' \, da_j. \end{aligned}$$

Multiply this expression for $\mathbf{v}_i$ by $f(x)(x - \mu_i)$, then sum the integrals over all the faces of ∂M_i to obtain the second contribution to the right-hand side of (6). □

These three differentiability lemmas justify the approximation (1). It is more convenient to express the approximation in terms of a sequence converging to $\boldsymbol{\mu}$; it will then apply directly to both the sequence $\{\mathbf{b}_n\}$ and the sequence $\{\boldsymbol{\mu} + n^{-1/2} \Gamma^{-1} \mathbf{Z}_n\}$.

LEMMA D. *Suppose $W(\cdot)$ has a minimum at $\boldsymbol{\mu}$. Let $\{\mathbf{a}_n\}$ be any sequence of random vectors in $\mathbb{R}^{kd}$ for which $\| \mathbf{a}_n - \boldsymbol{\mu} \| = o_p(1)$. Then, under the assumptions of Lemma C,*

$$W_n(\mathbf{a}_n) = W_n(\boldsymbol{\mu}) - n^{-1/2}\mathbf{Z}_n'(\mathbf{a}_n - \boldsymbol{\mu}) + \tfrac{1}{2}(\mathbf{a}_n - \boldsymbol{\mu})'\Gamma(\mathbf{a}_n - \boldsymbol{\mu}) + o_p(n^{-1/2}r_n) + o_p(r_n^2) \tag{8}$$

where $r_n = \| \mathbf{a}_n - \boldsymbol{\mu} \|$ and $\mathbf{Z}_n$ has an asymptotic $N(\mathbf{0}, V)$ distribution with V given by Equation (10) below.

PROOF. From Lemmas A and C,

$$P\phi(\cdot, \mathbf{a}) = P\phi(\cdot, \boldsymbol{\mu}) + (\mathbf{a} - \boldsymbol{\mu})'\gamma(\boldsymbol{\mu}) + \tfrac{1}{2}(\mathbf{a} - \boldsymbol{\mu})'\Gamma(\mathbf{a} - \boldsymbol{\mu}) + o(\| \mathbf{a} - \boldsymbol{\mu} \|^2).$$

The linear term must vanish because $\mathbf{a} = \boldsymbol{\mu}$ minimizes $W(\cdot)$. Set $\mathbf{a} = \mathbf{a}_n$ to find

$$P\phi(\cdot, \mathbf{a}_n) = P\phi(\cdot, \boldsymbol{\mu}) + \tfrac{1}{2}(\mathbf{a}_n - \boldsymbol{\mu})'\Gamma(\mathbf{a}_n - \boldsymbol{\mu}) + o_p(r_n^2). \tag{9}$$

Substitution from (9) and (5) into (2) and regrouping of terms gives

$$W_n(\mathbf{a}_n) = W_n(\boldsymbol{\mu}) + n^{-1/2}(\mathbf{a}_n - \boldsymbol{\mu})'X_n\boldsymbol{\Delta}(\cdot, \boldsymbol{\mu}) + o_p(n^{-1/2}r_n) + \tfrac{1}{2}(\mathbf{a}_n - \boldsymbol{\mu})'\Gamma(\mathbf{a}_n - \boldsymbol{\mu}) + o_p(r_n^2).$$

It remains only to set

$$\mathbf{Z}_n = -X_n\boldsymbol{\Delta}(\cdot, \boldsymbol{\mu}).$$

This vector has an asymptotic normal distribution with mean vector

$$-P\boldsymbol{\Delta}(\cdot, \boldsymbol{\mu}) = -\gamma(\boldsymbol{\mu}) = \mathbf{0},$$

and variance matrix

$$P\boldsymbol{\Delta}(\cdot, \boldsymbol{\mu})\boldsymbol{\Delta}(\cdot, \boldsymbol{\mu})'$$

whose (i, j)th block is

$$4P[(x - \mu_i)(x - \mu_j)'M_iM_j].$$

This vanishes for $i \neq j$, but reduces to

$$V_i = 4P[(x - \mu_i)(x - \mu_i)'M_i] \tag{10}$$

for $i = j$. Thus

$$\mathbf{Z}_n \rightsquigarrow N(\mathbf{0}, V)$$

where V is the $kd \times kd$ block diagonal matrix made up from the V_i's. □

3. The main theorem. Most of the assumptions needed to prove the central limit theorem for the vector of optimal cluster centres have already been introduced through the lemmas in Section 2. To these must be added the consistency conditions of Pollard (1981b, 1982b), to justify application of the local approximation (8).

THEOREM. *Let $\mathbf{b}_n$ be the vector of optimal k-means cluster centres for independent sampling from a distribution P on $\mathbb{R}^d$. Suppose*

(i) *the vector $\boldsymbol{\mu}$ that minimizes the population within cluster sum of squares $W(\cdot)$ is unique up to relabeling of its coordinates;*

(ii) $P\| x \|^2 < \infty$;

(iii) *the probability measure P has a continuous density f with respect to Lebesgue measure λ on $\mathbb{R}^d$;*

(iv) *there exists a dominating function $g(\cdot)$ with $f(x) \leq g(\| x \|)$, for all $x \in \mathbb{R}^d$, and $r^d g(r)$ integrable with respect to Lebesgue measures on $[0, \infty)$;*

(v) *the matrix Γ defined by evaluating (6) at $\mathbf{a} = \boldsymbol{\mu}$ is positive definite.*

Then $n^{1/2}(\mathbf{b}_n - \boldsymbol{\mu}) \rightsquigarrow N(\mathbf{0}, \Gamma^{-1}V\Gamma^{-1})$, *where* V *is the* $kd \times kd$ *diagonal matrix with*

$$V_i = 4P[M_i(x - \mu_i)(x - \mu_i)']$$

as its ith *diagonal block. Here* M_i *denotes the set of points in* $\mathbb{R}^d$ *closer to* μ_i *than to any other* μ_j.

PROOF. Conditions (i) and (ii) allow me to assume (Theorem 1 of Pollard 1982b) that $\mathbf{b}_n$ converges in probability to $\boldsymbol{\mu}$. Next I show that (iv) takes care of the continuity assumption needed in Lemma C. For convenience, take the m of Lemma C to equal zero, and write F for the face of the cluster instead of F_{ij}. Decompose the surface integral into the contributions made by the part of the face inside a ball B, of radius R and centre 0, and the part outside B. Use (iv) to bound the absolute value of each component of $\sigma(FB^c f(x)xx')$ by

$$\sigma(FB^c g(\| x \|) \| x \|^2) \leq \int [R, \infty) g(r) r^2 \cdot cr^{d-2}$$

where c denotes the surface area of a $(d-1)$-dimensional sphere. Integrability of $r^d g(r)$ enables me to choose R large enough to make all these contributions less than some given ε. Simple uniform continuity arguments prove continuity of the dependence of $\sigma(FBf(x)xx')$ on the location of the cluster centres. The rest is easy. (Undoubtedly condition (iv) could be improved upon.)

Put $\lambda_n = \| \mathbf{b}_n - \boldsymbol{\mu} \|$, then apply Lemma D with $\mathbf{a}_n = \mathbf{b}_n$. Since, by definition of b_n,

$$W_n(\mathbf{b}_n) \leq W_n(\boldsymbol{\mu}),$$

the representation (8) implies

$$-n^{-1/2}\mathbf{Z}_n'(\mathbf{b}_n - \boldsymbol{\mu}) + \tfrac{1}{2}(\mathbf{b}_n - \boldsymbol{\mu})'\Gamma(\mathbf{b}_n - \boldsymbol{\mu}) + o_p(n^{-1/2}\lambda_n) + o_p(\lambda_n^2) \leq 0. \tag{11}$$

Positive definiteness of Γ guarantees the existence of a positive constant κ such that

$$\mathbf{y}'\Gamma\mathbf{y} \geq \kappa \| \mathbf{y} \|^2$$

for all $\mathbf{y}$. Also $\mathbf{Z}_n$ is of order $O_p(1)$ because it converges in distribution. Thus inequality (11) leads to

$$\kappa\lambda_n^2 \leq O_p(n^{-1/2}\lambda_n) + o_p(n^{-1/2}\lambda_n) + o_p(\lambda_n^2)$$

which forces

$$\lambda_n = O_p(n^{-1/2}).$$

For simplicity set $n^{1/2}(\mathbf{b}_n - \boldsymbol{\mu}) = \boldsymbol{\theta}_n$. Then (8) becomes

$$\begin{aligned} W_n(\mathbf{b}_n) &= W_n(\boldsymbol{\mu}) - n^{-1}\mathbf{Z}_n'\boldsymbol{\theta}_n + \tfrac{1}{2}n^{-1}\boldsymbol{\theta}_n'\Gamma\boldsymbol{\theta}_n + o_p(n^{-1}) \\ &= W_n(\boldsymbol{\mu}) + \tfrac{1}{2}n^{-1} \| \Gamma^{1/2}\boldsymbol{\theta}_n - \Gamma^{-1/2}\mathbf{Z}_n \|^2 - \tfrac{1}{2}n^{-1}\mathbf{Z}_n'\Gamma^{-1}\mathbf{Z}_n + o_p(n^{-1}) \\ &= W_n(\boldsymbol{\mu} + n^{-1/2}\Gamma^{-1}\mathbf{Z}_n) + \tfrac{1}{2}n^{-1} \| \Gamma^{1/2}\boldsymbol{\theta}_n - \Gamma^{-1/2}\mathbf{Z}_n \|^2 + o_p(n^{-1}) \end{aligned}$$

as may be seen by setting $\mathbf{a}_n = \boldsymbol{\mu} + n^{-1/2}\Gamma^{-1}\mathbf{Z}_n$ in (8). Once again by definition of $\mathbf{b}_n$,

$$W_n(\mathbf{b}_n) \leq W_n(\boldsymbol{\mu} + n^{-1/2}\Gamma^{-1}\mathbf{Z}_n)$$

which forces the conclusion

$$\tfrac{1}{2}n^{-1} \| \Gamma^{1/2}\boldsymbol{\theta}_n - \Gamma^{-1/2}\mathbf{Z}_n \|^2 = o_p(n^{-1}),$$

or

$$\boldsymbol{\theta}_n = \Gamma^{-1}\mathbf{Z}_n + o_p(1).$$

The result follows. □

The positive definiteness required by (v) seems almost redundant. Because μ minimizes $W(\cdot)$, the matrix Γ must necessarily be nonnegative definite, but I can see no general method for ruling out possible singularity. The same sort of difficulty occurs in the asymptotic theory for maximum likelihood estimators; the information matrix is usually just assumed nonsingular. Some light is shed on the problem by the special case of two clusters ($k = 2$) in one dimension ($d = 1$).

Specification of the boundary point $m_n = \frac{1}{2}(b_{n1} + b_{n2})$ uniquely determines the two sample clusters; the population clusters are determined by $m = \frac{1}{2}(\mu_1 + \mu_2)$. With $r = |\mu_1 - \mu_2|$ and $p_1 = PM_1 = 1 - p_2$, the matrix Γ becomes

$$\begin{pmatrix} 2p_1 - \frac{1}{2}rf(m) & \frac{1}{2}rf(m) \\ \frac{1}{2}rf(m) & 2p_2 - \frac{1}{2}rf(m) \end{pmatrix}, \tag{12}$$

which is singular if and only if $rf(m) = 4p_1p_2$. I have not yet succeeded in constructing a distribution satisfying condition (i) for which this equality holds. Indeed condition (i) is difficult to check for most distributions. The only general criterion I know of is due to Fleischer (1964). For some standard distributions, ad hoc methods succeed though.

EXAMPLE. Consider fitting two clusters to a sample from a distribution P that is spread uniformly over the union of the two disjoint intervals $(-1 - h, -1 + h)$ and $(1 - h, 1 + h)$ on the real line. (Of course, $0 < h < 1$.) This density function is not everywhere continuous but that does not destroy the theorem—actually only continuity at the boundary point m is needed.

Because the split point m must lie equidistant from the conditional means of the two clusters it defines, the optimal population cluster centres must be $+1$ and -1.

From (12), the matrix Γ can be read off as I_2, the identity matrix. The diagonal elements of V equal twice the within-cluster variance components—that is, $2h^2/3$ for both clusters. The empirical cluster centres, suitably normalized, have asymptotically independent $N(0, 2h^2/3)$ distributions.

All this might have been expected for a population distribution consisting of two well-separated clusters. For large samples, the two population clusters are correctly identified with very high probability; the cluster centres behave almost like $n^{1/2}(X_n + 1)$ and $n^{1/2}(Y_n - 1)$, where X_n is the mean of a sample of $\frac{1}{2}n$ observations from the left cluster and Y_n is the mean of $\frac{1}{2}n$ observations from the right cluster. In the general case, the cluster boundaries cannot be so precisely located. Observations near the boundaries effectively contribute to the means of the observations in both clusters. This dependence between the means is the source of the surface integral contributions to (6); it generates the off-diagonal elements of Γ.

Acknowledgment. I thank Adrian Baddeley for help with Lemma C, and the referee, who detected an error in my proof of the Donsker class property needed for Lemma B.

REFERENCES

BADDELEY, A. (1977). Integrals on a moving manifold and geometrical probability. *Adv. Appl. Probability* **9** 588–603.

CHERNOFF, H. (1956). Large sample theory: parametric case. *Ann. Math. Statist.* **27** 1–22.

DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probability* **6** 899–929. Correction Ibid. (1979) **7** 909–911.

DUDLEY, R. M. (1981). Vapnik-Červonenkis Donsker classes of functions, Aspects Statistiques et aspects physiques des processus gaussiens, Proc. Colloque CNRS St. Flour, 1980. *CNRS*, Paris, 251–269.

FLEISCHER, P. (1964). Sufficient conditions for achieving minimum distortion in a quantizer. *IEEE Int. Conv. Rec.* 104–111.

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

HARTIGAN, J. A. (1978). Asymptotic distributions for clustering criteria. *Ann. Statist.* **6** 117–131.
LECAM, L. (1981). A remark on empirical measures (preprint).
POLLARD, D. (1981a). Limit theorems for empirical processes. *Z. Wahrsch. verw. Gebiete* **57** 181–195.
POLLARD, D. (1981b). Strong consistency of k-means clustering. *Ann. Statist.* **9** 135–140.
POLLARD, D. (1982a). A central limit theorem for empirical processes. *J. Australian Math. Soc.* (Series A) *33*(2).
POLLARD, D. (1982b). Quantization and the method of k-means. *IEEE Trans. Inform. Theory* **28** 199–205.

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