Journal of Classification 2:77-108 (1985)

Journal of
Classification

On Some Significance Tests in Cluster Analysis

H. H. Bock

Technical University Aachen

Abstract: We investigate the properties of several significance tests for distinguishing between the hypothesis H of a "homogeneous" population and an alternative A involving "clustering" or "heterogeneity," with emphasis on the case of multidimensional observations $x_1, \ldots, x_n \in \mathbb{R}^p$. Four types of test statistics are considered: the (s-th) largest gap between observations, their mean distance (or similarity), the minimum within-cluster sum of squares resulting from a k-means algorithm, and the resulting maximum F statistic. The asymptotic distributions under H are given for $n \to \infty$ and the asymptotic power of the tests is derived for neighboring alternatives.

Keywords: Significance test; Homogeneity; Heterogeneity; Gap test; Minimum within-cluster sum of squares; Maximum F statistics; Asymptotic normal distribution.

1. Introduction

When a clustering algorithm is applied to a set of data, a classification of objects is obtained whether or not the data exhibit a true or "natural" grouping structure. This fact causes no problems if clustering is done for obtaining a practical (even if somewhat artificial) stratification of the given set of objects, e.g., for organizational purposes. However, if interest lies more in the recognition of an unknown clustering structure of the data (data analysis), an artificial clustering is not acceptable, and therefore the classes resulting from the algorithm must, in addition, be investigated for their relevance and their validity. Apart from descriptive, graphical, or exploratory methods, this task can be performed by using probabilistic models and suitable statistical significance tests.

Our approach is to consider a set of n p-dimensional observation points $x_1, \ldots, x_n$ in Euclidean space $\mathbb{R}^p$. We describe a series of statistical

Author's Address: Dr. H. H. Bock, Institut für Statistik und Wirtschaftsmathematik, Technical University Aachen, Wüllnerstr. 3, D-5100 Aachen, West Germany.

methods for discriminating between, on the one hand, the hypothesis H that these observations are sampled from a "homogeneous" population, and an alternative A involving "heterogeneity" or a "clustering structure," on the other hand (Section 2). More specifically, we consider the following test statistics: the largest nearest neighbor distance (3.2) between the n observations ("gap" statistic with some modifications, Section 3); some kind of mean similarity (4.1) (Section 4); the minimum within-class sum of squares (5.1) resulting, e.g., from a k-means clustering algorithm; and, finally, the corresponding maximum F statistic (5.4) (Sections 5 and 6). If, for a given significance level (error probability) α, such a test statistic exceeds the corresponding critical value $c = c\ (\alpha)$, the hypothesis H of homogeneity is rejected (e.g., in favor of a clustering structure A).

For one-dimensional data ($p = 1$), these tests or modfications of them have been proposed and analyzed repeatedly in the literature. Some type of gap statistic has been used, for testing bimodality, in Weiss (1960) and Hartigan (1977), in Giacomelli et al. (1971) who proposes a "dip intensity" measure for the histogram, or in the papers of Newell (1963), Wallenstein and Naus (1973, 1974), del Pino (1979), Kuo and Rao (1981) who consider spacing or k-spacing methods in $\mathbb{R}^1$. Concerning the squared-error criterion and the F statistic, we may cite the paper of Engelman and Hartigan (1969), chapter 4.8 of Hartigan (1975), and various proposals and investigations by Sneath (1977a, 1977b, 1979a, 1979b) and Barnett et al. (1979). A more complete review is given in Bock (1981).

In the present paper, we shall emphasize the multidimensional case $p>1$. Then the crucial points are, in practice, the calculation of the critical threshold $c\ (\alpha)$ and the performance (power) of the tests under an alternative of clustering. We investigate these questions in an asymptotic sense for $n\to\infty$. We review and derive results on the asymptotic distribution of the test statistics under H, and characterize the asymptotic power of some tests under neighboring alternatives $A = A_n$ approaching H for $n\to\infty$. In particular, by recourse to a paper of Pollard (1982a), we obtain the asymptotic distribution of the maximum F statistic (5.4) from variance analysis and thereby generalize a theorem of Hartigan (1978) to the multivariate case (corollary 6.5 and theorem 6.6; these reults have been cited in Bock 1983). Moreover, it appears that the asymptotic power of the gap test may be characterized by a speed factor $(\log n)^{-1}$ (for A_n converging to H), and by a factor $n^{-1/4}$ for tests based on the mean similarity (4.1). Thus, the situation is worse than in classical likelihood procedures for parametric problems where the factor $n^{-1/2}$ occurs, and we cannot expect an excessive discrimination power from these tests.

A more thorough discussion of significance testing and evaluation methods in cluster analysis is provided by Dubes and Jain (1979), Bock (1981), and Perruchet (1982, 1983).

2. Probabilistic Models for Homogeneity and Heterogeneity

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be n p-dimensional observations which represent n objects under investigation and which are to be analyzed for homogeneity resp. a clustering structure. We shall use a probabilistic model and consider $x_1, \ldots, x_n$ as realizations of n independent p-dimensional random vectors $X_1, \ldots, X_n$, all with the same distribution density $f(x)$. Depending on the shape of $f(\cdot)$ we shall speak of a homogeneous or a clustered population. As for "homogeneity," an intuitive definition would be the

Uniformity hypothesis H_G:

$X_1, \ldots, X_n$ have a uniform distribution in a given bounded, open and connected set $G \subset \mathbb{R}^p$ with $|G| := vol(G) > 0$, i.e.,

$$f(x) = f_0(x) := \mathbf{1}_G(x)/|G| \quad . \tag{2.1}$$

Here $\mathbf{1}_G(\cdot)$ denotes the characteristic function of the set G: $\mathbf{1}_G(x) = 1$ or 0 if $x \in G$ resp. $x \notin G$. In practice, a relevant set G will be assumed to exhibit further "nice" properties, e.g., to be convex or to have an ellipsoidal or a rectangular form (see remark 3.2).

Another interpretation of "homogeneity" which is most suitable for many applications, underlies the

Unimodality hypothesis H_0:

We have $f(x) = h(x)$ with some unimodal density $h(\cdot)$.

Note that a density $h(\cdot)$ is called *unimodal* if there exists a point $\mu \in \mathbb{R}^p$ (the mode of h) such that for each unit vector (direction) $u \in \mathbb{R}^p$, the density $f(\mu + tu)$ is a strictly decreasing function of the real variable $t \geqslant 0$. In practice, there will be additional restrictions for f (e.g., continuity, convex contour lines, etc.).

In Figures 1.a and 1.b we have depicted two "homogeneous" samples from H_G resp. H_0; in contrast, Figures 1.c and 1.d illustrate two samples which indicate some typical clustering structure. In particular, Figure 1.c refers to the

Alternative A^*:

The common density $f(x)$ is multimodal, i.e., there exists a finite number $k \geqslant 2$ of distinct points $\bar{\mu}_1, \ldots, \bar{\mu}_k \in \mathbb{R}^p$ (the modes of f) where $f(\cdot)$ attains a strict relative maximum.

Thus, in a suitable graphical representation of $f(\cdot)$ or of the empirical point density, we may recognize "hills" and separating "valleys" where gaps between observations are much larger than is to be expected, e.g.,

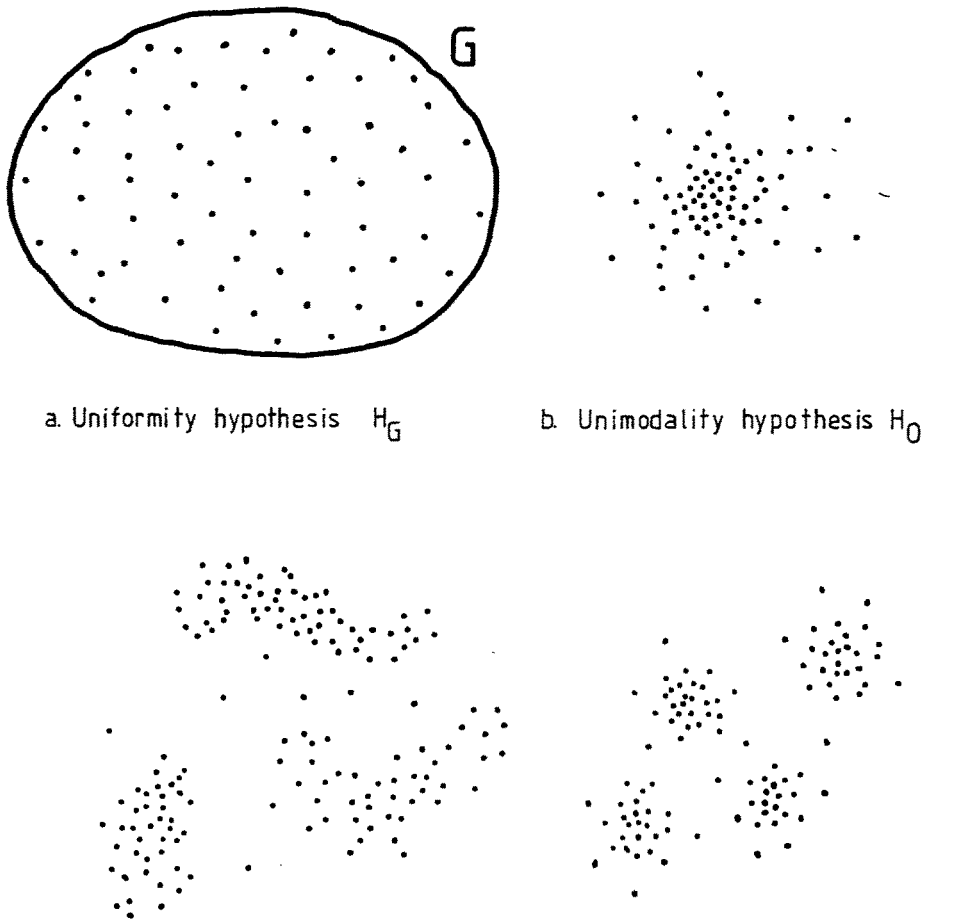

Figure 1. Samples under the models H_G, H_0 , A^*, and A.

under H_G. Therefore a "gap test" seems to be advisable in this case (see Section 3. A similar argumentation applies to some weaker definitions of unimodality or multimodality where the strictness condition is relaxed suitably). Another type of clustering is described by the

Mixture alternative A:

The common density $f(x)$ is a translation mixture of the form

$$f(x) = \sum_{i=1}^{k} p_i \cdot h(x - \mu_i) \tag{2.2}$$

with $k \geqslant 2$ different (subpopulation) centers $\mu_1, \ldots, \mu_k \in \mathbb{R}^p$, a density $h(x)$ describing the shape of the clusters, and k class proportions $p_1, \ldots, p_k > 0$ with $\Sigma_1^k p_i = 1$ (all μ_i, p_i, k unknown).

In typical situations where $h(\cdot)$ is continuous and unimodal and the class centers $\mu_1, \ldots, \mu_k$ are sufficiently distant, the density (2.2) will show several modes and can be subsumed under A^*. It will be made evident in Section 4 that the mixture case (2.2) can be tackled using some mean similarity criterion provided that the shape $h(\cdot)$ of the clusters is known. In other cases we may use the maximum F statistic in Section 5 for discriminating between A and H_0.

3. Gap Tests

For testing the uniformity hypothesis H_G, we consider, for each $j = 1, \ldots, n$, the minimum Euclidean distance U_{nj} from the sample point X_j to all other points X_ν $(\nu \neq j)$ resp. to the boundary ∂G of G:

$$U_{nj} := Min\{ \min_{\nu} \{||X_j - X_\nu||\}, ||X_j - \partial G||\} \qquad j = 1, \ldots, n \tag{3.1}$$

(a modified nearest neighbor distance). Then the gap statistic

$$D_n := Max\{U_{n1}, \ldots, U_{nn}\} \tag{3.2}$$

is the radius of the largest ball which can be centered at some X_j without containing, in its interior, some other point of $\{X_1, \ldots, X_n\}$ or ∂G. It is the basic criterion of the *gap test*: Reject H_G if and only if $D_n > c$. The threshold $c = c_n(\alpha)$ is to be calculated from $P_G(D_n > c) = \alpha$, the given error probability of the first kind.

Since the exact distribution of D_n under H_G is very intricate for finite n, we use, for calculating c, an asymptotic result given by Henze (1981, 1982):

Theorem 3.1 *For any bounded, open and connected set $G \subset \mathbb{R}^p$ with volume $|G|$, we have under H_G, for all $t \in \mathbb{R}$:*

$$\lim_{n \to \infty} P_G(n \nu_p \cdot D_n^p / |G| - \log n \leqslant t) = e^{-e^{-t}} =: L(t) \tag{3.3}$$

where $\nu_p := \pi^{p/2} / \Gamma(1 + p/2)$ is the volume of the unit ball in $\mathbb{R}^p$.

Thus, up to a linear transformation, the asymptotic distribution of D_n^p is Gumbel's extreme value distribution and for large n, we may approximately use the corresponding critical value

$$\tilde{c}_n(\alpha) := \{|G| \cdot [\log n - \log(-\log(1-\alpha))]/(n\, v_p)\}^{1/p}$$

for D_n.

Remark 3.1: Actually, at least for $p = 1$ and 2, the term $||X_j - \partial G||$ may be omitted in (3.1) without changing the asymptotic results given above (and below) provided that ∂G is a simply closed and twice differentiable curve; but no analogue seems to be valid for $p \geqslant 3$ (Henze 1981).

It is possible to characterize the asymptotic power of the gap test by using a sequence of neighboring ("contiguous") alternatives A_n approaching H_0 for $n \to \infty$. Let $X_1, \ldots, X_n$ each have the density $f_n(x) := f_0(x) + \Delta(x)/\log n$ with an arbitrary continuous and bounded function $\Delta(x)$ with $\int \Delta(x)\, dx = 0$. Then under a mild regularity assumption, namely $\sup\{|\Delta(x) - \Delta(y)| \mid ||x - y|| < \epsilon\} = o(1/(-\log \epsilon))$ for $\epsilon \to 0$, we have (Henze 1981):

$$\lim_{n\to\infty} P_{A_n}(n\, v_p \cdot D_n^p/|G| - \log n - \delta \leqslant t) = L(t) \quad \text{for } t \in \mathbb{R} \tag{3.4}$$

with a non-centrality parameter

$$\delta := \log\left\{\int_G \exp\left\{-\frac{\Delta(x)}{f_0(x)}\right\} \cdot f_0(x)\, dx\right\} \geqslant 0 \quad . \tag{3.5}$$

Thus the asymptotic error probability of the second kind is given by

$$\beta := \lim_{n\to\infty} P_{A_n}(D_n \leqslant c_n(\alpha)) = L(L^{-1}(1-\alpha) - \delta) = (1-\alpha)^{e^{\delta}} \quad .$$

The routine application of a gap test requires some practical comments:

Remark 3.2: The gap test can be performed only if the region G is known from the outset. It is evident from (3.1) that the choice of G has a great influence on the test statistic D_n. If G is not given beforehand, we must, in practice, resort to some approximate, conservative estimate G^* for G, e.g., a concentration ellipsoid or the convex hull of the observations (retaining only those in the interior of G^*). As an alternative, we may transform our data, linearly and componentwise, to the interval $[0,1]$ such that $G = [0,1]^p$ will be a suitable choice. However, such a transformation can be delicate since it introduces a different weighting of distances measured along the p original coordinate axes.

Remark 3.3: Since convergence is slow in (3.3), $\tilde{c}_n(\alpha)$ may be an inaccurate estimate for $c_n(\alpha)$ for finite n. However, we may profit from the fact that the relation (3.3) holds in the same form if the nearest neighbor

distances $U_{n1}, \ldots, U_{nn}$ were assumed to be independent (see Henze 1981). In this case, the equation for c reads:

$$P\,(D_n > c) = 1 - P\,(D_n \leqslant c) = 1 - \left[P_G\,(U_{n1} \leqslant c)\right]^n = \alpha \quad . \tag{3.6}$$

If the distribution of U_{n1} is known, solving for c gives another approximation for $c_n\,(\alpha)$ for medium-sized n. This approximation proves to be excellent, e.g., in the two-dimensional case $p = 2$ with $G = [0,1]^2$ the unit square, and

$$P_G\,(U_{n1} \leqslant c) = 1 - (1-2c)^2\,(1-\pi\,c^2)^{n-1}, \quad 0 \leqslant c \leqslant \tfrac{1}{2} \tag{3.7}$$

provided that $n \geqslant 20$ (Henze 1982).

There are several modifications and generalizations of the test statistic D_n and the corresponding gap test.

Given some integer r with $1 \leqslant r \leqslant n$, we may use $D_{nr} := U_{n(n+1-r)}$, the rth largest nearest neighbor distance, instead of D_n; here $U_{n(1)} \leqslant, \ldots, \leqslant U_{n(n)}$ denote the order statistics of $U_{n1}, \ldots, U_{nn}$. Then the asymptotic formulas (3.3) and (3.4) must be modified according to

$$\lim_{n\to\infty} P_G\,(n \cdot v_p \cdot D_{nr}^p / |G| - \log n \leqslant t) = L(t) \cdot \sum_{s=0}^{r-1} \frac{e^{-st}}{s!} \tag{3.3*}$$

$$\lim_{n\to\infty} P_{A_n}(n \cdot v_p \cdot D_{nr}^p / |G| - \log n - \delta \leqslant t) = L(t) \cdot \sum_{s=0}^{r-1} \frac{e^{-st}}{s!} \tag{3.4*}$$

where, on the right hand side, Smirnov's limit distribution is obtained (Henze 1982).

Remark 3.4: Gap statistics of this type are related to the well-known single linkage dendrogram from cluster analysis based on the $n(n-1)/2$ Euclidean distances $d_{ij} = ||x_i - x_j||$. The statistic D_{nr} (calculated without the term $||X_j - \partial G||$ in (3.1)) is just the level at which the rth last single object joins some other class in the single linkage dendrogram. Thus the resulting D_{nr}-test may be interpreted in the framework of hierarchies. In the one-dimensional case, a similar idea has been followed by Weiss (1960) and Hartigan (1977) who investigated the maximum gap in the minimum spanning tree for testing bimodality.

To obtain a non-trivial limiting error probability $0 < \beta < 1$, the contiguous densities f_n have to approach f_0 with the order $1/\log n$, which is very

slow. This result indicates some difficulty in distinguishing H_G from its (nonspecified omnibus) alternatives with the aid of a gap test (analogous difficulties are well known from nearest-neighbor density estimation). There is strong evidence that we can improve on this by using, in the definition of $D_n = D_{n1}$ or D_{nr}, the distance U_{nj}^* of x_j to its *s-th nearest* neighbor (respectively to ∂G) instead of U_{nj}, and to choose for the integer s a sequence $s = s_n$ approaching ∞ for $n \to \infty$. However, no exact results are known here, and in the multidimensional case, practical considerations (computer time and storage requirements) will restrict s to be small for large n.

For a fixed integer s, the modified statistic D_n^* (based on $U_{n1}^*, \ldots, U_{nn}^*$) has been investigated by Henze (1982, case a.). Similar tests for homogeneity (or goodness-of-fit) have been formulated using b., the empirical distribution of the transformed nearest-neighbor distances (Bickel and Breiman 1983), or in the one-dimensional case, c., the average of suitably transformed k-spacings (del Pino 1979, Kuo and Rao 1981). All papers show the same problems in performance as cited above. The tests are unable to detect alternatives approaching the uniform, respectively, a. at a rate faster than $1/\log n$, b. at a rate $1/\sqrt{n}$ (this may be cured by a suitable weighting depending on the alternative, as suggested by Schilling 1983a, 1983b), and c. at a rate faster than $n^{-1/4}$ (for symmetric statistics).

4. Tests Using Mean Similarity for Mixture Models

While in the last section the ranking of pairwise distances was the predominant feature, we will investigate, in this section, the use of a mean distance (or a mean similarity) in the clustering framework. More specifically, we will consider cases where "homogeneity" is understood in the sense of the unimodality hypothesis H_0, and "clustering" is described by a mixture alternative A, (2.2), where the within-group distribution density $h(\cdot)$ is the same as under H_0. From intuitive grounds (e.g., from a look at Figure 1.b and 1.d) it may be conjectured that H_0 and A can be distinguished by comparing the mean similarity of all observations with some standard threshold c (another motivation is given in remark 4.1).

To be specific, denote by $s = s(x,y)$ a similarity index describing the nearness of two points $x,y \in \mathbb{R}^p$, e.g., a decreasing function $s = q(\Delta)$ of the Euclidean distance $\Delta = ||x-y||$ or some symmetric function $s = \bar{q}(x-y)$ of the difference $x-y$ (see lemmas 4.1 and 4.2). The mean of all pairwise similarities $S_{jl} := s(X_j, X_l)$ of the observations $X_1, \ldots, X_n$ is defined by

$$T_n := \binom{n}{2}^{-1} \sum_{1=j<l=n} \sum S_{jl} \quad . \tag{4.1}$$

The *mean similarity test* rejects H_0 iff $T_n < c$ with some critical threshold $c = c_n(\alpha)$.

This test is motivated by the conjecture that, under the mixture alternative A, the mean T_n will be smaller in some sense than under H_0. For two quite general types of similarity indices this is verified by the following lemmas 4.1 and 4.2. As for the first one, denote by

$$g(\eta) := \int h(x)\, h(x+\eta)\, dx \tag{4.2}$$

the (symmetric) distribution density of the difference $Y := X_1 - X_2$ of two observations under H_0.

Lemma 4.1 *If $h(\eta)$ is a decreasing function $\gamma(||\eta||)$ of the Euclidean norm $||\eta||$, then the distance $\Delta := ||X_j - X_l||$ is stochastically larger under A than under H_0, i.e.,*

$$P_A(\Delta \leqslant d) \leqslant P_0(\Delta \leqslant d) \tag{4.3}$$

holds for all $d \geqslant 0$. This implies that for any decreasing function $q(\cdot)\colon \mathbb{R}_+ \to [0,1]$, the similarity $S := q(\Delta)$ is stochastically smaller under A than under H_0:

$$P_A(S \leqslant s) \geqslant P_0(S \leqslant s) \quad \text{for} \quad s \in \mathbb{R}_+ \tag{4.4}$$

and the same ordering holds for the expectation of S:

$$\lambda_A := E_A[S] \leqslant E_0[S] =: \lambda_H \tag{4.5}$$

Proof: For any $\tau \in \mathbb{R}^p$ and $d \geqslant 0$, consider balls $A := \{||\eta|| \mid ||\eta - \tau|| \leqslant d\}$ and $B := \{||\eta|| \mid ||\eta|| \leqslant d\}$ in $\mathbb{R}^p$ centered at τ resp. 0. Provided that $A \cap \bar{B} \neq \emptyset$ (i.e., $||\tau|| \leqslant 2d$), there is a one-to-one linear mapping of the cap $A \cap \bar{B}$ to the congruent cap $\bar{A} \cap B$ by reflecting it at the hyperplane passing through $\tau/2$ and orthogonal to the vector $\tau\colon y = \eta - 2\tau'(\eta - \tau/2)\cdot\tau/||\tau||^2$ (with Jacobian -1). For any $\eta \in \bar{B}$, $y \in B$ we have $||\eta|| \geqslant d \geqslant ||y||$ and $\gamma(||\eta||) \leqslant \gamma(d) \leqslant \gamma(||y||)$ since $\gamma(\cdot)$ is decreasing. Therefore, by a change of variable:

$$\int_{A \cap \bar{B}} \gamma(||\eta||)\, d\eta \leqslant \int_{A \cap \bar{B}} \gamma(d)\, d\eta = \int_{\bar{A} \cap B} \gamma(d)\, dy \leqslant \int_{\bar{A} \cap B} \gamma(||y||)\, dy$$

and we obtain:

$$\pi(\tau) := P_0(||Y-\tau|| \leqslant d) = \int_A h(\eta)\, d\eta$$

$$= \int_{A\cap B} \gamma(||\eta||)\, d\eta + \int_{A\cap \bar{B}} \gamma(||\eta||)\, d\eta$$

$$\leqslant \int_{A\cap B} \gamma(||\eta||)\, d\eta + \int_{\bar{A}\cap B} \gamma(||y||)\, dy$$

$$= \int_B \gamma(||\eta||)\, d\eta = P_0(||Y|| \leqslant d) = P_0(\Delta \leqslant d) = \pi(0) \ .$$

(A similar argument shows that $\pi(\tau)$ is a decreasing function of $||\tau||$.) Applying this inequality, we can prove (4.3):

$$P_A(\Delta \leqslant d) = \sum_i \sum_{i'} p_i\, p_{i'} \cdot P_0(||X_j - X_l + \mu_i + \mu_{i'}|| \leqslant d)$$

$$= \sum_i \sum_{i'} p_i\, p_{i'} \cdot P_0(||Y + \mu_i - \mu_{i'}|| \leqslant d)$$

$$= \sum_i \sum_{i'} p_i\, p_{i'} \cdot \pi(\mu_i - \mu_i)$$

$$\leqslant \pi(0) \cdot \sum_i \sum_{i'} p_i\, p_{i'} = \pi(0) = P_0(\Delta \leqslant d) \ .$$

Finally, the implication (4.3) => (4.4) => (4.5) follows by well-known theorems (e.g., Lehmann 1955). •

It is evident that we will obtain a quite analgous result when considering an *increasing* function $q(||X_j - X_l||)$ of pairwise distances and a corresponding "mean distance test." However, in this paper, we prefer a formulation using similarities because there is another similarity index whose investigation parallels the findings above, but which allows no obvious distance analog. This similarity index is generated by a kernel (a distribution density) $K(x)$ and the symmetric function $\bar{q}(y) := \int K(x)\, K(x-y)\, dx$ (a convolution density) according to the formula:

$$S = S_{jl} := \bar{q}(X_j - X_l) = \int K(x-X_j)\, K(x-X_l)\, dx \ . \qquad (4.6)$$

This definition is motivated by remark 4.1. We tacitly assume $K(\cdot)$ and $\bar{q}(\cdot)$ to be bounded. Note that if $K(\cdot)$ is spherically symmetric with center 0, then $\bar{q}(\cdot)$ has this property, too. If, additionally, $K(\cdot)$ is

unimodal then the convolution $\bar{q}\ (y)$ is again unimodal with mode 0 (this follows from theorem 4 of Wolfe 1975) and (4.6) reduces to the former type $S = q\ (\Delta)$.

Lemma 4.2 *For a similarity index $S = \bar{q}\ (X_j - X_l)$ of the type (4.6) the inequality (4.5) holds, i.e., S is, on the average, smaller under the mixture alternative A than under the hypothesis H_0.*

Proof: We introduce the function

$$k(\tau) := E_0\Big[K(\tau - X_1)\Big] = \int K(\tau - x)\ h(x)\ dx$$

and apply the relation $\bar{q}\ (\xi - n) = \int K(x-\xi)\ K(x-\mu)\ dx$. Then, using the independence of X_j and X_l for $j \neq l$:

$$\begin{aligned}
\lambda_A &= E_A\Big[\bar{q}(X_j - X_l)\Big] = \int E_A\Big[K(x-X_j)\ K(x-X_l)\Big]\ dx \\
&= \int E_A\Big[K(x-X_j)\Big] \cdot E_A\Big[K(x-X_l)\Big] dx = \int E_A\Big[K(x-X_1)\Big]^2 dx \\
&= \int \left\{\sum_{r=1}^{k} p_r\ k(x-\mu_r)\right\}^2 dx \\
&\leqslant \int \left\{\sum_r p_r\ k^2(x-\mu_r)\right\} dx = \sum_r p_r \int k^2(x-\mu_r)\ dx \\
&= \left(\sum_r p_r\right) \cdot \int k^2(x)\ dx = \int k^2(x)\ dx = \lambda_H
\end{aligned} \tag{4.7}$$

where the inequality sign follows, e.g., from Jensen's inequality. Equality is possible only in the trivial case when $k\ (x)$ is a constant. ●

Remark 4.1: Another motivation for the mean similarity test is provided by considering the number $N\ (\epsilon)$ of ϵ-neighboring pairs $\{X_j, X_l\}$ (i.e., with $||X_j - X_l|| \leqslant \epsilon$ and $j, l = 1, \ldots, n$, $j \neq l$). It has been shown by Eberl and Hafner (1971) that, for $n \to \infty$ and $\epsilon = \epsilon_n = n^{-1/p} \to 0$, $N\ (\epsilon_n)$ has an asymptotic Poisson distribution with expectation proportional to $I := \int f^2\ (x)\ dx$. Since neighboring pairs are conjectured to be more numerous under homogeneity, this integral is expected to be larger under H_0 than under A (the formal proof is similar to (4.7)). This remark may be the basis of a test procedure: First, the unknown density $f\ (\cdot\ )$ of our

observations is estimated by a kernel type estimator $\hat{f}_n(\cdot)$ (using the kernel $K(x)$); by insertion we obtain an estimate $\hat{I}_n := \int \hat{f}_n^2(x)\,dx$ for I; finally the hypothesis H_0 is rejected (in favor of A) iff $\hat{I}_n$ is smaller than some critical threshold. It turns out that this test is equivalent to the mean similarity test given above with the special similarity index (4.6) (Bock 1977).

According to these results, the mean similarity T_n seems to be an acceptable statistic for testing H_0 against A. However, its suitability and practicability depend critically on two assumptions: (i) that the density $h(\cdot)$ of the underlying distributions is known beforehand except for an unknown translation vector (since otherwise the critical level c for T_n cannot be obtained; see (4.16) below); (ii) that only translation mixture alternatives are involved. In practice, assumptions of this type are met when clustering results from a displacement of a known, homogeneous situation by intentional actions or accidental effects with unknown, group-specific shifts (e.g., the consumer behavior before and after a marketing strategy, or the dissemination of plant colonies from a common center). In particular, we must observe that any mean similarity T_n is sensitive to a change of scale (replacing $h(x)$ by $h(x/\sigma)/\sigma^p$ with some $\sigma > 1$ reduces, e.g., the mean λ_H of T_n). Therefore the mean similarity test is susceptible to confounding the effects of scale and clustering and can be applied only if the former possibility has been excluded from the outset. (In Section 5 we shall present a scale invariant test involving an optimal classification.)

Remark 4.2: At first look, the problem H_0 versus A is tempting to use a maximum likelihood ratio test (Wolfe 1970). However, for doing so, the exact class number k must be known additionally. Moreover, neither the exact nor the asymptotic distribution of the log likelihood ratio statistic is known in this case: the usual asymptotic theory (resulting in an asymptotic χ^2-type distribution) breaks down since the parameters in (2.2) are not identifiable under H_0 and Fisher's information matrix proves to be singular. (This question has been discussed by Wolfe (1971), Hartigan (1977b) ch. 2 and 6, Binder (1978) and others; some theoretical results have been obtained recently by Ghosh and Sen 1984). In constrast, for the mean similarity test, we are able not only to obtain the asymptotic distribution of the test statistic T_n (theorem 4.3), but also to characterize the asymptotic power of the test under local alternatives (see theorem 4.5 and (4.24)).--For both tests, the distribution density $h(\cdot)$ has to be specified; in practice, we will try, e.g., a multivariate normal distribution (example 4.1).

The mean similarity T_n is a U-statistic. Therefore its asymptotic distribution under H_0 or A can be inferred from the general theory of U-statistics (Silverman 1976, Randles and Wolfe 1979, theorem 3.3.13). For describing the results, we need the following auxiliary functions (conditional or unconditional means):

$$t(x) := E_0\left[\bar{q}(X_1-X_2)\,|\,X_2 = x\right] = \int \bar{q}(\xi-x)\; h(\xi)\; d\,\xi \tag{4.8}$$

$$t^*(x) := E_A\left[\bar{q}(X_1-X_2)\,|\,X_2 = x\right] = \sum_{r=1}^{k} p_r\; t(x-\mu_r) \tag{4.9}$$

$$\psi_1(y) := E_0\left[\bar{q}(X_1-X_2+y)\right] = \int \bar{q}(\zeta+y)\; g(\zeta)\; d\,\zeta \tag{4.10}$$

$$= \int k(x)\; k(x+y)\; dx$$

$$\psi_2(a,b) := E_0\left[\bar{q}(X_1-X_2+a)\;\bar{q}(X_1-X_3+b)\right] \tag{4.11}$$

$$= E_0\left[t(X_1+a)\;t(X_1+b)\right] = \int t(\xi+a)\;t(\xi+b)\;h(\xi)\;d\,\xi$$

$$\psi_3(y) := E_0\left[\bar{q}^2(X_1-X_2+y)\right] = \int \bar{q}^2(\zeta+y)\; g(\zeta)\; d\,\zeta\,. \tag{4.12}$$

(These definitions refer to the case $s\,(x,y) = \bar{q}\,(x-y)$ in (4.6); the similarity index $s\,(x,y) = q\,(||x-y||)$ may be treated analogously.)

The case of the hypothesis $H = H_0$ is summarized in theorem 4.3:

Theorem 4.3: *Under H_0 the expectation λ_H and the variance $\sigma^2_{H,n}$ of the mean similarity T_n are given by*

$$\lambda_H := E_0\left[T_n\right] = E_0\left[S_{12}\right] = \int k^2(x)\; dx = \psi_1(0) \tag{4.13}$$

$$\sigma^2_{H,n} := Var_0(T_n) = \left[\sigma^2_H + 2(n-2)\;\tau^2_H\right] / \binom{n}{2} \tag{4.14}$$

with

$$\sigma^2_H := Var_0(S_{12}) = \psi_3(0)-\psi_1^2(0)$$

$$\tau^2_H := Cov_0(S_{12},S_{13}) = Var_0(t(X_1)) = \psi_2(0,0)-\psi_1^2(0)$$

and

$$\lim_{n\to\infty}\; n\cdot\sigma^2_{H,n} = 4\tau^2_H\;. \tag{4.15}$$

Provided that $0<\sigma^2_H<\infty$ and $E_0\left[S_{12}^2\right]<\infty$, the standardized test statistic T_n has, under H_0 and for $n\to\infty$, an asymptotic normal distribution:

$$(T_n - \lambda_H)/\sigma_{H,n} \xrightarrow{L} N(0,1)$$

where $\xrightarrow{L}$ *means convergence in distribution.*

Thus the critical threshold $c = c_n(\alpha)$ for T_n may be approximated by

$$\hat{c}_n(\alpha) := \lambda_H + u_\alpha \cdot \sigma_{H,n} \tag{4.16}$$

with $\phi(u_\alpha) = \alpha$ and ϕ the distribution function of $N(0,1)$. We omit the details of the proof of theorem 4.3. Note that (4.15) implies that under the stated conditions:

$$\sqrt{n}\,(T_n - \lambda_H) \xrightarrow{L} N(0, 4\tau_H^2) \ .$$

Example 4.1: For illustration, let us consider the case where "homogeneity" is described by a p-dimensional normal distribution $N_p(0, \sigma^2 I_p)$ with a known variance σ^2 and the density $h(x) := (2\pi\sigma^2)^{-p/2} \cdot \exp\left\{-||x||^2/2\sigma^2\right\}$. Using the kernel $K(t) := (4\beta/\pi)^{p/4} \exp\left\{-2\beta\,||t||^2\right\}$ with some smoothing parameter $\beta > 0$, we obtain the similarity index $S = \bar{q}(X_j - X_l) = \exp\left\{-\beta\,\Delta^2\right\}$ with $\Delta := ||X_j - X_l||$ and the test statistic

$$T_n = \binom{n}{2}^{-1} \sum_{j<} \sum_{l} \exp\left\{-\beta\,||X_j - X_l||^2\right\} \ .$$

Under H_0, its expectation is given by:

$$\lambda_H = E_0[T_n] = \left(1 + 4\beta\sigma^2\right)^{-p/2}$$

and its variance is (4.14) with

$$\sigma_H^2 = \left(1 + 8\beta\sigma^2\right)^{-p/2} - \left(1 + 4\beta\sigma^2\right)^{-p}$$

$$\tau_H^2 = \left[\left(1 + 2\beta\sigma^2\right)\left(1 + 6\beta\sigma^2\right)\right]^{-p/2} - \left(1 + 4\beta\sigma^2\right)^{-p} .$$

Some simulation studies show that, e.g., for $1 \leqslant p \leqslant 4$, $15 \leqslant \beta\sigma^2 \leqslant 45$, $n = 100$, the normal approximation (4.16) of the 5% point $c_n(0.05)$ of T_n is satisfactory if $\lambda_H / \sigma_{H,n} \geqslant 2{,}5$ (thus avoiding skewness).

In the remainder of this section we investigate the asymptotic power of the mean similarity test. At first, we remark that a central limit theorem holds for T_n under any fixed mixture alternative A as well. Formally it is obtained from theorem 4.3 by substituting there the mixture density (2.2) for the density $h(\cdot)$. Some elementary, but tedious calculations lead to the following result (Bock 1977):

Corollary 4.4: *Under the mixture alternative (2.2), the expectation* λ_A *and the variance* $\sigma^2_{A,n}$ *of* T_n *are given by*

$$\lambda_A := E_A[T_n] = E_A[S_{12}] = \sum_{r=1}^{k} \sum_{s=1}^{k} p_r\, p_s\, \psi_1(\mu_r - \mu_s) \tag{4.17}$$

$$\sigma^2_{A,n} := Var_A(T_n) = \left[\sigma^2_A + 2(n-2)\tau^2_A\right] / \binom{n}{2} \tag{4.18}$$

with

$$\sigma^2_A := Var_A(S_{12}) = \sum_{r=1}^{k} \sum_{s=1}^{k} p_r\, p_s\, \psi_3(\mu_r - \mu_s) - \lambda^2_A$$

$$\tau^2_A := Cov_A(S_{12}, S_{13}) = Var_A(t^*(X_1)) \tag{4.19}$$

$$= \sum_{r=1}^{k} \sum_{s=1}^{k} \sum_{t=1}^{k} p_r\, p_s\, p_t \cdot \psi_3(\mu_r - \mu_s, \mu_r - \mu_t) - \lambda^2_A \quad .$$

This corollary will be used in the proof of theorem 4.5.

For characterizing the asymptotic power of the mean similarity test, we consider a sequence of mixture alternatives $A = A_n$ generated by (2.2) with class means of the form

$$\mu_i = \mu + n^{-1/4} \cdot z_i \quad i = 1, \ldots, k \tag{4.20}$$

which converge to some central point μ with the order $n^{-1/4}$ (the class proportions $p_1, \ldots, p_k$ are held fixed). We will prove:

Theorem 4.5: *Suppose that* $E_0\left[S^2_{12}\right] < \infty$, $0 < \tau^2_H < \infty$ *and that the function* $\psi_1(y)$, *(4.10), has continuous second derivatives (at* $y = 0$*). Then, under* A_n *and asymptotically for* $n \to \infty$, *we have:*

$$\sqrt{n}\ (T_n - \lambda_H) \xrightarrow{L} N(-\delta^2, 4\tau^2_H) \quad , \tag{4.21}$$

where the non-centrality parameter

$$\delta^2 = \sum_{i=1}^{k} p_i (z_i - \bar{z})'\, G\, (z_i - \bar{z}) \geqslant 0, \tag{4.22}$$

is calculated from the positive semi-definite $p \times p$ matrix of G of the second partial derivatives of $\psi_1(y)$ at $y = 0$:

$$G := -D_2\psi_1(0) = \int [grad\, k(\eta)][grad\, k(\eta)]'\, d\eta \tag{4.23}$$

and the weighted average $\bar{z} := \Sigma\, p_i\, z_i$.

By a simple application of (4.21), we are able now to calculate the asymptotic error probability of the second kind for the mean similarity test:

$$\begin{aligned}\beta &= \lim_{n\to\infty} P_{A_n}((T_n-\lambda_H)/\sigma_{H,n} \geqslant u_\alpha) \\ &= \lim_{n\to\infty} P_{A_n}(\sqrt{n}\ (T_n-\lambda_H)/(2\sigma_H) \geqslant u_\alpha) \\ &= 1-\phi(u_\alpha + \delta/2\tau_H) = \phi(-u_\alpha-\delta/2\tau_H)\ .\end{aligned} \tag{4.24}$$

It is seen from (4.20) that the value $\sqrt{n}\ \delta^2$ measures the spread of the class centers $\mu_1, \ldots, \mu_k$ around their (weighted) mean $\mu + \bar{z}/n^{1/4}$. In particular, for the normal density case of example 4.1 the non-centrality parameter is given by

$$\delta^2 = 2\beta\ (1 + 4\beta\ \sigma^2)^{-1-p/2} \cdot \sum_{r=1}^{k} p_r\ ||z_r-\bar{z}||^2\ .$$

The (low) rate of convergence $n^{-1/4}$ in (4.20) indicates some weakness of the mean similarity test in recognizing real spurious clusterings for finite n.

Proof of theorem 4.5. The proof relies on the representation

$$\sqrt{n}\ (T_n-\lambda_H) = \sqrt{n}\ (T_n-\lambda_A-Q_n) + \sqrt{n}\ Q_n + \sqrt{n}\ (\lambda_A-\lambda_H)$$

with the random variable

$$Q_n := \frac{2}{n}\sum_{i=1}^{n} (t^*(X_i)-\lambda_A) \tag{4.25}$$

containing n independent, centered terms calculated from the conditional mean t^* in (4.9) (for the sake of brevity we write A for A_n). We shall show that under this sequence for $n\to\infty$:

(a) $\sqrt{n}\ (\lambda_A - \lambda_H) \to -\delta^2$

(b) $d_n := n \cdot E_A[(T_n - \lambda_A - Q_n)^2] \to 0$ and therefore
$\sqrt{n}\ (T_n - \lambda_A - Q_n) \to 0$ in probability

(c) $\sqrt{n}\ Q_n \xrightarrow{L} N(0, 4\tau_H^2)$.

Then the result (4.21) follows from Slutzky's theorem (e.g., Randles and Wolfe 1978, p. 72).

(a) Suppose the function $\psi_1(y)$ to be sufficiently smooth to have the following Taylor expansion for $||y|| \to 0$:

$$\psi_1(y) - \psi(0) = y'\ grad\ \psi_1(0) - \tfrac{1}{2}\, y'\, Gy + o(||y||^3)$$

with the matrix $G := -D_2\, \psi_1(0)$. Inserting $y = \mu_r - \mu_s = (z_r - z_s)/\ n^{1/4} = O(n^{-1/4})$ and using the representations (4.13), (4.17) for λ_H, λ_A we obtain:

$$\begin{aligned}
\sqrt{n}\ (\lambda_A - \lambda_H) &= \sqrt{n} \sum_{r=1}^{k} \sum_{s=1}^{k} p_r\ p_s [\psi_1(\mu_r - \mu_s) - \psi_1(0)] \\
&= -\tfrac{1}{2} \sum_{r=1}^{K} \sum_{s=1}^{k} p_r\ p_s (z_r - z_s)'\ G\ (z_r - z_s) + o(n^{-1/4}) \\
&= -\sum_{r=1}^{k} p_r (z_r - \bar{z})'\ G (z_r - \bar{z}) + o(n^{-1/4}) \\
&= -\delta^2 + o(n^{-1/4}) \ ,
\end{aligned}$$

since the linear term $\Sigma_r\ \Sigma_s\ p_r\ p_s(\mu_r - \mu_s) = \Sigma_r\ p_r\ \mu_r - \Sigma_s\ p_s\ \mu_s = 0$ cancels out. The special representation (4.23) for G follows immediately by differentiation of

$$\begin{aligned}
grad\, \psi_1(y) &= grad \left(\int k(x)\ k(x+y)\, dx \right) \\
&= \int k(x) [grad\ k(x+y)]\, dx \\
&= \int k(\eta - y) [grad\ k(\eta)]\, d\eta
\end{aligned}$$

and inserting $y = 0$. This derivation holds e.g., if $||grad\ k(\eta)||$ is bounded; other representations for G are (under suitable regularity assumptions):

$$G = -\int D_2\, \bar{q}(\zeta) \cdot g(\zeta)\, d\zeta = -\int D_2\, g(\zeta) \cdot \bar{q}(\zeta)\, d\zeta \quad .$$

(b) Since $T_n - \lambda_A$ and Q_n are centered variables we have

$$d_n := n \cdot E_A\left[(T_n - \lambda_A - Q_n)^2\right] = n \cdot Var_A(T_n - \lambda_A - Q_n)$$

$$= n\left[Var_A(T_n) + Var_A(Q_n) - 2Kov_A(T_n, Q_n)\right]$$

where

$$Var_A(T_n) = \left[\sigma_A^2 + 2(n-2)\,\tau_A^2\right] / \binom{n}{2}$$

$$Var_A(Q_n) = \frac{4}{n} \cdot Var_A(t^*(X_1)) = \frac{4}{n}\,\tau_A^2$$

$$Cov_A(T_n, Q_n) = \frac{2}{n} \cdot \sum_{i<} \sum_{j} \sum_{l} Cov_A(S_{ij},\, t^*(X_l)) \,/\, \binom{n}{2}$$

$$= \frac{8}{n^2(n-1)} \sum_{i<} \sum_{j} Cov_A(S_{ij},\, t^*(X_i))$$

$$= \frac{4}{n} \cdot Cov_A(S_{12},\, t^*(X_1)) = \frac{4}{n}\, Var_A(t^*(X_1)) = \frac{4}{n}\,\tau_A^2 \quad .$$

Since $\sigma_{A_n}^2 \to \sigma_H^2$, $\tau_{A_n}^2 \to \tau_H^2$ for $n \to \infty$, we conclude

$$d_n = \frac{2}{n-1}\left[\sigma_{A_n}^2 - 2\tau_{A_n}^2\right] \to 0.$$

(c) The asymptotic normality of $\sqrt{n}\; Q_n$ follows from Liapounov's central limit theorem for the bounded case (see Loève 1963, p. 277). It is applied to the centered i.i.d. variables $Y_{nl} := t^*(X_l) - \lambda_{A_n}$ $(l = 1, \ldots, n)$. Since the function $\bar{q}(\cdot)$, (4.6), is bounded by assumption, a look at (4.8), (4.9), (4.10) and (4.17) shows that the variables Y_{nl} are bounded, too. Moreover, the variance $Var\,(\Sigma_l\, Y_{nl}) = n \cdot \tau_{A_n}^2$ approaches ∞ since $\tau_{A_n}^2 \to \tau_H^2 > 0$ for $n \to \infty$. Thus the assumptions of Liapounov's theorem are fulfilled and $\Sigma_l\, Y_{nl} / (\sqrt{n}\; \tau_{A_n})$ is asymptotically $N(0,1)$. This implies that $\sqrt{n}\; Q_n / (2\tau_H) = \Sigma_l\, Y_{nl} / (\sqrt{n}\; \tau_H)$ is $N(0,1)$, too, for $n \to \infty$. ●

5. Tests Using the Least Squared Error Criterion (Maximum F Test)

Since the tests given above do not use any clustering at all, they can be applied, at least in principle, without or before starting a clustering algorithm. However, in practice, the main interest is often more in finding a suitable classification, and only subsequently to test if this classification is more marked than for random, homogeneous data. Therefore we need methods for simultaneously testing and constructing a classification of data. It is to be expected that methods of this type will extract more information from the data concerning a prospective clustering structure and that the corresponding tests will exhibit a better power performance than a gap or a mean similarity test which is tailored more to the hypothesis of homogeneity than to the clustering alternative. But since their performance depends on both the clustering model and the clustering algorithm, it is evident that both must be adapted in a suitable sense from the outset.

We treat this problem in the framework of partition type classifications using the well-known within cluster sum of squares criterion: Given some fixed class number $k \geqslant 2$, we look for a partition $\mathbf{C} = (C_1, \ldots, C_k)$ of the given objects $1, \ldots, n$ (respectively of the sample $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$) with k nonoverlapping classes $C_1, \ldots, C_k$ such that the "variance criterion"

$$g_n(\mathbf{C}) := \frac{1}{n} \sum_{i=1}^{k} \sum_{j \in C_i} ||x_j - \bar{x}_{C_i}|| \to \min =: g_n^* \tag{5.1}$$

is minimized over all k-partitions $\mathbf{C}$. Here $\bar{x}_{C_i}$ denotes the mean of all x_j belonging to the class C_i of $\mathbf{C}$. An optimal partition $\mathbf{C}_n^* = (C_{n1}^*, \ldots, C_{nk}^*)$, i.e., with $g_n^* = g_n(\mathbf{C}_n^*)$, is usually approximated by a k-means, minimum-distance, or exchange algorithm (see, e.g., Bock 1974, Hartigan 1975, Späth 1982,1983).

Intuitively, we expect that the minimum criterion value g_n^* tends to be small if the random vectors $X_1, \ldots, X_n$ have a multimodal density (e.g., a translation mixture density (2.2)) with k well separated modes (see Figure 1.d) and to be large if a homogeneity hypothesis H_0 holds with some unimodal density h. Therefore, for testing H_0 and simultaneously for assessing the relevance of an optimal classification $\mathbf{C}_n^*$, the following *minimum-variance test* seems to be reasonable:

(A) Reject H_0 and accept $\mathbf{C}_n^*$ iff $g_n^* := \min_{\mathbf{C}} g_n(\mathbf{C}) < \gamma$

where $\gamma = \gamma_n(\alpha)$ is some suitably chosen threshold.

However, this test lacks an important invariance property which seems to be indispensable for practical purposes: Whilst any scale transformation

of the type $\tilde{X}_j := \beta X_j$ $(1 \leqslant j \leqslant n$; with some arbitrary constant $\beta \neq 0)$ preserves unimodality as well as multimodality and thus leaves the test problem invariant, the test statistic g_n^* changes its value into $\beta^2 g_n^*$. Since often, in practice, only the type of the distribution of X_j can be specified (involving an unknown scale factor or standard deviation), a scale invariant test will be advisable.

A corresponding test statistic can be obtained from the remark that for any partition $\mathbf{C}$ the total sum of squares s_n^2 can be decomposed into two parts:

$$s_n^2 := \frac{1}{n} \sum_{j=1}^{n} ||x_j - \bar{x}||^2 = g_n(\mathbf{C}) + b_n(\mathbf{C}) \tag{5.2}$$

where

$$b_n(\mathbf{C}) := \frac{1}{n} \sum_{i=1}^{k} |C_i| \cdot ||\bar{x}_{C_i} - \bar{x}||^2 \tag{5.3}$$

is the SSQ between the classes of $\mathbf{C}$ and $\bar{x}$ the mean of all $x_1, \ldots, x_n$. Therefore the minimization problem (5.1) is equivalent to the maximization problem:

$$k_n(\mathbf{C}) := \frac{b_n(\mathbf{C})}{g_n(\mathbf{C})} = \frac{\sum_{i=1}^{k} |C_i| \cdot ||\bar{x}_{C_i} - \bar{x}||^2}{\sum_{i=1}^{k} \sum_{j \in C_i} ||x_j - \bar{x}_{C_i}||^2} \to \max =: k_n^* \tag{5.4}$$

where the maximum value k_n^* may be expressed by

$$k_n^* = k_n(\mathbf{C}_n^*) = \frac{b_n(\mathbf{C}_n^*)}{g_n(\mathbf{C}_n^*)} = \frac{s_n^2}{g_n^*} - 1 \ . \tag{5.5}$$

Thus the test (A) is equivalent to the *maximum F test.*

(B) Reject H_0 and accept $\mathbf{C}_n^*$ iff $k_n^* := \max_{\mathbf{C}} k_n(\mathbf{C}) > \kappa$

with some critical level $\kappa = \kappa_n(\alpha)$ and a scale invariant test statistic k_n^*. The name derives from the fact that $((n-k)/(k-1)) \cdot k_n(\mathbf{C})$ is just the well-known F-ratio statistic when testing the hypothesis $\mu_1 = \ldots = \mu_k$ in the multivaraiate variance analysis model $H_C : X_j \sim N_p(\mu_i, \sigma^2 I_p)$ for all $j \in C_i$ and $i = 1, \ldots, k$, assuming $\mathbf{C}$ to be *known.* Thus k_n^* is, essentially, the maximum F-ratio which can be obtained by searching for the "most significant" partition $\mathbf{C}$.

Under normal distribution assumptions, both tests (A) and (B) were shown to be optimal in some Bayesian sense (Bock 1972, 1974 ch. 13).

6. The Asymptotic Distribution of the Test Statistics g_n^* and k_n^*

In the past, a main obstacle for generally applying the minimum variance and maximum F tests (A), (B) in Section 5 has been the non-availability of the sampling distributions of the test statistics g_n^* and k_n^*, hence of the thresholds $\gamma_n(\alpha)$ and $\kappa_n(\alpha)$. Only for the one-dimensional case have some simulation results been published by Engelman and Hartigan (1969), and Hartigan (1978) obtained the asymptotic normal distribution for g_n^* and k_n^* in this case. For two dimensions some single simulation values may be found in Lee (1979).

In this section we shall derive the asymptotic distribution of the test statistics g_n^* and k_n^* in the general p-dimensional case. Throughout we assume that $X, X_1, X_2, \ldots$ are independent random vectors of $\mathbb{R}^p$, each with the same distribution P such that $E[\,||X||^2\,]<\infty$ and the support of P has cardinality at least k, the given class number. For example, P may be induced by some distribution density $f(x)$ which describes, as the case may be, a hypothesis or an alternative from Section 2.

Our derivation is based on the well-known result that the minimization problem (5.1) for $\mathbf{C}$ is equivalent to the following *best-location problem.*

Look for k locations $z_1, \ldots, z_k \in \mathbb{R}^p$ such that the mean minimum squared deviation

$$W(z, P_n) := \frac{1}{n}\sum_{j=1}^{n} \min_{1 \leqslant i \leqslant k}\left\{||x_j - z_i||^2\right\} \tag{6.1}$$

$$= \frac{1}{n}\sum_{j=1}^{n} \Delta(x_j, z) \xrightarrow[z]{} \inf =: W_n^*$$

is minimized over all systems $z = (z_1, \ldots, z_k)$ with k locations of $\mathbb{R}^p$. Here P_n is the empirical distribution of the sample $x_1, \ldots, x_n$ and

$$\Delta(x,z) := \min_{1 \leqslant j \leqslant k}\left\{||x - z_i||^2\right\}$$

the distance between x and the nearest location of z.

To be more specific, let us denote, for a given partition $\mathbf{C} = (C_1, \ldots, C_k)$ of $\{x_1, \ldots, x_n\}$, by $z(\mathbf{C}) := (\bar{x}_{C_1}, \ldots, \bar{x}_{C_k})$ the set of corresponding class means. Inversely, for a given location system $z = (z_1, \ldots, z_k)$, denote by $\mathbf{B}(z) := (B_1(z), \ldots, B_k(z))$ the corresponding minimum distance partition of $\mathbb{R}^p$ with k classes

$$B_i(z) := \{x | x \in \mathbb{R}^p,\ \|x-z_i\| = \min_{1 \leq j \leq k} \|x-z_j\|\} \tag{6.2}$$

($i = 1, \ldots, k$; ties may be resolved arbitrarily), and by $\mathbf{C}(z) := (C_1(z), \ldots, C_k(z))$ the corresponding minimum distance partition *of the sample* $\{x_1, \ldots, x_n\}$, i.e., with classes $C_i(z) := B_i(z) \cap \{x_1, \ldots, x_n\}$. Then the equivalence of both problems (5.1) and (6.1) is made precise by:

Theorem 6.1 (Bock 1974):

a. *If* $\mathbf{C}_n^* = (C_{n1}^*, \ldots, C_{nk}^*)$ *is an optimum partition of* $\{x_1, \ldots, x_n\}$ *for (5.1) then its center system* $z(\mathbf{C}_n^*)$ *is optimum for (6.1).*

b. *If* $z_n^* = (z_{n1}^*, \ldots, z_{nk}^*)$ *is an optimum location system for (6.1), then its minimum distance partition* $\mathbf{C}(z_n^*)$ *of* $\{x_1, \ldots, x_n\}$ *is optimum for (5.1).*

c. *We have, necessarily, the stationarity conditions* $z_n^* = z(\mathbf{C}(z_n^*))$ *and* $\mathbf{C}_n^* = \mathbf{C}(z(\mathbf{C}_n^*))$, *(e.g.,* $z_{ni}^* = \bar{x}_{C_{ni}^*}$ *for* $i = 1, \ldots, k$ *if* $\mathbf{C}_n^*$ *is unique).*

d. *The minimum criterion values in (5.1) and (6.1) coincide:*

$$g_n^* := g_n^*(\mathbf{C}_n^*) = W(z_n^*, P_n) =: W_n^* \ . \tag{6.3}$$

This last relation shows that the asymptotic distribution of g_n^* is obtained by considering the asymptotics of the solution z_n^* of (6.1) and its minimum criterion value $W_n^* = W(z_n^*, P_n)$.

Since for $n \to \infty$ the empirical distribution P_n converges to P (in the weak sense) we expect that the asymptotic behavior of the optimal location system $z_n^* = (z_{n1}^*, \ldots, z_{nk}^*)$ is related to the soluton $\zeta^* = \zeta^*(P) = (\zeta_1^*, \ldots, \zeta_k^*)$ of the following continuous version of (6.1):

$$W(z,P) := \int_{\mathbb{R}^p} \min_{1 \leq i \leq k} \left\{ \|x-z_i\|^2 \right\} dP(x) = E\left[\Delta\ (X,z)\right]$$

$$= \sum_{i=1}^{k} \int_{B_i(z)} \|x-z_i\|^2\ dP(x) \to \inf_z =: W^* \tag{6.4}$$

and, moreover, to the partition $\mathbf{B}^* = (B_1^*, \ldots, B_k^*)$ of $\mathbb{R}^p$ which solves the continuous analogue of (5.1):

$$g(\mathbf{B}) := \sum_{i=1}^{k} \int_{B_i} ||x-\mu_{B_i}||^2 \, dP(x) \to \inf_{\mathbf{B}} = : g^* \quad . \tag{6.5}$$

Here minimization is over all partitions $\mathbf{B} = (B_1, \ldots, B_k)$ of $\mathbb{R}^p$ with k Borel sets $B_1, \ldots, B_k \subset \mathbb{R}^p$, and $\mu_{B_i} := E[X|X \in B_i]$ denotes the conditional mean of X in B_i $(1 \leqslant i \leqslant k)$. Evidently, the continuous version of theorem 6.1 holds as well (Bock 1974).

Theorem 6.2:

a. *If $\mathbf{B}^* = (B_1^*, \ldots, B_k^*)$ is an optimal partition of $\mathbb{R}^p$ for (6.5), then its center system $\zeta(\mathbf{B}^*) := (\mu_{B_1^*}, \ldots, \mu_{B_k^*})$ is optimal for (6.4).*

b. *If the location system $\zeta^* = (\zeta_1^*, \ldots, \zeta_k^*)$ is optimal for (6.4), then each minimum-distance partition $\mathbf{B}(\zeta^*)$ of $\mathbb{R}^p$ is optimal for (6.5).*

c. *Necessarily, the stationarity conditions $\mathbf{B}^* = \mathbf{B}(\zeta(\mathbf{B}^*))$ resp. $\zeta^* = \zeta(\mathbf{B}(\zeta^*))$ hold. In most cases (e.g., if $\mathbf{B}^*$ is essentially unique) this means that $\mathbf{B}^* = \mathbf{B}(\zeta^*)$ and*

$$\zeta_i^* = E[X|X \in B_i^*] \quad i = 1, \ldots, k. \tag{6.6}$$

d. *The minimum critrion values are equal for both problems (6.4) and (6.5):*

$$g^* := g(\mathbf{B}^*) = W(\zeta^*, P) = W^* \quad . \tag{6.7}$$

Remark 5.1: It is evident from their definition that all criteria $g_n(\mathbf{C})$, $W(z,P)$ and $G(\mathbf{B})$ remain invariant under a relabeling of the classes $C_1, \ldots, C_k$ of $\mathbf{C}$, the locations $z_1, \ldots, z_k$ of z, or the regions $B_1, \ldots, B_k$ of $\mathbf{B}$. Therefore, the solutions $\mathbf{C}_n^*$, z_n^*, ζ_n^* and $\mathbf{B}^*$ of the corresponding minimum problems (5.1), (6.1), (6.4) and (6.5) cannot be unique in the strong sense, but only up to $k!$ permutations. This will be tacitly understood when referring to a "unique" solution, as well as the fact that the special assignment of class boundaries is irrelevant here.

Note that the existence of a solution $\mathbf{B}^*$ is guaranteed for (6.4) since $E[||X||^2] < \infty$, but that there may be different solutions. For example, if $p \geqslant 2$ and X has a spherically symmetric distribution P, then any rotation of an optimal center system $\zeta^* = (\zeta_1^*, \ldots, \zeta_k^*)$ is another solution of (6.4). For some special cases involving normal distributions see Gray and Karnin (1982) and Baubkus (1985). On the other hand, in the one-dimensional case, uniqueness holds, e.g., if $\log f(x)$ is concave (Fleischer 1964, Trushkin 1982, Kieffer 1983). The corresponding solution for the normal distribution may be found in Ogawa (1951, 1962), Cox (1957), or Bock (1974, § 15.f).

Now we may formulate the following consistency property which has been proved, with different methods and under different assumptions, by Degens (1978), Bryant and Williamson (1978) and Pollard (1981, 1982b):

Theorem 6.3: *Let $L(P) \subset \mathbb{R}^{kp}$ be the set of all solutions ζ^* of (6.4) and $z_n^* = (z_{n1}^*, \ldots, z_{nk}^*)$ any solution for (6.1), e.g., the empirical center system $z_n^* := z(\mathbf{C}_n^*)$. Then, for $n \to \infty$, we have almost surely $z_n^* \to L(P)$ in the sense that the minimum distance $\inf\{|||z_n - \zeta||| \mid \zeta \in L(P)\}$ converges to 0 (where $|||\cdot|||$ denotes the Euclidean norm in $\mathbb{R}^{kp}$). In particular, if (6.4) has a unique solution $\zeta^* = (\zeta_1^*, \ldots, \zeta_k^*)$ then, under an appropriate relabeling of the classes of $\mathbf{C}_n^*$, we have convergence of all centers: $z_{ni}^* = \bar{x}_{C_{ni}^*} \to \zeta_i^* = E[X \mid X \in B_i^*]$ a.s. for $i = 1, \ldots, k$.*

The asymptotic distribution of the center system $z_n^* = z(\mathbf{C}_n^*)$ has been obtained by Pollard (1982a) in the case where the solution ζ^* of (6.4) is unique and the random vectors X_j are distributed with a density $f(x)$. For later use, the following theorem 6.4 combines several final and intermediate results of his paper. We introduce the following notation:

(i) The random vector $Y_n := (Y'_{n1}, \ldots, Y'_{nk})'$ of $\mathbb{R}^{kp}$ whose ith subvector

$$Y_{ni} := 2n^{-1/2} \cdot \sum_{j=1}^{n} (X_j - \zeta_i^*) \cdot \mathbf{1}_{B_i^*}(X_j)$$

$(i = 1, \ldots, k)$ is, essentially, the centered mean of all X_j lying in the polyhedron B_i^* (remember that $\mathbf{1}_B(x)$ denotes the characteristic function of a set B). By the central limit theorem, Y_n has an asymptotic normal distribution:

$$Y_n \xrightarrow{L} N(0, V) \quad \text{with} \quad V := \mathit{diag}\,(V_1, \ldots, V_k) \tag{6.8}$$

whose covariance matrix has a block diagonal form with

$$\begin{aligned} V_i &:= p_i \cdot Cov(X \mid X \in B_i^*) && i = 1, \ldots, k \\ p_i &:= P(X \in B_i^*) && i = 1, \ldots, k \ . \end{aligned} \tag{6.9}$$

(ii) The $kp \times kp$ matrix $\Gamma = (\Gamma_{il})$ of all partial second derivatives of the deviation function $W(z, P)$ with respect to z at $z = \zeta^*$. Γ is made up of $p \times p$ blocks Γ_{il} given by

$$\Gamma_{il} := \frac{2}{||\zeta_i^* - \zeta_l^*||} \int_{F_{il}} (x-\zeta_i^*)\,(x-\zeta_l^*)'\, f(x)\, d\sigma(x) \qquad i \neq l \tag{6.10}$$

$$\Gamma_{ii} := 2p_i\, I_p - 2\sum_{l \neq i} \frac{1}{||\zeta_i^* - \zeta_l^*||} \int_{F_{il}} (x-\zeta_i^*)\,(x-\zeta_i^*)'\, f(x)\, d\sigma(x)$$

$(i, l = 1, \ldots, k)$. The integrals extend over the common boundary $F_{il} = \partial B_i^* \cap \partial B_l^*$ of the polytopes B_i^*, B_l^* (possibly $F_{il} = \varnothing$) and $d\sigma(x)$ is the $(p-1)$-dimensional Lebesque measure on F_{il}. The surface integrals occur because in (6.4) the domain of integration is dependent on z.

Theorem 6.4 (Pollard 1982a): *Assume that the density $f(x)$ of the random vectors $X_1, X_2, \ldots$ is regular in some sense: $f(x)$ is continuous and dominated by some function $\rho(||x||)$ with $\int r^p \rho(r)\, dr < \infty$, $E||X||^2| < \infty$ and Γ is positive definite. Moreover, suppose that the solution $\zeta^* = (\zeta_1^*, \ldots, \zeta_k^*)$ of (6.4) is unique, and that the optimal empirical class centers z_{ni}^* are labeled such that $z_{ni}^* \to \zeta_i^*$ a.s. for all i (see theorem 6.3). Then, for $n \to \infty$:*

a.
$$\sqrt{n}\,(z_n^* - \zeta^*) = \Gamma^{-1}\, Y_n + o_p(1) \tag{6.11}$$

has an asymptotic normal distribution $N(O, \Gamma^{-1} V \Gamma^{-1})$.

b.
$$g_n^* = W_n^* = W(z_n^*, P_n) \tag{6.12}$$
$$= W(\zeta^*, P_n) - (1/2n) \cdot Y'_n\, \Gamma^{-1}\, Y_n + o_p(1/n)$$

where the approximating term

$$\tilde{W}_n := W(\zeta^*, P_n) = \frac{1}{n} \sum_{j=1}^{n} \Delta(x_j, \zeta^*) \xrightarrow{L} N(W^*, \tau^2/n) \tag{6.13}$$

is, by the central limit theorem, asymptotically normally distributed with expection

$$W^* = W(\zeta^*, P) = g(\mathbf{B}^*) \quad \text{and} \tag{6.14}$$

$$\tau^2 := Var(\Delta(X, \zeta^*)) = \sum_{i=1}^{k} p_i\, E\left[||X - \zeta_i^*||^4 \mid X \in B_i^*\right] - g^{*2}\ . \tag{6.15}$$

Remark 5.2: The approximations (6.11) and (6.12) may be found on page 924 of Pollard (1982a): While (6.12) is identical with the last line there, the relation (6.11) results from inserting the second-last line into the seventh-last line.

In the sequel, we shall assume without further reference that the results of theorem 6.4 are valid. An immediate consequence is:

Corollary 6.5: *The minimum within cluster sum of squares* $g_n^* = W_n^*$ *has an asymptotic normal distribution given by*

$$\sqrt{n}\ (g_n^* - g^*) = \sqrt{n}\ (W_n^* - W^*) \xrightarrow{L} N(0,\tau^2). \tag{6.16}$$

Proof: From (6.12) we obtain

$$\sqrt{n}\ (W_n^* - W^*) = \sqrt{n}\ (\tilde{W}_n - W^*) - 1/\ (2\sqrt{n}) \quad Y'_n\, \Gamma^{-1}\, Y_n + o_p(1/\sqrt{n})$$

where the first term is asymptotically $N\ (0, \tau^2)$ according to (6.13). Moreover, since Y_n has the limiting distribution (6.8), the second term is $o_p(1)$. Thus (6.16) follows. ●

The asymptotic distribution of the maximum F-ratio k_n^* needs some further notation. First let us introduce some unconditional and conditional moments of X:

$$\begin{array}{ll} \mu := E[X] & \zeta_i^* = E[X|X\in B_i^*] \\ \sigma_t := E[\,||X-\mu||^t\,] & \sigma_{ti} := E[\,||X-\zeta_i^*||^t \,|\, X\in B_i^*] \text{ for } t=2,4 \\ \mu_3 := E[(X-\mu)\cdot||X-\mu||^2\,] & \mu_{3i} := E[(X-\zeta_i^*)\cdot||X-\zeta_i^*||^2 \,|\, X\in B_i^*]\,, \end{array}$$

for $i = 1, \ldots, k$. Then

$$b^* := \sum p_i \cdot ||\zeta_i^* - \mu||^2 \tag{6.17}$$

is the population SSQ between and $W^* = g^* = G(\zeta^*) = \Sigma\ p_i\ \sigma_{2i}$ the population SSQ within the classes of $\mathbf{B}^*$. In analogy to (5.2), they sum to $\sigma_2 = g^* + b^*$ such that the population SSQ ratio is given by

$$k^* := b^*/\ g^* = \sigma_2/\ g^* - 1\ . \tag{6.18}$$

Finally, let $\gamma := \sigma_2/\ g^* = k^* + 1$ and

$$\begin{aligned} \kappa^2 &:= \sigma_4 + 2\gamma \sum p_i[(\gamma/2)\sigma_{4i} - 2\ (\zeta_i^* - \mu)'\,\mu_{3i} - \sigma_{2i}\cdot||\zeta_i^* - \mu||^2] \\ &= \sum p_i[||\zeta_i^* - \mu||^2 - k^*\sigma_{2i}]^2 + 4\sum p_i(\zeta_i^* - \mu)'\ V_i(\zeta_i^* - \mu) \\ &\quad - 4k^*\cdot\sum p_i(\zeta_i^* - \mu)'\,\mu_{3i} + k^{*2}\sum p_i\{\sigma_{4i} - \sigma_{2i}^2\}\ . \end{aligned} \tag{6.19}$$

Now we formulate the multivariate generalization of theorem 2 in Hartigan (1978) and, by the way, prove the conjecture in chapter 7 of Hartigan (1977):

Theorem 6.6 *The maximum F-ratio* $k_n^* = s_n^*/\, g_n^* - 1$ *has an asymptotic normal distribution given by*

$$\sqrt{n}\ (k_n^* - k^*) \xrightarrow{L} N(0, \kappa^2/\, g^{*2}) \quad . \tag{6.20}$$

Proof. By definition, we can write

$$\sqrt{n}\ (k_n^* - k^*) = \sqrt{n}\ (s_n^2/\, g_n^* - \sigma_2/\, g^*) = \sqrt{n}\ (s_n^2 - \gamma g_n^*)/\, g_n^* \quad .$$

Since, by theorem 6.5, the denominator g_n^* converges to g^* in probability, the assertion (6.20) will be proved if we show that for $n \to \infty$:

$$h_n := \sqrt{n}\ (s_n^2 - \gamma\ g_n^*) \xrightarrow{L} N(0, \kappa^2) \quad . \tag{6.21}$$

Now, write $s_n^2 = \sum ||X_j - \mu||^2/\, n - ||\bar{X}_n - \mu||^2$ with the mean vector $\bar{X}_n := \sum X_j\, /\, n$, and use Pollard's approximation (6.12) for g_n^*; then you obtain

$$h_n = \sqrt{n} \left(\frac{1}{n} \sum_{j=1}^{n} ||X_j - \mu||^2 - ||\bar{X}_n - \mu||^2 - \gamma\ \tilde{W}_n \right) +$$

$$\left(\gamma/\, 2n^{1/2} \right) Y'_n\, \Gamma^{-1}\, Y_n + o_p\left(n^{-1/2} \right) \quad .$$

Since the last two terms are $o_p(1)$ because of (6.8), we have to prove that the first term, H_n say, is asymptotically $N(0, \kappa^2)$. Bearing in mind the definition (6.13) of $\tilde{W}_n$, we see that H_n may be expressed by the $(p+2)$ dimensional mean vector

$$\bar{U}_n := n^{-1} \sum_{j=1}^{n} U_j \quad \text{with summands} \quad U_j := \begin{pmatrix} ||X_j - \mu||^2 \\ X_j - \mu \\ \Delta\ (X_j, \zeta^*) \end{pmatrix} \tag{6.22}$$

$(j = 1, \ldots, n)$. To be specific, we have

$$H_n = \sqrt{n}\ (h(\bar{U}_n) - h(u^*)) \tag{6.23}$$

with a function h: $\mathbb{R}^{p+2} \to \mathbb{R}$ defined by $h(u) := u_1 - ||u_2||^2 - \gamma u_3$ (here he vector $u = (u_1, u'_2, u_3)' \in \mathbb{R}^{p+2}$ has been suitably split into three parts) and $u^* := E[U_1] = (\sigma_2, 0', g^*)'$. Note that $h(u^*) = \sigma_2 - \gamma g^* = 0$.

We pause to state a simple transformation lemma (see e.g., Witting and Nölle 1970):

Lemma 6.7 *Suppose that some k dimensional random vectors $\bar{U}_n$ converge, for $n \to \infty$, to a point $u^* \in \mathbb{R}^k$ and satisfy $\sqrt{n}\ (\bar{U}_n - u^*) \xrightarrow{L} N(0,M)$ with some covariance matrix M. Consider a differentiable real-valued function h: $\mathbb{R}^k \to \mathbb{R}$ whose gradient vector $\lambda(u) := (\partial h/\partial u_1, \ldots, \partial h/\partial u_k)'$ is continuous at $u = u^*$. Then, asymptotically for $n \to \infty$, we have*

$$H_n := \sqrt{n}\ (h(\bar{U}_n) - h(u^*)) \xrightarrow{L} N(0,\kappa^2)$$

with the variance $\kappa^2 := \lambda(u^*)' M \lambda(u^*)$.

By the law of large numbers and the central limit theorem, our mean vectors $\bar{U}_n$ in (6.22) satisfy the stated conditions, with $u^* = E[U_1]$ and $M = cov(U_1) := E[(U_1 - u^*)(U_1 - u^*)']$. Moreover, the special function $h(\cdot)$ given above has a continuous gradient vector $\lambda(u) = (1, -2u'_2, -\gamma)'$ with $\lambda^* := \lambda(u^*) = (1, 0', -\gamma)'$. Thus, by the lemma, the variable H_n (6.23) is asymptotically $N(0,\kappa^2)$ with $\kappa^2 := \lambda^{*\prime} M \lambda^*$.

Since U_1 is a partitioned vector, its covariance matrix $M = (M_{st})$ is partitioned, too, with blocks M_{st} $(1 \leqslant s,t \leqslant 3)$. Since $\lambda^* = (-1, 0', -\gamma)'$, we obtain

$$\kappa^2 = \lambda^{*\prime} M \lambda^* = M_{11} - 2\gamma\ M_{13} + \gamma^2 M_{33}\ . \tag{6.24}$$

After some straightforward calculation we obtain:

$$M_{11} := Var(||X-\mu||^2) = \sigma_4 - \sigma_2^2$$

$$= \sum_{i=1}^{k} p_i [\sigma_{4i} + 4(\zeta_i^* - \mu)'\ V_i(\zeta_i^* - \mu) + 4(\zeta_i^* - \mu)'\mu_{3i}$$

$$+ 2\sigma_{2i}||\zeta_i^* - \mu||^2 + ||\zeta_i^* - \mu||^4] - \sigma_2^2$$

$$M_{13} := Kov(||X-\mu||^2, \Delta(X, \zeta_{sp^*}))$$

$$= \sum_{i=1}^{k} p_i \left[\sigma_{4i} + 2(\zeta_i^* - \mu)'\mu_{3i} + \sigma_{2i}\ ||\zeta_i^* - \mu||^2\right] - \sigma_2\ g^*$$

$$M_{33} := Var\ (\Delta\ (X,\zeta^*)) = \sum_{i=1}^{k} p_i\ \sigma_{4i} - g^{*2} = \tau^2\ .$$

Inserting these expressions in (6.24) gives just the formula (6.19) for κ^2. ●

In practice we will use the maximum F test for deciding if a k-clustering C_n^* found by the k-means method, is more marked than a clustering obtained for some random, homogeneous data. Before applying this test, we have to specify a density $f\ (x)$ describing the hypothesis H_0 of "homogeneity," to determine the optimum partition $\mathbf{B}^* = (B_1^*, \ldots, B_k^*)$ of $\mathbb{R}^p$ with its k centers ζ_i^*, and to calculate the asymptotic mean k^* and the variance κ^2 / g^{*2} in (6.20). In the case of a multivariate normal distribution, $\mathbf{B}^*$ has been found from simulations and geometrical considerations for small values of k and p (Baubkus 1985). Since the test criteria g_n^*, k_n^* are defined as minimum values, their exact distribution is supposed to be skewed and the approximation (6.20) may be satisfactory only for vary large n. This question has to be investigated more thoroughly as well as the asymptotic behavior in the case of multiple solutions.

7. Conclusions

Four types of tests for "homogeneity" versus "clustering" have been investigated in the foregoing sections. It is obvious that no one of these tests is the "ideal" one or can cope with all situations of clustering. The main difficulty for practical applications will be the need to specify, in some sense or the other, the type of clustering or homogeneity to be detected. On the one hand, the size and shape of the domain G or the type $h(\cdot)$ of a (unimodal) density are to be chosen beforehand while, on the other hand, "valleys" and "modes" respectively mixtures are supposed to be indicators of the alternative. In particular, the maximum F test will be tailored mainly to normal density situations with spherical clusters of the same variance.

These problems reflect the multidimensional aspects met in cluster analysis and are typical for situations where no simple, clear-cut structure is given at the beginning (in contrast to most one-dimensional testing problems). This necessitates some care when applying any test for clustering, bearing in mind that different types of clusters may be present simultaneously in the data and that the number of clusters is, in some sense, dependent on the intended level of information compression or coarsening (e.g., in a hierarchy of clusters). Thus, a global application of a cluster test to a large or high-dimensional data set will not be advisable in most cases. However, a "local" application (e.g., of the maximum F test with $k = 2$ classes) to a specified part of the data will often be useful for providing evidence for or against a prospective clustering tendency.

References

BARNETT, V., KAY, R., and SNEATH, P.H.A. (1979), "A Familiar Statistic in an Unfamiliar Guise - A Problem in Clustering," *The Statistician, 28*, 185-191.

BAUBKUS, W. (1985), "Minimizing the Variance Criterion in Cluster Analysis: Optimal Configurations in the Multidimensional Normal Case," Diplomarbeit, Institute of Statistics, Technical University Aachen, 117 p.

BICKEL, P.J., and BREIMAN, L. (1983), "Sums of Functions of Nearest Neighbor Distances, Moment Bounds, Limit Theorems and a Goodness of Fit Test," *Annals of Probability, 11*, 185-214.

BINDER, D.A. (1978), "Bayesian Cluster Analysis," *Biometrika, 65*, 31-38.

BOCK, H.H. (1972), "Statistische Modelle und Bayes'sche Verfahren zur Bestimmung einer unbekannten Klassifikation normalverteilter zufälliger Vektoren," *Metrika, 18*, 120-132.

BOCK, H.H. (1974), *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Clusteranalyse)*, Göttingen: Vandenhoeck & Ruprecht, 480 p.

BOCK, H.H. (1977),"On Tests Concerning the Existence of a Classification," in *Proceedings First International Symposium on Data Analysis and Informatics*, Le Chesnay, France, Institut de Recherche en Informatique et en Automatique (IRIA), 449-464.

BOCK, H.H. (1981), "Statistical Testing and Evaluation Methods in Cluster Analysis," in *Proceedings on the Golden Jubilee Conference in Statistics: Applications and New Directions*, December 1981, Calcutta, Indian Statistical Institute, 1984, 116-146.

BOCK, H.H. (1983), "Statistische Testverfahren im Rahmen der Clusteranalyse," *Proceedings of the 7th Annual Meeting of the Gesellschaft für Klassifikation e.V.*, in *Studien zur Klassifikation*, Vol. 13, ed. M. Schader, Frankfurt: Indeks-Verlag, 161-176.

BRYANT, P., and WILLIAMSON, J.A. (1978), "Asymptotic Behavior of Classification Maximum Likelihood Estimates," *Biometrika, 65*, 273-281.

COX, D.R. (1957), "Note on Grouping," *Journal of the American Statistical Association, 52*, 543-547.

DAVID, H.A. (1981), *Order Statistics*, New York: Wiley, chap. 9.3, 9.4.

DEGENS, P.O. (1978), "Clusteranalyse auf topologisch-masstheoretischer Grundlage," Dissertation, Fachbereich Mathematik, Universitaet Muenchen.

DEL PINO, G.E. (1979), "On the Asymptotic Distribution of k-spacings with Applications to Goodness-of-Fit Tests," *Annals of Statistics, 7*, 1058-1065.

DUBES, R., and JAIN, A.K. (1979), "Validity Studies in Clustering Methodologies," *Pattern Recognition, 11*, 235-254.

EBERL, W., and HAFNER, R. (1971), "Die asymptotische Verteilung von Koinzidenzen," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 18*, 322-332.

ENGELMAN, L., and HARTIGAN, J.A. (1969), "Percentage Points of a Test for Clusters," *Journal of the American Statistical Association, 64*, 1647-1648.

FLEISCHER, P.E. (1964), "Sufficient Conditions for Achieving Minimum Distortion in a Quantizer," *IEEE Int. Conv. Rec.*, part 1, 104-111.

GHOSH, J.K., and SEN, P.K. (1984), "On the Asymptotic Distribution of the Log Likelihood Ratio Statistic for the Mixture Model and Related Results," Preprint, Calcutta: Indian Statistical Institute.

GIACOMELLI, F., WIENER, J., KRUSKAL, J.B., v. POMERANZ, J., and LOUD, A.V. (1971), "Subpopulations of Blood Lymphocytes Demonstrated by Quantitative Cytochemistry," *Journal of Histochemistry and Cytochemistry, 19*, 426-433.

GRAY, R.M., and KARNIN, E.D. (1982), "Multiple Local Optima in Vector Quantizers," *IEEE Trans. Information Theory, IT-28*, 256-261.

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

HARTIGAN, J.A. (1977), "Distribution Problems in Clustering," in *Classification and Clustering*, ed. J. van Ryzin, New York: Academic Press, 45-72.

HARTIGAN, J.A. (1978), "Asymptotic Distributions for Clustering Criteria," *Annals of Statistics, 6*, 117-131.

HENZE, N. (1981), "An Asymptotic Result on the Maximum Nearest Neighbor Distance Between Independent Random Vectors with an Application for Testing Goodness-of-Fit in $\mathbb{R}^p$ on Spheres," Dissertation, University of Hannover, published in *Metrika, 30*, 245-260.

HENZE, N. (1982), "The Limit Distribution for Maxima of Weighted r-th Nearest Neighbor Distances," *Journal of Applied Probability, 19*, 334-354.

KIEFFER, J.C. (1983), "Uniqueness of Locally Optimal Quantizer for Log-concave Density and Convex Error Weighting Function," *IEEE Trans. Infromation, IT-29*, 42-27.

KUO, M., and RAO, J.S. (1981), "Limit Theory and Efficiences for Tests Based on Higher Order Spacings," in *Proceedings on the Golden Jubilee Conference in Statistics: Applications and New Directions*, December 1981, Calcutta: Indian Statistical Instiute, 1984.

LEE, K.L. (1979), "Multivariate Tests for Clusters," *Journal of the American Statistical Association, 74*, 708-714.

LEHMANN, E.L. (1955), "Ordered Families of Distributions," *Annals of Mathematical Statistics, 26*, 399-419.

LOEVE, M. (1963), *Probability Theory*, Princeton, NJ: van Nostrand.

NEWELL, G.F. (1963), "Distribution for the Smallest Distance Between any Pair of the k-th Nearest Neighbor Random Points on a Line," in *Proc. Symp. Time Series Analysis*, ed. M. Rosenblatt, New York: Wiley, 89-103.

OGAWA, J. (1951), "Contributions to the Theory of Systematic Statistics I," *Osaka Mathematical Journal, 3*, 175-213.

OGAWA, J. (1962), "Determination of Optimum Spacings in the Case of Normal Distribution," in *Contributions to Order Statistics*, eds. A.E. Sarhan and B.G. Greenberg, New York: Wiley, p. 277 ff.

PERRUCHET, C. (1982), "Les Epreuves de Classifiabilité en Analyse des Données," Note technique NT/PAA/ATR/MTI/810, Issy-les-Moulineaux, France: Centre National d'Etudes de Télécommunications, September 1982.

PERRUCHET, C. (1983), "Significance Tests for Clusters: Overview and Comments," in *Numerical Taxonomy*, ed. J. Felsenstein, Berlin: Springer, 199-208.

POLLARD, D. (1981), "Strong Consistency of k-means Clustering," *Annals of Statistics, 9*, 135-140.

POLLARD, D. (1982a), "A Central Limit Theorem for k-means Clustering," *Annals of Probability, 10*, 919-926.

POLLARD, D. (1982b), "Quantization and the Method of k-means," *IEEE Trans. Information Theory, IT-28*, 119-205.

RANDLES, R.H., and WOLFE, D.A. (1979), *Introduction to the Theory of Non-parametric Statistics*, New York: Wiley.

SCHILLING, M.F. (1983a), "Goodness of Fit Testing in $\mathbb{R}^m$ Based on the Weighted Empirical Distribution of Certain Nearest Neighbor Statistics," *Annals of Statistics, 11*, 1-12.

SCHILLING, M.F. (1983b), "An Infinite-dimensional Approximation for Nearest Neighbor Goodness of Fit," *Annals of Statistics, 11*, 13-24.

SILVERMAN, B.W. (1976), "Limit Theorems for Dissociated Random Variables," *Advances in Applied Probability, 8*, 806-819.

SNEATH, P.H.A. (1977a), "A Method for Testing the Distinctness of Clusters: A Test of the Disjunction of Two Clusters in Euclidean Space as Measured by their Overlap," *Jour. Int. Assoc. Math. Geol., 9*, 123-143.

SNEATH, P.H.A. (1977b), "Cluster Significance Tests and Their Relation to Measures of Overlap," in *Proceedings First International Symposium on Data Analysis and Informatics*,

Versailles, September 1977, Institut de Recherche d'Informatique et d'Automatique (IRIA), Le Chesnay, France, 1, 15-36.

SNEATH, P.H.A. (1979a), "The Sampling Distribution of the W Statistic of Disjunction for the Arbitrary Division of a Random Rectangular Distribution," *Journal. Int. Assoc. Math. Geol., 11*, 423-429.

SNEATH, P.H.A. (1979b), "Basic Program for a Significance Test for 2 Clusters in Euclidean Space as Measured by Their Overlap," *Computers and Geosciences, 5*, 143-155.

SPAETH, H. (1982), *Cluster Analysis Algorithms*, Chichester: Horwood.

SPAETH, H. (1983), *Cluster-Formation und -Analyse*, München-Wien: Oldenbourg.

TRUSHKIN, A.V. (1982), "Sufficient Conditions for Uniqueness of a Locally Optimal Quantizer for a Class of Convex Error Weighting Functions," *IEEE Trans. Information Theory, IT-28*, 187-198.

WALLENSTEIN, S.R., and NAUS, J.I. (1973), "Probabilities for a k-th Nearest Neighbor Problem on the Line," *Ann. Probab., 1*, 188-190.

WALLENSTEIN, S.R., and NAUS, J.I. (1974), "Probabilities of the Size of Largest Clusters and Smallest Intervals," *Journal of the American Statistical Association, 69*, 690-697.

WEISS, L. (1960), "A Test of Fit Based on the Largest Sample Spacing," *SIAM Journal of the Society for Industrial and Applied Mathematics, 8*, 295-299.

WITTING, H., and NOELLE, G. (1979), *Angewandte Mathematische Statistik*, Stuttgart: B.G. Teubner, theorem 2.10.

WOLFE, J.H. (1970), "Pattern Clustering by Multivariate Mixture Analysis," *Multivariate Behavioral Research, 5*, 329-350.

WOLFE, J.H. ((1981), "A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixture of Multinormal Distribution," Technical Bulletin STB 72-2, San Diego: U.S. Naval Personal and Training Research Laboratory.

WOLFE, S.J. (1975), "On the Unimodality of Spherically Symmetric Stable Distribution Functions," *Journal of Multivariate Analysis, 5*, 236-242.