Statistics & Probability Letters 1 (1982) 46–50
North-Holland Publishing Company

July 1982

Tests for Simultaneously Determining Numbers of Clusters and Their Shape with Multivariate Data

Ashis Sen Gupta

Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.

Received May 1982; revised version received May 1982

Abstract. Given a set of data, very little is known about tests to determine number of clusters and/or elements of the clusters. Even in the simplest case of detecting between only one or two clusters with multivariate normal data, theoretically the number of tests needed seems to be infinite. Alternatively, suppose N independent estimates of generalized variances (GVs) are computed from a given set of p-dimensional vector observations. Assuming multivariate normality, tests based on GVs are proposed which objectively and uniquely determine, simultaneously, the number of clusters and their corresponding elements. Only a reasonably small number of tests are required for this stepwise procedure. The exact percentage points are either available from existing tables or can be computed from a result presented.

Keywords. Cluster analysis, generalized variance, likelihood ratio test, union intersection test.

1. Introduction

Given a set of points, clustering techniques based on different criteria, e.g., similarity measures, distance functions, correlations, etc. have received considerable attention. However, statistical tests for determining clusters for the given set of points are scarce and need detailed research. In the present paper, we deal with the problem of clustering in a more general set-up. Most often, the characteristics that need to be clustered have repeated data. For example, in clustering of schools by performance (Hartigan, 1975), scores on various students for each school are collected. In clustering of cities by demand of some specified articles, one usually collects data from different shops in each city. In such situations one has already some natural subgroupings, to be termed 'subclusters' of observations. For example, scores on various students for a specific school, sales of the articles in various shops for a specific city, etc., constitute the elements of a subcluster. Usually, the mean of each subcluster is used as a 'point' for forming clusters (e.g., see Hartigan, 1975; p. 118, Mezzich and Solomon, 1980, p. 109). In many cases, the means may be nearly equal and as such clustering on the basis of means will not be adequate. The variations within each subcluster need then be taken into consideration to provide precise differentiation among the clusters (see Mezzich and Solomon, 1980, p. 63]. For our analysis, we will start with the original set of data. We will emphasize the case of nearly equal means. Further, if multivariate normal distributions with equal means for the original observations are assumed, clustering on the basis of equal multivariate scatter is a natural choice. In such cases, given the multi-dimensional observation vectors in each subcluster, a measure of multi-dimensional scatter will be used as a criterion for determining the clusters. One can use the dispersion matrix Σ or its determinant, a scalar, the generalized variance $|\Sigma|$.

For a k-dimensional random vector, grouping by equal covariance matrices demands componentwise equality for all the $\frac{1}{2}k(k+1)$ distinct elements of the matrices. If k is not small, this choice is quite

restrictive and may not be desirable. However, in such situations, the generalized variance is more convenient to work with and is being proposed here as a criterion. Note first that it is reasonable to expect that any scalar function of Σ, used as a measure of multi-dimensional statistical scatter, should take into consideration the magnitude of the correlations among the variables. Secondly, it is known that the expected volume of the simplex formed by the $k+1$ random points in k dimensions or k random points and the mean vector is equal to the generalized variance, $|\Sigma|$. This is a natural generalization of the fact that the expected distance between two points or one point and the mean is the variance in one dimension. Thirdly, if the probability that a random point will lie in a k-dimensional ellipsoid of unit volume is large, then the population is well concentrated about the mean. For multivariate normal populations, $|\Sigma|$ has a further interesting interpretation. For such populations, the smaller the generalized variance, the higher the 'concentration' of the variable around $X=\mu$, since the density function at $X=\mu$ is a monotonically decreasing function of $|\Sigma|$. For more details, the reader is referred to Wilks (1967). Hence, using generalized variance, elements of different clusters will have different degrees of multivariate concentration as reflected by the volume $|\Sigma|$.

2. Clustering by generalized variance

Suppose that the means of the given k subclusters are nearly equal. Assuming multivariate normality for the p-dimensional observations with population dispersion matrix Σ_i for the ith subcluster, one may attempt to differentiate the populations by Σ_i. In principle, one needs to consider all possible groups of subclusters, the total number being the sum of Stirling numbers of the second kind – an incredibly large number, even for small k. Let $\Delta_i^{2p}=|\Sigma_i|$ denote the population generalized variance (GV) and so Δ_i^2 will be called standardized generalized variance (SGV). With Δ_i^{2p} or Δ_i^2 it is justifiable to consider only contagious partitions, i.e., if $d_i^2<d_j^2<d_k^2$ and d_i^2 and d_k^2 are included in the same cluster, then so should d_j^2 where the d_s^2 are sample SGVs. This reduces, to a great extent, the number of partitions to be considered to $\sum_{g=1}^{k}\binom{k-1}{g-1}=2^{k-1}$ – still quite prohibitive. However, if the number of groups, g, is specified beforehand, then only $\binom{k-1}{g-1}$ partitions need be considered. In any case, since the sample SGVs may be the same for different sample sizes, a test based on, say for $g=2$, the ratio of two SGVs will lead to different conclusions with differing sample sizes for the same SGVs. Hence, we restrict ourselves to the case of equal sample sizes. Further, we will propose a simple alternative testing procedure to obtain clusters in the absence of any knowledge of g.

3. Statistical tests

We will consider two approaches to the testing problem with contagious partitions only.

3.1. Likelihood ratio test

Result 3.1. (i) *The LR test for* H_{01}: $\underline{X}_{il}$, $l=1,\ldots,T$, $i=1,\ldots,c$, *c specified, are observations from p-variate normal populations with SGVs* Δ_i^2 *all equal, against* H_{11}: *the given sets of observations come from c p-variate normal populations with SGVs not all equal, is given by*

$$L(c)=\max_{p(c)}\prod_{\substack{i=1\\ m_i\in p(c)}}^{c}\prod_{j=1}^{m_i}\left(d_j^2/\hat{\sigma}_0^2\right)^{Tp/2}<L_0$$

where the maximization is over all contagious partitions p(c) of the k subclusters into c clusters; for a fixed partition, say p'(c), $(m_1,\ldots,m_c)\in p'(c)$; d_j^2 *and* $\hat{\sigma}_0^2$ *are defined below and* L_0 *is a constant to be determined from the specified level of the test.*

(ii) *The LR test for* H_{02}: *same as* H_{01}, *c not specified, against* H_{12}: *same as* H_{11}, *c not specified, is given by*

$$L = \max_{c} L(x) < L_{00}$$

where L_{00} *is a constant to be determined from the specified level of the test.*

Proof. Let η_m denote the LR test for $H_0: \Delta_i^2$ all equal, $i = 1,\ldots,m$, against H_1: not H_0. Surely,

$$L(c) = \max_{p(c)} \prod_{m_i \in p(c)} \eta_{m_i}.$$

We derive η_{m_i} below in a more general form (Sen Gupta, 1981) where the number of observations can be possibly different for different populations.

Test for the equality of SGVs of c (> 2) *independent multivariate normal populations.* Let x_{il}, $l = 1,\ldots,N_i$, $i = 1,\ldots,c$ denote c random samples from c independent populations $N_{p_i}(\mu_i, \Sigma_i)$, $i = 1,\ldots,c$, respectively. We are interested in testing $H_0: \Delta_i^2$, $i = 1,\ldots,c$, all equal, against the alternative H_1: at least one inequality.

Under both H_0 and H_1, $\hat{\mu}_i = \bar{x}_i$, $i = 1,\ldots,c$. Let θ_{ij}, $i = 1,\ldots,c$, $j = 1,\ldots,p_i$, be the characteristic roots of $\Sigma_i^{-1}S_i$ respectively where S_i, $i = 1,\ldots,c$, are the sample sums of products matrices for X_i, $i = 1,\ldots,c$, respectively. For finding the MLEs of Σ_i, $i = 1,\ldots,c$, under H_0, it suffices to consider the objective function

$$\Phi = \kappa + \sum_{i=1}^{c} \sum_{j=1}^{p_i} \left(\tfrac{1}{2}N_i \ln \theta_{ij} - \tfrac{1}{2}\theta_{ij}\right) + \sum_{i=1}^{c} \tfrac{1}{2}\lambda_{ii+1}\left[\left(\sum_{j=1}^{p_i} \frac{1}{p_i} \ln \theta_{ij} - \ln s_i^2\right) - \left(\sum_{j=1}^{p_i} \frac{1}{p_{i+1}} \ln \theta_{i+1j} - \ln s_{i+1}^2\right)\right]$$

where κ is a constant and λ_{ii+1} are undetermined Lagrange multipliers with $c+1$ being replaced by 1 in the suffixes. Differentiating Φ with respect to the θ_{ij}'s and equating to zeros we have that

$$p_i N_i + (\lambda_{ii+1} - \lambda_{i-1i}) = p_i \theta_{ij}, \quad i = 1,\ldots,c, j = 1,\ldots,p_i, \lambda_{01} = \lambda_{c1}$$

$$\Rightarrow \theta_{ij} = \theta_{ij'} \Rightarrow \theta_{ij} = s_i^2/\hat{\sigma}_0^2, \quad i = 1,\ldots,c$$

where $\hat{\sigma}_0^2$ is the MLE of σ_0^2, the common unknown value of Δ_i^2, $i = 1,\ldots,c$. So

$$\sigma_0^2 \sum_{i=1}^{c} p_i N_i + \hat{\sigma}_0^2 \left[\sum_{i=1}^{c} (\lambda_{ii+1} - \lambda_{i-1i})\right] = \sum_{i=1}^{c} p_i s_i^2 \quad \Rightarrow \quad \hat{\sigma}_0^2 = \sum_{i=1}^{c} p_i s_i^2 / \sum_{i=1}^{c} p_i N_1.$$

Note that this agrees with the MLE for σ_0^2 of the univariate case. Hence, we get the following lemma.

Lemma 3.2. *The LR test for* $H_0: \Delta_i^2$, *all equal, against* H_1: *at least one of the* Δ_i^2, $i = 1,\ldots,c$, *different is given by*

$$\text{reject } H_0 \quad \textit{if and only if} \quad \eta = \prod_{i=1}^{c} \left(d_i^2/\hat{\sigma}_0^2\right)^{N_i p_i/2} < \eta_0$$

where η_0 *is a constant to be determined from the specified level of the test.*

Consider now the LR test based on the Σ_i's. Let $S(m_i) = \Sigma_{j=1}^{m_i} S_j$.

Result 3.3. *The LR test for* H_{0v}^* *against* H_{1v}^*, *where* H_{uv}^* *is the same as* H_{uv}, $u = 0, 1$, $v = 1, 2$, *above with* Δ_i^2

replaced by Σ_i is given by

$$\text{reject } H_{01}^* \quad \text{if} \quad L^*(c) = \max_{p(c)} \prod_{\substack{i=1 \\ m_i \in p(c)}}^{c} \left\{ \prod_{j=1}^{m_i} \left(|S_j|^{1/m_i} / |S(m_i)| \right) \right\} < L_0^*(c),$$

$$\text{reject } H_{02}^* \quad \text{if} \quad L^* = \max_c L_1^*(c) < L_{10}^*$$

where $L_{10}^*(c)$ and L_{10}^* are constants to be determined from the specified levels of the tests.

Proof. LR criterion for $H_0: \Sigma_j$'s all equal, $j = 1, \ldots, m_i$ against H_1: at least one of the Σ_j's different is given by the term within the second bracket. The rest follows as in Result 3.1.

3.2. Union intersection test

For some vector $\boldsymbol{a}$, consider the linear compound $\boldsymbol{a}'X$. We use the heuristic approach of Roy to produce the test by the union intersection (UI) principle.

Result 3.4. *The UI tests for H_{0v}^* against H_{1v}^*, defined in Result* 3.3, *$v = 1, 2$ are given by*

$$\text{reject } H_{01}^* \quad \text{if} \quad U(c) = \max_{p(c)} \prod_{\substack{i=1 \\ m_i \in p(c)}}^{c} \prod_{j=1}^{m_i} \left\{ \lambda_1 \left(S(m_i) S_j^{-1} \right) \right\}^{1/c} < U_0(c),$$

$$\text{reject } H_{02}^* \quad \text{if} \quad U = \max_c U(c) > U_0$$

where $\lambda_1(V)$ denotes the maximum characteristic root of V and $U_0(c)$, U_0 denote some constants to be determined.

Proof. It suffices to note that, after some simplifications, it follows that the term within the square bracket corresponds to the UI criterion for testing equality of Σ_j's, $j = 1, \ldots, m_i$ with equal sample sizes for the m_i populations.

4. Distributions of the criteria

The exact distributions for the above test criteria seem intractable. When c, the number of clusters, is specified, the percentage points may be available through simulation. For $c = 2$, percentage points for the LR test, based on equal means assuming equal covariance matrices, were obtained through simulation by Engelman and Hartigan (1969) for $p = 1$ and by Lee (1979) for $p = 2$. For $p > 2$ and additionally when c is not specified, the problem becomes compounded. In this case, one may attempt to exploit advantageously a test due to Birnbaum (1974) which does not require explicit knowledge of the critical values of the test statistics. It will be of interest to consider the distribution of η_m for the above tests and also for a test to be proposed in Section 5. The exact distribution of η_2 for any p and η_m for any m, $p = 2$ are available from Sen Gupta (1981).

Result 4.1. *The null distribution of η_m for any m, $p = 2$ is that of Bartlett's test statistic for homogeneity with parameters* $(\mathrm{d}f)n_i = p(N-p) + 1$, *$i = 1, \ldots, m$, N being the common sample size.*

Proof. The exact distribution of Bartlett's test statistic is available from Theorem 1 by Chao and Glaser (1978). Consider their generalization $L_{k,a}$. Also the exact distribution of η_m for $p = 1$ can be deduced from

Theorem 4.4.1 by Sen Gupta (1981). It then suffices to note that, for our case, $a_j = 1/m$, $\nu = mp(N-p)$ and hence the exact null density of η_m is the same as that of $L_{p(N-p)/2,a}$.

The above result enables us to get the percentage points for η_m via the use of tables by Dyer and Keating (1980).

When N is large, a multivariate $F_{\max}$ criterion proposed by Sen Gupta (1981), whose large sample distribution is conveniently tabulated, may also be used in stead of η_m.

5. Alternative test

A statistical test to obtain clusters is proposed which is simple and also requires considerably less computations than the above procedures. The cluster configuration will be unique for the given procedure.

Procedure 5.1. Arrange the sample GVs based on an equal number of observations, N, in, say, a decreasing order, $d_1^2 \geq \cdots \geq d_k^2$. With k subclusters, go to k^*, the 'mid-point' of the subclusters. Test, $H_0: \Delta_1^2 = \cdots = \Delta_{k^*}^2$. If H_{01} is accepted, proceed to test if $d_{k^*+1}^2$ can also be included in the cluster, i.e., $H_0^*: \Delta_1^2 = \cdots = \Delta_{k^*}^2 = \Delta_{k^*+1}^2$. If H_{01} is rejected, test $H_0^{**}: \Delta_1^2 = \cdots = \Delta_{k^*-1}^2$. Continue this process. This will give one cluster, say $(d_1^2, \ldots, d_l^2)$. Repeat the same technique with $d_{l+1}^2, \ldots, d_k^2$.

The above process will need only about $\frac{1}{4}k(k+1)$ operations and thus achieves tremendous saving.

References

Birnbaum, Z.W. (1974), Computers and unconventional test-statistics, in: F. Proschan and R.J. Serfling, eds., *Reliability and Biometry: Statistical Analysis of Lifelength* (Soc. Indust. Appl. Math., Philadelphia) pp. 441–458.

Chao, M. and R.E. Glaser (1978), The exact distribution of Bartlett's test statistic for homogeneity of variances with unequal sample sizes, *J. Amer. Statist. Assoc.* **73**, 422–426.

Dyer, D.C. and J.P. Keating (1980), On the determination of critical values for Bartlett's test, *J. Amer. Statist. Assoc.* **75**, 313–319.

Engelman, L. and J.A. Hartigan (1969), Percentage points of a test for clusters, *J. Amer. Statist. Assoc.* **64**, 1647–1648.

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

Lee, K.L. (1979), Multivariate tests for clusters, *J. Amer. Statist. Assoc.* **74**, 708–814.

Mezzich, J.E. and H. Solomon (1980), *Taxonomy and Behavioral Science* (Academic Press, New York).

Sen Gupta, A. (1981), Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions, Tech. Rept. 50, Department of Statistics, Stanford University, also submitted for publication.

Wilks, S.S. (1967), Multidimensional statistical scatter, in: T.W. Anderson, ed., *Collected Papers: Contributions to Mathematical Statistics*, pp. 597–614.