CLUSTERING METHODS BASED ON LIKELIHOOD RATIO CRITERIA

By

A. J. Scott
and
M. J. Symons

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 710

September 1970

CLUSTERING METHODS BASED ON LIKELIHOOD RATIO CRITERIA

A. J. Scott[1] and M. J. Symons

Department of Biostatistics
University of North Carolina at Chapel Hill

SUMMARY

The standard classification model with several normal populations is extended to the cluster analysis situation where little or no previous information about the population parameters is available. Some common clustering procedures are shown to be extensions of likelihood ratio methods of classification. The analysis suggests that the procedures may have a tendency to partition the sample into groups of about the same size. This suggestion is examined in an example.

1. INTRODUCTION

Clustering methods that split a large number of multivariate observations into a smaller number of relatively homogeneous groups are important in biological applications and many other fields. There are a wide variety of techniques available and some useful comparisons are contained in papers by Gower [1968] and Friedman and Rubin [1967]. The techniques seem to be applied in two rather different situations. In one case, the purpose of the analysis is purely descriptive. There are no assumptions, implicit or otherwise, about the form of the underlying population and the grouping is simply

---

[1]This work was done while on leave from the London School of Economics during 1969-70.

a useful condensation of the data. In the other case, it is felt that the population is composed of several distinct sub-category and the purpose of the analysis is to group together all those observations belonging to the same subcategory. We are concerned with this second type of problem here.

As a model for this situation, we suppose that each observation in the sample may come from any one of a small number of different distributions. This would be the standard classification problem if the distributions were known, or there was a substantial amount of information about them from previous samples (Anderson [1958], Ch. 6), but little or no prior knowledge about the component distributions is available in most situations where clustering techniques are used. In either case, classification or clustering, we want to group together all the observations from the same distribution. Let γ denote the set of identifying labels, i.e., if there are n sample observations, γ is an unknown parameter with n components where the i^{th} component indicates the distribution from which the i^{th} observation came. We derive the maximum likelihood estimate of γ under the assumption that the underlying distributions are multivariate normal and this turns out to be equivalent to several standard clustering methods with different assumptions about the covariance structure. These methods are shown to be natural extensions of standard classification rules based on the likelihood ratio criterion.

A related approach has been considered by Wolfe [1967, 1969] and Day [1969] who suppose that the observations are drawn independently from a mixture of multivariate normal distributions. This is equivalent to the model above with the additional assumption that γ is an (unobservable) random variable whose components are the outcomes of n independent multinomial trials. An indirect estimate of γ is obtained by estimating the parameters of the

mixture and using standard classification methods with these estimates in place of the unknown parameter. It is possible to go a step further than this and consider a Bayesian approach in which all the parameters are random variables. A very short sketch of this approach is given in section 4 and the results are compared with the maximum likelihood results. The comparison suggests that the methods based on maximum likelihood will perform best when the populations are represented in about equal proportions. The well-known Fisher Iris data is used in section 5 to explore this suggestion empirically.

2. THE MODEL

The sample consists of n observations $\underset{\sim}{Y} = (\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n)$ where $\underset{\sim}{y}_i$ represents measurements on p characteristics. Suppose that the observations are independent and that each may come from any one of G possible p-variate normal distributions with means $\underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_G$ and covariance matrices $\Sigma_1, \ldots, \Sigma_G$. For generality we allow for the possibility of a previous sample of independent observations $\underset{\sim}{x}_{g1}, \ldots, \underset{\sim}{x}_{gm_g}$ from each distribution. Then the joint distribution of $\underset{\sim}{Y}$ and the previous observations is completely determined by $\underset{\sim}{\mu}_g$, Σ_g $(g=1,\ldots,G)$ and the grouping or classification parameter $\underset{\sim}{\gamma} = (\gamma_1, \ldots, \gamma_n)$ where $\gamma_i = g$ when y_i comes from the g^{th} sub-population. If $\Theta = (\underset{\sim}{\gamma}, \underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_G, \Sigma_1, \ldots, \Sigma_G)$ denotes the collection of all the parameters, the log likelihood function, $\ell(\Theta)$, is given by

$$\ell(\Theta) = -\tfrac{1}{2} \sum_{g=1}^{G} \Big[\sum_{i=1}^{m_g} (\underset{\sim}{x}_{gi} - \underset{\sim}{\mu}_g)' \Sigma_g^{-1} (\underset{\sim}{x}_{gi} - \underset{\sim}{\mu}_g) + \sum_{C_g} (\underset{\sim}{y}_i - \underset{\sim}{\mu}_g)' \Sigma_g^{-1} (\underset{\sim}{y}_i - \underset{\sim}{\mu}_g) + (m_g + n_g) \log |\Sigma_g| \Big] \tag{1}$$

where C_g is the set of $\underset{\sim}{y}_i$'s assigned to the g^{th} group or cluster by $\underset{\sim}{\gamma}$, n_g is

the number of observations in C_g, and $|\Sigma_g|$ denotes the determinant of Σ_g.

The classification or clustering problem is to estimate $\underset{\sim}{\gamma}$ and hence the clusters $C_1,\ldots,C_G$. If the means and covariances are known, or there are a large number of previous observations from each sub-population, this is the classical model for the classification problem. When there is little or no prior information about the components, the problem becomes one of cluster analysis.

3. MAXIMUM LIKELIHOOD ESTIMATES

For a given partition of $\underset{\sim}{Y}$ into G groups $C_1,\ldots,C_G$ the likelihood is maximized by substituting the ordinary maximum likelihood estimates of $\underset{\sim}{\mu}_g$ and Σ_g. The estimate of $\underset{\sim}{\mu}_g$, whatever the assumptions about the Σ_g, is

$$\hat{\underset{\sim}{\mu}}_g(\underset{\sim}{\gamma}) = \frac{m_g \bar{\underset{\sim}{x}}_g + n_g \bar{\underset{\sim}{y}}_g}{m_g + n_g}$$

where $\bar{\underset{\sim}{y}}_g$ is the mean of the n_g observations in C_g. When $\hat{\underset{\sim}{\mu}}_g(\underset{\sim}{\gamma})$ is substituted for $\underset{\sim}{\mu}_g$ in expression (1) it follows that the maximum likelihood estimate, $\hat{\underset{\sim}{\gamma}}$, of $\underset{\sim}{\gamma}$ can be found by minimizing

$$\sum_{g=1}^{G} \{\mathrm{tr}[(W_{gx}+W_{gy}+W_{gxy})\Sigma_g^{-1}] + (m_g+n_g)\log|\Sigma_g|\} \qquad (2)$$

where

$$W_{gx} = \sum_{i=1}^{m_g} (\underset{\sim}{x}_{gi}-\bar{\underset{\sim}{x}}_g)(\underset{\sim}{x}_{gi}-\bar{\underset{\sim}{x}}_g)',$$

$$W_{gy} = \sum_{C_g} (\underset{\sim}{y}_i-\bar{\underset{\sim}{y}}_g)(\underset{\sim}{y}_i-\bar{\underset{\sim}{y}}_g)',$$

and

$$W_{gxy} = \frac{m_g n_g}{m_g+n_g} (\bar{y}_g - \bar{x}_g)(\bar{y}_g - \bar{x}_g)' .$$

3.1 Equal Covariance Matrices

If $\Sigma_g = \Sigma (g=1,\ldots,G)$ then expression (2) reduces to

$$\mathrm{tr}[(W_x+W_y+W_{xy})\Sigma^{-1}] + (m+n) \log |\Sigma| , \tag{3}$$

where $W_x = \Sigma W_{gx}$ is the within-groups sum of squares matrix for the x's, $W_y = \Sigma W_{gy}$ is the within-groups sum of squares matrix for the y's and $W_{xy} = \Sigma W_{gxy}$ is the contribution due to the differences between y_g and x_g.

If Σ is known, (3) reduces further. The assumption of known Σ is reasonable when each y_i is actually the mean of many observations which can be used to provide a good estimate of Σ. Edwards and Cavalli-Sforza [1965] discuss such an example. In this situation $\hat{\gamma}$ is the grouping that minimizes

$$\mathrm{tr}[(W_y+W_{xy})\Sigma^{-1}] \tag{4}$$

Two cases are of particular interest. In the limit as the m_g's become large, expression (4) is minimized by assigning each y_i independently to the group with the smallest value of $(y_i - \bar{x}_g)'\Sigma^{-1}(y_i - \bar{x}_g)$. This, of course, is the standard classification procedure when the costs of misclassification are assumed equal and no prior information about γ is available. It is also the minimax procedure. (Anderson [1958], §6.6)

At the other extreme there is no previous information at all so that

$m_g=0$ $(g=1,\ldots,G)$. Then expression (4) reduces to

$$\mathrm{tr}(W_y \Sigma^{-1}) = \sum_{g=1}^{G} \sum_{C_g} (\underset{\sim}{y}_i-\bar{\underset{\sim}{y}}_g)' \Sigma^{-1} (\underset{\sim}{y}_i-\bar{\underset{\sim}{y}}_g) \ . \tag{5}$$

Equivalently, $\hat{\underset{\sim}{\gamma}}$ maximizes the weighted between-groups sum of squares

$$\mathrm{tr}(B_y \Sigma^{-1}) = \sum_{g=1}^{G} n_g (\bar{\underset{\sim}{y}}_g-\bar{\underset{\sim}{y}})' \Sigma^{-1} (\bar{\underset{\sim}{y}}_g-\bar{\underset{\sim}{y}}) \ , \tag{6}$$

where

$$B_y = \sum_{y=1}^{G} n_g (\bar{\underset{\sim}{y}}_g-\bar{\underset{\sim}{y}}) (\bar{\underset{\sim}{y}}_g-\bar{\underset{\sim}{y}})' \ .$$

It would be natural to pre-standardize in this case so that $\Sigma=I$, the identity matrix, and the maximum likelihood partition minimizes

$$\mathrm{tr}(W_y) = \sum_{g=1}^{G} \sum_{C_g} (\underset{\sim}{y}_i-\bar{\underset{\sim}{y}}_g) \ (\underset{\sim}{y}_i-\bar{\underset{\sim}{y}}_g)', \tag{7}$$

the total within-groups sum of squares, or maximizes $\mathrm{tr}(B_y)$, the between-groups sum of squares. This has been widely used as a criterion for cluster analysis and forms the basis of the method suggested by Edwards and Cavilli-Sforza [1965]. They use the criterion to partition the set first into two groups, then to subdivide each group, and so on.

If Σ is not known, its maximum likelihood estimate for fixed $\underset{\sim}{\gamma}$ is equal to $(W_x+W_y+W_{xy})/(m+n)$. When this is substituted in expression (3) it follows that $\underset{\sim}{\gamma}$ is the grouping that minimizes.

$$|W_x + W_y + W_{xy}| \ . \tag{8}$$

Again some extreme cases are of special interest. For a single new observation $\underset{\sim}{y}$ it can be shown that this reduces to minimizing

$$\frac{m_g}{m_g+1} \; (\underset{\sim}{y}-\bar{\underset{\sim}{x}}_g)' \; W_x^{-1} (\underset{\sim}{y}-\bar{\underset{\sim}{x}}_g) \tag{9}$$

which is a natural extension of the likelihood ratio method for classifying an observation into one of two populations given previous samples from each. (Anderson [1958], §6.5.5)

When there is no previous sample information about any of the populations, $\hat{\underset{\sim}{\gamma}}$ is the grouping that minimizes $|W_y|$, the determinant of the within-groups sum of squares matrix. This is another widely used criterion, suggested by Friedman and Rubin [1967]. They have found methods based on $|W_y|$ to work well in several empirical studies.

The special case of two groups has been considered by John [1970] in a slightly different context. He shows that minimizing $|W_y|$ is equivalent to maximizing

$$\mathrm{tr}(B_y \, T^{-1}) = \sum_{g=1}^{2} n_g (\bar{\underset{\sim}{y}}_g - \bar{\underset{\sim}{y}})' \; T^{-1} (\bar{\underset{\sim}{y}}_g - \bar{\underset{\sim}{y}}) \; , \tag{10}$$

where

$$T = \sum_{1}^{n} (\underset{\sim}{y}_i - \bar{\underset{\sim}{y}}) (\underset{\sim}{y}_i - \bar{\underset{\sim}{y}})'$$

is the total scatter matrix. This is a weighted between-groups sum of squares like expression (6) for known Σ but the weighting depends on the sample quantity T rather than Σ. Use of expression (10) makes computation easier but unfortunately the result does not extend in a simple way to more than two groups. It can be shown in the same way that for G=2 minimizing $|W_y|$ is equivalent to

maximizing $\mathrm{tr}(W_y^{-1} B_y)$, the Hotelling Trace. This has also been considered as a criterion for cluster analysis by Friedman and Rubin [1967].

3.2 Unequal Covariance Matrices

If Σ_g the $(g=1,\ldots,G)$ are specified then $\hat{\underset{\sim}{\gamma}}$ minimizes

$$\sum_{g=1}^{G} \{\mathrm{tr}[(W_{gx}+W_{gy}+W_{gxy})\ \Sigma_g^{-1}] + n_g \log |\Sigma_g|\} \tag{11}$$

There seems to be little of interest to say about this case.

If the Σ_g are not known, the maximum likelihood estimates for given $\underset{\sim}{\gamma}$ are equal to $(W_{gn}+W_{gy}+W_{gny})/(m_g+n_g)$ for $g=1,\ldots,G$, which can be substituted in expression (2). In this case, $\underset{\sim}{\gamma}$ is the grouping that minimizes

$$\prod_{g=1}^{G} |W_{gx} + W_{gy} + W_{gxy}|^{m_g+n_g} . \tag{12}$$

Again the extreme cases have a fairly simple form. When a single observation $\underset{\sim}{y}$ is to be classified on the basis of previous samples, this reduces to assigning $\underset{\sim}{y}$ to the population with the smallest value of

$$|W_{gx}|\ [1 + \frac{m_g}{m_g+1} (\underset{\sim}{y}-\underset{\sim}{x}_g)'\ W_{gx}^{-1}(\underset{\sim}{y}-\underset{\sim}{x}_g)]^{n_g+1} . \tag{13}$$

This is a natural extension of the likelihood ratio method (Anderson [1958], §6.5.5) to unequal covariance matrices.

When there is no previous sample information, $\underset{\sim}{\gamma}$ is equivalent to choosing groups so that

$$\prod_{g=1}^{G} |W_{gy}|^{n_g} \tag{14}$$

maximizing $\mathrm{tr}(W_y^{-1} B_y)$, the Hotelling Trace. This has also been considered as a criterion for cluster analysis by Friedman and Rubin [1967].

3.2 Unequal Covariance Matrices

If Σ_g the (g=1,...,G) are specified then $\hat{\underset{\sim}{\gamma}}$ minimizes

$$\sum_{g=1}^{G} \{\mathrm{tr}[(W_{gx}+W_{gy}+W_{gxy})\ \Sigma_g^{-1}] + n_g \log |\Sigma_g|\} \tag{11}$$

There seems to be little of interest to say about this case.

If the Σ_g are not known, the maximum likelihood estimates for given $\underset{\sim}{\gamma}$ are equal to $(W_{gn}+W_{gy}+W_{gny})/(m_g+n_g)$ for g=1,...,G, which can be substituted in expression (2). In this case, $\underset{\sim}{\gamma}$ is the grouping that minimizes

$$\prod_{g=1}^{G} |W_{gx} + W_{gy} + W_{gxy}|^{m_g+n_g} . \tag{12}$$

Again the extreme cases have a fairly simple form. When a single observation $\underset{\sim}{y}$ is to be classified on the basis of previous samples, this reduces to assigning $\underset{\sim}{y}$ to the population with the smallest value of

$$|W_{gx}|\ [1 + \frac{m_g}{m_g+1} (\underset{\sim}{y}-\underset{\sim}{x}_g)'\ W_{gx}^{-1}(\underset{\sim}{y}-\underset{\sim}{x}_g)]^{n_g+1} . \tag{13}$$

This is a natural extension of the likelihood ratio method (Anderson [1958], §6.5.5) to unequal covariance matrices.

When there is no previous sample information, $\underset{\sim}{\gamma}$ is equivalent to choosing groups so that

$$\prod_{g=1}^{G} |W_{gy}|^{n_g} \tag{14}$$

is minimized. (The restriction that at least (p+1) observations must be assigned to each group avoids the degenerate case of infinite likelihood.) As far as we know, this has not been used as a criterion for cluster analysis, although Chernoff [1970] has suggested using the individual cluster scatter matrices W_{gy} rather than the pooled matrix W_y when the cluster shapes are very different.

4. BAYES ESTIMATES

An alternative approach is to specify a prior distribution for $\Theta = (\underset{\sim}{\gamma}, \underset{\sim}{\mu}_1, \ldots, \underset{\sim}{\mu}_G, \Sigma_1, \ldots, \Sigma_G)$ which can be combined with the likelihood defined by (1) to obtain the posterior distribution of Θ given the sample and previous observations. We can then obtain the marginal posterior probability of any grouping $\underset{\sim}{\gamma}$ by integrating out the nuisance parameters $\underset{\sim}{\mu}_1, \ldots \underset{\sim}{\mu}_G$ and $\Sigma_1, \ldots, \Sigma_G$. The obvious choice of a single summary statistic for the distribution is the mode since there is no natural ordering of the G^n possible values of $\underset{\sim}{\gamma}$, and we take the posterior mode as our point estimate of $\underset{\sim}{\gamma}$.

Results very similar to those of the previous section are obtained when there is little prior information about $\underset{\sim}{\mu}_g$ and Σ_g beyond that contained in the previous samples. If $\Sigma_g = \Sigma$ (g=1,...,G), for example, this situation might be approximated roughly by taking the prior density of Θ to have the form

$$p(\Theta) \propto p(\underset{\sim}{\gamma}) \; |\Sigma|^{-\frac{p+1}{2}} \tag{15}$$

over the region of interest. (Geisser and Cornfield [1963]) In this case the posterior probability of $\underset{\sim}{\gamma}$ is proportional to

$$|W_x+W_y+W_{xy}|^{-\frac{(m+n-g)}{2}} \; p(\underset{\sim}{\gamma}) \prod_{m_g+n_g>0} (m_g+n_g)^{-\frac{1}{2}} \tag{16}$$

The grouping with the highest posterior probability corresponds exactly to the maximum likelihood estimate of $\underset{\sim}{\gamma}$ if

$$p(\underset{\sim}{\gamma}) \propto \prod_{m_g+n_g>0} (m_g+n_g)^{\frac{1}{2}} . \tag{17}$$

When the previous samples are large this is approximately constant which implies that each individual is regarded as being equally likely to come from any of the G populations. This seems a reasonable choice if there are only a few new observations to be classified, but the prior probability that the groups are of equal size becomes large as n increases. In the cluster analysis situtation with $m_g=0$ the density (17) becomes

$$p(\underset{\sim}{\gamma}) \propto \prod_{n_g>0} n_g^{\frac{1}{2}} \tag{18}$$

which puts even heavier weight on groups of equal size. The fact that the maximum likelihood estimate of $\underset{\sim}{\gamma}$ corresponds to the Bayes estimate for such extreme weights suggest that it may have a tendency to force the data into a balanced split. The suggestion is examined in an example in the next section.

If there is no extraneous information about individual $\underset{\sim}{y}_i$'s a natural way of generating $p(\underset{\sim}{\gamma})$ is to suppose that each $\underset{\sim}{y}_i$ has the same probability, Π_g say, of coming from the $g^{\underline{th}}$ population, and to specify a distribution for $(\Pi_1,\ldots,\Pi_{G-1})$. This is exactly the same as supposing that the $\underset{\sim}{y}_i$'s are independent observations from a mixture of G normal distributions, as in Wolfe [1969] and Day [1969], and then specifying a prior distribution for the mixture probabilities. Thus, from a Bayesian viewpoint, there is no difference between the two models: The mixture model simply requires a special, but

very natural, type of prior distribution for $\underset{\sim}{\gamma}$. A difficulty with this approach is the choice of a distribution for $(\Pi_1,\ldots,\Pi_{G-1})$ to reflect rather vague knowledge. As indicated above, the results that are closest to the maximum likelihood partition are obtained by taking $\Pi_i=G^{-1}$ with certainty. Another possibility is to take a uniform density on $(\Pi_1,\ldots,\Pi_{G-1})$ where $\Pi_i\geq 0$ and $\sum \Pi_i=1$.

This leads to

$$p(\underset{\sim}{\gamma}) \propto \left[\prod_{g=1}^{G} n_g!\right]^{-1} \tag{19}$$

which puts much more emphasis on uneven splits than the first choice. These two alternatives correspond to the Maxwell-Boltzmann and Einstein-Bose distribution of physics. (Feller [1950])

5. NUMERICAL RESULTS

Friedman and Rubin [1967] looked at the performance of a number of clustering methods, including those based on $|W_y|$ and $\mathrm{tr}(W_y)$, when applied to 3 sets of data containing groups of observations from several populations. In two of the data sets, the groups were exactly equal in size while the third set had 4 groups containing 23, 18, 21, and 35 observations respectively. The $|W_y|$ criterion gives excellent results for all three sets, but the discussion of the previous section suggest that the results might be less satisfactory with more disparate groups.

We explored the consequences of unequal group sizes by looking at various subsets of the second set of data, the well-known Iris data published by Fisher [1936] in his paper introducing the linear discriminant function. The data are reproduced on p. 318 of Kendall and Stuart [1966]. There are four

measurements on 50 plants from each of three species of Iris: Iris setosa, Iris versicolor, and Iris virginica. The Setosa plants are so well separated from the others that any reasonable method can isolate them, but there is some overlap between the other two species. Most of the separation is in the measurements on petal length and petal width and these components are plotted in Figure 1. The split into 3 groups reported by Friedman and Rubin for the $|W_y|$ criterion recovered the three species completely except for two Versicolor plants which were placed in the Virginica cluster and one Virginica plant which was grouped with the other Versicolor plants.

[Insert Figure 1.]

In practice, it is impossible to find the minimum value of $|W_y|$ by searching over all possible partitions of a set of n observations unless n is quite small. We used an approximate routine constructed by D. J. MacRae [1970], which incorporates techniques suggested by Forgy [1965], MacQueen [1967], and Friedman and Rubin [1967]. This produces a relative minimum for $|W_y|$ in the sense that any reassignment of one or two observations results in a larger value, but does not guarantee an absolute minimum. To supplement the program, we examined individually all the partitions in the neighborhood of the true split as well as particular partitions that were suggested by a visual inspection of the two-dimensional scatter diagrams.

We first looked at subsets of the Versicolor and Virginica species. As long as the two groups were of equal size, the $|W_y|$ criterion gave good results. Even when size was reduced to 10 observations from each species, the method consistently yielded clusters that could be identified clearly with a single species with only about 10% of the observations misclassified. Next, 5

unbalanced data sets were created by splitting the Versicolor plants into 5 equal groups and combining each group with all 50 Virginica plants. In each case there was one relative minimum near the true split and another close to a partition into 2 equal groups of 30 observations each. The results are summarized in Table 1. In 4 of the 5 cases the near-equal split had the lowest value of $|W_y|$ that was found and only in one case was anything like the actual partition recovered.

[Insert Table 1.]

Five more sets were produced in the same way by splitting the Virginica into 5 equal groups and combining each group with all 50 Versicolor plants. These led to very similar results with a near-equal split having the smallest value of $|W_y|$ in 4 of the 5 sets. The effect of increasing the size of the smaller group was investigated next. The results were very little better when 15 observations from one species was combined with all 50 observations from the other. There was a substantial improvement, however, when the smaller group was increased to 20 observations. Six such sets were constructed and clusters clearly identifiable with a single species were produced with every set. About 11% of the observations were misclassified.

Finally, subsets of the Setosa plants were combined with the other species. The separation is much more clear-cut here and the results were very satisfactory. The Setosa observations were always isolated perfectly even when as few as 10 Setosa plants were combined with all 50 Versicolor plants.

On the whole, the results confirm the value of $|W_y|$ as a criterion for cluster analysis. It led to meaningful clusters whenever the separation was large or the groups were of roughly the same size. However, those results do

seem to confirm that there is a tendency to divide the data into even groups if the separation between the sub-populations is not large. This suggests that some care should be taken in interpreting an analysis based on the criterion of Section 3 if the resulting clusters are of about the same size.

6. CONCLUDING NOTE

So far we have not touched on one of the most basic questions of cluster analysis: How many clusters are there? Although we have only set an upper bound to the number of clusters in theory, in practice the maximum likelihood methods will always partition the data into the maximum number of partitions allowed. One way of approaching the question is to rephrase it as a testing problem. For example, the fundamental question of whether there is more than one cluster can be considered as a test of the null hypothesis:

$$H_o: \quad \gamma_1=\gamma_2= \cdots =\gamma_n$$

against the alternative that not all the γ_i's are equal. If λ denotes the likelihood ratio statistic then, in the case $\Sigma_g=\Sigma(g=1,\ldots,G)$ with Σ unknown,

$$-2 \log \lambda = n \log \left[\max_{\underset{\sim}{\gamma}} (|T|/|W_y|)\right] \tag{20}$$

where $T = \sum(\underset{\sim}{y}_i-\bar{\underset{\sim}{y}})(\underset{\sim}{y}_i-\bar{\underset{\sim}{y}})'$ is the total scatter matrix. The assumptions of the usual asymptotic theory for likelihood ratio tests do not hold here and the problem of finding the null distribution appears to be intractable. Friedman and Rubin [1967] noted that the log-Max in (17) gives an indication of the number of clusters.

REFERENCES

Anderson, T. W. [1958]. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Chernoff, H. [1970]. Metric considerations in the k-means method of cluster analysis. Classification Society meetings. Ohio University, Columbus, April 8-9.

Day, N. E. [1969]. Estimating the components of a mixture of normal distributions. Biometrika 56, 463-74.

Edwards, A. W. F. and Cavalli-Sforza, L. L. [1965]. A method for cluster analysis. Biometrics 21, 362-75.

Feller, W. [1957]. An Introduction to Probability Theory and Its Applications. 2nd Edition, Wiley, New York.

Fisher, R. A. [1936]. The use of multiple measures in taxonomic problems. Annals of Eugenetics 8, 376-86.

Forgy, E. W. [1965]. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. WNAR meetings. University of California, Riverside, June 22-23.

Friedman, H. P. and Rubin, J. [1967]. On some invariant criterion for grouping data. Journal of the American Statistical Association 62, 1159-1178.

Geisser, S. and Cornfield, J. [1963]. Posterior distributions for multivariate normal parameters. Journal of the Royal Statistical Society, Series B 25, 368-76.

Gower, J. C. [1967]. A comparison of some methods of cluster analysis. Biometrics 23, 623-637.

John, S. [1970]. On identifying the population of origin of each observation in a mixture of observations from two normal populations. Technometrics 12, 553-63.

Kendall, M. G. and Stuart, A. [1966]. The Advanced Theory of Statistics Vol. 3. Hafner, New York.

Mac Queen, J. [1965]. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. I, 281-297.

McRae, D. J. [1969]. MIKCA: a FORTRAN IV iterative K-means cluster analysis program. Research Memo. No. 34. Psychometric Laboratory, University of North Carolina at Chapel Hill.

Wolfe, J. H. [1967]. NORMIX: computation methods for estimating the parameters of multivariate normal mixtures of distributions. Research Memo. SRM 68-2. U. S. Naval Personnel Research Activity, San Diego, California.

Wolfe, J. H. [1969]. Pattern clustering by multivariate mixture analysis. Research Memo. SRM 69-17. U. S. Naval Personnel Research Activity, San Diego, California.

LIST OF FIGURES

FIGURE 1

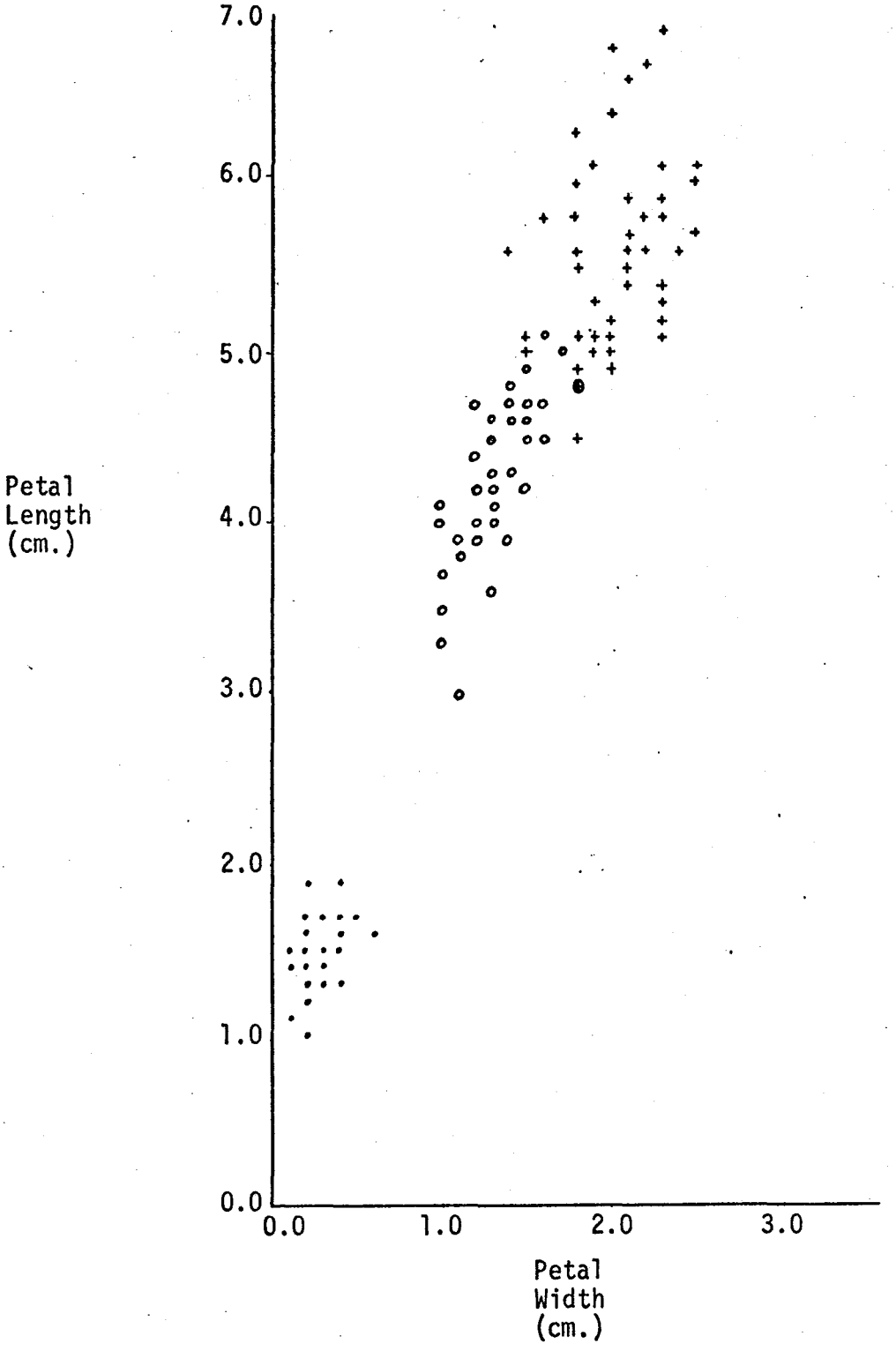

TABLE 1

Values of $|W_y|$ at relative minima and the corresponding cluster sizes

Observations	1^{st} 10 Versicolor 50 Virginica		2^{nd} 10 Versicolor 50 Virginica		3^{rd} 10 Versicolor 50 Virginica		4^{th} 10 Versicolor 50 Virginica		5^{th} 10 Versicolor 50 Virginica	
$\|W_y\|$	1153.0	1217.1	1404.3	1441.4	884.1	1225.3	112.4	1306.8	1086.2	1069.5
Cluster Sizes	26,34	11,49	30,30	10,50	30,30	10,50	30,30	10,50	29,31	10,51