九州大学学術情報リポジトリ
Kyushu University Institutional Repository

MULTI-SAMPLE CLUSTER ANALYSIS AS AN ALTERNATIVE TO MULTIPLE COMPARISON PROCEDURES

Bozdogan, Hamparsum
Department of Mathematics, University of Virginia

https://doi.org/10.5109/13380

出版情報：Bulletin of informatics and cybernetics. 22 (1/2), pp.95-130, 1986-03. Research Association of Statistical Sciences
バージョン：
権利関係：

KYUSHU UNIVERSITY

Bulletin of Informatics and Cybernetics Vol. 22, No. 1~2, 1986

MULTI-SAMPLE CLUSTER ANALYSIS AS AN ALTERNATIVE TO MULTIPLE COMPARISON PROCEDURES*

By

Hamparsum BOZDOGAN**

Abstract

This paper studies *multi-sample cluster analysis*, the problem of grouping samples, as an alternative to multiple comparison procedures through the development and the introduction of *model-selection criteria* such as those: *Akaike's Information Criterion* (*AIC*) and its extension CAIC also known as *Schwarz's Criterion* (*SC*), as new procedures for comparing *means, groups, or samples*, and so forth, in identifying and selecting the homogeneous groups or samples from the heterogeneous ones in multi-sample data analysis problems.

An enumerative clustering technique is presented to generate all possible choices of clustering alternatives of groups, or samples on the computer using efficient combinatorial algorithms without forcing an arbitrary choice among the clustering alternatives, and to find all sufficiently simple groups or samples consistent with the data and a parsiidentify the best clustering among the alternative clusterings.

Numerical examples are carried out and presented on a real data set on grouping the samples into fewer than K groups. Through a Monte Carlo study, an application of multi-sample cluster analysis is shown in designing optimal decision tree classifiers in reducing the dimensionality of remotely sensed heterogeneous data sets to achieve a parsimonious grouping of samples.

The results obtained demonstrate the utility and versatility of model-selection criteria which avoid the notorious choice of levels of significance and which are free from the ambiguities inherent in the application of conventional hypothesis testing procedures.

1. Introduction

Many practical situations require the presentation of multivariate data from several structured samples for *comparative inference* and the grouping of the *heterogeneous* samples into *homogeneous* sets of samples.

For example, in analysis of variance, to describe any variations of the treatment

* This research was supported by Army Research Office Contract DAAG 29-82-K-0155, and was presented at the 9 th German Classification Society's Annual Meeting at the University of Karlsruhe, West Germany, June 26-28, 1985.

** Department of Mathematics, University of Virginia, Charlottesville, Virginia, 22903 U.S.A.

means, we partition the treatment means into groups with hopefully the same mean for all treatments in the same group to find a more parsimonious grouping of treatments (Cox and Spjøtvoll [13]).

In remote sensing technology, see, e.g., Argentiero et al. [7], we classify or group different samples of large dimensional remotely sensed heterogeneous data sets into homogeneous sets of samples to reduce the dimensionality of these data sets and to design optimal decision tree classifiers. Decision tree classifiers are popular and useful to study the underlying data structure which have the property that samples are subjected to a sequence of decision rules before they are assigned to a unique class to identify and determine the number of types that the classes originally might have been consisted. Such an approach, providing that it is well designed, will give us a classification scheme which is accurate, flexible, and computationally efficient.

The purpose of this paper is, therefore, to propose and to study Multi-Sample Cluster Analysis (MSCA), the problem of grouping samples, developed by this author (see, e.g., Bozdogan [8], Bozdogan and Sclove [12], as an alternative to Multiple Comparison Procedures (MCP's) through the development and introduction of model-selection criteria such as those of Akaike ([1], [2], [4]), Schwarz [33], and Bozdogan [11], as new procedures for the comparisons and identification of various collections of *groups, samples, treatments, experimental conditions* or *diagnostic classifications,* and so forth, in multi-sample data analysis problems.

In the statistical literature, the Analysis of Variance (ANOVA) is a widely used model for comparing two or more univariate samples, where the familiar Student's t and F statistics are used for formal comparisons among two or more samples. In the multi-sample case the Multivariate Analysis of Variance (MANOVA) is a widely used model for comparing two or more multivariate samples. In the MANOVA model, the likelihood ratio principle leads to Wilks' [42] lambda, or in short Wilks' Λ criterion as the test statistic. It plays the same role in the multivariate analysis that F-ratio statistic plays in the univariate case.

Often, however, the formal analyses involved in ANOVA or in MANOVA are not revealing or informative. For this reason, in any problem where a set of parameters is to be partitioned into groups, it is reasonable to provide a practically useful statistical procedure or procedures that would use some sort of statistical model to aid in comparisons of various collections of comparable groups, samples, etc., and identify the homogeneous groups from the heterogeneous ones, or vice versa, and tell us which groups (or samples) should be clustered together and which groups (or samples) should not be clustered together.

The object of this paper is to point out an enumerative clustering technique to generate all possible choices of clustering alternatives of groups or samples on the computer using efficient combinatorial algorithms without forcing an arbitrary choice among the clustering alternatives.

Thus the central idea is that through Multi-Sample Cluster Analysis (MSCA) as an alternative to Multiple Comparison Procedures (MCP's) and through the use of model-selection criteria we shall find all sufficiently simple partitions of groups or samples consistent with the data and identify the best clustering among the alternative

clusterings. We achieve this by utilizing a new information-theoretic approach to the multi-sample conventional tests of homogeneity models discussed in Bozdogan [10]. This approach unifies the conventional test procedures without the worry of what level of significance α one needs to use. In a conventional pre-test situation, it has become customary to fix the level of significance a priori at, for example, 1%, 5%, or 10% levels regardless of the number of parameters estimated within a model. This is essentially arbitrary and no rational basis exists for making such an arbitrary choice. Model-selection criteria adapt themselves to the number of parameters estimated within a model to achieve parameter parsimony, and the significance level is adjusted accordingly from one model to the next.

In Section 2, we shall briefly discuss the Multiple Comparison Procedures (MCP's) and present their formulation in the multivariate case. Then we shall outline the existing problems inherent with the MCP's. In Section 3, we shall propose Multi-Sample Cluster Analysis (MSCA) as an alternative to conventional Multiple Comparison Procedures (MCP's). We shall define the general MSCA problem, and discuss how to obtain the total number of clustering alternatives for a given K, the number of groups or samples in detail for both MCP's and MSCA. In the subsequent section, that is, in Section 4, we shall briefly give the formal definitions of model-selection criteria and present the three most commonly used multivariate multi-sample models, that is, multi-sample hypotheses, and give their model-selection-replacements. For more on this, we refer the reader to Bozdogan [10]. In Section 5, we shall give numerical examples on a real data set, and show an application of MSCA in designing optimal decision tree classifiers.

Finally, in Section 6, we shall present our conclusions, and give a listing of the combinatorial subroutines in the Appendix.

2. Multiple Comparison Procedures (MCP's)

In the univariate analysis of variance (ANOVA) model for testing the equality of K population means, as we mentioned in the introduction of this paper, the test statistic $F=S_B^2/S_w^2$ is used for comparing several population means. If we compute the value of F for the sample data, and if it is larger than the critical value of F obtained from standard F-tables at some prescribed α level, then we reject the overall, "omnibus", *null hypothesis*

$$H_0: \mu_1=\mu_2=\cdots=\mu_K \tag{2.1}$$

in favor of the *alternative hypothesis* given by

$$H_1: \text{the } K \text{ population means are not all equal.}$$

While rejecting the null hypothesis gives us some information about the population means, namely the heterogeneity of the means, we do not know which means differ from each other. Hence, both ANOVA or MANOVA do not pinpoint exactly where the significant differences lie, and an F test alone, generally falls short of satisfying all of the practical requirements involved (Duncan [14]). For example, if $K=3$ and $H_0: \mu_1=\mu_2=\mu_3$ is rejected, then we do not know whether the main differences are

between μ_1 and μ_2, or between μ_1 and μ_3 and so on. Therefore, we are faced with many new problems, and we may ask the following simple and yet important questions: Does μ_1 differ from μ_2?, Does μ_1 differ from μ_3?, Which of the samples are considered coming from common populations, which are not?

As in the univariate ANOVA model, the same problems arise in the multivariate analysis of variance (MANOVA) model also. That is, rejection of the null hypothesis does not indicate which groups, samples, or treatments, or any combinations of them are different and which should be considered as coming from common populations, which are not.

Therefore, it is important to obtain some idea where the differences in the means or mean vectors are when we reject the null hypothesis and establish some relationships among the unequal means or mean vectors.

2.1 Formulation of MCP's

In the univariate case, i.e., in the case of one response variable, there exists a multitude of Multiple Comparison Procedures (MCP's) available in the literature. However, in the multivariate case, even only two variables, there seems to be a few applicable techniques have been developed for MCP's in practice. The problem of comparing the means of two Multivariate Normal (MVN) populations, assuming a common covariance matrix $\boldsymbol{\Sigma}$, can easily be extended to the case of comparing K normal populations when there are n_g independent p-dimensional observations from the g-th population.

Following Seber ([35], p. 433), we now recapitulate the formulation of MCP's in the multivariate case.

Let y_{gi} be the ith sample observation $(i=1, 2, \cdots, n_g)$ from the ith MVN distribution $N_p(\mu_g, \boldsymbol{\Sigma})(g=1, 2, \cdots, K)$ so that we have the following MANOVA model for comparing g population mean vectors.

$$y_{gi}=\mu_g+\varepsilon_{gi} \qquad (g=1, 2, \cdots, K;\ i=1, 2, \cdots, n_g) \tag{2.2}$$

where the ε_{gi} are i.i.d. $N_p(0, \boldsymbol{\Sigma})$. Then

$$\begin{pmatrix} y'_{11} \\ y'_{12} \\ \vdots \\ y'_{1n_1} \\ \hline \vdots \\ \hline y'_{K1} \\ y'_{K2} \\ \vdots \\ y'_{Kn_K} \end{pmatrix} = \begin{bmatrix} 1_{n_1} & 0 & \cdots 0 \\ 0 & 1_{n_2} & \cdots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \cdots 1_{n_K} \end{bmatrix} \begin{bmatrix} \mu'_1 \\ \mu'_2 \\ \vdots \\ \mu'_K \end{bmatrix} + \begin{pmatrix} \varepsilon'_{11} \\ \varepsilon'_{12} \\ \vdots \\ \varepsilon'_{1n_1} \\ \hline \vdots \\ \hline \varepsilon'_{K1} \\ \varepsilon'_{K2} \\ \vdots \\ \varepsilon'_{Kn_K} \end{pmatrix}$$

$$\boldsymbol{y}=\boldsymbol{XB}+\boldsymbol{E}\,. \tag{2.3}$$

The hypothesis of equal population means, $H_0: \mu_1=\mu_2=\cdots=\mu_K$, can be written in the form

$$\boldsymbol{CB}=\begin{bmatrix} 1 & 0\cdots 0 & -1 \\ 0 & 1\cdots 0 & -1 \\ \vdots & \vdots & \vdots \\ 0 & 0\cdots 0 & -1 \end{bmatrix}\begin{bmatrix} \mu_1' \\ \mu_2' \\ \vdots \\ \mu_K' \end{bmatrix}=\mathbf{0}\,. \tag{2.4}$$

When the hypothesis of equal population means, H_0, is rejected, those means, or linear combinations of them, that led to the rejection of the hypothesis are of interest. We, therefore, use Roy's [31] maximum root test to construct simultaneous intervals for the mean differences since it has the advantage of being linked directly to a set of simultaneous confidence intervals, although it is not as powerful as the likelihood ratio test.

For a general linear model we have

$$\begin{aligned} 1-\alpha &= \Pr\,[\phi_{\max}\leqq\phi_\alpha] \\ &= \Pr\,[a'\boldsymbol{CB}b\in a'\boldsymbol{C}\hat{\boldsymbol{B}}b\pm\{\phi_\alpha a'\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'a\cdot b'\boldsymbol{W}b\}^{1/2} \quad \text{for all } a,\, b]. \end{aligned} \tag{2.5}$$

Thus a set of multiple confidence intervals for all linear combinations is given by

$$a'\boldsymbol{C}\hat{\boldsymbol{B}}b\pm\{\phi_\alpha\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'a\cdot b'\boldsymbol{W}b\}^{1/2}\,, \tag{2.6}$$

and the set has an overall confidence of $100(1-\alpha)\%$, and where $\hat{\boldsymbol{B}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$, and $\boldsymbol{W}$ is the "within-groups" or "within-samples" SSP matrix. Applying this to our one way model in (2.4), we have

$$\boldsymbol{CB}=\begin{bmatrix} (\mu_1-\mu_K)' \\ (\mu_2-\mu_K)' \\ \vdots \\ (\mu_{K-1}-\mu_K)' \end{bmatrix} \tag{2.7}$$

so that

$$\begin{aligned} a'\boldsymbol{CB} &= a_1(\mu_1-\mu_K)'+a_2(\mu_2-\mu_K)'+\cdots+a_{K-1}(\mu_{K-1}-\mu_K)' \\ &= \sum_{g=1}^{K} c_g\mu_g'\,, \end{aligned}$$

where $c_g=a_g$ $(g=1, 2, \cdots, K-1)$, $c_K=-\sum_{g=1}^{K-1}a_g$ and $\sum_{g=1}^{K}c_g=0$. We therefore find the class of all contrasts of the μ_g. Furthermore, since in the one-way classification model $\bar{y}_{g\cdot}$ is the least squares estimate of μ_g,

$$a'\boldsymbol{C}\hat{\boldsymbol{B}}=\sum_{g=1}^{K}c_g\bar{y}_{g\cdot}\,,$$

and, when $p=1$

$$\mathrm{Var}\,[\sum_g c_g\bar{y}_{g\cdot}]=\sigma^2\sum_g\frac{c_g^2}{n_g}\,.$$

Hence, the probability is $1-\alpha$ that [with $\phi_\alpha=\theta_\alpha/(1-\theta_\alpha)$]

$$\sum_g c_g\mu_g'b\in\sum_g c_g\bar{y}_{g\cdot}'b\pm\left\{\phi_\alpha\left[\sum_g\frac{c_g^2}{n_g}\right]b'\boldsymbol{W}b\right\}^{1/2} \tag{2.8}$$

simultaneously for all contrasts and all b.

In general, we would be interested in pairwise contrasts $\mu_r-\mu_s$ and the correspond-

ing subset of (2.8), namely,

$$(\mu_r-\mu_s)'b\in(\bar{y}_{r.}-\bar{y}_{s.})'b\pm\left\{\phi_\alpha\left(\frac{1}{n_r}+\frac{1}{n_s}\right)b'\boldsymbol{W}b\right\}^{1/2}. \tag{2.9}$$

If the maximum root test of $H_0: \boldsymbol{CB}=\boldsymbol{0}$ is significant, then at least one of the intervals given by (2.8) does not contain zero and we can use the simultaneous intervals to look at any contrast suggested by the data.

Krishnaiah ([20], [21], [22]), based on a multivariate analogue of Tukey's Studentized range, proposed a set of simultaneous intervals for all linear combinations $(\mu_r-\mu_s)'b$. Writing $H_0: \mu_1=\mu_2=\cdots=\mu_K$ as

$$H_0=\bigcap_{\substack{r\\ r<s}}\bigcap_{s} H_{ors}$$

where $H_{ors}: \mu_r-\mu_s=0$, we can test each of the $I=\begin{bmatrix}K\\2\end{bmatrix}=K(K-1)/2$ hypotheses H_{ors} using a Hotelling's T^2 statistic, T_k^2, say, based on a pooled estimate $\boldsymbol{S}_w=\boldsymbol{W}/\nu$ of Σ where $\nu=\sum_g(n_g-1)=n-K$. We can test H_0 using

$$T^2_{\max}=\max_{1\leq k\leq I} T_k^2.$$

If c_α satisfies $\Pr[T^2_{\max}\leqq c_\alpha|H_0]=1-\alpha$, then $\Pr[T_k^2\leqq\alpha,\ k=1, 2, \cdots, I|H_0]=1-\alpha$, and the probability is $1-\alpha$ that

$$(\mu_r-\mu_s)'b\in(\bar{y}_{r.}-\bar{y}_{s.})'b\pm\left\{c_\alpha\left(\frac{1}{n_r}+\frac{1}{n_s}\right)b'\boldsymbol{W}b\right\}^{1/2}$$

simultaneously for all r, s $(r\neq s)$ and for all b. These intervals are the same as (2.9), except that ϕ_α is replaced by c_α. Since the intervals of (2.9) are a subset of (2.8), the overall probability exceeds $1-\alpha$ and $\phi_\alpha>c_{\alpha/\nu}$. Unfortunately, extensive tables of c_α are not available.

If we are interested in just a certain number, say, m, of the elements of $\boldsymbol{B}$, we can use the Bonferroni method of constructing m conservative confidence intervals with an overall confidence of at least $100(1-\alpha)\%$, and a t-value $t_\nu^{\alpha/2m}$.

For more details on MCP's, we refer the reader to Duncan [14], Gabriel ([16], [17]), Miller ([25], [26]), O'Neill and Wetherill [30], Thomas [40], Spjøtvoll [38], Seber [35], and many other authors since the literature is quite rich in this area.

2.2 Problems with MCP's

While many MCP's have been proposed in many different papers in the univariate case, including the ones referred to as above, unfortunately, there are still some serious drawbacks of these procedures, and there are a few MCP's available in practice in the multivariate case which are operational.

The major problems with MCP's in general can briefly be summarized as follows.

(i) MCP's either reject or accept the hypothesis of homogeneity, that is equality of means, or equivalently, an MCP declares each set of means as heterogeneous (rejected) or homogeneous (accepted). MCP decision rule is not transitive; i. e., model M_1 may be preferred to M_2, M_2 to M_3, and M_3 to M_1, etc.

(ii) The decision to accept or reject a model depends on a given significance level

α to maximize the power of the test. In an MCP, it is not clear how the level of the test should be defined, and it is not clear how to control the overall error rate. Also, it is not clear what should be optimized.

(iii) Running all $\begin{bmatrix} K \\ 2 \end{bmatrix} = K(K-1)/2$ pairwise MCP's increases the number of null hypothesis to be tested, and more likely we would reject one of them if all the null hypothesis are actually true, and thus increasing the probability of incorrectly rejecting *at least one* H_0.

(iv) Existing MCP's in general are all devised to handle pairwise comparisons. They need to be extended to handle all $\begin{bmatrix} K \\ k \end{bmatrix}$ k-subsets of a K-set hypothesis.

(v) In MCP's, arbitrary assumptions are made on the parameters of the models. For example, in the formulation of MCP's in Section 2.1, for the MANOVA model we assumed a common dispersion matrix $\boldsymbol{\Sigma}$. For unequal $\boldsymbol{\Sigma}_g$, i.e., for covariance heterogeneity, Olson [29] from a large scale simulation study, reported a high inflation in Type I error and excessive rejections of H_0 on the basis of Roy's maximum root test, $\phi_{\max}$. In this respect the other two statistics Wilks' Λ and Hotelling's T_g^2 behave like $\phi_{\max}$. Therefore, it might be expected that in the presence of covariance heterogeneity, MCP's might also give erroneous results.

(vi) In the multivariate literature there does not exist any simple MCP to handle the case where both mean vectors and the covariance matrices in a model might vary and we still want to carry out comparative simultaneous inference. The same holds in the case of complete homogeneity, that is, when data are assumed to have come from identical populations and we still want to carry out comparative simultaneous inference.

Clearly, there are many problems connected with the existing MCP's. Therefore, for this reason, in the next section, we shall introduce and utilize a general methodology called Multi-Sample Cluster Analysis (MSCA) as an alternative to Multiple Comparison Procedures (MCP's). MSCA depends on fast and efficient combinatorial algorithms. The analysis is done under the best fitting model, and therefore, no arbitrary assumptions are made on the parameters of the model. The only assumption made is the multivariate normality on the data, which can be tested by using the multivariate measures of *skewness* and *kurtosis* (Mardia et al. [24]).

3. Multi-Sample Cluster Analysis as an Alternative to Multiple Comparison Procedures

The problem of MCP's can be viewed as one of *clustering* of means, groups, samples, or treatments. The possibility of using cluster analysis in place of an MCP appears to be originally suggested by Plackett in his discussion of the review paper by O'Neill and Wetherill [30].

In the literature, Scott and Knott [34] used a cluster analysis approach for grouping means; Cox and Spjøtvoll [13] used simple partitioning of means into groups based on the standard F statistic, to mention a few. Their procedures in the spirit are similar to ours, but in general our method is completely different and new. Therefore, here we shall propose MSCA or what Gower [18] calls it, "K-Group Classification" or equiv-

alently what we also call "K-Sample Cluster Analysis" as an alternative to MCP's.

Next, we discuss the general MSCA problem.

3.1 The Multi-Sample Cluster (MSC) problem

The problem of Multi-Sample Cluster Analysis (MSCA) arises when we are given a collection of *groups, profiles, samples, treatments,* etc., whether these are formed naturally or experimentally, and our goal is to cluster these into homogeneous groups. Thus the problem here is to cluster "groups" or "samples" rather than "individuals" or "objects" as in the single-sample case.

Suppose each individual, object, or case, has been measured on p response or outcome measures (dependent variables) simultaneously in K independent groups or samples (factor levels). Let

$$\boldsymbol{X}(n\times p)=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \vdots\\ \boldsymbol{X}_K\end{bmatrix}\begin{matrix}(n_1\times p)\\ (n_2\times p)\\ \vdots\\ (n_K\times p)\end{matrix} \tag{3.1}$$

be a single data matrix of K groups or samples, where $\boldsymbol{X}_g(n_g\times p)$ represents the observations from the gth group or sample, $g=1, 2, \cdots, K$, and $n=\sum_{g=1}^{K} n_g$. The goal of cluster analysis is to put the K groups or samples into k homogeneous groups, samples, or classes where k is unknown and varying, but $k\leqq K$.

Thus we obtain a smallest number k such that the data are consistent with K groups, and from a robustness viewpoint of test statistics, K should generally be as small as possible. Surprisingly, robustness properties do not always improve with increasing group size: small groups are preferred to achieve a parsimonious grouping of samples, and in general to reduce the dimensionality of multi-sample data sets.

We generate all possible clustering alternatives of groups or samples on the computer using efficient combinatorial algorithms and we assemble information from all the different groupings of size K without forcing an arbitrary choice among the clustering alternatives.

We next discuss how to obtain the total number of clustering alternatives for a given K, the number of groups or samples.

3.2 Determining the number of clustering alternatives

Let K be the number of samples, and let k be the number of clusters of samples. If we use MCP's, and if all *pairwise* comparisons among the K groups were desired, then this would require in general ${}_KC_2=\begin{bmatrix}K\\2\end{bmatrix}=K(K-1)/2$ tests. On the other hand, if we consider the combinations of K groups or samples taken k at a time, where $k\leqq K$, then there are $\begin{bmatrix}K\\k\end{bmatrix}$ k-subsets of a K-set altogether. In Appendix A.1, we give a simple algorithm called MCP which constructs all the possible alternatives sequentially in "lexicographic", i.e., in "alphabetical order". A listing of the output from the subroutine of MCP is shown in Table 3.1.

We note that the existing conventional MCP's are not devised to handle the case of $\begin{bmatrix}K\\k\end{bmatrix}$, i.e., k-subsets of a K-set hypotheses or tests for $k>2$. They all need to be

Table 3.1 A Simple Pattern of Clustering Alternatives of Multiple Comparison for Different Combinations of K Samples Taken k at a Time

Combinations	Alternatives	Clustering 2-Subsets	3-Subsets	4-Subsets
$\begin{bmatrix}3\\2\end{bmatrix}$	1	(1, 2)		
	2	(1, 3)		
	3	(2, 3)		
$\begin{bmatrix}4\\2\end{bmatrix}, \begin{bmatrix}4\\3\end{bmatrix}$	1	(1, 2)	(1, 2, 3)	
	2	(1, 3)	(1, 2, 4)	
	3	(1, 4)	(1, 3, 4)	
	4	(2, 3)	(2, 3, 4)	
	5	(2, 4)		
	6	(3, 4)		
$\begin{bmatrix}5\\2\end{bmatrix}, \begin{bmatrix}5\\3\end{bmatrix}, \begin{bmatrix}5\\4\end{bmatrix}$	1	(1, 2)	(1, 2, 3)	(1, 2, 3, 4)
	2	(1, 3)	(1, 2, 4)	(1, 2, 3, 5)
	3	(1, 4)	(1, 2, 5)	(1, 2, 4, 5)
	4	(1, 5)	(1, 3, 4)	(1, 3, 4, 5)
	5	(2, 3)	(1, 3, 5)	(2, 3, 4, 5)
	6	(2, 4)	(1, 4, 5)	
	7	(2, 5)	(2, 3, 4)	
	8	(3, 4)	(2, 3, 5)	
	9	(3, 5)	(2, 4, 5)	
	10	(4, 5)	(3, 4, 5)	

modified accordingly.

If we use the complete enumeration technique, then the number of clustering of K groups or samples into k nonempty clusters of samples is given by the following theorem.

THEOREM 3.2.1. *The number of ways of clustering K samples or groups into k-sample clusters where $k \leqq K$ such that none of the k-sample clusters is empty is given by*

$$\sum_{g=0}^{k}(-1)^g\begin{bmatrix}k\\g\end{bmatrix}(k-g)^K, \tag{3.2}$$

when the order of samples (or groups) within each cluster is irrelevant.

PROOF. Duran and Odell ([15], p. 26).

In this theorem the k-sample clusters are assumed to be distinct. However, in clustering or partitioning K samples into k subsets, none of which is empty, the order of k-sample clusters or k-subsets is irrelevant. Consequently, from this fact and Theorem 3.2.1, it follows that the *total number of ways of clustering K samples into k-sample clusters (or subsets)* is given by

$$w=S(K,\ k)=\frac{1}{k\,!}\sum_{g=0}^{k}(-1)^g\begin{bmatrix}k\\g\end{bmatrix}(k-g)^K \tag{3.3}$$

which is known as the Stirling Number of the Second Kind, which gives us the *number of clustering alternatives.*

If k, the number of clusters of samples is known in advance, then the total number of

clustering alternatives is given by $S(K, k)$. However, if k is not specified a priori and is unknown, but $k \leqq K$, then the total number of clustering alternatives is given by

$$\sum_{k=1}^{K} S(K, k). \tag{3.4}$$

$S(K, k)$ can be written in terms of the recursive formula

$$\begin{aligned} S(K, k) &= kS(K-1, k)+S(K-1, k-1) \quad \text{with } S(1, 1)=1 \\ &\text{and } S(1, k)=0 \text{ for } k \neq 1, \text{ and } S(K, 2)=2^{K-1}-1. \end{aligned} \tag{3.5}$$

For detailed explanations and proofs, see, e.g., Duran and Odell [15], and Späth [37]. Table 3.2 gives $S(K, k)$ for values of K and k up to 10 which is generated from the subroutine STIRN2 in Appendix A.2. This subroutine constructs a table of total number of clustering alternatives for various values of K, number of samples, and k varying number of clusters of samples.

Table 3.2 Number of Clustering Alternatives for Various Values of K and k

K \ k	1	2	3	4	5	6	7	8	9	10	Total
1	1										1
2	1	1									2
3	1	3	1								5
4	1	7	6	1							15
5	1	15	25	10	1						52
6	1	31	90	65	15	1					203
7	1	63	301	350	140	21	1				877
8	1	127	966	1701	1050	266	28	1			4140
9	1	255	3025	7770	6951	2646	462	36	1		21147
10	1	511	9330	34105	42525	22827	5880	750	45	1	115975

Consider, for example, $K=4$ samples. We now wish to cluster $K=4$ groups or samples first into $k=1$ group or sample, $k=2$ groups or samples, $k=3$ groups or samples, and $k=4$ groups or samples in a hierarchical fashion. In order to be able to generate all possible clustering alternatives, we utilize Table 3.2. We have the total number of ways of clustering $K=4$ groups or samples into $k=1$ homogeneous group or sample is 1. The total number of ways of clustering $K=4$ groups or samples into $k=2$ homogeneous groups or samples is 7. The total number of ways of clustering $K=4$ groups or samples into $k=3$ homogeneous groups or samples is 6, and finally, the total number of ways of clustering $K=4$ groups or samples into $k=4$ homogeneous groups or samples is 1. Thus adding up these results, we obtain, in total 15 clustering alternatives as the total for $K=4$ groups or samples into $k=1$, 2, 3, and 4 homogeneous groups. We note that 15 is nothing but the sum of the values of row 4 in Table 3.2.

In general, clustering alternatives can be classified according to their *representation forms* to make it easy to list all possible clustering alternatives. The subroutine REPFM in Appendix A.3 gives the partition of K (number of samples) which is a positive integer, into a specified k number of parts. For example, the representation forms of $K=4$ groups or samples into $k=1$, 2, 3, and 4 groups or samples are:

$$
\begin{aligned}
4 &= \{4\} \\
&= \{3\}+\{1\} \\
&= \{2\}+\{2\} \\
&= \{2\}+\{1\}+\{1\} \\
&= \{1\}+\{1\}+\{1\}+\{1\},
\end{aligned}
$$

where each of the components in a representation $\{g\}$ denotes the number g, of groups or samples in the corresponding cluster. The components of a representation form will always be written in a hierarchical order to depict the patterns of clustering alternatives. In our example there are 15 clustering alternatives but only 5 representation forms. In general the number of representation forms is much smaller than the number of clustering alternatives.

To generate and list the clustering alternatives corresponding to their representation forms, we use the subroutine ALLSUB given in Appendix A.4. This subroutine generates and lists all the simple patterns of clustering alternatives for a specified number of samples K for Multi-Sample Cluster Analysis (MSCA). For example, Table

Table 3.3 A Simple Pattern of Clustering Alternatives of Multi-Sample Cluster Analysis of K Samples into k Varying Number of Clusters of Samples

No. of Clustering Alternatives	Alternatives	Clustering	k
$\sum_{k=1}^{K=3} S(3, k)=5$	1	(1, 2, 3)	1
	2	(1, 2) (3)	2
	3	(1, 3) (2)	2
	4	(2, 3) (1)	2
	5	(1) (2) (3)	3
$\sum_{k=1}^{K=4} S(4, k)=15$	1	(1, 2, 3, 4)	1
	2	(2, 3, 4) (1)	2
	3	(1, 3, 4) (2)	2
	4	(1, 2, 4) (3)	2
	5	(1, 2, 3) (4)	2
	6	(1, 4) (2, 3)	2
	7	(1, 3) (2, 4)	2
	8	(1, 2) (3, 4)	2
	9	(3, 4) (1) (2)	3
	10	(2, 4) (1) (3)	3
	11	(2, 3) (1) (4)	3
	12	(1, 4) (2) (3)	3
	13	(1, 3) (2) (4)	3
	14	(1, 2) (3) (4)	3
	15	(1) (2) (3) (4)	4

3.3 gives a simple pattern of clustering alternatives when $K=3$ and $K=4$ groups or samples, and we wish to cluster them into $k=1, 2, 3$ and $k=1, 2, 3$, and 4 homogeneous groups, respectively.

Looking at Table 3.3 for $K=4$ groups or samples, we see that, in alternative one, the group or sample 1, 2, 3 and 4 are all clustered together. In terms of a hypothesis on means, this corresponds to $\mu_1=\mu_2=\mu_3=\mu_4$, all being equal. Hence, indicating that group 1, 2, 3, and 4 are all homogeneous or identical. On the other hand, in alternative fifteen, the group or sample 1, 2, 3, and 4 are clustered as singletons. In terms of hypothesis on means, this corresponds to μ_1, μ_2, μ_3, and μ_4 all being different, and therefore, we have 4-sample clusters. Hence, indicating that groups 1, 2, 3, and 4 are all heterogeneous. In a similar fashion, we interpret the other clustering alternatives continuing down the line of Table 3.3.

In concluding this section, we see that in general the total number of ways of clustering K groups or samples into k homogeneous groups or samples is given by equation (3.3), and the total number of possible clustering alternatives is given by the expression (3.4). Furthermore, the listings of the necessary combinatorial subroutines are presented in the Appendix.

Having discussed how to determine the number of clustering alternatives, we might ask more questions as follows which need to be answered.

(i) How do we identify the best fitting or approximating model?
(ii) Which clustering alternative do we choose?
(iii) Is it fair to compare different models at the same risk level?
(iv) Should we assume common or varying variance-covariance matrices in clustering samples?
(v) How do we interpret the results?, and so on.

3.3 Splitting algorithm for Multi-Sample Cluster Analysis (MSCA)

When the cardinality of samples to be clustered is more than $K=10$ groups or samples, to save computer time and cost of computation, we use the following *Splitting Algorithm* to search for an optimal clustering alternative for $k=1, 2, \cdots, 10$, groups or samples, stage-wise.

STAGE-1: Start with $k=1$-Sample Cluster, that is, when all the groups or samples are all together in their own cluster, and compute the AIC and CAIC.

STAGE-2: K-Samples in the root node is split into $k=2$-Sample Clusters by using the Stirling Number of the Second Kind (STIRN2) subroutine. The AIC's and CAIC's are computed for all the clustering alternatives and the best clustering alternative is chosen by the minimum value of AIC or CAIC to be split next.

STAGE-3: The best clustering alternative in STAGE-2 based on the value of the criteria is now split into $k=3$-Sample Clusters by STIRN2, and the AIC's and CAIC's are computed to choose the best $k=3$-Sample Clusters.

STAGE-4: The process in STAGE-3 is repeated until all the groups are clustered in their own singleton clusters.

In this manner, the Splitting Algorithm moves from one optimal stage to the next

instead of generating all possible clustering alternatives at once, and then searching for the best clustering alternative as k (number of clusters of samples) varies. This requires enormous storage space on the computer and it is very prohibitive, but nevertheless, is not impossible to do. Our approach is very effective in the sense that it is more advantageous over the Binary Splitting Algorithms used in the literature since one can see and construct the stage-wise optimal decision trees as one walks through the algorithm.

In the next section, Section 4, we shall present our proposed new approach, namely model-selection criteria such as Akaike's Information Criterion (AIC) and CAIC, as new procedures for comparisons and identification of groups or samples under three different but linked multivariate models, and give also their AIC-replacements.

4. Model-Selection Criteria and Multivariate Models

4.1 Model-selection criteria

The "classical" or "conventional" approach to the model selection problem has its basic roots in statistical hypothesis testing problems. Hypothesis testing problems are always based on the assumption that available data are actually generated from one type of model with a known structure, and the goal is to select this model by analyzing the given data set.

On the contrary, in recent years, the literature has placed more and more emphasis on model selection criteria or procedures. The necessity of introducing the concept of *model selection* or *model identification* has been recognized and the problem is posed on the choice of the "best" approximating model among a class of competing models by a suitable model selection criteria given a data set. Model selection criteria are figures of merit for competing models. That model, which optimizes the criterion, is chosen to be the best model.

Suppose there are K alternative models M_k, $k=1, 2, \cdots, K$, represented by the densities $f_1(\cdot\,|\,\theta_1), f_2(\cdot\,|\,\theta_2), \cdots, f_K(\cdot\,|\,\theta_K)$ for the explanation of a random vector X and given n observations, and for the identification, comparison, and the choice of the models $\{M_k : k \in K\}$ with different number of parameters. Akaike, in his pioneering work in a very important sequence of papers, including Akaike ([1], [2], [3], [5]), developed a model selection criterion for the identification of an optimal and a parsimonious model in data analysis from a class of models, which takes model complexity into account. His approach is based on the Kullback-Liebler Information (KLIC) and the asymptotic properties of maximum likelihood ratio statistic. The AIC statistic is an estimator of the risk of a model under the maximum likelihood estimation and it is defined as follows.

DEFINITION 4.1.1. Let $\{M_k : k \in K\}$ be a set of competing models indexed by $k=1, 2, \cdots, K$. Then, the criterion

$$\mathrm{AIC}\,(k) = -2 \log_e L[\hat{\theta}(k)] + 2m(k) \tag{4.1}$$

which is minimized to choose a model M_k over the set of models is called *Akaike's Information Criterion* (AIC).

In (4.1), $L[\theta(k)]$ is the likelihood function of observations, $\hat{\theta}(k)$ is the maximum likelihood estimate of the parameter vector θ under the model M_k, and $m(k)$ is the number of independent parameters estimated when M_k is the model. According to AIC, inclusion of an additional parameter is appropriate if $\log_e L[\hat{\theta}(k)]$ increases by one unit or more, i.e., if $\log_e L[\hat{\theta}(k)]$ increases by a factor of e ($\sim$2.718) or more.

Akaike ([3], [4]), and Schwarz [33] by taking different approaches, developed a new Bayesian model-selection criterion called by the generic name, BIC. More recently, Bozdogan [11], without violating Akaike's main theory, extended AIC in several ways to make AIC asymptotically consistent to penalize overparameterization more stringently to pick only the simplest of the "true" models whenever there is nothing to be lost by doing so. In this case, model selection criterion is called CAIC, a generic name dubbed by Bozdogan [11], and one of its forms is defined as follows.

DEFINITION 4.1.2. Let $\{M_k : k \in K\}$ be a set of competing models indexed by $k = 1, 2, \cdots, K$. Then the criterion

$$\text{CAIC}(k) = -2 \log_e L[\hat{\theta}(k)] + m(k)[\log_e n + 1], \tag{4.2}$$

or simply

$$\text{CAIC}(k) = -2 \log_e L[\hat{\theta}(k)] + m(k) \log_e n \tag{4.3}$$

which is minimized to choose a model M_k over the set of models.

Essentially, the criterion in (4.2) has the same components as Schwarz's Criterion (SC), but CAIC in (4.2) is obtained directly from Akaike's proof of AIC without violating his original set-up in estimating *minus twice the negentropy*, and without resorting to first-order approximations to the posterior probability of M_k, over a set of models which is the case for SC. Therefore, from a theoretical point of view in estimating minus twice the negentropy, AIC and CAIC are entropy based criteria. Schwarz's Criterion (SC) is Bayesian, and it is not entropy based. For more on this, see, e.g., Akaike [5].

We note that for large sample sizes, AIC and CAIC, or SC differ from one another in the manner in which they adjust the usual likelihood ratio statistic, taking into account the difference in dimensionality between the models. Both criteria choose parsimonious model or models, but CAIC for large n, becomes more stringent and favors lower dimensional models and achieves a fully automatic Occam's Razor, that is, choosing the simplest of the "true" models whatever they might be.

In the literature, there exists other Akaike-type model-selection criteria which can be generalized and be put into what we call *Generalized Information Criterion* (GIC) defined by

$$\text{GIC}(k) = -2 \log_e L[\hat{\theta}(k)] + a(n)m(k) + b(k), \tag{4.4}$$

where n is the sample size, $\log_e = \ln$ denotes the natural logarithm, $L[\hat{\theta}(k)]$ denotes the maximum of the likelihood over the parameters, and $m(k)$ is the number of independent parameters in the k-th model. For a given criterion, $a(n)$ is the *cost* of fitting an additional parameter and $b(k)$ is an additional term depending upon the criterion and the model k. For example, *Kashyap's* [19] *Criterion* (KC) falls under the expression for GIC given in (4.4). Kashyap's Criterion (KC) is based on reasoning similar to

BIC and SC, but contains an extra term, and it could be expected to perform better. However, it is not conveniently usable in applications, especially in the type of problems we are looking at in this paper. In KC, the extra term $b(k)=\log_e[\det B(k)]$ where det denotes determinant and $B(k)$ is the negative of the matrix of second partials of $\log_e L[\theta(k)]$, evaluated at the maximum likelihood estimates. Therefore, for our purposes, it is prohibitively expensive to compute KC and its extra term. For this reason, we have chosen only to work with AIC and CAIC which are sufficient for all practical purposes. Hence, we have chosen not to introduce Kashyap's Criterion (KC) here.

We next derive the forms of AIC's only for three linked but different multivariate models for the convenience of the readers. For more details on this, we refer the reader to Bozdogan ([8], [10]), Bozdogan and Sclove [12]. Derivations of CAIC's or SC's follow similarly.

4.2 Multivariate models and their AIC's

Throughout this section we shall suppose that we may have independent data matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_K$, where the rows of $\boldsymbol{X}_g(n_g \times p)$ are independent and identically distributed (i.i.d.) $N_p(\mu_g, \boldsymbol{\Sigma}_g)$, $g=1, 2, \cdots, K$. In terms of the parameters $\theta=(\mu_1, \mu_2, \cdots, \mu_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_K)$ the models we are going to consider are as follows:

(i) $\theta=(\mu_1, \mu_2, \cdots, \mu_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_K)$ $[m=kp+kp(p+1)/2$ parameters$]$

(ii) $\theta=(\mu_1, \mu_2, \cdots, \mu_K, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}, \cdots, \boldsymbol{\Sigma})$ $[m=kp+p(p+1)/2$ parameters$]$

(iii) $\theta=(\mu, \mu, \cdots, \mu, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}, \cdots, \boldsymbol{\Sigma})$ $[m=p+p(p+1)/2$ parameters$]$.

In this section we shall derive the forms of AIC for these models. Recall the definition of AIC in (4.1).

$$\begin{aligned}\mathrm{AIC} &= -2\log_e L(\hat{\theta})+2m \\ &= -2\log_e(\text{maximized likelihood})+2m,\end{aligned}$$

where m denotes the number of free parameters within the model.

4.2.1 AIC for the test of homogeneity of covariances model: AIC$(\{\mu_g, \boldsymbol{\Sigma}_g\})\equiv$AIC (varying μ and $\boldsymbol{\Sigma}$)

Consider K normal populations with different mean vectors μ_g and different covariance matrices $\boldsymbol{\Sigma}_g$, $g=1, 2, \cdots, K$. Let $\boldsymbol{X}_{gi}$, $i=1, 2, \cdots, n_g$, be a random sample of observations from the g-th population $N_p(\mu_g, \boldsymbol{\Sigma}_g)$.

Now, we derive the form of Akaike's Information Criterion (AIC) to test the hypothesis that the covariance matrices of these populations are equal. The likelihood function of all the sample observations is given by

$$L(\mu_g, \boldsymbol{\Sigma}_g; \boldsymbol{X})=\prod_{g=1}^{K} L_g(\mu_g, \boldsymbol{\Sigma}_g; \boldsymbol{X}_g), \tag{4.5}$$

or by

$$\begin{aligned}L=&(2\pi)^{-np/2}\prod_{g=1}^{K}|\Sigma_g|^{-n_g/2} \\ &\times\exp\left\{-1/2\,\mathrm{tr}\sum_{g=1}^{K}\boldsymbol{\Sigma}_g^{-1}\boldsymbol{A}_g-1/2\,\mathrm{tr}\sum_{g=1}^{K}n_g\boldsymbol{\Sigma}_g^{-1}(\bar{X}_g-\mu_g)(\bar{X}_g-\mu_g)'\right\},\end{aligned} \tag{4.6}$$

where $n=\sum_{g=1}^{K} n_g$ and $\boldsymbol{A}_g=\sum_{i=1}^{n_g}(X_{gi}-\bar{X}_g)(X_{gi}-\bar{X}_g)'$.

The log likelihood function is

$$\begin{aligned} l(\mu_g, \boldsymbol{\Sigma}_g; \boldsymbol{X}) &\equiv \log_e L(\mu_g, \boldsymbol{\Sigma}_g; \boldsymbol{X}) \\ &= -(np/2)\log_e(2\pi) - 1/2\sum_{g=1}^{K} n_g \log_e|\boldsymbol{\Sigma}_g| \\ &\quad -1/2\,\mathrm{tr}\sum_{g=1}^{K}\boldsymbol{\Sigma}_g^{-1}\boldsymbol{A}_g - 1/2\,\mathrm{tr}\sum_{g=1}^{K} n_g\boldsymbol{\Sigma}_g^{-1}(\bar{X}_g-\mu_g)(\bar{X}_g-\mu_g)'. \end{aligned} \tag{4.7}$$

The maximum likelihood estimates (MLE's) of μ_g and $\boldsymbol{\Sigma}_g$ are

$$\hat{\mu}_g=\bar{X}_g, \tag{4.8}$$

and

$$\hat{\boldsymbol{\Sigma}}_g=\boldsymbol{A}_g/n_g, \quad g=1, 2, \cdots, K. \tag{4.9}$$

Since the K populations are independent, the likelihood of all the sample observations is simply the product of the separate likelihoods, and so maximizing (4.7) is equivalent to as maximizing the individual sample likelihoods, separately. This, thus, yields the MLE's given in (4.8) and (4.9) above.

Substituting the MLE's into (4.7) and simplifying, the maximized log likelihood becomes

$$\begin{aligned} l(\{\hat{\mu}_g, \hat{\boldsymbol{\Sigma}}_g\}; \boldsymbol{X}) &\equiv \log_e L(\{\hat{\mu}_g, \hat{\boldsymbol{\Sigma}}_g\}; \boldsymbol{X}) \\ &= -(np/2)\log_e(2\pi) - 1/2\sum_{g=1}^{K} n_g \log_e|n_g^{-1}\boldsymbol{A}_g| - (np/2). \end{aligned} \tag{4.10}$$

Since

$$\mathrm{AIC}=-2\log_e L(\hat{\theta})+2m, \tag{4.11}$$

where $m=kp+kp(p+1)/2$ is the number of parameters, the AIC becomes

$$\begin{aligned} \mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}) &\equiv \mathrm{AIC}\ (\text{varying } \mu \text{ and } \boldsymbol{\Sigma}) \\ &= np\log_e(2\pi) + \sum_{g=1}^{K} n_g \log_e|n_g^{-1}\boldsymbol{A}_g| + np + 2[kp + kp(p+1)/2]. \end{aligned} \tag{4.12}$$

Since the constants do not affect the result of comparison of models, we could ignore them and reduce the form of AIC to a much simpler form

$$\begin{aligned} \mathrm{AIC}^*(\{\mu_g, \boldsymbol{\Sigma}_g\}) &\equiv \mathrm{AIC}^*\ (\text{varying } \mu \text{ and } \boldsymbol{\Sigma}) \\ &= \sum_{g=1}^{K} n_g \log_e|\boldsymbol{A}_g| + 2[kp + kp(p+1)/2], \end{aligned} \tag{4.13}$$

where n_g=sample size of group or sample $g=1, 2, \cdots, K$,
$|\boldsymbol{A}_g|$=the determinant of sum of squares and cross-products (SSCP) matrix for group or sample $g=1, 2, \cdots, K$,
k=number of groups or samples compared, and
p=number of variables.

However, for purposes of comparison we retain the constants and use AIC given by (4.12).

4.2.2 AIC for the multivariate analysis of variance (MANOVA) model: AIC($\{\mu_g, \boldsymbol{\Sigma}\}$)≡AIC (varying μ and common $\boldsymbol{\Sigma}$)

Consider in this case, K normal populations with different mean vectors μ_g, $g=1, 2, \cdots, K$, but each population is assumed to have the same covariance matrix. Let X_{gi}, $g=1, 2, \cdots, K$; $i=1, 2, \cdots, n_g$, be a random sample of observations from the g-th population $N_p(\mu_g, \boldsymbol{\Sigma})$.

To derive Akaike's Information Criterion (AIC) in this case, we use the log likelihood function given in (4.7). Since each population is assumed to have the same covariance matrix $\boldsymbol{\Sigma}$, the log likelihood function becomes

$$\begin{aligned} l(\{\mu_g\}, \boldsymbol{\Sigma}; \boldsymbol{X}) &\equiv \log L(\{\mu_g\}, \boldsymbol{\Sigma}; \boldsymbol{X}) \\ &= -(np/2)\log(2\pi) - (n/2)\log|\boldsymbol{\Sigma}| - 1/2\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}\sum_{g=1}^{K}\boldsymbol{A}_g \\ &\quad -1/2\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}\sum_{g=1}^{K} n_g(\bar{x}_g-\mu_g)(\bar{x}_g-\mu_g)', \end{aligned} \tag{4.14}$$

and the maximum-likelihood estimates (MLE's) of μ_g, and $\boldsymbol{\Sigma}$ are

$$\hat{\mu}_g = \bar{X}_g, \quad g=1, 2, \cdots, K, \tag{4.15}$$

and

$$\hat{\boldsymbol{\Sigma}} = n^{-1}\boldsymbol{W}, \tag{4.16}$$

where

$$\boldsymbol{W} = \sum_{g=1}^{K}\boldsymbol{A}_g.$$

Substituting these back into (4.14) and simplifying, the maximized log likelihood becomes

$$\begin{aligned} l(\{\hat{\mu}_g\}, \hat{\boldsymbol{\Sigma}}; \boldsymbol{X}) &\equiv \log L(\{\hat{\mu}_g\}, \hat{\boldsymbol{\Sigma}}; \boldsymbol{X}) \\ &= -(np/2)\log(2\pi) - (n/2)\log|n^{-1}\boldsymbol{W}| - (np/2), \end{aligned} \tag{4.17}$$

where $\boldsymbol{W}$ is the "within-groups" SSP matrix.

Since

$$\mathrm{AIC} = -2\log_e L(\hat{\theta}) + 2m, \tag{4.18}$$

where $m = kp + p(p+1)/2$ is the number of parameters, then AIC becomes

$$\begin{aligned} \mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\}) &\equiv \mathrm{AIC}\ (\text{varying } \mu \text{ and common } \boldsymbol{\Sigma}) \\ &= np\log_e(2\pi) + n\log_e|n^{-1}\boldsymbol{W}| + np + 2\left[kp + \frac{p(p+1)}{2}\right]. \end{aligned} \tag{4.19}$$

Since the constants do not affect the result of comparison of models, we could ignore them and reduce the form of AIC to a much simpler form

$$\begin{aligned} \mathrm{AIC}^*(\{\mu_g, \boldsymbol{\Sigma}\}) &\equiv \mathrm{AIC}^*\ (\text{varying } \mu \text{ and common } \boldsymbol{\Sigma}) \\ &= n\log_e|\boldsymbol{W}| + 2\left[kp + \frac{p(p+1)}{2}\right], \end{aligned} \tag{4.20}$$

where

$n = \sum_{g=1}^{K} n_g$ = the total sample size,

$|\boldsymbol{W}|$ = the determinant of "within-groups" SSP matrix,
k = number of groups or samples compared,
p = number of variables.

However, for purposes of comparison we retain the constants and use AIC given by (4.19).

4.2.3 AIC for the test of complete homogeneity model: AIC$(\{\mu, \boldsymbol{\Sigma}\}) \equiv$ AIC (common μ and $\boldsymbol{\Sigma}$)

Consider again K normal populations with the same mean vector μ and the same covariance matrix $\boldsymbol{\Sigma}$. To derive the form of AIC for the test of complete homogeneity model, we set all μ_g's equal to μ and all the $\boldsymbol{\Sigma}_g$'s equal to $\boldsymbol{\Sigma}$ in (4.5) in Section 4.2.1, and obtain the log likelihood function which is given by

$$\begin{aligned} l &\equiv \log L(\mu, \boldsymbol{\Sigma}; \boldsymbol{X}) \\ &= -(np/2)\log(2\pi) - (n/2)\log|\boldsymbol{\Sigma}| - 1/2\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}(\boldsymbol{W}+\boldsymbol{B}) \\ &\quad -(n/2)\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1}(\bar{X}-\mu)(\bar{X}-\mu)'. \end{aligned} \tag{4.21}$$

The MLE's of μ and $\boldsymbol{\Sigma}$ are

$$\hat{\mu} = \bar{X}, \tag{4.22}$$

and

$$\hat{\boldsymbol{\Sigma}} = 1/n(\boldsymbol{W}+\boldsymbol{B}) = \boldsymbol{T}/n. \tag{4.23}$$

Substituting these back into (4.20), we have the maximized log likelihood

$$\begin{aligned} l(\hat{\mu}, \hat{\boldsymbol{\Sigma}}) &\equiv \log L(\hat{\mu}, \hat{\boldsymbol{\Sigma}}; \boldsymbol{X}) \\ &= -(np/2)\log(2\pi) - (n/2)\log|n^{-1}\boldsymbol{T}| - (np/2). \end{aligned} \tag{4.24}$$

Thus, using the equation of AIC in (4.11) again, where $m = p + p(p+1)/2$ is the number of parameters this time, the AIC becomes

$$\begin{aligned} \mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\}) &\equiv \mathrm{AIC}\ (\text{common } \mu \text{ and } \boldsymbol{\Sigma}) \\ &= np\log_e(2\pi) + n\log_e|n^{-1}\boldsymbol{T}| + np + 2\left[p + \frac{p(p+1)}{2}\right]. \end{aligned} \tag{4.25}$$

After ignoring the constants, AIC takes the simplified form

$$\begin{aligned} \mathrm{AIC}^*(\{\mu, \boldsymbol{\Sigma}\}) &\equiv \mathrm{AIC}^*\ (\text{common } \mu \text{ and } \boldsymbol{\Sigma}) \\ &= n\log_e|\boldsymbol{T}| + 2\left[p + \frac{p(p+1)}{2}\right], \end{aligned} \tag{4.26}$$

where $|\boldsymbol{T}|$ = the determinant of the "total" SSP matrix. However, for purposes of comparison we retain the constants and use AIC given in (4.25).

4.3 AIC-replacements for multi-sample conventional tests of homogeneity models

In Section 4.2, having derived the exact analytical forms of Akaike's Information Criterion (AIC) for each of the multivariate models, in this section, we shall give the AIC-replacements for the multivariate multi-sample conventional tests of homogeneity models and establish the relationship of AIC-replacements with that of the conventional procedures. For more details on this, we shall refer the reader to Bozdogan [10].

We next state the following very important theorem which we shall utilize in establishing the relationships of the AIC-replacements and the conventional procedures.

THEOREM 4.3.1. *If Ω_1 is a parameter space in R^K, and if Ω_0 is a k-dimensional subspace of Ω_1, then under suitable regularity conditions, for each $\theta \in \Omega_0$, $-2\log\lambda$ has an asymptotic χ^2_{K-k} distribution as $n \to \infty$.*

PROOF. See, for example, Wilks [42], and Silvey ([36], p. 113).

We are now in a position to give the AIC-replacements for the multivariate multi-sample conventional tests of homogeneity models.

4.3.1 AIC-replacement for Box's M for testing homogeneity of covariances

As an alternative to Box's M test for testing the equality of covariance matrices for which extensive tables are not readily available, we may summarize the *condition* for rejecting

$$H_{oa}: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_K. \tag{4.27}$$

against

$$H_{1a}: \text{Not all } K \text{ covariance matrices are equal,}$$

as follows:

RELATION 4.3.1. We *reject H_{oa} (test of homogeneity of covariances)* if

$$\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\}) > \mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}), \tag{4.28}$$

or if

$$\Delta\mathrm{AIC}(H_{ob}; H_{oa}) = \mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\}) - \mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}) > 0 \tag{4.29}$$

iff

$$n\log_e|n^{-1}\boldsymbol{W}| - \sum_{g=1}^{K} n_g \log_e|n_g^{-1}\boldsymbol{A}_g| > (k-1)p(p+1) \tag{4.30}$$

iff

$$-2\log\lambda_{oa} > (k-1)p(p+1), \tag{4.31}$$

where $\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})$ is given in (4.19) and $\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\})$ is given is (4.12), and where $-2\log\lambda_{oa}$ has an asymptotic chi-squared distribution with $1/2(k-1)p(p+1)$ degrees of freedom by Theorem 4.3.1. Using this fact, we establish the following:

RELATION 4.3.2. For comparing *pairs of models,*

$$\chi^2 \cong \mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\}) - \mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}) + 2\left[\frac{1}{2}(k-1)p(p+1)\right], \tag{4.32}$$

where χ^2 is tested as a chi-square with degrees of freedom $d.f. = 1/2(k-1)p(p+1)$.

4.3.2 AIC-replacement for Wilks' Λ criterion for testing the equality of mean vectors

As an alternative to Wilks' Λ Criterion, Bartlett's V statistic, and other conventional procedures for testing the equality of mean vectors given a common covariance matrix between the groups or samples, we may summarize the *condition* for rejecting

$$H_{ob}: \mu_1 = \mu_2 = \cdots = \mu_K \tag{4.33}$$

against

$$H_{1b}: \text{Not all } \mu_K \text{ are equal,}$$

as follows:

RELATION 4.3.3. We *reject H_{ob} (one-way multivariate analysis of variance hypothesis)* if

$$\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})>\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\}), \tag{4.34}$$

or if

$$\Delta\,\mathrm{AIC}(H_{oc}\,;H_{ob})=\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})-\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})>0 \tag{4.35}$$

iff

$$n\log_e|n^{-1}\boldsymbol{T}|-n\log_e|n^{-1}\boldsymbol{W}|>2p(k-1) \tag{4.36}$$

iff

$$-2\log\lambda_{ob}>2p(k-1), \tag{4.37}$$

because this test is done under the assumption of a common $\boldsymbol{\Sigma}$.

$\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})$ is given in (4.19) and $\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})$ is given in (4.25), and where $-2\log\lambda_{ob}$ has an asymptotic chi-squared distribution with $p(k-1)$ degrees of freedom by Theorem 4.3.1. Again, using this fact, we establish the following:

RELATION 4.3.4. For comparing *pairs of models,*

$$\chi^2\cong\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})-\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})+2[p(k-1)], \tag{4.38}$$

where χ^2 is tested as a chi-square with degrees of freedom $d.f.=p(k-1)$.

4.3.3 AIC-replacement for testing complete homogeneity

Combining the results in Section 4.3.1 and 4.3.2, we may summarize the *condition* for rejecting

$$H_{oc}:\ \mu_1=\mu_2=\cdots=\mu_K \quad\text{and}\quad \boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\cdots=\boldsymbol{\Sigma}_K \tag{4.39}$$

against

H_{1c}: Not all mean K vectors and covariance matrices are equal,

as follows:

RELATION 4.3.5. We *reject* H_{oc} (*test of complete homogeneity*) if

$$\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})>\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\}) \tag{4.40}$$

and

$$\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})>\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}),$$

or if

$$\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})>\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}), \tag{4.41}$$

since $\hat{L}_{ob}=\hat{L}_{oc}$. That is, reject H_{oc} if

$$\Delta\,\mathrm{AIC}(H_{oc}\,;H_{oa})=\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})-\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\})>0 \tag{4.42}$$

iff

$$n\log_e|n^{-1}\boldsymbol{T}|-\sum_{g=1}^{K}n_g\log_e|n_g^{-1}\boldsymbol{A}_g|>p(p+3)(k-1) \tag{4.43}$$

iff

$$-2\log\lambda_{oc}>p(p+3)(k-1), \tag{4.44}$$

where $\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\})$ is given in (4.12), $\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})$ is given in (4.19), and $\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})$ is given in (4.25).

We note that $-2\log\lambda_{oc}$ has an asymptotic chi-squared distribution with $1/2p(p+3)(k-1)$ degrees of freedom by Theorem 4.3.1. Thus, we now establish our final relation as follows:

RELATION 4.3.6. For comparing *pairs of models,*

$$\chi^2\cong\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})-\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\})+2\left[\frac{1}{2}p(p+3)(k-1)\right], \tag{4.45}$$

where χ^2 is tested as a chi-square with degrees of freedom $d.f.=1/2p(p+3)(k-1)$.

5. Numerical Examples

In this section we shall give two different numerical examples and study Multi-Sample Cluster Analysis (MSCA) as an alternative to Multiple Comparison Procedures (MCP's). In Section 5.1, we shall study tests of homogeneity from model-selection viewpoint for the *varieties of rice* data set given in Srivastava and Carter ([39], p. 100), where the detailed conventional analysis of this data set is discussed and treated. In Section 5.2, we shall show the application of MSCA in designing optimal decision tree classifiers which are popular in remote sensing technology.

5.1 Multi-sample clustering of varieties of rice

Suppose four *varieties of rice* (see, e.g., Srivastava and Carter [39]), namely variety A, B, C and D are shown in 20 plots, where each variety of rice is assigned at random to five plots. Two variables were measured six weeks after transplanting: x_1, *the height of the plant*, and x_2, *the number of tillers per plant*. Thus for this data set we have $p=2$ characteristics, $n_g=5$, $g=1,2,3,4$, and $n=\sum n_g=20$.

We next study tests of homogeneity for this data set by using our procedure, show step-by-step analysis and compare our results with that of the conventional tests.

(i) Identification of the best fitting parametric model:

We present the summary of the AIC-values under the three parametric multivariate normal models as follows:

$$\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\})\equiv\mathrm{AIC}(\text{varying } \mu \text{ and } \boldsymbol{\Sigma})=186.324 \tag{5.1}$$

$$\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})\equiv\mathrm{AIC}(\text{varying } \mu \text{ and common } \boldsymbol{\Sigma})=178.290 \tag{5.2}$$

$$\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})\equiv\mathrm{AIC}(\text{common } \mu \text{ and } \boldsymbol{\Sigma})=185.440\,. \tag{5.3}$$

The minimum AIC occurs under the MANOVA model in (5.2). Therefore, according to the definition of AIC, the MANOVA model is the best fitting model for the analysis of the varieties of rice data set. In other words, we are accepting the equality of covariance matrices for this data set. In fact, if we perform a conventional multivariate test for the homogeneity of covariance matrices, we obtain Box's $M=7.97272$ or $\chi^2_9=6.173$ with P-value$=.722$ (approximately). Hence, the acceptance of the test of homogeneity of covariance matrices clears the way for a test on the homogeneity of the variety mean vectors which is the MANOVA null hypothesis. As we saw above, the minimum AIC procedure already picked the MANOVA model as the best fitting model for this data set.

(ii) Test of homogeneity of mean vectors:

Having determined the best fitting model, that is, the MANOVA model, we now test the null hypothesis:

$$H_{ob}:\ \mu_{(A)}=\mu_{(B)}=\mu_{(C)}=\mu_{(D)}$$

against the alternative hypothesis which negates H_{ob}. Using Relation 4.3.3, since

$$\mathrm{AIC}(\{\mu, \boldsymbol{\Sigma}\})=185.440>\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\})=178.290\,, \tag{5.4}$$

we reject H_{ob}, and claim that there is a difference in varieties.

(iii) Multiple comparisons under the best fitting model:

Now we need to compare four varieties of rice simultaneously under the best fitting model, that is, under $\{\mu_g, \boldsymbol{\Sigma}\}$, in terms of the parameters. For this we proceed to use AIC$(\{\mu_g, \boldsymbol{\Sigma}\})$ to compare the four varieties of rice *pairwise*. Our results are presented in Table 5.1.

Table 5.1 Pairwise Comparisons of Four Varieties of Rice on All Variables Under the MANOVA Model

Alternative	Varieties	k	$2m$	AIC$(\{\mu_g, \boldsymbol{\Sigma}\})$
1	(A, B)	1	10	164.322*
2	(A, C)	1	10	139.046[a]
3	(A, D)	1	10	153.147
4	(B, C)	1	10	153.705
5	(B, D)	1	10	145.015[b]
6	(C, D)	1	10	146.396

NOTE: $n=20$ observations; $p=2$ variables; $m=kp+p(p+1)/2$ parameters
AIC$(\{\mu_g, \boldsymbol{\Sigma}\}) = np \log_e(2\pi) + n \log_e |n^{-1}\boldsymbol{W}| + np + 2m$.
[a] First minimum AIC; i.e., best homogeneous pair
[b] Second minimum AIC; i.e., second best homogeneous pair
* Indicates that there is a difference between varieties A and B.

Looking at Table 5.1, we see that, using all the variables simultaneously, the first minimum AIC occurs at the alternative submodel 2 where we have (A, C) as one homogeneous pair. Second best homogeneous pair is (B, D). We never choose the pair (A, B), that is, submodel 1, since its AIC value is quite large indicating the inferiority of this submodel, or indicating that there is a difference between varieties A and B, and that they should not be put together as one homogeneous group.

Although the pairwise comparison is the most commonly used Multiple Comparison Procedure (MCP) in the literature, it is not general, and informative. It only considers the variabilities in pairs of groups or samples, and it ignores the variabilities in other groups. Therefore, for this reason, we shall next propose our new methodology, that is, Multi-Sample Cluster Analysis (MSCA), as an alternative to Multiple Comparison Procedures (MCP's).

(iv) Multi-Sample Cluster Analysis (MSCA) of varieties:

We now cluster $K=4$ samples (varieties of rice) into $k=1, 2, 3$, and 4-Sample Clusters on the basis of all the variables, where $p=2$ in this case. We obtain in total fifteen possible clustering alternatives by using STIRN2 subroutine in Appendix A.2. Using a newly developed statistical computer software by this author called AICPARM: A General Purpose Program for Computing AIC's and CAIC's for Univariate and Multivariate Normal Parametric Models, and using the MANOVA model as our best fitting model, we obtained the results given in Table 5.2.

Looking at Table 5.2, we see that, the minimum AIC and CAIC clustering occurs at alternative submodel 7, that is, $k=2$-Sample Clusters is (1, 3) (2, 4)$\equiv(A, C)$ (B, D), indicating that there seems to be two types of varieties of rice rather than four varie-

Table 5.2 Multi-Sample Cluster Analysis of $K=4$ Varieties of Rice into $k=1, 2, 3$, and 4-Sample Clusters.
The AIC's and CAIC's on All Variables

Alternative	Clustering	k	m	AIC($\{\mu_g, \boldsymbol{\Sigma}\}$)	CAIC($\{\mu_g, \boldsymbol{\Sigma}\}$)
1	(1, 2, 3, 4)	1	5	185.440125*	190.418762
2	(2, 3, 4) (1)	2	7	180.937897	187.907990
3	(1, 3, 4) (2)	2	7	183.446991	190.417084
4	(1, 2, 4) (3)	2	7	187.684692	194.654785
5	(1, 2, 3) (4)	2	7	183.596893	190.566986
6	(1, 4) (2, 3)	2	7	185.540253	192.510345
7	(1, 3) (2, 4)	2	7	175.777649*	182.747742*
8	(1, 2) (3, 4)	2	7	188.730988	195.701080
9	(3, 4) (1) (2)	3	9	181.179932	190.141510
10	(2, 4) (1) (3)	3	9	177.366486	186.328064
11	(2, 3) (1) (4)	3	9	180.485565	189.447144
12	(1, 4) (2) (3)	3	9	185.593323	194.554901
13	(1, 3) (2) (4)	3	9	176.836731*	185.798309*
14	(1, 2) (3) (4)	3	9	187.137421	196.098999
15	(1) (2) (3) (4)	4	11	178.289734*	189.242767*

NOTE: $A=1$, $B=2$, $C=3$, and $D=4$; $n=20$ observations
$p=2$ variables; $m=kp+p(p+1)/2$ parameters
$\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}\}) = np \log_e(2\pi) + n \log_e |n^{-1}\boldsymbol{W}| + np + 2m$
$\mathrm{CAIC}(\{\mu_g, \boldsymbol{\Sigma}\}) = n \log_e(2\pi) + n \log_e |n^{-1}\boldsymbol{W}| + np + m \log_e(n)$
* Minimum AIC's and CAIC's for $k=1, 2, 3$ and 4-sample clusters, respectively

ties. The second minimum AIC and CAIC occur at the alternative submodel 13 and at $k=3$-Sample Clusters where we have (1, 3) (2) (4)$\equiv$(A, C) (B) (D) as our clustering, telling us that if we were to cluster any one of the two varieties of rice, we should cluster varieties A and C together as one homogeneous group, and we should cluster varieties B and D completely separately. We note that the larger values of AIC and CAIC are indications of the inferiority of the submodels. Furthermore, we can see the effect of clustering each variety by looking at the differences of AIC's and CAIC's across each clustering alternative. According to AIC and CAIC, the most inferior submodel is 8 where we have (1, 2) (3, 4)$\equiv$(A, B) (C, D) as our clustering.

In comparing our results in Table 5.1 and 5.2, we see that Multi-Sample Cluster Analysis (MSCA) is much more general and informative than the pairwise Multiple Comparison Procedures (MCP's) to be used for simultaneous comparative inference.

(v) Determining the variables contributing most to the differences in varieties:

Since there is heterogeneity in the mean vectors (or locations) of the four varieties of rice, we further proceed on the basis of univariate theory to study the behaviour of the variety data on each of the $p=2$ variables. Our results are given in Table 5.3.

Interpreting the results in Table 5.3, we note that x_2=*number of tillers per plant* shows significant homogeneity across four varieties of rice, and in fact, is the best variable according to the minimum AIC value. The first variable, that is, x_1=*height of plant*, on the basis of the AIC value indicates that there is a difference in heights

Table 5.3 Univariate AIC's on $p=2$ Variables for Four Varieties of Rice

Variables	AIC($\{\mu_g, \sigma^2\}$)
1. Height of Plant	111.70*
2. Number of Tillers Per Plant	65.94

NOTE: $\mathrm{AIC}(\{\mu_g, \sigma^2\}) = n \log_e(2\pi) + n \log_e\left(\frac{SSW}{n}\right) + n + 2(k+1)$

* Indicates that there is a difference in heights between the varieties.

between the varieties. The general conclusion is that there exists more heterogeneity in means on variable x_1 than x_2 across the four types of varieties of rice.

5.2 Application of Multi-Sample Cluster Analysis in designing decision trees in remote sensing

In remote sensing technology, the *decision tree classifier* has been widely used in various problems in *geoscience* and *remote sensing, speech analysis, biomedical applications*, etc., and in many other areas. For more on this, we refer the reader to Argentiero et al. [7], Kulkarni and Kanal [23], Mui and Fu [27], Wang and Suen [41], and others.

Using a *decision tree classifier* over a *single stage classifier*, we have an advantage in the sense that, a *complex global decision* can be made via a series of simple and local decisions. This enables us to use a decision tree classifier in two main types of applications:

(i) recognition of pattern classes, and

(ii) tree classifier can make a decision much more quickly compared to single stage classifier.

For example, in remote sensing problems one is faced with an image (or scene) which is a rectangular array with I-rows (scan lines), and J-columns (the number of resolution elements per scan line of one resolution element (an individual)). Each cell (individual or pixel) generates a $p \times 1$ measurement vector x_{ij}, $i=1, 2, \cdots, I$, and $j=1, 2, \cdots, J$. We denote the features by

$$X_1, X_2, \cdots, X_p.$$

The vector feature is

$$x=(X_1, X_2, \cdots, X_p).$$

The observed digital image is

$$\{x_{ij}: i=1, 2, \cdots, I, j=1, 2, \cdots, J\},$$

where

$$x_{ij}=(x_{1ij}, x_{2ij}, \cdots, x_{pij})$$

is the vector of numerical values of the p features at pixel (i, j). For more on this, see also Sclove [32].

In order to recognize an image (or scene), we need to perform classification, that is, grouping of pixels, to check the homogeneity of large dimensional Multispectral Scanner (MSS) data sets with a view toward identifying objects, and recognizing the pattern classes, and so forth. This is the major task of cluster analysis techniques.

After the features are extracted, a decision rule is then applied to assign the

reduced dimensional samples to available classes by merging and subjecting these samples to a sequence of decision rules before they are assigned to a unique class. Such an approach further reduces the dimensionality of these large dimensional data sets, and it results in an optimal decision tree classifier which is computationally efficient, accurate, and flexible. Argentiero et al. [7], give an example on how to design an optimal decision tree classifier by using a conventional statistical procedure, namely the multivariate F-test, and give the associated table look-up decision tree classifier on a simulated heterogeneous multivariate data set where both the mean vectors and the covariance matrices among the five classes are varying. It seems that such an approach is primitive and the decision rule at each stage depends upon a given significance level α. Also it is not clear how they controlled the overall error rate in their study. We cannot simply use the usual F-tables in the presence of covariance heterogeneity without testing the equality of covariance matrices.

To provide an example of Multi-Sample Cluster Analysis (MSCA) for the classification of large dimensional data sets arising from the merging of remote sensing data, we reconstructed the data structure presented in Argentiero et al. [7] with different sample sizes. That is, we simulated 100 different $p=4$ variate multivariate normal samples from the $K=5$ classes using the IMSL procedure GGNSM. The simulated data was based on the class statistics given in Table 5.4 which were obtained from a Landsat-2 satellite over a midwestern county. The five classes were consisted of two

Table 5.4 The Class Statistics of Landsat-2 Multispectral Scanner (MSS) Signatures

Class Type	Channel	Mean Vector	Covariance Matrix
(1) Non-Wheat $n_1=50$	1 2 3 4	$\mu_1=\begin{bmatrix}27.7\\24.5\\75.1\\37.4\end{bmatrix}$	$\Sigma_1=\begin{bmatrix}12.7 & 25.0 & -51.4 & -30.8\\25.0 & 63.4 & -140.7 & -84.2\\-51.4 & -140.7 & 415.5 & 242.1\\-30.8 & -84.2 & 242.1 & 143.4\end{bmatrix}$
(2) Non-Wheat $n_2=75$	1 2 3 4	$\mu_2=\begin{bmatrix}34.7\\40.4\\47.0\\19.7\end{bmatrix}$	$\Sigma_2=\begin{bmatrix}12.7 & 17.2 & 8.8 & 0.6\\17.2 & 30.0 & 9.9 & -1.2\\8.8 & 9.9 & 27.3 & 10.4\\0.6 & -1.2 & 10.4 & 6.0\end{bmatrix}$
(3) Non-Wheat $n_3=100$	1 2 3 4	$\mu_3=\begin{bmatrix}33.3\\38.5\\44.1\\18.7\end{bmatrix}$	$\Sigma_3=\begin{bmatrix}2.6 & 2.6 & 4.3 & 1.9\\2.6 & 7.2 & 2.5 & 0.3\\4.3 & 2.5 & 41.2 & 19.9\\1.9 & 0.3 & 19.9 & 11.1\end{bmatrix}$
(4) Winter Wheat $n_4=125$	1 2 3 4	$\mu_4=\begin{bmatrix}28.5\\27.5\\51.2\\24.0\end{bmatrix}$	$\Sigma_4=\begin{bmatrix}5.8 & 7.4 & -6.0 & -4.3\\7.4 & 16.2 & -14.4 & -8.9\\-6.7 & -14.4 & 26.7 & 14.1\\-4.3 & -8.9 & 14.1 & 9.0\end{bmatrix}$
(5) Winter Wheat $n_5=150$	1 2 3 4	$\mu_5=\begin{bmatrix}21.5\\16.7\\54.9\\29.1\end{bmatrix}$	$\Sigma_5=\begin{bmatrix}7.3 & 10.3 & 4.1 & -1.0\\10.3 & 18.0 & 4.9 & -2.8\\4.1 & 4.9 & 26.0 & 11.4\\-1.0 & -2.8 & 11.4 & 8.1\end{bmatrix}$

Table 5.5 Multi-Sample Cluster Analysis of $K=5$ Simulated Class Types of Different Crops into $k=1, 2, 3, 4$, and 5-Sample Clusters, The AIC's and CAIC's on All Variables

Alternative	Clustering	k	m	AIC($\{\mu_g, \boldsymbol{\Sigma}_g\}$)	CAIC($\{\mu_g, \boldsymbol{\Sigma}_g\}$)
1	(1, 2, 3, 4, 5)	1	14	11650.509766*	11709.513672*
2	(1, 2, 3, 4) (5)	2	28	10434.767578*	10552.775391*
3	(1, 2, 3, 5) (4)	2	28	11100.183594	11218.191406
4	(1, 2, 4, 5) (3)	2	28	11102.822266	11220.830078
5	(1, 3, 4, 5) (2)	2	28	11223.681641	11341.689453
6	(1) (2, 3, 4, 5)	2	28	10954.841797	11072.849609
7	(1, 2, 3) (4, 5)	2	28	10753.361328	10871.369141
8	(1, 2, 4) (3, 5)	2	28	11195.396484	11313.404297
9	(1, 2, 5) (3, 4)	2	28	11245.294922	11363.302734
10	(1, 3, 4) (2, 5)	2	28	11121.404297	11239.412109
11	(1, 3, 5) (2, 4)	2	28	11340.630859	11458.638672
12	(1, 4, 5) (2, 3)	2	28	10475.867188	10593.875000
13	(1, 5) (2, 3, 4)	2	28	10612.294922	10730.302734
14	(1, 4) (2, 3, 5)	2	28	10763.261719	10881.269531
15	(1, 3) (2, 4, 5)	2	28	11045.414063	11163.421875
16	(1, 2) (3, 4, 5)	2	28	11122.408203	11240.416016
17	(1) (2, 5) (3, 4)	3	42	10669.160156	10846.171875
18	(1) (2, 4) (3, 5)	3	42	10737.947266	10914.958984
19	(1) (2, 3) (4, 5)	3	42	9931.666016	10108.677734
20	(1, 5) (2) (3, 4)	3	42	10321.367188	10498.378906
21	(1, 5) (2, 4) (3)	3	42	10301.919922	10478.931641
22	(1, 5) (2, 3) (4)	3	42	9684.330078	9861.341797
23	(1, 4) (2) (3, 5)	3	42	10377.035156	10554.046875
24	(1, 4) (2, 5) (3)	3	42	10288.802734	10465.814453
25	(1, 4) (2, 3) (5)	3	42	9378.460938*	9455.472656*
26	(1, 3) (2) (4, 5)	3	42	10464.267578	10641.279297
27	(1, 3) (2, 5) (4)	3	42	10564.726563	10741.738281
28	(1, 3) (2, 4) (5)	3	42	10171.974609	10348.986328
29	(1, 2) (4, 5) (3)	3	42	10418.652344	10595.664063
30	(1, 2) (3, 5) (4)	3	42	10607.343750	10784.355469
31	(1, 2) (3, 4) (5)	3	42	10145.806641	10322.818359
32	(1, 2, 3) (4) (5)	3	42	9788.210938	9965.222656
33	(1, 2, 4) (3) (5)	3	42	10041.552734	10218.564453
34	(1, 2, 5) (3) (4)	3	42	10552.988281	10730.000000
35	(1, 3, 4) (2) (5)	3	42	10055.792969	10232.804688
36	(1, 3, 5) (2) (4)	3	42	10667.771484	10844.783203
37	(1, 4, 5) (2) (3)	3	42	10420.597656	10597.609375
38	(1) (2, 3, 4) (5)	3	42	9894.478516	10071.490234
39	(1) (2, 3, 5) (4)	3	42	10451.314453	10628.326172
40	(1) (2, 4, 5) (3)	3	42	10457.542969	10634.554688
41	(1) (2) (3, 4, 5)	3	42	10580.152344	10757.164063
42	(1) (2) (3) (4, 5)	4	56	9876.396484	10112.414063

Alternative	Clustering	k	m	AIC$(\{\mu_g, \boldsymbol{\Sigma}_g\})$	CAIC$(\{\mu_g, \boldsymbol{\Sigma}_g\})$
43	(1) (2) (3, 5) (4)	4	56	10065.087891	10301.105469
44	(1) (2) (3, 4) (5)	4	56	9603.550781	9839.568359
45	(1) (2, 5) (3) (4)	4	56	9976.855469	10212.873047
46	(1) (2, 4) (3) (5)	4	56	9584.103516	9820.121094
47	(1) (2, 3) (4) (5)	4	56	8966.513672*	9202.531250*
48	(1, 5) (2) (3) (4)	4	56	9629.060547	9865.078125
49	(1, 4) (2) (3) (5)	4	56	9223.191406	9459.208984
50	(1, 3) (2) (4) (5)	4	56	9499.115234	9735.132813
51	(1, 2) (3) (4) (5)	4	56	9453.501953	9689.519531
52	(1) (2) (3) (4) (5)	5	70	8911.246094*	9206.267578*

NOTE: $n=500$ total number of observations; $p=4$ variables; $m=kp+kp(p+1)/2$ parameters;

$$\mathrm{AIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}) = np \log_e(2\pi) + \sum_{g=1}^{k} n_g \log_e |n_g^{-1}\boldsymbol{A}_g| + np + 2m$$

$$\mathrm{CAIC}(\{\mu_g, \boldsymbol{\Sigma}_g\}) = np \log_e(2\pi) + \sum_{g=1}^{k} n_g \log_e |n_g^{-1}\boldsymbol{A}_g| + np + m \log_e(n)$$

* Minimum AIC's and CAIC's for $k=1, 2, 3, 4$ and 5-sample clusters, respectively.

types of winter wheat and the three confusion crops, or non-wheat crops. The four channels, that is, $p=4$, are those of the Multispectral Scanner on board of the Landsat-2. The number of observations in each class are as follows: $n_1=50$, $n_2=75$, $n_3=100$, $n_4=125$, and $n_5=150$ in total of $n=\sum n_g=500$ observations. A *priori* class probabilities are assumed to be equal.

We note that the correct parametric model for the simulated data is varying mean vectors and the varying covariance matrices which was checked by our procedure.

Each of the 100 different samples of multivariate data from each of the five normal populations were then analyzed using the AICPARM program of Bozdogan [9]. The *results of one such sample* is given in Table 5.5 for clustering $K=5$ simulated class types of different groups into $k=1$, 2, 3, 4 and 5-Sample Clusters on all variables and the corresponding AIC's and CAIC's are shown for each of the clustering alternatives.

Looking at Table 5.5, we see that *for this particular sample* AIC picks $k=5$ as being the correct number of classes (submodel 52), and then among the $k=4$-Sample Clusters it picks alternative submodel 47; among the $k=3$-Sample Clusters it picks submodel 25; and finally among the $k=2$-Sample Clusters it picks submodel 2 as the best clustering alternative, respectively in a hierarchical fashion. According to AIC, we never cluster the five class types as one homogeneous group (submodel 1). A typical design of the decision tree classifier from the results obtained according to AIC and CAIC is shown in Figure 5.1. This also turns out to be the optimal decision tree classifier which we shall see shortly.

Looking at the same results for CAIC's in Table 5.5, we see that this is a special run for CAIC's in that it is the only sample for which CAIC picks a structure different from that of AIC. Namely, CAIC picks $k=4$-Sample Clusters in submodel 47 first as the best clustering structure showing its tendency toward the lower dimensional model, and then it picks $k=5$-Sample Clusters in submodel 52 as the second best clustering structure.

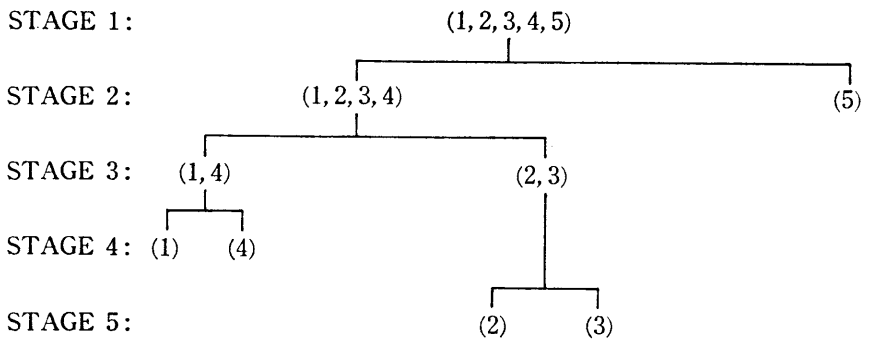

Figure 5.1. A typical design of decision tree classifier according to AIC and CAIC.

Often in constructing a decision tree classifier, it is assumed that the decision rules applied at each node are statistically independent and has no influence on decision tree design. Such an assumption, however, does influence the validity of the overall probability of correct classification (Argentiero, et al. [7]). To test the validity assumption among all the 100 different samples from each of the five normal populations in 100 repetitions of our Monte Carlo experiment, we also selected the *best clustering alternatives* for $k=1$, 2, 3, 4 and 5-Sample Clusters on the basis of AIC and CAIC. The results of the best clusters for $k=1$, 2, 3, 4, 5 as chosen by the minimum AIC procedure is shown in Table 5.6.

Table 5.6 Results of Monte Carlo Experiment Using AIC

k-Sample Clusters	Clustering	% Selected by AIC
$k=1$	(1, 2, 3, 4, 5)	100
$k=2$	(1, 2, 3, 4) (5)	75
	(1, 4, 5) (2, 3)	25
$k=3$	(1, 4) (2, 3) (5)	100
$k=4$	(1) (4) (2, 3) (5)	100
$k=5$	(1) (2) (3) (4) (5)	100

Looking at Table 5.6, we see that using AIC, we obtain a value of 0.75 for the *global probability, P_c, of correct classification* for the decision tree shown in Figure 5.1 which is the structure of our optimal decision tree classifier. Argentiero, et al. [6], reported that the theoretically optimal full-dimensional Bayes decision rule provided an accuracy of 0.79 when applied to this problem, which shows that our approach is nearly optimum.

CAIC chose the same optimal clusterings as above in Table 5.6 for each k, $k=$ 1, 2, 3, 4 and 5. However, in *only one* of the runs results of which are reported in Table 5.5, as we discussed before, CAIC picks $k=4$ first and then $k=5$ second. In other words, the order of selection is reversed in this particular sample. Therefore, CAIC in total picks 5 populations 99% of the time.

Thus, in general, if we already know a priori what k should be, AIC and CAIC agree over all the samples on which clustering is optimal. In our Monte Carlo experiment, AIC always chose the optimal value of k as five, the correct number of underlying heterogeneous normal populations from which each of the samples were taken. CAIC, on the other hand, picks 5 populations 99% of the time.

We emphasize the fact that this estimated structure chosen by CAIC ($k=4$-Sample Clusters) should not be construed to be inferior to that structure chosen by AIC ($k=5$-Sample Clusters). What this means is that the transformed p-dimensional feature space is such that the p-dimensional region contains at least 99% of the probability associated with each class density function under the multivariate normality assumption. In this sense, we use AIC to obtain the *upper bound* for the associated class probability and use CAIC to determine the *lower bound* on the associated class probability. Certainly, it is clear that CAIC or SC are developed for the analysis of the asymptotic property of the decision when the sample size n gets larger and larger (i.e., it tends to infinity), while AIC is designed for finite-sample modeling situations which is the case here. Moreover, it is applicable even when the sample size $n=1$, if only the necessary distributional property of AIC is assured.

6. Conclusions

The results of the above method clearly illustrate the flexibility of the minimum AIC and CAIC procedures over the classical hypothesis testing. We see that AIC and CAIC can indeed identify the best clustering alternatives when we cluster samples into homogeneous sets of samples under the best fitting model. We can detect the source of heterogeneity without any lengthly calculations or subjectivity, and we can measure the amount of homogeneity and heterogeneity in clustering samples. With this new approach it is now possible to determine *a priori* whether we should use equal or varying covariance matrices in the analysis of a data set. We can reduce the dimensionality of data sets as shown on the variety of rice data set, and we do not need to assume any arbitrary level of significance α and table look-up.

The model selection by AIC and CAIC is also more satisfying since all the possible clustering alternatives are considered.

Thus, from the results presented in this paper, we see that both AIC and CAIC *unify* the conventional test procedures and avoid the existing ambiguities inherent in these procedures. They avoid any restriction on K, the number of classes or groups, and p, the number of variables. The use of AIC and CAIC show how to combine the information in the likelihood with an appropriate function of the parameters to obtain estimates of the information provided by competing alternative models. Therefore, the definition of AIC and CAIC give clear formulation of the principle of parsimony in statistical model building or comparison as we demonstrated by numerical examples.

In concluding, the new approach presented in this paper will provide the researcher with a concise, efficient, and a more refined way of studying simultaneous comparative inference for a particular multi-sample data set. The ability of AIC and CAIC to allow the researcher to extract global information from the results of fitting several models is a unique characteristic that is not shared by the conventional procedures nor is it

realized by conventional significance tests.

Therefore, for these reasons the use of model-selection criteria is recommended in conjunction with Multi-Sample Cluster Analysis (MSCA) as an alternative to Multiple Comparison Procedures (MCP's).

Appendix: Combinatorial Subroutines

Here we give a listing of major combinatorial subroutines which we implemented in a newly developed statistical computer software by this author called AICPARM: A General Purpose Program for Computing AIC's for Univariate and Multivariate Normal Parametric Models. For a lucid discussion and details on combinatorial algorithms, we refer the reader to Nijenhuis and Wilf [28].

A.1 MCP: Combination of K samples taken k at a time for MCP's in lexicographic order

This subroutine generates and lists different combinations of K groups or samples taken k at a time sequentially. There are $\begin{bmatrix} K \\ k \end{bmatrix}$ k-subsets of a K-set altogether, and MCP is a simple algorithm which constructs the all possible alternatives in "lexicographic", that is, in "alphabetical order". A listing of the output from this program is shown in Table 3.1.

```
      PROGRAM MCP

      PROGRAM MCP
      INTEGER A(100), N, K, H, M2
      LOGICAL MTC
      MTC=.FALSE.
      PRINT *, 'K CHOOSE k'
      PRINT *, 'WHAT IS K?'
      READ *, N
      PRINT *, 'WHAT IS k?'
      READ *, K
   10 CONTINUE
      CALL NEXKSB(N, K, A, MTC, H, M2)
      PRINT 2, (A(I), I=1, K)
    2 FORMAT(30(1X, I1))
      IF(MTC) GOTO 10
      END
      SUBROUTINE NEXKSB(N, K, A, MTC)
      INTEGER A(K)
      LOGICAL MTC
      INTEGER H, M2
      SAVE H, M2
   30 IF(MTC) GOTO 40
   20 M2=0
```

```
      H=K
      GOTO 50
   40 IF(M2.LT.N-H) H=0
      H=H+1
      M2=A(K+1-H)
   50 DO 51 J=1,H
   51 A(K+J-H)=M2+J
      MTC=A(1).NE.N-K+1
      RETURN
      END
```

A.2 STIRN2: Stirling number of the second kind

This subroutine constructs a table of the total number of clustering alternatives for various values of K, number of samples, and k varying number of clusters of samples. A listing of the output from this program is shown in Table 3.2.

```
      PROGRAM STIRN2

      PROGRAM STIRN2
      REAL S(20,20),T
      INTEGER N,K
C
      S(1,2)=0.
      DO 5 I=1,20
         S(I,1)=1.
    5 CONTINUE
C
      PRINT 30,'TOTAL',(I,I=1,20)
   30 FORMAT(13X,A,6(I14,1X),3(:/T19,6(I14,1X)))
   40 FORMAT(I2,1X,7(I14,1X),3(:/T19,6(I14,1X)))
C
      PRINT 40,1,1,1
      DO 20 N=2,20
         T=1.
         DO 10 K=2,N
            S(N,K)=K*S(N-1,K)+S(N-1,K-1)
            T=T+S(N,K)
   10    CONTINUE
         PRINT 40,N,T,(S(N,I),I=1,N)
   20 CONTINUE
C
C
      END
```

A.3 REPFM: Representation forms of clustering alternatives

Clustering alternatives can be classified according to their *representation forms* to

make it easy to list all possible clustering alternatives. The subroutine REPFM gives the partition of K (number of samples) which is a positive integer, into a specified k number of parts. For example, the representation forms of $K=6$ samples into $k=3$ parts are:

$$\begin{aligned} 6 &= \{4\}+\{1\}+\{1\} \\ &= \{3\}+\{2\}+\{1\} \\ &= \{2\}+\{2\}+\{2\}. \end{aligned}$$

PROGRAM REPFM

```
      PROGRAM REPFM
      INTEGER N, D, I, K
      INTEGER R(100), M(100)
      LOGICAL MTC, FIRST
      EXTERNAL NEXPAR
      MTC=.TRUE.
      FIRST=.TRUE.
      PRINT *, 'WHAT IS N?'
      READ *, N
   10 CONTINUE
      IF(MTC) CALL NEXPAR(N, R, M, D, MTC, FIRST)
      PRINT 2, ((R(I), K=1, M(I)), I=1, D)
    2 FORMAT (30(1X, I1))
      IF(MTC) GOTO 10
      STOP
      END
      SUBROUTINE NEXPAR(N, R, M, D, MTC, FIRST)
      INTEGER N, M, R, S, D, SUM, F
      LOGICAL MTC, FIRST
      DIMENSION R(N), M(N)
      INTRINSIC MOD
      SAVE
      IF(.NOT. FIRST) GOTO 20
      FIRST=.FALSE.
   30 S=N
      D=0
   50 D=D+1
      R(D)=S
      M(D)=1
   40 MTC=M(D). NE. N
      RETURN
   20 IF(.NOT. MTC) GOTO 30
      SUM=1
      IF(R(D). GT. 1) GOTO 60
```

```
      SUM=M(D)+1
      D=D-1
   60 F=R(D)-1
      IF(M(D).EQ.1) GOTO 70
      M(D)=M(D)-1
      D=D+1
   70 R(D)=F
      M(D)=1+SUM/F
      S=MOD(SUM,F)
      IF(S) 40,40,50
      END
```

A.4 ALLSUB: All possible partitioning of *K*-samples into *k*-sample clusters

This subroutine generates and lists all the simple patterns of clustering alternatives for a specified number of samples K for Multi-Sample Cluster Analysis. A listing of the output from this program is shown in Table 3.3.

```
      PROGRAM ALLSUB

      PROGRAM ALLSUB
      INTEGER N,NC
      INTEGER P(100),Q(100)
      CHARACTER*80 LIST
      EXTERNAL NEXEQU
      LOGICAL MTC
      PRINT *,'HOW MANY GROUPS?'
      READ *,N
      MTC=.FALSE.
   10 CALL NEXEQU(N,NC,P,Q,MTC)
      CALL NEXLST(N,P,Q,NC)
      IF(MTC) GOTO 10
      END
      SUBROUTINE NEXEQU(N,NC,P,Q,MTC)
      INTEGER N,NC
      INTEGER P(N),Q(N)
      LOGICAL MTC
      SAVE
      IF(MTC) GOTO 20
   10 NC=1
      DO 11 I=1,N
   11     Q(I)=1
      P(1)=N
   60 MTC=NC.NE.N
      RETURN
   20 M=N
```

```
30 L=Q(M)
   IF(P(L). NE. 1) GOTO 40
   Q(M)=1
   M=M-1
   GOTO 30
40 NC=NC+M-N
   P(1)=P(1)+N-M
   IF(L. NE. NC) GOTO 50
   NC=NC+1
   P(NC)=0
50 Q(M)=L+1
   P(L)=P(L)-1
   P(L+1)=P(L+1)+1
   GOTO 60
   END
   SUBROUTINE NEXLST(N, P, Q, NC)
   INTEGER P(N), Q(N)
   CHARACTER*80 CLIST
   INTEGER ILAST, I, J
 1 FORMAT(I2)
   CLIST=' '
   ILAST=1
   DO 10 I=1, NC
       CLIST(ILAST : ILAST)='('
       ILAST=ILAST+1
       NCLI=P(I)
       DO 20 J=1, N
           IF(Q(J). EQ. I) THEN
               WRITE(CLIST(ILAST : ILAST+1), 1) J
               ILAST=ILAST+2
               NCLI=NCLI-1
               IF(NCLI. EQ. 0) GOTO 21
               CLIST(ILAST+1 :)=','
               ILAST=ILAST+1
           ENDIF
20     CONTINUE
21     CLIST(ILAST : ILAST)=')'
       ILAST=ILAST+1
10 CONTINUE
   PRINT *, CLIST
   RETURN
   END
```

Acknowledgments

The author extends his appreciation to Professor S. Arikawa, the Associate Editor, the Editorial Board, and the referee for the comments and suggestions provided during the revision of this paper. I also wish to thank Professor Donald E. Ramirez for reading and commenting on some parts of the paper. This research was supported by Army Research Office Contract DAAG29-82-K-055, at the University of Illinois at Chicago during the past three summers as a Faculty Associate. The author is indebted to ARO, and a special thanks goes to my thesis advisor, Professor Stanley L. Sclove, and to Professor Hirotugu Akaike for his continued support and encouragement in my work.

References

[1] AKAIKE, H.: *Information Theory and an Extension of the Maximum Likelihood Principle*, Second International Symposium on Information Theory, B.N. Petrov and F. Csaki (Eds.), Budapest: Akademiai Kiado, (1973), 267-281.

[2] AKAIKE, H.: *A New Look at the Statistical Model Identification*, IEEE Transactions on Automatic Control, **AC-19**, (1974), 716-723.

[3] AKAIKE, H.: *On Entropy Maximization Principle*, in P.R. Krishnaiah (Ed.), Proceedings on Applications of Statistics, Amsterdam, North-Holland, (1977), 27-47.

[4] AKAIKE, H.: *A Bayesian Analysis of the Minimum AIC Procedure*, Annals of Institute of Statistical Mathematics, **30**, Part A, (1979), 9-14.

[5] AKAIKE, H.: *Likelihood of a Model and Information Criteria*, Journal of Econometrics, **16**, (1981), 3-14.

[6] ARGENTIERO, P., STRONG, J., and KOCH, D.: *Inventory Estimation on the Massively Parallel Processor*, in Proc. 1980 Symp. Mach. Process Remotely Sensed Data, Prude University, West Lafayette, IN, June 1980.

[7] ARGENTIERO, P., CHIN R., and BEAUDET, P.: *An Automated Approach to the Design of Decision Tree Classifier*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **PAMI-4**, 1, (1982), 51-57.

[8] BOZDOGAN, H.: *Multi-Sample Cluster Analysis and Approaches To Validity Studies in Clustering Individuals*, Unpublished Ph. D. Thesis, Department of Mathematics, University of Illinois at Chicago, Chicago, IL 60680, 1981.

[9] BOZDOGAN, H.: *AICPARM: A General Purpose Program for Computing AIC's for Univariate and Multivariate Parametric Models*, Technical Paper #1 in Statistics, Department of Mathematics, University of Virginia, Charlottesville, VA., 22903, 1983.

[10] BOZDOGAN, H.: *AIC-Replacements for Multivariate Multi-Sample Conventional Tests of Homogeneity Models*, Technical Paper #4 in Statistics, Department of Mathematics, University of Virginia, Charlottesville, VA., 22903, 1984.

[11] BOZDOGAN, H.: *Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Extensions*, Technical Paper #7 in Statistics, Department of Mathematics, University of Virginia, Charlottesville, VA., 22903, 1985.

[12] BOZDOGAN, H., and SCLOVE, S.L.: *Multi-Sample Cluster Analysis Using Akaike's Information Criterion*, Annals of Institute of Statistical Mathematics, **36**, Part B, (1984), 243-253.

[13] COX, D.R. and SPJØTVOLL, E.: *On Partitioning Means into Groups"*, Scandinavian Journal of Statistics, **9**, (1982), 147-152.

[14] DUNCAN, D.B.: *Multiple Range and Multiple* F *Tests*, Biometrics, **11**, (1955), 1-42.

[15] DURAN, B.S., and ODELL, P.L.: *Cluster Analysis: A Survey*, New York, Springer-Verlag, 1974.

[16] GABRIEL, K.R.: *A Procedure for Testing the Homogeneity of All Sets of Means in Analysis of Variance*, Biometrics, **20**, (1964), 459-477.

[17] GABRIEL, K.R.: *Simultaneous Test Procedures in Multivariate Analysis of Variance*, Biometrika, **55**, (1968), 489-504.
[18] GOWER, J.: *Is Classification Statistical?*, Key Note Lecture, Twelfth Annual Meeting, The Classification Soceity (North American Branch), Toronto, Canada, May 31-June 2, 1981.
[19] KASHYAP, R.I.: *Optimal Choice of AR and MA Parts in Autoregressive Moving Average Models*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **4**, (1982), 99-104,
[20] KRISHNAIAH, P.R.: *Multiple Comparison Tests in Multiresponse Experiments*, Sankhya *A*, **27**, (1965), 65-72.
[21] KRISHNAIAH, P.R.: *Simultaneous Test Procedures Under General MANOVA Models*, P.R. Krishnaiah (Ed.), Multivariate Analysis, Vol. II, New York, Academic Press, (1969), 121-143.
[22] KRISHNAIAH, P.R.: *Some Developments on Simultaneous Test Procedures*, P.R. Krishnaiah (Ed.), Developments in Statistics, **2**, New York, Academic Press. (1979), 157-201.
[23] KULKARNI, A., and KANAL, L.: *An Optimization Approach to Hierarchical Classifier Design*, Proceedings of 3rd IJCPR, Coronado, CA, 1976.
[24] MARDIA, K.V., KENT, J.T., and BIBBY, J.M.: *Multivariate Analysis*, New York, Academic Press, 1979.
[25] MILLER, R.G., JR.: *Simultaneous Statistical Inference*, New York, McGraw-Hill, 1966.
[26] MILLER, R.G. JR.: *Simultaneous Statistical Inference*, New York, Springer-Verlag, 1981.
[27] MUI, J., and FU, K.: *Automated Classification of Nucleated Blood Cells Using a Binary Tree Classifier*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **PAMI-2**, (1980), 429-443.
[28] NIJENHUIS, A., and WILF, H.S.: *Combinatorial Algorithms*, Second Edition, New York, Academic Press, 1978.
[29] OLSON, C.L.: *Comparative Robustness of Six Tests in Multivariate Analysis of Variance*, Journal of American Statistical Association, **69**, (1974), 894-908.
[30] O'NEILL, R. and WETHERILL, G.B.: *The Present State of Multiple Comparison Methods*, Journal of Royal Statistical Society, Series B, **33**, (1971), 218-241.
[31] ROY, S.N.: *Some Aspects of Multivariate Analysis*, New York, John Wiley, 1957.
[32] SCLOVE, S.L.: *Application of the Conditional Population-Mixture Model to Image Segmentation*, Technical Report No. **A82-1**, August 15, 1982, ARO, Contract DAAG29-82-0155, Quantititave Methods Department, University of Illinois at Chicago, ILL. 60680,
[33] SCHWARZ, G.: *Estimating the Dimension of a Model*, Annals of Statistics, **6**, (1978), 461-464.
[34] SCOTT, A.J., and KNOTT, M.: *A Cluster Analysis Method for Grouping Means in the Analysis of Variance*, Biometrics, **30**, (1974), 507-512.
[35] SEBER, G.A.F.: *Multivariate Observations*, New York, John Wiley, 1984.
[36] SILVEY, S.D.: *Statistical Inference*, Baltimore, Penguin, 1970.
[37] SPÄTH, H.: *Cluster Analysis Algorithms: For Data Reduction and Classification of Objects*, New York, John Wiley, 1980.
[38] SPJØTVOLL, E.: *Multiple Testing in the Analysis of Variance*, Scandinavian Journal of Statistics, **1**, (1974), 97-114.
[39] SRIVASTAVA, M.S., and CARTER, E.M.: *An Introduction to Applied Multivariate Statistics*, New York, North Holland, 1983.
[40] THOMAS, D.A.H.: *Multiple Comparisons Among Means. A Review*, The Statistician, **22**, (1973), 16-42.
[41] WANG, Q.R., and SUEN, C.Y.: *Analysis and Design of a Decision Tree Based on Entropy Reduction and Its Application to Large Character Set Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **PAMI-6**, **No. 4**, (1984), 406-417.
[42] WILKS, S.S.: *Certain Generalization in the Analysis of Variance*, Biometrika, **24**, (1932), 471-494.

Communicated by H. Akaike
Received January 8, 1985
Revised September 9, 1985