Ann. Inst. Statist. Math.
36 (1984), Part B, 163-180

# MULTI-SAMPLE CLUSTER ANALYSIS USING AKAIKE'S INFORMATION CRITERION*

HAMPARSUM BOZDOGAN AND STANLEY L. SCLOVE

(Received Mar. 23, 1982; revised Jan. 28, 1983)

## Summary

Multi-sample cluster analysis, the problem of grouping samples, is studied from an information-theoretic viewpoint via Akaike's Information Criterion (AIC). This criterion combines the maximum value of the likelihood with the number of parameters used in achieving that value. The multi-sample cluster problem is defined, and AIC is developed for this problem. The form of AIC is derived in both the multivariate analysis of variance (MANOVA) model and in the multivariate model with varying mean vectors and variance-covariance matrices. Numerical examples are presented for AIC and another criterion called $w$-square. The results demonstrate the utility of AIC in identifying the best clustering alternatives.

## 1. Introduction

In this paper, we shall develop Akaike's Information Criterion (AIC) for multi-sample cluster analysis with common and also with varying variance-covariance matrices, since often in practice the assumption of equal variance-covariance matrices is a rather dubious requirement. The problem of multi-sample cluster analysis arises when we are given a collection of samples (groups, treatments), to be clustered into homogeneous groups. Many practical situations require the presentation of multivariate data from several structured samples for comparative purposes and the grouping of the heterogeneous samples into homogeneous sets of samples. Thus, it is reasonable to provide a practically useful statistical procedure that would use some sort of statistical model to

* This research was supported by Office of Naval Research Contract N00014-80-C-0408, Task NR042-443 and Army Research Office Contract DAAG 29-82-K-0155, at the University of Illinois at Chicago.

Key words and phrases: Multi-sample cluster analysis, $w$-square criterion, Akaike's Information Criterion (AIC), MANOVA model, multivariate model with varying mean vectors and variance-covariance matrices, maximum likelihood.

aid in comparisons of various collections of samples, identify homogeneous groups of samples, telling us which samples should be clustered together and which should not. Examples of multi-sample clustering situations are abundant in practice. We shall give two of these examples later and illustrate numerically. The concept of multi-sample cluster analysis presented in this paper is relatively new. It has not been definitively studied before either using the conventional simultaneous test procedures (STP's) which are based on inference for the multivariate analysis of variance (MANOVA) model, or from an information-theoretic viewpoint, which we shall adopt in this paper via Akaike's Information Criterion (AIC).

Multivariate analysis of variance (MANOVA) is a widely used model for comparing two or more multivariate samples with a common covariance matrix. In this model, the likelihood ratio principle leads to *Wilks' lambda* [17], or in short *Wilks' $\Lambda$ Criterion* as the test statistic. It plays the same role in multivariate analysis that $F$-ratio statistic plays in the univariate case. Often, however, the formal analyses involved in MANOVA are not revealing or informative. Moreover, the test statistics used under this model are derived under the assumption of equal covariance matrices. If we have a reason to doubt equality of covariances, then we may first want to test the equality of covariances. In the multivariate case the equality of covariance matrices is certainly more hazardous. If the covariance matrices are unequal, a bias occurs in the test for equality of mean vectors. Therefore, for this reason we may want to first test the equality of covariance matrices instead of immediately leaping to the MANOVA hypothesis. This is an important option to use in clustering groups or samples when we are not willing to assume equal covariance matrices between the samples or groups in the multi-sample data. Once the MANOVA hypothesis of equality of mean vectors is rejected at some prescribed significance level $\alpha$, then it is necessary to study in detail the discrepancies between the null hypothesis and the data. In the statistical literature, in the MANOVA case, there are a variety of conventional multiple comparison procedures for studying the discrepancies between the null hypothesis and the data. These test procedures are: Step-down Methods, Union Intersection Tests, and Simultaneous Confidence Intervals. For more details on these test procedures refer to Gabriel [7], Krishnaiah [10], [11], Srivastava [16], and others. As noted in Consul [4], the exact distributions of these conventional test procedures are either unknown or are known for some particular cases only. Moreover, the problem of finding the percentage points of these statistics has become rather difficult. For these reasons, and for our purposes, these test procedures have little practical use. Furthermore, they create additional problems in terms of how to control the overall error rate $\alpha$, since we can no longer use the same $\alpha$ to discover where the discrepancies between the null hypothesis and the data might occur.

In the case of testing the equality of covariance matrices, we find ourselves in the same situation as in the MANOVA model. For this problem also, there are in the statistical literature several test procedures. For example, one of the most commonly used tests is *Box's M test* despite the fact that it is very restrictive. For instance, Box's approximation seems to be only good if each sample size, $n_g$ exceeds 20, and if the number of samples, $K$, and the number of variables, $p$, exceed 5. It is also very expensive to compute it on a high speed computer, even on an IBM 370. Once the hypothesis of equality of covariance matrices is rejected at some prescribed significance level $\alpha$, then again it is necessary to study in detail the discrepancies between the null hypothesis and the data. Further reviewing the statistical literature, we see that there are no conventional simultaneous test procedures (STP's) in this case in studying the discrepancies between the null hypothesis and the data. One can perhaps construct a sequential likelihood ratio type test, but as is mentioned in Muirhead ([14], pp. 296), the likelihood ratio test in testing the equality of covariance matrices has the defect that, when the sample sizes $n_1, n_2, \cdots, n_K$ are not all equal, it is biased. Therefore, in the multi-sample cluster problem with varying parameters, carrying out a sequence of likelihood ratio tests leaves much to be desired in identifying homogeneous groups of samples.

More recently, however, in the statistical literature, we see a likelihood based approach, called *w-square criterion* given in Mardia et al. [12] to aid in comparing various collections of samples, identifying homogeneous groups of samples, and telling which should be clustered together. For normal samples with equal covariance matrices, the *w-square criterion* is defined by

$$w_a^2=\sum_{g=1}^{K}\sum_{\bar{X}_j\in C_g} n_j(\bar{X}_j-\bar{\bar{X}}_g)'\hat{\Sigma}^{-1}(\bar{X}_j-\bar{\bar{X}}_g) \tag{1.1}$$

where

$C_g$=the set of $\bar{X}_j$ assigned to the $g$th group, $g=1, 2, \cdots, K$,

$\bar{\bar{X}}_g$=the weighted mean vector of the means in the $g$th group, or the cluster set $C_g$ of groups,

$\hat{\Sigma}=W/(n-K)$, the pooled estimate of $\Sigma$,

$W=\sum_{g=1}^{K} A_g$ is the within-samples SSP matrix,

$n=\sum_{g=1}^{K} n_g$, and

$K$ = the number of groups or samples to be clustered.

If the matrix of Mahalanobis distances $D_{ij}$ given by

$$D_{ij}^2 = (\bar{X}_i - \bar{X}_j)' \hat{\Sigma}^{-1} (\bar{X}_i - \bar{X}_j)$$

is available, then for computational convenience, $w_a^2$ can be written as

$$w_a^2 = 1/2 \sum_{g=1}^{K} N_g^{-1} \sum_{C_g} n_i n_j D_{ij}^2 \tag{1.2}$$

where

$$N_g = \sum_{\bar{X}_j \in C_g} n_j \,.$$

Thus, when we are given multi-sample data and wish to cluster the samples, we compute $w_a^2$ in (1.1) or (1.2) for some or all of the alternative groupings of samples, and choose the minimum of $w_a^2$ to be the "best" alternative clustering of samples. This is appropriate, since maximizing the likelihood implies minimizing $w_a^2$. Even though the $w_a^2$ criterion is a step forward in identifying homogeneous groups of samples and evaluating multi-sample clusters, it has some disadvantages. For instance, it does not make any allowance for $m$, the number of parameters estimated within the model and the subsequent alternative submodels. It is always zero when the groups or samples are clustered as singletons, as we shall see later in Section 4. As it is given in (1.1), we can only work with $w_a^2$ criterion when we assume equal covariance matrices.

For the above stated reasons, and the problems encountered in the conventional test procedures which we discussed above, in this paper we shall propose Akaike's Information Criterion (AIC) as a new and unifying procedure for evaluating multi-sample clusters, and use it to identify the best clustering alternatives.

In 1971, Akaike first introduced an information criterion, referred to as a Model Identification Criterion or Akaike's Information Criterion (AIC), for the identification and comparison of statistical models in a class of competing models with different numbers of parameters. It is defined by

$$\begin{aligned} \text{AIC} = {} & (-2)\log_e(\text{maximized likelihood}) \\ & +2\,(\text{number of free parameters within the model})\,. \end{aligned} \tag{1.3}$$

It was obtained by Akaike [2], [3] based on the recognition that the classical method of maximum likelihood could be viewed as a method of identification of a statistical model realized by maximizing an estimate of the generalized entropy, or the expected log likelihood, of the model being fitted. It estimates minus twice the expected log likelihood of the model whose parameters are determined by the method of maximum likelihood. When several competing models are being compared or fitted, AIC is a simple procedure which measures the *badness of fit* or the *discrepancy* of the estimated model from the true model when a set of data is given. The first term in (1.3) stands for the penalty of *badness of fit* when the maximum likelihood estimators of the parameters of the model are used. The second term in the definition of AIC, on the other hand, stands for the penalty of increased *unreliability* or *compensation for the bias* in the first term as a consequence of increasing number of parameters. If more parameters are used to describe the data, it is natural to get a larger likelihood, possibly without improving the true goodness of fit by penalizing the use of additional parameters. Thus, when there are several competing models, the parameters within the models are estimated by the method of maximum likelihood and the AIC-values are computed and compared to find a model with the minimum value of AIC. This procedure is called the *minimum AIC procedure.* The model with the minimum AIC is called the *minimum AIC estimate* (MAICE) and is designated as the *best model.* Thus, in applying AIC the emphasis is on comparing the "goodness of fit" of various models with an allowance made for *parsimony.*

In Section 2, we shall define the general multi-sample cluster problem. In Section 3, we shall derive the AIC procedure both for the multivariate analysis of variance (MANOVA) model, and for the multivariate model with varying covariance matrices. We shall, in Section 4, give different numerical examples of multi-sample cluster analysis on different real data sets to demonstrate our results from applying minimum AIC procedures in different computer analyses. Finally, in Section 5, we shall present our conclusions and discussion.

## 2. The multi-sample cluster problem

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

$$X(n\times p)=\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix} \tag{2.1}$$

be a single data matrix of $K$ groups or samples, where $X_g(n_g\times p)$ represents the observations from the $g$th 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, but $k \leqq K$.

Often individuals or objects have been sampled from $K>1$ populations. The data matrix may be represented in partitioned form as above. Let $n_g$ represent the number of individuals in the $g$th (random) sample, $g=1, 2, \cdots, K$. The $n_g$ are not resticted to being equal or proportional to other $n_g$'s. The total number of observations is $n=\sum_{g=1}^{K} n_g$. Let $\boldsymbol{X}_{gi}$ be the $p\times 1$ vector of observations in group $g=1, 2, \cdots, K$, and for individual $i=1, 2, \cdots, n_g$.

## 3. Derivation of AIC for two multivariate models

### 3.1 *AIC for the multivariate analysis of variance (MANOVA) model: AIC (common $\Sigma$)*

We now turn our attention to consider situations with several multivariate normal samples.

For example, we may have multi-sample data with sample sizes $n_1, n_2, \cdots, n_K$ which are assumed to come from $K$ populations, the first with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$, the second with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\Sigma, \cdots$, the $K$th with mean vector $\boldsymbol{\mu}_K$ and covariance matrix $\Sigma$. Therefore, throughout this section we shall suppose that we may have independent data matrices $X_1, X_2, \cdots, X_K$, where the rows of $X_g(n_g\times p)$ are independent and identically distributed (i.i.d.) according to a multivariate normal distribution, $N_p(\boldsymbol{\mu}_g, \Sigma)$, $g=1, 2, \cdots, K$. We may want to compare the $K$ sample mean vectors given that all $K$ distributions have a common covariance matrix $\Sigma$. This is the well known multivariate analysis of variance (MANOVA) model. In terms of the parameters the MANOVA model is $\boldsymbol{\theta}=(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K, \Sigma)$ with $m=kp+p(p+1)/2$ parameters, where $k$ is the number of groups, and $p$ is the number of variables.

We shall derive the form of AIC for this model. Recall the definition of AIC from Section 1,

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

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

Consider $K$ normal populations with different mean vectors $\boldsymbol{\mu}_g$, $g=1, 2, \cdots, k, \cdots, K$. Let $\boldsymbol{X}_{gi}$, $g=1, 2, \cdots, K$; $i=1, 2, \cdots, n_g$, be a random sample of observations from the $g$th population $N_p(\boldsymbol{\mu}_g, \Sigma)$. If the groups or samples can differ only in their mean vectors, we can write the multivariate one-way analysis variance (MANOVA) model as

$$(3.1.1) \qquad \boldsymbol{X}_{gi}=\boldsymbol{\mu}_g+\boldsymbol{\varepsilon}_{gi}\,, \qquad g=1, 2, \cdots, K;\ i=1, 2, \cdots, n_g\,,$$

where $\boldsymbol{X}_{gi}$ is the $(p\times 1)$ response or outcome vector in the $g$th group for the $i$th individual or object, $\boldsymbol{\mu}_g$ are vector parameters, and $\boldsymbol{\varepsilon}_{gi}$ are independent $N_p(\boldsymbol{0}, \Sigma)$ random vector errors.

Thus, the basic *null hypothesis* we usually are interested in testing is given by

$$(3.1.2) \qquad H_0:\ \boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_K\,.$$

The alternative hypothesis is given by

$$H_1:\ \text{Not all } \boldsymbol{\mu}_K \text{ are equal}\,.$$

Wilks' lambda is a *general* statistic for handling this problem. Although there are several other conventional statistics for this purpose, they all can be viewed as special cases of Wilks' $\Lambda$ which we shall not discuss here.

For notational purposes, we shall denote by $T$ the "total" sum of squares and products (SSP) matrix, by $W$ the "within-group" or "within-sample" SSP matrix, and by $B$ the "between-group" SSP matrix. Hence, it can be shown that

$$(3.1.3) \qquad T=W+B\,,$$

where

$$(3.1.4) \qquad T=\sum_{g=1}^{K}\sum_{i=1}^{n_g}(\boldsymbol{X}_{gi}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{gi}-\bar{\boldsymbol{X}})'\,,$$

$$(3.1.5) \qquad W=\sum_{g=1}^{K}\sum_{i=1}^{n_g}(\boldsymbol{X}_{gi}-\bar{\boldsymbol{X}}_g)(\boldsymbol{X}_{gi}-\bar{\boldsymbol{X}}_g)'\,,$$

and

$$(3.1.6) \qquad B=\sum_{g=1}^{K}n_g(\bar{\boldsymbol{X}}_g-\bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_g-\bar{\boldsymbol{X}})'\,,$$

with

$$\bar{\boldsymbol{X}}_g=\frac{1}{n_g}\sum_{i=1}^{n_g}\boldsymbol{X}_{gi}\,, \qquad g=1, 2, \cdots, K\,,$$

$$\bar{\boldsymbol{X}}=\frac{1}{n}\sum_{g=1}^{K}\sum_{i=1}^{n_g}\boldsymbol{X}_{gi}\,, \qquad n=\sum_{g=1}^{K}n_g\,.$$

Therefore, we can present multivariate one-way analysis of variance (MANOVA) table as follows.

Table 3.1. MANOVA table

| Source | d.f. | SSP matrix | Wilks' criterion |
|---|---|---|---|
| Between samples | $K-1$ | $B$ | $\|W\|/\|T\|$ |
| Within samples | $n-K$ | $W$ | $\sim\Lambda(p;\, n-K;\, K-1)$ |
| Total | $n-1$ | $T$ | |

Now, we derive the form of Akaike's Information Criterion (AIC) for the MANOVA model given in (3.1.1), subject to the constraint given in (3.1.2). The likelihood function of all the sample observations is given by

$$L(\boldsymbol{\mu}_g, \Sigma_g; X)=\prod_{g=1}^{K} L_g(\boldsymbol{\mu}_g, \Sigma_g; X_g)\,, \tag{3.1.7}$$

or by

$$\begin{aligned} L=&(2\pi)^{-np/2}\prod_{g=1}^{K}|\Sigma_g|^{-n_g/2} \\ &\times\exp\left\{-1/2\operatorname{tr}\sum_{g=1}^{K}\Sigma_g^{-1}A_g-1/2\operatorname{tr}\sum_{g=1}^{K}n_g\Sigma_g^{-1}(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)'\right\}, \end{aligned} \tag{3.1.8}$$

where

$$n=\sum_{g=1}^{K}n_g \quad\text{and}\quad A_g=\sum_{i=1}^{n_g}(X_{gi}-\bar{X}_g)(X_{gi}-\bar{X}_g)'\,.$$

The log likelihood function is

$$\begin{aligned} l(\boldsymbol{\mu}_g, \Sigma_g; X)&\equiv\log L(\boldsymbol{\mu}_g, \Sigma_g; X) \\ &=-(np/2)\log(2\pi)-1/2\sum_{g=1}^{K}n_g\log|\Sigma_g|-1/2\operatorname{tr}\sum_{g=1}^{K}\Sigma_g^{-1}A_g \\ &\quad-1/2\operatorname{tr}\sum_{g=1}^{K}n_g\Sigma_g^{-1}(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)'\,. \end{aligned} \tag{3.1.9}$$

Since the common covariance matrix is $\Sigma$, the log likelihood function becomes

$$\begin{aligned} l(\{\boldsymbol{\mu}_g\}, \Sigma; X)&\equiv\log L(\{\boldsymbol{\mu}_g\}, \Sigma; X) \\ &=-(np/2)\log(2\pi)-(n/2)\log|\Sigma|-1/2\operatorname{tr}\Sigma^{-1}\sum_{g=1}^{K}A_g \\ &\quad-1/2\operatorname{tr}\Sigma^{-1}\sum_{g=1}^{K}n_g(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)'\,, \end{aligned} \tag{3.1.10}$$

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

$$\hat{\boldsymbol{\mu}}_g=\bar{X}_g\,, \qquad g=1, 2, \cdots, K\,, \tag{3.1.11}$$

and

$$(3.1.12) \qquad \hat{\Sigma}=n^{-1}W,$$

where $W=\sum_{g=1}^{K} A_g$.

Substituting these back into (3.1.10) and simplifying, the maximized log likelihood becomes

$$(3.1.13) \quad l(\{\hat{\mu}_g\}, \hat{\Sigma}; X) \equiv \log L(\{\hat{\mu}_g\}, \hat{\Sigma}; X)$$
$$= -(np/2)\log(2\pi)-(n/2)\log|n^{-1}W|-(np/2),$$

where $W$ is the "within-group" SSP matrix.

Since

$$(3.1.14) \qquad \text{AIC}=-2\log_e L(\hat{\theta})+2m,$$

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

$$(3.1.15) \quad \text{AIC (common } \Sigma)$$
$$=np\log_e(2\pi)+n\log_e|n^{-1}W|+np+2\left[kp+\frac{p(p+1)}{2}\right].$$

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

$$(3.1.16) \qquad \text{AIC* (common } \Sigma)=n\log_e|W|+2\left[kp+\frac{p(p+1)}{2}\right]$$

where

$n=\sum_{g=1}^{K} n_g=$the total sample size,<br>
$|W|=$the determinant of "within-group" SSP matrix,<br>
$k=$number of groups or samples compared,<br>
$p=$number of variables.

However, for purposes of comparison we retain the constants and use AIC (common $\Sigma$).

### 3.2 *AIC for the multivariate model with varying parameters: AIC (varying $\mu$ and $\Sigma$)*

As we mentioned in Section 1, the assumption of equality of covariance matrices in MANOVA can cause serious problems. For this reason we may want first test the equality of covariance matrices against the alternative that not all covariance matrices are equal, given no restriction on the population mean vectors. Therefore, throughout this section we shall suppose that we may have independent data matrices $X_1, X_2, \cdots, X_K$, where the rows of $X_g(n_g\times p)$ are independent and identically distributed (i.i.d.) $N_p(\boldsymbol{\mu}_g, \Sigma_g)$, $g=1, 2, \cdots, K$. In terms of the parameters with varying mean vectors and covariance matrices, the multivariate model we shall consider is

$$\boldsymbol{\theta}=(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_K, \Sigma_1, \Sigma_2, \cdots, \Sigma_K)$$

with $m=kp+kp(p+1)/2$ parameters, where $k$ is the number of groups, and $p$ is the number of variables.

Thus, the basic *null hypothesis* we usually are interested in testing is given by

$$(3.2.1) \qquad H_0\colon \ \Sigma_1=\Sigma_2=\cdots=\Sigma_K\,.$$

The *alternative hypothesis* is given by

$$H_1\colon \ \text{Not all } K \text{ covariance matrices are equal.}$$

In multivariate analysis this is known as the *test of homogeneity of covariance matrices.*

To derive Akaike's Information Criterion (AIC) in this case the log likelihood function is given by

$$(3.2.2) \qquad \begin{aligned} l(\{\boldsymbol{\mu}_g, \Sigma_g\}; X) &\equiv \log L(\{\boldsymbol{\mu}_g, \Sigma_g\}; X) \\ &= -(np/2)\log(2\pi) - 1/2 \sum_{g=1}^{K} n_g \log|\Sigma_g| \\ &\quad -1/2 \sum_{g=1}^{K} n_g \operatorname{tr} \Sigma_g^{-1} A_g - 1/2 \sum_{g=1}^{K} n_g(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)'(\bar{\boldsymbol{x}}_g-\boldsymbol{\mu}_g)\,. \end{aligned}$$

The MLE's of $\boldsymbol{\mu}_g$ and $\Sigma_g$ are

$$(3.2.3) \qquad \hat{\boldsymbol{\mu}}_g=\bar{X}_g\,, \qquad g=1, 2, \cdots, K\,,$$

and

$$(3.2.4) \qquad \hat{\Sigma}_g=A_g/n_g\,.$$

Substituting these back into (3.2.2) and simplifying, the maximized log likelihood becomes

$$(3.2.5) \qquad \begin{aligned} l(\{\hat{\boldsymbol{\mu}}_g, \hat{\Sigma}_g\}; X) &\equiv \log L(\{\hat{\boldsymbol{\mu}}_g, \hat{\Sigma}_g\}; X) \\ &= -(np/2)\log(2\pi) - 1/2 \sum_{g=1}^{K} n_g \log|n_g^{-1} A_g| - (np/2)\,. \end{aligned}$$

Since

$$(3.2.6) \qquad \mathrm{AIC} = -2\log_e L(\hat{\boldsymbol{\theta}}) + 2m\,,$$

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

$$\text{(3.2.7)}\quad \text{AIC (varying } \boldsymbol{\mu} \text{ and } \Sigma)=np\log_e(2\pi)+\sum_{g=1}^{K} n_g \log_e|n_g^{-1}A_g| \\ +np+2[kp+kp(p+1)/2]\,.$$

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

$$\text{(3.2.8)}\quad \text{AIC* (varying } \boldsymbol{\mu} \text{ and } \Sigma)=\sum_{g=1}^{K} n_g \log_e|A_g|+2[kp+kp(p+1)/2]\,,$$

where

$n_g$ =sample size of group or sample $g=1, 2, \cdots, K$,
$|A_g|$=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 (3.2.7).

## 4. Numerical examples of multi-sample cluster analysis on real data sets

In this section we shall give two different numerical examples of multi-sample cluster analysis, cluster the samples, and choose the best clusterings by using Akaike's Information Criterion (AIC) as derived in Sections 3.1 and 3.2. In Example 4.1 we shall also present the numerical results of using the $w$-square criterion as an alternative approach. We shall briefly discuss the relative merits of AIC over $w$-square criterion. One should note that these criteria are qualitatively and quantitatively different.

Our computations were carried out for all the examples we shall present here on an IBM 4341, configured as a 370, by using a newly developed statistical library software by the first author for multi-sample cluster analysis using AIC, called AICPARM.

We shall illustrate our results first on the Fisher [5] iris data.

*Example* 4.1. *Clustering of irises by groups:*
The iris data set is composed of 150 iris species belonging to three groups or species, namely *Iris setosa* ($S$), *Iris versicolor* ($Ve$), and *Iris virginica* ($Vi$) measured on sepal and petal length and width. Each group is represented by 50 plants.

This data set has been quite extensively studied in classification and cluster analysis since it was published by Fisher [5], and still today, is being used as a " testing ground " for classification and clustering methods proposed by many investigators such as Friedman and Rubin [6], Kendall [8], Solomon [15], Mezzich and Solomon [13], and many others, including the present authors.

For each of the 150 plants we already know the group structure of the iris species, namely $K=3$ groups or samples. Even though the two species, *Iris setosa* and *Iris versicolor* were found growing in the same colony, and *Iris virginica* was found growing in a different colony, Fisher reports in his linear discriminant analysis the separation of *I. setosa* completely from *I. versicolor* and *I. virginica*. Since then other investigators have shown similar results in their studies such as the ones we mentioned above.

With this in mind, we cluster $K=3$ samples (species) into $k=1$, 2, and 3 groups on the basis of all the four variables. We obtain in total five possible clustering alternatives. (In general, the total number of possibilities is a Stirling Number of the Second Kind; see, e.g., Abramowitz and Stegun [1]). Denoting *I. setosa* by $S$, *I. versicolor* by $Ve$, and *I. virginica* by $Vi$, we have $(S)$ $(Ve)$ $(Vi)$, $(S, Ve)$ $(Vi)$, $(S, Vi)$ $(Ve)$, $(Ve, Vi)$ $(S)$, and $(S, Ve, Vi)$ as the possible clustering alternatives. Using the MANOVA model and the multivariate model with varying parameters discussed in Sections 3.1 and 3.2 as our underlying models for clustering these three iris species, we obtained the following results.

Looking at Tables 4.1 and 4.2, we see that, using all four variables simultaneously under both models, the MAICE clustering is $(S)$ $(Ve)$ $(Vi)$. This indicates that indeed there are three types of species. Not surprisingly, the second minimum AIC occurs at the alternative submodel 4 $(Ve, Vi)$ $(S)$, under both models, telling us that if we were to cluster any one of the two iris groups, we should cluster *I. versicolor* and *I. virginica* together as one homogeneous group, and we should cluster *I. setosa* completely separately. We note that the AIC

Table 4.1. The AIC's for irises by groups on all variables under MANOVA model

| Alternative | Clustering | $n \log_e(2\pi)$ | $n \log_e \lvert n^{-1}W \rvert$ | $np$ | $k$ | $2m$ | AIC (common $\Sigma$) |
|---|---|---|---|---|---|---|---|
| 1 | $(S)$ $(Ve)$ $(Vi)$ | 1,102.724 | −1,504.2 | 600 | 3 | 44 | 242.524*a* |
| 2 | $(S, Ve)$ $(Vi)$ | 1,102.724 | −1,085.9 | 600 | 2 | 36 | 652.824 |
| 3 | $(S, Vi)$ $(Ve)$ | 1,102.724 | − 988.39 | 600 | 2 | 36 | 750.334 |
| 4 | $(Ve, Vi)$ $(S)$ | 1,102.724 | −1,299.6 | 600 | 2 | 36 | 439.124*b* |
| 5 | $(S, Ve, Vi)$ | 1,102.724 | − 941.73 | 600 | 1 | 28 | 788.994 |

$n=150$ plants, $p=4$ variables

$m=kp+p(p+1)/2$ parameters

AIC (common $\Sigma$)$=np \log_e(2\pi)+n \log_e |n^{-1}W|+np+2m$

*a*: First minimum AIC

*b*: Second minimum AIC

Table 4.2. The AIC's for irises by groups on all variables under the model with varying parameters

| Alternative | Clustering | $np \log_e(2\pi)$ | $\sum_{g=1}^{K} n_g \log_e \lvert n_g^{-1} A_g \rvert$ | $np$ | $k$ | $2m$ | AIC (varying $\mu$ and $\Sigma$) |
|---|---|---|---|---|---|---|---|
| 1 | $(S)$ $(Ve)$ $(Vi)$ | 1,102.724 | −1,653.895 | 600 | 3 | 84 | 132.829*a* |
| 2 | $(S, Ve)$ $(Vi)$ | 1,102.724 | −1,251.675 | 600 | 2 | 56 | 507.049 |
| 3 | $(S, Vi)$ $(Ve)$ | 1,102.724 | −1,144.480 | 600 | 2 | 56 | 614.244 |
| 4 | $(Ve, Vi)$ $(S)$ | 1,102.724 | −1,463.770 | 600 | 2 | 56 | 294.954*b* |
| 5 | $(S, Ve, Vi)$ | 1,102.724 | − 941.580 | 600 | 1 | 28 | 789.144 |

$n=150$ plants, $p=4$ variables

$m=kp+kp(p+1)/2$ parameters

AIC (varying $\mu$ and $\Sigma$)$=np \log_e(2\pi)+\sum_{g=1}^{K} n_g \log_e \lvert n_g^{-1}A_g \rvert+np+2m$

*a*: First minimum AIC

*b*: Second minimum AIC

values under submodel 2 and 3 are quite large indicating the inferiority of these submodels. We can see the effect of clustering *I. setosa* with *I. versicolor* in submodel 2, and also with *I. virginica* in submodel 3, by comparing the differences of AIC's in these submodels with that of submodel 4 in which *I. versicolor* and *I. virginica* were clustered together and *I. setosa* was clustered as a separate cluster on its own. According to AIC, we never cluster three iris species as one homogeneous group (submodel 5). Again by comparing the differences of AIC's of submodel 5 with that of submodels 4, 3, and 2, respectively, we can measure the amount of heterogeneity contributed by *I. setosa, I. versicolor* and *I. virginica,* respectively, in each clustering alternative under the MANOVA model and the multivariate model with varying mean vectors and covariance matrices. The larger this difference, the greater the heterogeneity or separation of that group or sample from that of homogeneous groups or samples in each clustering alternative.

In comparing the AIC's in Tables 4.1 and 4.2, we further notice that AIC (varying $\mu$ and $\Sigma$) values are much less than the AIC (common $\Sigma$) values for each of the clustering alternatives except for the last clustering alternative (i.e., alternative 5) in clustering the iris groups or species. Since according to the definition of AIC, the model with the minimum AIC is chosen to be the *best model,* then the above results suggest that when we are clustering iris data, and in general, we should use different covariance matrices rather than using equal covariance matrices.

We now present our results on the iris data by using the $w$-square criterion given by (1.1) in Section 1, when we assume equal covariance matrices between the iris groups or species. We should note here that in $w$-square criterion given by (1.1) and in Mardia et al. ([12], pp. 367), the estimated pooled-within groups covariance matrix of $\Sigma$ is computed only once across all the groups or samples to be clustered regardless of the number of clustering alternatives. In our version of $w$-square criterion we follow the same procedure, but we recompute the estimate of $\Sigma$ in each clustering alternative when we vary the number of clusters of groups or samples, $k$, when we are given, $K$, the number of groups or samples to be clustered. We do this both under the assumption of equal and separate covariance matrices between the iris groups. Therefore, our numerical values on $w$-square criterion are quite different then the original $w$-square criterion given in Mardia et al. [12], despite the fact that we get the same results.

We give the computational results as follows.

Table 4.3. The values of $w_a^2$ for irises by groups on all variables

| Alternative | Clustering | $w_a^2$ (common $\Sigma)^a$ | $w_a^2$ (common $\Sigma)^b$ | $w_a^2$ (varying $\Sigma)^c$ |
|---|---|---|---|---|
| 1 | $(S)$ $(Ve)$ $(Vi)$ | —* | —* | —* |
| 2 | $(S, Ve)$ $(Vi)$ | 2246.6046 | 137.9722 | 94.4149 |
| 3 | $(S, Vi)$ $(Ve)$ | 4484.6178 | 142.3345 | 96.0706 |
| 4 | $(Ve, Vi)$ $(S)$ | 430.0267** | 109.5511** | 76.8212** |
| 5 | $(S, Ve, Vi)$ | 4774.1661 | 175.2091 | 175.2091 |

$n=150$ plants, $p=4$ variables
$a$ : Original $w_a^2$ given in (1.1)
$b, c$: Our version of $w_a^2$
* : $w_a^2$ cannot be computed (always equal to zero)
** : Minimum of $w_a^2$

Hence, we interpret the results in Table 4.3 in the same manner as we did for AIC's. We see that at the alternative submodel 1, $w_a^2$ cannot be computed and is always equal to zero when the iris groups are clustered as singletons. This is always the case in general. Certainly this is a definite disadvantage of $w_a^2$ as compared to AIC which has a value even if the iris groups are clustered as singletons, so that AIC can aid us in determining and understanding the amount of heterogeneity or separation of the groups on a unique scale. The minimum of $w_a^2$ occurs at the alternative submodel 4, telling us again that, if we were to cluster any one of the two iris groups, we should cluster *I. versicolor* and *I. virginica* together as one homogeneous group, and we should cluster *I. setosa* completely separate as one heterogeneous group.

In short, $w$-square criterion gives the same results as AIC does, but as we mentioned in Section 1, it does not make any allowance for $m$, the number of parameters estimated within the clustering alternatives. AIC makes such an allowance to achieve a parsimony when we compare "the goodness of fit" of various models as we do in comparing different clustering alternatives. $W$-square criterion is short of having this important feature. Also as we saw, when we have singleton clusters, it cannot be computed.

Therefore, in our next example, we shall only give our results on AIC, since our purpose is to introduce AIC in this paper as a new approach to be used in evaluating multi-sample clusters.

*Example* 4.2. *Clustering graduate students by their classification groups:*

A data set for applicants to admission to a Graduate School of Business given in Johnson and Wichern ([9], pp. 528) is composed of data for 85 applicants who were classified by the admissions officer as *Admit* ($A$), *Not Admit* ($NA$), and *Borderline* ($B$), based on undergraduate grade point average (GPA) and graduate management aptitude test (GMAT) scores. The group sizes are $n_1=31$, $n_2=28$, and $n_3=26$ applicants.

With this in mind, we cluster $K=3$ groups of applicants into $k=$ 1, 2 and 3 homogeneous groups on the basis of the two variables. Using the MANOVA model and the multivariate model with varying parameters, our results are as follows.

Hence, looking at Tables 4.4 and 4.5, we see that, under both models, the first minimum AIC occurs at the alternative submodel 1, that is, when ($A$) ($NA$) ($B$) are all clustered separately. This indicates that indeed there are three groups of applicants. Therefore, the MAICE is submodel 1. The second minimum AIC occurs at the alternative submodel 4 again under both models, telling us that if we were to cluster any one of the two groups, then we should cluster *Borderline* ($B$) and *Not Admit* ($NA$) groups together as one homogeneous group, and we should cluster *Admit* ($A$) group completely separate as one heterogeneous group. On the other hand, if we want to make a third

Table 4.4. The AIC's for applicants by their classification groups under MANOVA model

| Alternative | Clustering | $n \log_e (2\pi)$ | $n \log_e \lvert n^{-1}W \rvert$ | $np$ | $k$ | $2m$ | AIC (common $\Sigma$) |
|---|---|---|---|---|---|---|---|
| 1 | $(A)\ (NA)\ (B)$ | 312.4391 | 406.1716 | 170 | 3 | 18 | 906.6107$a$ |
| 2 | $(A, NA)\ (B)$ | 312.4391 | 566.7477 | 170 | 2 | 14 | 1063.1868 |
| 3 | $(A, B)\ (NA)$ | 312.4391 | 491.7043 | 170 | 2 | 14 | 988.1434$c$ |
| 4 | $(B, NA)\ (A)$ | 312.4391 | 474.0420 | 170 | 2 | 14 | 970.4811$b$ |
| 5 | $(A, B, NA)$ | 312.4391 | 581.9931 | 170 | 1 | 10 | 1074.4322 |

$n=85$ applicants, $p=2$ variables

$m=kp+p(p+1)/2$ parameters

AIC (common $\Sigma$)$=np \log_e (2\pi)+n \log_e \lvert n^{-1}W \rvert+np+2m$

$a$: First minimum AIC

$b$: Second minimum AIC

$c$: Third minimum AIC

Table 4.5. The AIC's for applicants by their classification groups under the model with varying parameters

| Alternative | Clustering | $n \log_e (2\pi)$ | $\sum_{g=1}^{K} n_g \log_e \lvert n_g^{-1} A_g \rvert$ | $np$ | $k$ | $2m$ | AIC (varying $\mu$ and $\Sigma$) |
|---|---|---|---|---|---|---|---|
| 1 | $(A)\ (NA)\ (B)$ | 312.4391 | 388.7472 | 170 | 3 | 30 | 901.1863$a$ |
| 2 | $(A, NA)\ (B)$ | 312.4391 | 509.4198 | 170 | 2 | 20 | 1011.8589 |
| 3 | $(A, B)\ (NA)$ | 312.4391 | 480.2378 | 170 | 2 | 20 | 982.6769$c$ |
| 4 | $(B, NA)\ (A)$ | 312.4391 | 465.7116 | 170 | 2 | 20 | 968.1507$b$ |
| 5 | $(A, B, NA)$ | 312.4391 | 581.9931 | 170 | 1 | 10 | 1074.4332 |

$n=85$ applicants, $p=2$ variables

$m=kp+kp(p+1)/2$ parameters

AIC (varying $\mu$ and $\Sigma$)$=np \log_e (2\pi)+\sum_{g=1}^{K} n_g \log_e \lvert n_g^{-1} A_g \rvert + np + 2m$

$a$: First minimum AIC

$b$: Second minimum AIC

$c$: Third minimum AIC

choice, then the third minimum of AIC occurs at the alternative submodel 3, indicating to us the closeness of the *Admit* ($A$) group to the *Borderline* ($B$) group as one homogeneous cluster, and leaving *Not Admit* ($NA$) group on its own as a singleton cluster. Therefore, this way, we can check the significance of each of the clustering alternatives in the decision making process. In this example, we also never cluster the three groups as one homogeneous group (submodel 5).

In comparing the AIC's in Tables 4.4 and 4.5 for this example we also notice that, AIC (varying $\mu$ and $\Sigma$) values are less than the AIC (common $\Sigma$) values for each of the clustering alternatives except for the last clustering alternative (i.e., alternative 5) in clustering the applicant groups. These results suggest that we should use different covariance matrices. However, the values of AIC (varying $\mu$ and $\Sigma$) and AIC (common $\Sigma$) are significantly closer to one another that if we were to assume equal covariance matrices between the applicant groups *a priori*, it would not have been a dubious assumption for this particular data set.

Thus, it should be noted that via AIC we can now easily check the validity of our assumptions in terms of using equal covariance matrices as opposed to separate covariance matrices in a particular data set which is important in the multi-sample clustering situation, and in general.

## 5. Conclusions and discussion

From our numerical results in Section 4, we see that AIC and consequently minimum AIC procedures can indeed successfully identify the best clustering alternatives when we cluster samples into homogeneous sets of samples both in the MANOVA model and the multivariate model with varying covariance matrices. We can measure the amount of homogeneity and heterogeneity in clustering samples. We can determine *a priori* whether we should use equal or varying covariance matrices in the analysis of a data set.

The fact that AIC does not require the table look-up, which is the case in conventional procedures, adds to the importance of the results obtained. This is one of the important virtues that AIC breaks away from conventional procedures which try to test whether a parameter is "significant" or not using a significance level $\alpha$ which is essentially arbitrary. The other important virtue of AIC is that the penalty represented by the term 2×(number of free parameters) clearly demonstrates the necessity of choosing a class of models, at least one of which will be able to provide a good approximation to the distribution of data without adjusting too many parameters.

Thus, in concluding, we see that the use of AIC shows how to combine the information in the likelihood with an appropriate function of the number of parameters to obtain estimates of the information provided by competing alternative models. Therefore, the definition of MAICE gives a clear formulation of the principle of parsimony in statistical model building or comparison as we demonstrated by numerical examples. And MAICE provides a versatile procedure for statistical model identification which is free from the ambiguities inherent in the application of conventional statistical procedures.

## Acknowledgements

This paper is a summary of some of the results in the Ph.D. thesis of the first author in the Department of Mathematics at the University of Illinois at Chicago, under the supervision of the second author. The first author is grateful to Professor Stanley L. Sclove for his valuable and useful comments. The suggestions of the Associate Editor, Dr. Noboru Ohsumi, and the referees were very helpful and are gratefully acknowledged. Appreciation is also extended to Mr. Ronald Ball for his programming assistance.

UNIVERSITY OF ILLINOIS*

## REFERENCES

[1] Ambramowitz, M., and Stegun, I. A. (1968). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* (Nat. Bur. of Stand. Appl. Math. Ser., No. 55),

* Now at University of Virginia.

7th printing, U.S. Govt. Printing Office, Washington, D.C.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory*, (eds. B. N. Petrov and F. Csaki), Akademiai Kiado, Budapest, 267-281.

[3] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, **19**, 716-723.

[4] Consul, P. C. (1969). The exact distributions of likelihood criteria for different hypotheses, *Multivariate Analysis* II, (ed. P. R. Krishnaiah), Academic Press, New York.

[5] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**, 179-188.

[6] Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data, *J. Amer. Statist. Ass.*, **62**, 1159-1178.

[7] Gabriel, K. R. (1969). A comparison of some methods of simultaneous inference in MANOVA, *Multivariate Analysis* II, (ed. P. R. Krishnaiah), Academic Press, New York.

[8] Kendall, M. G. (1966). Discrimination and classification, *Muitivariate Analysis*, (ed. P. R. Krishnaiah), Academic Press, New York.

[9] Johnson, R. A. and Wichern, D. W. (1982). *Applied Multivariate Statistical-Analysis*, Prentice-Hall, Englewood Cliffs.

[10] Krishnaiah, P. R. (1969). Simultaneous test procedures and general MANOVA models, *Multivariate Analysis* II, (ed. P. R. Krishnaiah), Academic Press, New York.

[11] Krishnaiah, P. R. (1979). Some developments on simultaneous test procedures, *Developments in Statistics*, (ed. P. R. Krishnaiah), Vol. 2, Academic Press, New York.

[12] Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, New York.

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

[14] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

[15] Solomon, H. (1971). Numerical taxonomy, *Mathematics in the Archaeological and Historical Sciences*, Edinburgh University Press, Edinburgh, 62-81.

[16] Srivastava, J. N. (1969). Some studies on intersection tests in multivariate analysis of variance, *Multivariate Analysis* II, (ed. P. R. Krishnaiah), Academic Press, New York.

[17] Wilks, S. S. (1932). Certain generalizations in the analysis of variance, *Biometrika*, **24**, 471-494.

Ann. Inst. Statist. Math.
36 (1984), Part B, 539

# CORRECTION TO "MULTI-SAMPLE CLUSTER ANALYSIS USING AKAIKE'S INFORMATION CRITERION"

HAMPARSUM BOZDOGAN AND STANLEY L. SCLOVE

(This volume, pp. 163-180)

The affiliations of the authors were incorrectly listed. The correct affiliations are as follows:

| | | |
|---|---|---|
| Hamparsum Bozdogan* | and | Stanley L. Sclove** |
| * University of Virginia | | ** University of Illinois at Chicago |

The Editor deeply regrets for this editorial mistake.