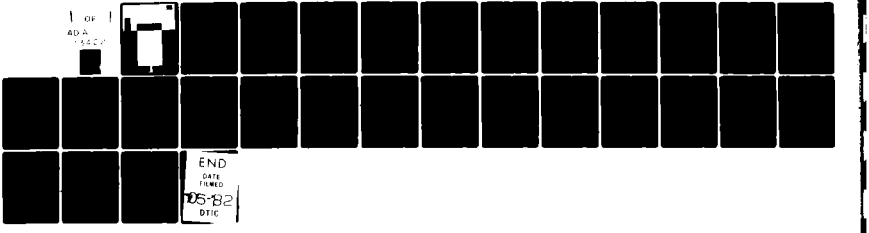
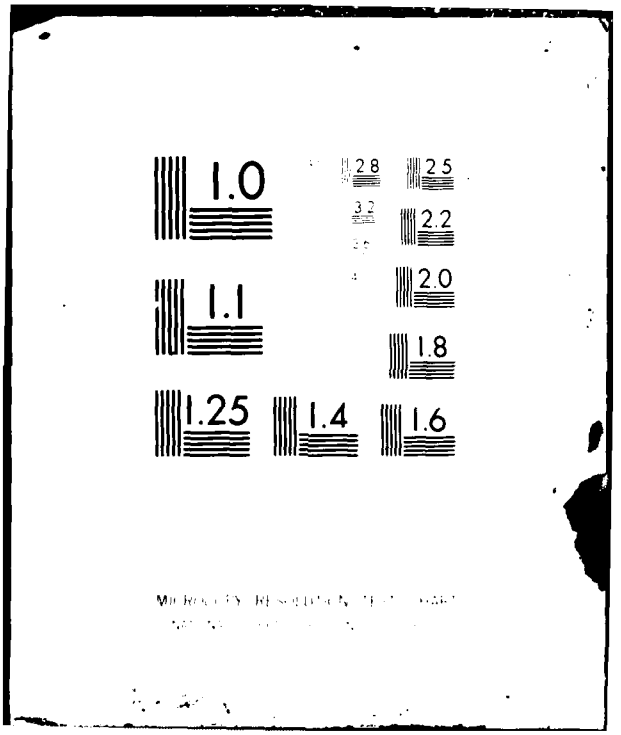
AD-A113 422 ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE --ETC F/G 12/1
MULTI-SAMPLE CLUSTER ANALYSIS WITH VARYING PARAMETERS USING AKA--ETC(U)
MAR 82 H BOZDOGAN, S L SCLOVE N00014-80-C-0408
UNCLASSIFIED TR-82-2 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2

College of Business Administration, University of Illinois at Chicago Circle

UICC

MULTI-SAMPLE CLUSTER ANALYSIS WITH VARYING PARAMETERS USING AKAIKE'S INFORMATION CRITERION*

by

HAMPARSUM BOZDOGAN
Quantitative Methods Department
University of Illinois at Chicago Circle

and

STANLEY L. SCLOVE
Departments of Mathematics and Quantitative Methods
University of Illinois at Chicago Circle

TECHNICAL REPORT NO. 82-2
March 8, 1982

PREPARED FOR THE
OFFICE OF NAVAL RESEARCH
UNDER
CONTRACT N00014-80-C-0408,
TASK NR042-443
with the University of Illinois at Chicago Circle

Principal Investigator: Stanley L. Sclove

QUANTITATIVE METHODS DEPARTMENT
UNIVERSITY OF ILLINOIS AT CHICAGO CIRCLE
CHICAGO, ILLINOIS 60680

---

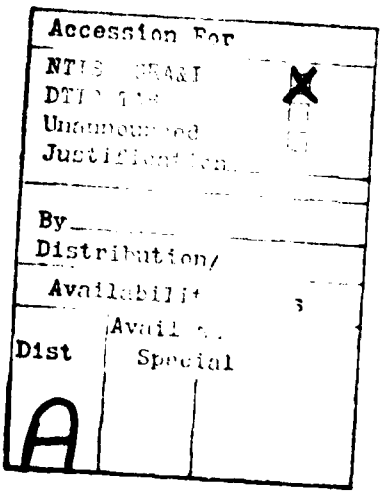
*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

VV88

MULTI-SAMPLE CLUSTER ANALYSIS WITH VARYING PARAMETERS USING AKAIKE'S INFORMATION CRITERION*

Hamparsum Bozdogan and Stanley L. Sclove
University of Illinois at Chicago Circle

CONTENTS



*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

MULTI-SAMPLE CLUSTER ANALYSIS WITH VARYING PARAMETERS USING AKAIKE'S INFORMATION CRITERION*

Hamparsum Bozdogan and Stanley L. Sclove
University of Illinois at Chicago Circle

ABSTRACT

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 the univariate model with varying means and variances, and in the multivariate model with varying mean vectors and variance-covariance matrices. Numerical examples are presented and results are shown to demonstrate the utility of AIC in identifying the best clustering alternatives.

Key Words and Phrases: Multi-sample cluster analysis; Akaike's Information Criterion (AIC); Univariate model with varying means and variances, Multivariate model with varying mean vectors and variance-covariance matrices; maximum likelihood.

---

*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

MULTI-SAMPLE CLUSTER ANALYSIS WITH VARYING PARAMETERS USING AKAIKE'S INFORMATION CRITERION*

Hamparsum Bozdogan and Stanley L. Sclove
University of Illinois at Chicago Circle

1. Introduction

In a previous paper, we introduced and developed Akaike's Information Criterion (AIC) for multi-sample cluster analysis, the problem of grouping samples. We derived the form of AIC in both univariate and multivariate analysis of variance models, where the assumptions of independence, univariate and multivariate normality, equal variances and variance-covariance matrices were fundamental to the analysis. We gave numerical examples and results which demonstrated the utility of AIC in identifying the best clustering alternatives. (See, Bozdogan and Sclove [5]).

In this paper, we shall continue to study the multi-sample cluster problem. However, here we shall develop Akaike's Information Criterion (AIC) for multi-sample cluster analysis with varying means and variances in the univariate model, and with varying mean vectors and variance-covariance matrices in the multivariate model, since often in practice the assumption of equal parameters within the model is a rather dubious requirement.

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 in which parameters might vary. For this reason it is reasonable to provide a practically useful statistical procedure that would use some sort of statistical model to aid in comparisons of various collections of samples, identify homogeneous groups of samples, and tell us which should be clustered together and which samples

---

*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

should not. For examples of multi-sample clustering situations, we refer the reader to Bozdogan and Sclove [5].

In the statistical literature several conventional test procedures are available for testing whether or not several populations have equal variances, as required by the analysis of variance (ANOVA) model. If we have a reason to doubt this is the case, then we may first want to test the equality of variances. In the multivariate case the equality of covariance matrices is certainly more hazardous. Therefore, for this reason we may want first to test the equality of variances in the univariate case, and the equality of covariance matrices in the multivariate case. This is an important option to use in clustering groups or samples, and in general.

In the literature, however, there are several test procedures for testing the equality of variances, and covariance matrices. For example, in the multivariate case, the most commonly used test is Box's M test despite the fact that it is very expensive to compute it on a high speed computer, even on an IBM 370. Moreover, as in the case of MANOVA, these test procedures are not revealing or informative in multi-sample clustering problems. Therefore, in this paper we shall propose again Akaike's Information Criterion (AIC) as a new procedure for comparing the clusters, and use it to identify the best clustering alternatives.

In 1971, Akaike first introduced an information criterion, referred to as an Automatic (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

(1.1) AIC = -2 $\log_e$ (maximized likelihood)

+2 (number of independently adjusted parameters within the model).

It was obtained with the aid of an information theoretic interpretation of the method of maximum likelihood by Akaike ([2], [3]). 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 (1.1) stands for the penalty of badness of fit or downward bias 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. The AIC avoids this spurious improvement 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.

In Section 2, we shall define the general multi-sample cluster problem, and in Section 3, we shall briefly discuss the number of clustering alternatives for a given K groups or samples into k nonempty clusters. In the

subsequent sections, that is, in Section 4 and in 5, we shall derive the AIC procedure for the univariate model with varying means and variances, and for the multivariate model with varying mean vectors and covariance matrices. In Section 6, we shall give numerical examples for both univariate and multivariate multi-sample cluster analysis on the same real data set to demonstrate our results of AIC and minimum AIC procedures obtained from different computer analyses.

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

$$\underset{\sim}{X}\ (n \times p) = \begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{X}_K \end{bmatrix} \tag{2.1}$$

be a single data matrix of K groups or samples, where $\underline{X}_g$ $(n_g \times p)$ represents the observations from the g-th group or sample, g=1,2,...,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 $k \leq K$.

Often individuals or objects have been sampled from K>1 populations. For multi-samples or multiple groups of individuals or objects 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,...,K. The n_g are not

restricted to being equal or proportional to other n_g's. The total number of observations is $n = \sum_{g=1}^{K} n_g$. Let X_{gi} be the px1 vector of observations in group g=1,2,...,K, and for individual i=1,2,...,n_g.

3. The Number of Clustering Alternatives for a Given K Samples into k Nonempty Clusters

In this section, we shall just briefly mention how to obtain the total number of clustering alternatives for a given K, the number of groups or samples. For details we again refer the reader to Bozdogan and Sclove [5].

In general, the total number of ways of clustering K groups or samples into k clusters is given by

$$(3.1) \qquad S(K,k) = \frac{1}{k!} \sum_{g=0}^{k} \binom{k}{g} (-1)^g (k-g)^K$$

which is known as the Stirling Number of the Second Kind (see, e.g., Abramowitz and Stegun [1]) and also called the number of clustering alternatives.

If k, the number of clusters of groups or 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 varies, then the total number of clustering alternatives for a given K, the number of groups or samples, is given by

$$(3.2) \qquad \sum_{k=1}^{K} S(K,k) \quad .$$

For example, K=4 samples, if k is not specified a priori and varies, then there are in total 15 possible clustering alternatives to cluster K=4 groups or samples first into k=4 groups or samples, then k=3 groups or samples, k=2 groups or samples, and k=1 group or sample by using the equation (3.1)

respectively, and summing the results by using the expression (3.2) to obtain 15 as the total number of possible clustering alternatives.

Therefore, the total number of ways of clustering K groups or samples into k homogeneous groups or samples is given by equation (3.1), and the total number of possible clustering alternatives is given by the expression (3.2).

4. AIC For The Univariate Model

We now turn our attention to consider situations with several univariate normal samples. For example, we may have multi-sample data with samples of sizes $n_1, n_2, \ldots, n_K$ which are assumed to have come from K populations, the first with mean μ_1 and variance σ_1^2, the second with mean μ_2 and variance $\sigma_2^2, \ldots$, the Kth with mean μ_K and variance σ_K^2. We may want to decide in this case if the variances of these K samples will be treated as equal or not, given no restriction on the population means. In terms of the parameters the univariate model is $\underset{\sim}{\theta} = (\mu_1, \mu_2, \ldots, \mu_K, \sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2)$ with m=2k parameters, where k is the number of groups.

Recall the definition of AIC from Section 1,

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

where m denotes the number of independently adjusted parameters within the model.

Suppose there are K independent samples of independent observations, with n_g, g=1,2,...,K, observations in the g-th group and $n = \sum_{g=1}^{K} n_g$. Denote the unknown means of the groups by $\mu_1, \mu_2, \ldots, \mu_K$, and the unknown variances of the groups by $\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2$. Assume that the samples $(z_{11}, z_{12}, \ldots, z_{1n_1}; \ldots;$

$z_{K_1},\ldots,z_{Kn_K}$) are drawn randomly from K populations which are $N(\mu_g,\sigma_g^2)$. The basic null hypothesis of interest is given by

$$(4.1) \qquad H_0: \ \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2.$$

The alternative hypothesis is given by

H_1: Not all K variances are equal.

In the statistical literature, this is also known as the test of homogeneity of variances or Bartlett's test.

To derive the form of AIC subject to this constraint, we call the common unknown value of variances σ^2. The likelihood function in this case is given by

$$(4.2) \qquad L(\{\mu_g,\sigma_g^2\};\underset{\sim}{z}) = (2\pi)^{-n/2} \prod_{g=1}^{K} (\sigma_g^2)^{-n_g/2} \exp\{-\sum_{g=1}^{K} (\frac{1}{2\sigma_g^2}) \sum_{i=1}^{n_g} (z_{gi} - \mu_g)^2.$$

The log likelihood is

$$(4.3) \qquad l(\{\mu_g,\sigma^2\}) \equiv \log L(\{\mu_g,\sigma_g^2\};\underset{\sim}{z})$$

$$= -\frac{n}{2}\log(2\pi) - 1/2 \sum_{g=1}^{K} n_g \log \sigma_g^2 - \sum_{g=1}^{K} \frac{1}{2\sigma_g^2} \sum_{i=1}^{n_g} (z_{gi} - \mu_g)^2,$$

and the MLE's of μ_g and σ_g^2 are given by

$$(4.4) \qquad \hat{\mu}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} z_{gi} = \bar{z}_{g\cdot},$$

and

$$(4.5) \qquad \hat{\sigma}_g^2 = \frac{1}{n} \sum_{i=1}^{n_g} (z_{gi} - \bar{z}_{g.})^2 = s_g^2 \ , \ g=1,2,\ldots,K.$$

Substituting these back into (4.3) and simplifying, the maximized log likelihood becomes

$$(4.6) \qquad l(\{\hat{\mu}_g,\hat{\sigma}_g^2\};\underset{\sim}{z}) \equiv \log L(\{\hat{\mu}_g,\hat{\sigma}_g^2\};\underset{\sim}{z})$$

$$= -\frac{n}{2} \log(2\pi) - 1/2 \sum_{g=1}^{K} n_g \log s_g^2 - \frac{n}{2} \ .$$

Since

$$(4.7) \qquad AIC = -2 \log_e L(\hat{\underset{\sim}{\theta}}) + 2m,$$

where m is the number of parameters, and since

$$(4.8) \qquad -2 \log L(\{\hat{\mu}_g,\hat{\sigma}_g^2\}) = n \log(2\pi) + \sum_{g=1}^{K} n_g \log s_g^2 + n,$$

then AIC becomes

$$(4.9) \qquad AIC \text{ (varying } \mu \text{ and } \sigma^2) = n \log(2\pi) + \sum_{g=1}^{K} n_g \log s_g^2 + n + 2(2k).$$

Since the constants do not affect the result of comparison of models, we could ignore them and use the simplified version

$$(4.10) \qquad AIC^* = \text{(varying } \mu \text{ and } \sigma^2) = \sum_{g=1}^{K} n_g \log_e S_g^2 + 2(2k)$$

where $\quad S_g^2 = \frac{1}{n_g} \sum_{i=1}^{n_g} (z_{gi} - \bar{z}_{g.})^2, \ g=1,2,\ldots,K,$

k = number of groups or samples compared, or the number of independently adjusted parameters within the model.

However, for purposes of comparison we retain the constants and use AIC given by (4.9).

5. AIC For the Multivariate Model

As we mentioned in Section 1, that the assumption of equality of variances in one-way ANOVA, causes serious problems when we are testing the equality of several means. Parallel to this assumption, in the multivariate case the equality of covariance matrices even causes more serious problems. For this reason we may want first to test the equality of covariance matrices against the alternative that not all covariance matrices are equal. Therefore, throughout this section we shall suppose that we may have independent data matrices $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_K$, where the rows of $\underset{\sim}{X}_g$ $(n_g \times p)$ are independent and identically distributed (i.i.d.) $N_p(\underset{\sim}{\mu}_g, \underset{\sim}{\Sigma}_g)$, $g=1,2,\ldots,K$. In terms of the parameters the multivariate model we shall consider is

$$\underset{\sim}{\theta} = (\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots, \underset{\sim}{\mu}_K, \underset{\sim}{\Sigma}_1, \underset{\sim}{\Sigma}_2, \ldots, \underset{\sim}{\Sigma}_K)$$

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

$$(5.1) \qquad H_0: \ \underset{\sim}{\Sigma}_1 = \underset{\sim}{\Sigma}_2 = \ldots = \underset{\sim}{\Sigma}_K.$$

The alternative hypothesis is given by

H_1: Not all K 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

$$(5.2)\qquad l(\{\underline{\mu}_g,\underline{\Sigma}_g\};\underline{Z}) \equiv \log L(\{\underline{\mu}_g,\underline{\Sigma}_g\};\underline{Z})$$

$$= -\frac{np}{2}\log(2\pi) - 1/2\sum_{g=1}^{K} n_g \log|\underline{\Sigma}_g| - 1/2\sum_{g=1}^{K} n_g \mathrm{tr}\underline{\Sigma}_g^{-1}\underline{A}_g$$

$$- 1/2\sum_{g=1}^{K} n_g(\bar{\underline{z}}_g - \underline{\mu}_g)'(\bar{\underline{z}}_g - \underline{\mu}_g) .$$

The MLE's of $\underline{\mu}_g$ and $\underline{\Sigma}_g$ are

$$(5.3)\qquad \hat{\underline{\mu}}_g = \bar{\underline{z}}_g \ , \ g=1,2,\ldots,K,$$

and

$$(5.4)\qquad \hat{\underline{\Sigma}}_g = \underline{A}_g/n_g.$$

Substituting these back into (5.2) and simplifying, the maximized log likelihood becomes

$$(5.5)\qquad l(\{\hat{\underline{\mu}}_g,\hat{\underline{\Sigma}}_g\};\underline{Z}) \equiv \log L(\{\hat{\underline{\mu}}_g,\hat{\underline{\Sigma}}_g\};\underline{Z})$$

$$= -\frac{np}{2}\log(2\pi) - 1/2\sum_{g=1}^{K} n_g \log|n_g^{-1}\underline{A}_g| - \frac{np}{2} .$$

Since

$$(5.6) \qquad AIC = -2 \log_e L(\hat{\underset{\sim}{\theta}}) + 2m ,$$

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

$$(5.7) \qquad AIC(\text{varying } \underset{\sim}{\mu} \text{ and } \underset{\sim}{\Sigma}) = np\log(2\pi) + \sum_{g=1}^{K} n_g \log|n_g^{-1}\underset{\sim}{A}_g| + np + 2[kp + kp(p+1)/2].$$

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

$$(5.8) \qquad AIC^*(\text{varying } \underset{\sim}{\mu} \text{ and } \underset{\sim}{\Sigma}) = \sum_{g=1}^{K} n_g \log_e|\underset{\sim}{A}_g| + 2[kp + kp(p+1)/2],$$

where n_g = sample size of group or sample $g=1,2,\ldots K$,

$|\underset{\sim}{A}_g|$ = the determinant of sum of squares and cross-products (SSCP) matrix for group or sample $g=1,2,\ldots,K$,

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 (5.7).

6. Numerical Examples of Multi-Sample Cluster Analysis on Fisher Iris Data

In this section we shall give numerical examples of both univariate and multivariate multi-sample data, and cluster the groups or samples, and choose the best clusterings by using Akaike's Information Criterion (AIC) as derived in Sections 4 and 5.

Our computations were carried out for all the examples we shall present here on an IBM 370, using various statistical software packages such as

MINITAB, SPSS, and SPEAKEASY (VM/CMS version).

6.1. A Univariate Examples

For the univariate numerical examples we shall illustrate our results on Fisher [6] iris data.

Example 6.1. 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. The data set for the 150 irises are given in Table 6.1.

This data set has been quite extensively studied in classification and cluster analysis since it was published by Fisher [6], and still today, is being used as a "testing ground" for classification and clustering methods proposed by many investigators such as Friedman and Rubin [7], Kendall [8], Solomon [10], Mezzich and Solomon [9], 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, let us take K=3 groups or species on each of the variables separately and cluster them into k=1,2, and 3 homogeneous groups. Since we are dealing with K=3 groups, by using equation (3.1) and the

TABLE 6.1 MEASUREMENTS ON THREE TYPES OF IRIS

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.5	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

expression (3.2) in Section 3, we obtain in total 5 possible clustering alternatives. 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 all the possible clustering alternatives of three iris species. In terms of the parameters, using the univariate model $\underline{\theta} = (\mu_1,\mu_2,\ldots,\mu_K,\sigma_1^2,\sigma_2^2,\ldots,\sigma_K^2)$ as our underlying model with varying means and variances for clustering the iris groups, from a simple run on the computer by using MINITAB package, we obtained the AIC's for each of the 5 clustering alternatives of each of the four variables separately. We report our results on each of the four variables, respectively, as follows.

TABLE 6.2. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE SEPAL LENGTH

Alternative	Clustering	$n\log_e(2\pi)$	$\sum_{g=1}^{K} n_g \log_e S_g^2$	n	k	2(2k)	AIC
1	(S) (Ve) (Vi)	275.681	-218.710	150	3	12	218.971[a]
2	(S, Ve) (Vi)	275.681	-136.019	150	2	8	297.662
3	(S, Vi) (Ve)	275.681	- 79.394	150	2	8	354.287
4	(Ve, Vi) (S)	275.681	-188.536	150	2	8	245.145[b]
5	(S, Ve, Vi)	275.681	- 57.603	150	1	4	372.078

TABLE 6.3. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE SEPAL WIDTH

Alternative	Clustering	$n\log_e(2\pi)$	$\sum_{g=1}^{K} n_g \log_e S_g^2$	n	k	2(2k)	AIC
1	(S) (Ve) (Vi)	275.681	-329.102	150	3	12	108.579[a]
2	(S, Ve) (Vi)	275.681	-262.503	150	2	8	171.178
3	(S, Vi) (Ve)	275.681	-292.416	150	2	8	141.265
4	(Ve, Vi) (S)	275.681	-319.093	150	2	8	114.588[b]
5	(S, Ve, Vi)	275.681	-250.132	150	1	4	179.549

TABLE 6.4. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE PETAL LENGTH

Alternative	Clustering	$n\log_e(2\pi)$	$\sum_{g=1}^{K} n_g \log_e S_g^2$	n	k	2(2k)	AIC
1	(S) (Ve) (Vi)	275.681	-313.055	150	3	12	124.626[a]
2	(S, Ve) (Vi)	275.681	12.795	150	2	8	446.476
3	(S, Vi) (Ve)	275.681	70.394	150	2	8	504.075
4	(Ve, Vi) (S)	275.681	-215.414	150	2	8	218.267[b]
5	(S, Ve, Vi)	275.681	169.888	150	1	4	599.569

TABLE 6.5. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE PETAL WIDTH

Alternative	Clustering	$n\log_e(2\pi)$	$\sum_{g=1}^{K} n_g \log_e S_g^2$	n	k	2(2k)	AIC
1	(S) (Ve) (Vi)	275.681	-519.344	150	3	12	-81.663[a]
2	(S, Ve) (Vi)	275.681	-245.374	150	2	8	188.307
3	(S, Vi) (Ve)	275.681	-181.176	150	2	8	252.505
4	(Ve, Vi) (S)	275.681	-398.271	150	2	8	35.410[b]
5	(S, Ve, Vi)	275.681	- 82.454	150	1	4	347.227

$$\text{AIC(varying } \mu \text{ and } \sigma) = n\log_e(2\pi) + \sum_{g=1}^{K} n_g \log_e S_g^2 + n + 2(2k)$$

[a]First Minimum AIC

[b]Second Minimum AIC

Looking at each of the tables above, we see that on each of the variables the first minimum AIC occurs at the alternative submodel 1, namely (S) (Ve) (Vi). That is, the MAICE is submodel 1 indicating that indeed there are three types of species across all the variables. But the second minimum AIC is at the alternative submodel 4 again across all the variables indicating that if we were to cluster any iris species, we should cluster I. versicolor and I. virginica together, as one homogeneous group.

Thus our minimum AIC results for each of the variables confirm other investigators' findings, including Fisher's results on the iris data. Moreover, if we

were to choose among the submodels then we would choose the one with smallest minimum AIC as the best submodel. Examining the Tables 6.2, 6.3, 6.4, and 6.5, we see that the smallest minimum AIC occurs at the submodel 1 in Table 6.5 on variable petal width. This indicates that petal width alone separates the three iris species with virtual certainty, confirming again Fisher's results (see, e.g., Fisher [6]).

Hence, we note here that we are clustering the irises by groups or species under a more general model rather than using the ANOVA model as our underlying model which we cosidered in a previous paper on multi-sample cluster analysis.

6.2. A Multivariate Example

Now we consider Fisher iris data again and this time we cluster K=3 groups or species into k=1, 2, and 3 homogeneous groups on the basis of all the four variables, assuming the multivariate model given in terms of the parameters $\underset{\sim}{\theta} = (\mu_1, \mu_2, \ldots, \underset{\sim}{\Sigma}_1, \underset{\sim}{\Sigma}_2, \ldots, \underset{\sim}{\Sigma}_K)$ as the underlying model for clustering these three iris groups or species. On the iris data, running SPSS MANOVA program, we obtain the following sum of squares and products (SSCP) matrices for each of the clustering alternatives. These are:

(1) (S) (VE) (VI)

$$\underset{\sim}{A}_{(S)} = \begin{bmatrix} 6.0882 & 4.8616 & .8014 & .5062 \\ 4.8616 & 7.0408 & .5732 & .4556 \\ .8014 & .5732 & 1.4778 & .2974 \\ .5062 & .4556 & .2974 & .5442 \end{bmatrix}$$

$$|50\ \underset{\sim}{A}^{-1}_{(S)}| = 1.949E\text{-}6$$

$$\log_e (1.949E\text{-}6) = -13.148$$

$$\underline{A}_{(VE)} = \begin{bmatrix} 13.561 & 4.362 & 9.066 & 2.7436 \\ 4.362 & 4.825 & 4.05 & 2.019 \\ 9.066 & 4.05 & 10.82 & 3.582 \\ 2.7436 & 2.019 & 3.582 & 1.9162 \end{bmatrix}$$

$$|50^{-1}\ \underline{A}_{(VE)}| = 1.8053E\text{-}5$$

$$\log_e (1.8053E\text{-}5) = -10.922$$

$$\underline{A}_{(VI)} = \begin{bmatrix} 19.813 & 4.5944 & 14.861 & 2.4056 \\ 4.5944 & 5.0962 & 3.4976 & 2.3338 \\ 14.861 & 3.4976 & 14.925 & 2.3924 \\ 2.4056 & 2.3338 & 2.3924 & 3.6962 \end{bmatrix}$$

$$|50^{-1}\ \underline{A}_{(VI)}| = 1.2244E\text{-}4$$

$$\log_e (1.2244E\text{-}4) = -9.0079$$

(2) (S, VE) (VI)

$$\underline{A}_{(S,\ VE)} = \begin{bmatrix} 40.901 & -5.9433 & 74.361 & 28.144 \\ -5.9433 & 22.69 & -41.404 & -15.291 \\ 74.361 & -41.404 & 208.02 & 79.425 \\ 28.144 & -15.291 & 79.425 & 31.62 \end{bmatrix}$$

$$|100^{-1}\ \underline{A}_{(S,\ VE)}| = 3.3118E\text{-}4$$

$$\log_e (3.3118E\text{-}4) = -8.0128$$

(3) (S, VI) (VE)

$$\underline{A}_{(S,\ VI)} = \begin{bmatrix} 88.469 & -8.4997 & 177.42 & 73.311 \\ -8.4997 & 17.29 & -42.351 & -17.414 \\ 177.42 & -42.351 & 434.61 & 184.69 \\ 73.311 & -17.414 & 184.69 & 83.45 \end{bmatrix}$$

$$|100^{-1}\ \underline{A}_{(S,\ VI)}| = .0025193$$

$$\log_e (.0025193) = -5.9838$$

(4) (VE, VI) (S)

$$\underline{A}_{(VE,\ VI)} = \begin{bmatrix} 44.264 & 12.322 & 45.245 & 16.699 \\ 12.322 & 10.962 & 14.137 & 7.9228 \\ 45.245 & 14.137 & 67.476 & 28.584 \\ 16.699 & 7.9228 & 28.584 & 17.862 \end{bmatrix}$$

$$|100\ \underline{A}^{-1}_{(VE,\ VI)}| = 3.1476E\text{-}4$$

$$\log_e (3.1476E\text{-}4) = -8.0637$$

(5) (S, VE, VI)

$$\underline{A}_{(S,\ VE,\ VI)} = \begin{bmatrix} 102.6 & -6.0197 & 189.78 & 76.884 \\ -6.0197 & 28.307 & -49.119 & -18.124 \\ 189.78 & -49.119 & 464.33 & 193.05 \\ 76.884 & -18.124 & 193.05 & 86.57 \end{bmatrix}$$

$$|150\ \underline{A}^{-1}_{(S,VE,VI)}| = .0018787$$

$$\log_e (.0018787) = -6.2772$$

After carrying out all our computations for each of the clustering alternatives (using the Matrix Algebra Routines in the SPEAKEASY interactive computer package), we obtain the AIC's from (5.7). The results are shown in Table 6.6.

TABLE 6.6. THE AIC'S FOR IRISES BY GROUPS ON ALL VARIABLES

Alternative	Clustering	$np\log_e(2\pi)$	$\sum_{g=1}^{K} n_g \log_e \lvert n_g^{-1} \underline{A}_g \rvert$	np	k	2m	AIC
1	(S) (Ve) (Vi)	1,102.724	-1,653.89 5	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 $\underline{\mu}$ and $\underline{\Sigma}$) = $np\log_e (2\pi) + \sum_{g=1}^{K} n_g \log_e \lvert n_g^{-1} \underline{A}_g \rvert + np + 2m$

[a]First Minimum AIC

[b]Second Minimum AIC

Hence, looking at the Table 6.6, we see that, using all four variables simultaneously the first minimum AIC occurs at the alternative submodel 1, that is, when (S) (Ve) (Vi) are all clustered separately. This indicates again that indeed there are three types of species. Therefore, the MAICE is submodel 1. Not surprisingly, the second minimum AIC occurs at the alternative submodel 4 telling us that if we were to cluster any one of the two iris groups, we should cluster I. veriscolor and I. virginica together as one homogeneous group, and we should cluster I. setosa completely separate as one heterogeneous group.

Here, it is important to note that we obtained also the same results when we used the four variables separately in our computation of AIC in the previous section, which is encouraging.

7. Conclusions and Discussion

From our numerical results in Section 6, we see that AIC and consequently minimum AIC procedures can successfully indeed identify the best clustering alternatives when we cluster samples into homogeneous sets of samples both in the univariate and the multivariate models with varying parameters.

In our previous paper on multi-sample cluster analysis (Bozdogan and Sclove [5]), we considered ANOVA and MANOVA as our two underlying models where the assumption of equal variances and covariances were used to cluster the groups or samples for multi-sample data. There, we used AIC also in identifying the best clustering alternatives in clustering the iris groups or species. We obtained the same results in determining the three types of iris species and 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 summarizing the results of AIC-values for the multivariate case only from the previous and this paper, we obtain the following table.

TABLE 6.7. THE AIC'S FOR IRISES BY GROUPS ON ALL VARIABLES UNDER TWO MULTIVARIATE MODELS

Alternative	Clustering	AIC(varying $\underset{\sim}{\mu}$ and $\underset{\sim}{\Sigma}$)	AIC(common $\underset{\sim}{\Sigma}$)
1	(S) (Ve) (Vi)	132.829[a]	242.524[a]
2	(S, Ve) (Vi)	507.049	652.824
3	(S, Vi) (Ve)	614.244	750.334
4	(Ve, Vi) (S)	294.954[b]	439.124[b]
5	(S, Ve, Vi)	789.144	788.994

[a]First Minimum AIC

[b]Second Minimum AIC

Comparing the AIC's in Table 6.7 above, we see that AIC(varying $\underset{\sim}{\mu}$ and $\underset{\sim}{\Sigma}$) values are much less than the AIC(common $\underset{\sim}{\Sigma}$) 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 in data analysis.

As we mentioned in the introduction of this paper, in practice the assumption of equal covariance matrices within the model is a rather dubious requirement.

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.

Acknowledgement

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.

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), 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, Budapest: Akademiai Kiado, 267-281.

[3] _____ (1974), "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, AC-19, 716-723.

[4] Anderson, T.W., and Sclove, S.L. (1978), An Introduction to the Statistical Analysis of Data, Boston: Houghton Mifflin Company.

[5] Bozdogan, H., and Sclove, S.L. (1982), "Multi-Sample Cluster Analysis Using Akaike's Information Criterion," Technical Report No. 82-1, ONR Contract N00014-80-C-0408 (NR042-443), Quantitative Methods Department, University of Illinois at Chicago Circle, Chicago, Illinois 60680.

[6] Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, 179-188.

[7] Friedman, H.P., and Rubin, J. (1967), "On Some Invariant Criteria for Grouping Data," Journal of the American Statistical Association, 62, 1159-1178.

[8] Kendall, M.G. (1966), "Discrimination and Classification," in P.R. Krishnaiah (Ed.), Multivariate Analysis, New York: Academic Press.

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

[10] Solomon, H. (1971), "Numerical Taxonomy," Mathematics in the Archaeological and Historical Sciences, Edinburgh: Edinburgh University Press, 62-81.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Techical Report 82-2	2. GOVT ACCESSION NO. AD-A113 422	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Multi-Sample Cluster Analysis With Varying Parameters Using Akaike's Information Criterion		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Hamparsum Bozdogan and Stanley L. Sclove		8. CONTRACT OR GRANT NUMBER(s) N00014-80-C-0408
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Illinois at Chicago Circle Box 4348, Chicago, IL 60680		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS —		12. REPORT DATE March 8, 1982 13. NUMBER OF PAGES 23+ii
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Office of Naval Research Statistics and Probability Branch Arlington, VA 22217		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Unlimited distribution

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multi-sample cluster analysis; Akaike's Information Criterion (AIC); Univariate model with varying means and variances; Multivariate model with varying mean vectors and variance-covariance matrices; maximum likelihood.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

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 the univariate model with varying means and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

variances, and in the multivariate model with varying mean vectors and variance-covariance matrices. Numerical examples are presented and results are shown to demonstrate the utility of AIC in identifying the best clustering alternatives.


D
FI
05