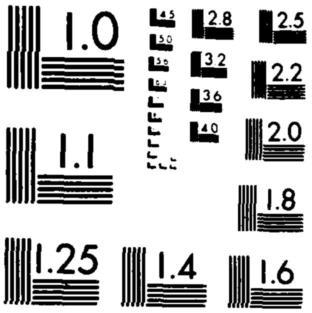
AD-A130 020 DETERMINING THE NUMBER OF COMPONENT CLUSTERS IN THE STANDARD MULTIVARIATE..(U) ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE METHODS H BOZDOGAN 1/1

UNCLASSIFIED 16 JUN 83 UIC/DQM/A83-1 ARO-19085.5-MA F/G 12/1 NL

END
DATE
FILMED
7 83
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

DETERMINING THE NUMBER OF COMPONENT CLUSTERS IN THE STANDARD MULTIVARIATE NORMAL MIXTURE MODEL USING MODEL-SELECTION CRITERIA*

by

HAMPARSUM BOZDOGAN

TECHNICAL REPORT NO. UIC/DQM/A83-1
June 16, 1983

PREPARED FOR THE
ARMY RESEARCH OFFICE
UNDER
CONTRACT DAAG29-82-K-0155
with the University of Illinois at Chicago

Statistical Models and Methods for
Cluster Analysis and Image Segmentation

Principal Investigator: Stanley L. Sclove

QUANTITATIVE METHODS DEPARTMENT
COLLEGE OF BUSINESS ADMINISTRATION
UNIVERSITY OF ILLINOIS AT CHICAGO
BOX 4348, CHICAGO, IL 60680

*Presented at the Fourteenth Annual Meeting of the Classification Society (North American Branch), Philadelphia, PA, May 29-31, 1983.

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER Technical Report No. UIC/DQM/A83-1	2. GOVT ACCESSION NO. AD-A130020	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Determining the Number of Component Clusters in the Standard Multivariate Normal Mixture Model Using Model-Selection Criteria		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Hamparsum Bozdogan		8. CONTRACT OR GRANT NUMBER(s) DAAG29-82-K-0155
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Illinois at Chicago Box 4348, Chicago, IL 60680		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE June 16, 1983 13. NUMBER OF PAGES 42 + iii
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		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)

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

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

Standard multivariate normal mixture model; Akaike's Information Criterion (AIC); Schwarz's Criterion (SC)

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

The problem of clustering individuals is considered within the context of a multivariate normal mixture using model-selection criteria. Often, the number k of components in the mixture is not known. In practical problems, the question arises as to the appropriate choice of k. The problem is to

cont

(Abstract, continued)

cont decide how many components are in the mixture, a difficult multiple decision problem.

In the statistical literature, several hypothesis testing variety of criteria have been proposed for this purpose. However, all these criteria possess sampling distributional problems. What the null distribution of the criterion is if the data actually contain k clusters is not known, and remains largely unresolved still.

Two well known model-selection criteria, namely Akaike's Information Criterion (AIC) and Schwarz's Criterion are proposed for the first time as two new approaches to the problem of what the appropriate choice of k in the mixture multinormal model should be. The forms of these two model-selection criteria are obtained in the standard multivariate normal mixture model. Analyses are carried out on the same data set by applying the model-selection criteria for different choices of k using the ~~mixture~~ algorithm under two assumptions with common covariance matrices between the component normals, and with varying covariance matrices in determining the appropriate number of types or clusters. The results are obtained when data initially partitioned into equal size groups; when data initially reordered; when data initialized by k-means algorithm; when data initialized by special initialization scheme; and when special initialization scheme is used on reordered data.

DETERMINING THE NUMBER OF COMPONENT CLUSTERS IN THE STANDARD MULTIVARIATE NORMAL MIXTURE MODEL USING MODEL-SELECTION CRITERIA*

HAMPARSUM BOZDOGAN
Department of Quantitative Methods
University of Illinois

CONTENTS

*Presented at the Fourteenth Annual Meeting of the Classification Society (North American Branch), Philadelphia, PA, May 29-31, 1983.

DETERMINING THE NUMBER OF COMPONENT CLUSTERS IN THE STANDARD MULTIVARIATE NORMAL MIXTURE MODEL USING MODEL-SELECTION CRITERIA*

HAMPARSUM BOZDOGAN
Department of Quantitative Methods
University of Illinois at Chicago

ABSTRACT

The problem of clustering individuals is considered within the context of a multivariate normal mixture using model-selection criteria. Often, the number k of components in the mixture is not known. In practical problems, the question arises as to the appropriate choice of k. The problem is to decide how many components are in the mixture, a difficult multiple decision problem.

In the statistical literature, several hypothesis testing variety of criteria have been proposed for this purpose. However, all these criteria possess sampling distributional problems. What the null distribution of the criterion is if the data actually contain k clusters is not known, and remains largely unresolved still.

Two well known model-selection criteria, namely Akaike's Information Criterion (AIC) and Schwarz's Criterion are proposed for the first time as two new approaches to the problem of what the appropriate choice of k in the mixture multinormal model should be. The forms of these two model-selection criteria are obtained in the standard multivariate normal mixture model. Analyses are carried out on the same data set by applying the model-selection criteria for different choices of k using the mixture algorithm under two assumptions with common covariance matrices between the component normals, and with varying covariance matrices in determining the appropriate number of types or clusters. The results are obtained when data initially partitioned into equal size groups; when data initially reordered; when data initialized by k-means algorithm; when data initialized by special initialization scheme; and when special initialization scheme is used on reordered data.

Key Words and Phrases: Standard multivariate normal mixture model; Akaike's Information Criterion (AIC); Schwarz's Criterion (SC).

*Presented at the Fourteenth Annual Meeting of the Classification Society (North American Branch), Philadelphia, PA, May 29-31, 1983.

DETERMINING THE NUMBER OF COMPONENT CLUSTERS IN THE STANDARD MULTIVARIATE NORMAL MIXTURE MODEL USING MODEL-SELECTION CRITERIA*

HAMPARSUM BOZDOGAN
University of Illinois at Chicago

1. Introduction

What is the most appropriate number of clusters for a set of data? How do we decide the number of clusters present in the data? Which cluster or clusters do we choose? These are some fundamental questions confronting practitioners and research workers in classification and clustering. The importance and the difficulty of this problem have been noted by many authors such as Beale (1969), Marriott (1971), Calinski and Harabasz (1974), Maronna and Jacovkis (1974), Matusita and Ohsumi (1980), and others. For a good discussion on some of the test procedures used in deciding and determining the number of clusters, we refer the reader to Milligan (1981), Dubes and Jain (1979), and Everitt (1979, 1974).

It is reasonable for an investigator to discover whether there is any structure in the data, or whether they indicate just a single cluster or group. If there is only one group, that is, no cluster structure, then most investigators would decide that clustering techniques were not needed. Discovering the structure in the data has its own practical importance. For example, in studying medical and psychological syndromes; processing remotely sense data for target identification or for predicting crop yields; in problems of taxonomy; and in many other applications we might want to find out whether the observations fall into natural groups or not. If they do, then we might want to discover how many groups or clusters there might be, and how do we identify and interpret them?

In the literature, numerous attempts have been made to devise reasonable

*Presented at the Fourteenth Annual Meeting of the Classification Society (North American Branch), Philadelphia, PA, May 29-31, 1983.

indicators for the problem of choosing the number of clusters present, identification and interpretation of clustering results by many investigators. Still today, however, there is no satisfactory solution and a unified flexible approach. The major difficulties with deriving formal significance tests similar to those of ordinary "t" and "F" test statistics in cluster analysis, appear to be the difficulty of determining the sampling distribution of the proposed test statistic [Everitt (1979)]. The problem of deriving a sampling distribution is formidable, and the choice of a fixed level of significance for comparison of different number of clusters with various number of parameters is wrong since this does not take into account the increase of the variability of the estimates when the number of parameters is increased. Therefore, the theoretical difficulties faced in deriving sampling distribution of a proposed test statistic, in the context of cluster analysis, are rather involved and not practical. This point has been advocated by Gnanadesikan and Wilk (1969), and others in the literature.

This suggests that, if we use the formal signficance test type indicators or statistics in conjunction with the clustering algorithms or techniques, then we must devise a criterion (or criterions) which will combine both the estimation problem and the testing together to decide on the number of clusters present in a data set.

Therefore, in this paper we shall propose and establish two theoretically-based procedures in deciding and determining the number of clusters present, identifying the best clustering alternative or alternatives. We shall achieve this by introducing two well known model-selection criteria, namely, Akaike's Information Criterion (AIC), and its derivative, Schwarz' Criterion (SC) as two new and unifying procedures.

Thus, the main focus of this paper will be to show how to use these two model-selection criteria in deciding and determining the number of component clusters present in the standard multivariate normal mixture model without knowing a priori classification of the observations.

In Section 2, we shall discuss the standard multivariate normal mixture model and the clustering criteria used under this model, namely, the maximum likelihood approach. In Section 3, we shall discuss and review the use of fitting the mixture model to determine the number of component clusters, and its corresponding unresolved problems. We shall, in Section 4, present the two model-selection criteria, and list their important general characteristics. In Section 5, we shall give the forms of the model-selection criteria to be used in standard normal mixture model approach to clustering. We shall apply these two model-selection criteria in Section 6 in deciding and identifying the number of components or clusters present in the Fisher iris data and present the numerical results. Finally, in Section 7, we shall present conclusions and discussion.

2. The Standard Multivariate Normal Mixture Model

2.1. The Model

As has been suggested before [see, e.g., Fleiss and Zubin (1969)], often when we consider clustering problems it seems relevant and logical to consider the sample as arising from several different populations rather than a single population since the individuals within a class or group differ from one another. That is, each individual in the sample is assumed to have come from one of several populations (types).

Given a sample from the overall mixed population, or assuming that the sample has come from a mixture population, the problem from a clustering viewpoint is to determine and describe the number of subpopulations or groups, the parameters of the distribution characterizing each subpopulation or group, and which group each individual belongs to.

Therefore, the problem of clustering individuals, objects, or cases, to be considered here, will be studied within the context of a mixture of multivariate normal distributions.

More specifically, we shall consider a multivariate <u>normal</u> mixture model,

$$(2.1.1) \qquad f(\underset{\sim}{X}) \equiv f(\underset{\sim}{X};\underset{\sim}{\Pi},\underset{\sim}{\mu},\underline{\Sigma}) = \sum_{k=1}^{K} \Pi_k f_k(\underset{\sim}{X};\underset{\sim}{\mu}_k,\underline{\Sigma}_k)$$

where $\underset{\sim}{\Pi} = (\Pi_1,\Pi_2,\ldots,\Pi_{K-1})$ are K - 1 independent mixing propor and are such that

$$0 < \Pi_k < 1 \qquad\qquad \Pi_K = 1 - \sum_{k=1}^{K} \Pi_k ,$$

and where $f_k(\underset{\sim}{X};\underset{\sim}{\mu}_k,\underline{\Sigma}_k)$ is the k-th component multivariate normal density, given by

$$(2.1.2) \qquad f_k(\underset{\sim}{X};\underset{\sim}{\mu}_k,\underline{\Sigma}_k) = (2\Pi)^{-p/2}|\underline{\Sigma}_k|^{-1/2}\exp\{-1/2(\underset{\sim}{X} - \underset{\sim}{\mu}_k)'\underline{\Sigma}_k^{-1}(\underset{\sim}{X} - \underset{\sim}{\mu}_k)\}.$$

The model given by the p.d.f. in (2.1.1) is called the standard multivariate normal mixture model to distinguish it from the modified conditional mixture model considered by Symons (1981), Sclove (1977,1982), Scott and Symons (1971), and John (1970).

In the statistical literature, several authors, including Wolfe (1970), Day (1969), Binder (1978), Hartigan (1977), and others, have considered clustering problems in which a standard mixture of multivariate normals is used as a statistical model given by (2.1.1).

2.2. Clustering Criteria: The Maximum Likelihood Approach

Wolfe (1970) has considered clustering based on the standard normal mixture model. He uses the maximum likelihood (ML) approach to estimate the mixing proportions Π_k, the mean vector $\underset{\sim}{\mu}_k$ and the covariance matrices $\underline{\Sigma}_k$. The maximum likelihood estimators (MLE's) nave well known desirable properties and it is natural to consider the ML approach for estimating the parameters in a mixture of multivariate normal distributions. To estimate the parameters the likelihood of the data is required which is given by

$$(2.2.1) \qquad L(\underline{X}|\underset{\sim}{\Theta}) = \prod_{i=1}^{n} \{\sum_{k=1}^{K} \Pi_k f_k(\underset{\sim}{X}_i;\underset{\sim}{\mu}_k,\underline{\Sigma}_k)\} ,$$

or the log of the likelihood is

$$(2.2.2) \qquad l \equiv \log_e L(\underline{X}|\underset{\sim}{\Theta}) = \sum_{i=1}^{n} \log_e\{\sum_{k=1}^{K} \Pi_k f_k(\underset{\sim}{X}_i;\underset{\sim}{\mu}_k,\underline{\Sigma}_k)\} .$$

It is the likelihood in (2.2.1) or the log likelihood in (2.2.2) that is maximized with respect to $\underset{\sim}{\Theta} = (\Pi_1,\Pi_2,\ldots,\Pi_K,\underset{\sim}{\mu}_1,\underset{\sim}{\mu}_2,\ldots,\underset{\sim}{\mu}_K,\underline{\Sigma}_1,\underline{\Sigma}_2,\ldots,\underline{\Sigma}_K)$, the vector of parameters, by Wolfe (1970) and Day (1969). The maximum likelihood equations are obtained by equating the first partial derivatives of (2.2.2) witn respect to the Π_k, the elements of each vector $\underset{\sim}{\mu}_k$, and those of each matrix $\underline{\Sigma}_k$, to zero. These equations are solved iteratively by a modified Newton-Raphson method. The iterative MLE's are given by

$$(2.2.3) \qquad \hat{\Pi}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{P}(k|\underset{\sim}{X}_i) \qquad k=1,2,\ldots,K-1$$

$$(2.2.4) \qquad \hat{\underset{\sim}{\mu}}_k = \frac{1}{n\hat{\Pi}_k} \sum_{i=1}^{n} \underset{\sim}{X}_i \hat{P}(k|\underset{\sim}{X}_i) \qquad k=1,2,\ldots,K$$

(2.2.5) $$\hat{\underset{\sim}{\Sigma}}_k = \frac{1}{n\hat{\Pi}} \sum_{i=1}^{n} \hat{P}(k|\underset{\sim}{X}_i)(\underset{\sim}{X}_i - \hat{\underset{\sim}{\mu}}_k)(\underset{\sim}{X}_i - \hat{\underset{\sim}{\mu}}_k)' \qquad k=1,2,\ldots,K$$

where

$\hat{\Pi}_k$ = the mixing proportion for type or cluster k,

$\hat{\underset{\sim}{\mu}}_k$ = vector of means for cluster k,

$\hat{\underset{\sim}{\Sigma}}_k$ = covariance matrix for cluster k,

$\underset{\sim}{X}_i$ = vector of observations for the i-th point in the sample, and

$$\hat{P}(k|\underset{\sim}{X}_i) = \frac{\hat{\Pi}_k f_k(\underset{\sim}{X}_i;\hat{\underset{\sim}{\mu}}_k,\hat{\underset{\sim}{\Sigma}}_k)}{\sum_{k=1}^{K} \hat{\Pi} f_k(\underset{\sim}{X}_i;\hat{\underset{\sim}{\mu}}_k,\hat{\underset{\sim}{\Sigma}}_k)}$$ = posterior probability of group membership of $\underset{\sim}{X}_i$ in cluster k.

If the clusters have a common covariance matrix, then we use

(2.2.6) $$\hat{\underset{\sim}{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{X}_i \underset{\sim}{X}_i' - \sum_{k=1}^{K} \hat{\Pi}_k \hat{\underset{\sim}{\mu}}_k \hat{\underset{\sim}{\mu}}_k' \ .$$

Since the iterative process is used to solve the equations, actually, several sets of values may satisfy the equations, and the results may depend on the initial values for the iteration process. Since mixture analysis attempts to find maximum-likelihood estimates of the parameters, the best solution for our purposes is the one with the greatest likelihood, or the greatest log likelihood.

Once the MLE's are known, we can regard each distribution as indicating a separate cluster, and individuals are then assigned by the Bayes allocation rule. That is, assign $\underset{\sim}{X}_i$ to the k-th distribution when

$$(2.2.7) \qquad \hat{\pi}_l f(\underset{\sim}{X}_i;\hat{\underset{\sim}{\mu}}_l,\underline{\Sigma}_l) < \hat{\pi}_k f(\underset{\sim}{X}_i;\hat{\underset{\sim}{\mu}}_k,\underline{\Sigma}_k) \text{ for all } l \neq k \ .$$

This process is repeated, increasing the log likelihood at each stage, until no further reallocation of the $\underset{\sim}{X}$'s occurs. Another way to put it, individual i is assigned to that component (or group) k for which the estimated posterior probability of group membership, $\hat{P}(k|\underset{\sim}{X}_i)$, is largest. Therefore, for a particular individual i, the optimal $\{P(k|\underset{\sim}{X}_i)\}$ will be $P(k|\underset{\sim}{X}_i) = 1$ when the individual i is from component (or group) k and zero otherwise.

It should be noted here that, this is one of the points where the standard normal mixture model considered here, differs from that of the conditional mixture model. That is, in the conditional mixture model, the individual i is assigned to group k for which the estimated density is largest rather than the estimated posterior probability of group membership which happens to be the case in the standard normal mixture model. For more details on this, refer to Sclove (1979).

3. Fitting the Mixture Model to Determine the Number of Component Clusters: Unresolved Problems

As we mentioned in Section 1, we may want to ask whether there really is a mixture or whether there is just a single underlying component cluster. In practice, this could be the sort of question we might be interested in since fitting the standard normal mixture model to determine the number of component clusters has many practical importance and use. For example, we may want to determine the number of disease types in the study of disease patterns, the blood pressure types, and psychiatric disorder types. In reliability analysis, we may want to determine the number of laser types on the basis of mean laser life. Lasers are employed in telephone communication systems in which coherent laser light is used to transmit telephone communications. In image processing, we may want to determine the number of classes of segments, etc.

As it was noted by Sokal (1977), the problems of inference on the number of clusters "actually" present in a set of data, and of testing for model fit, have not yet received much successful attention but more and more are recognized as important.

Thus, the standard mixture problem will be to decide how many component clusters are in the mixture, a difficult multiple decision problem. A simpler problem is to decide whether k=r or k=r+1 component clusters are necessary. In practice, it is common to specify a larger hypothesized number of clusters, say k, and create sequence of k=1,2,...,K component clusters by using the mixture algorithm.

In the literature, several methods have been proposed in determining the number of component clusters when the technique of fitting standard normal mixture model is used. One type of these techniques are informal graphical techniques, and the other type is more formal hypothesis testing variety of technique..

When the technique of fitting mixture of distributions is used as a clustering technique, likelihood ratio test is a more natural criterion for testing the number of component clusters or groups in this context. However, as we shall see, it has its thorny problems.

Let L_k denote the maximized likelihood, for given k. Then

$$(3.1) \qquad \lambda = L_k/L_{k'}$$

is the likelihood ratio statistic for testing k clusters against k' clusters (k < k'). From a Monte Carlo investigation, Wolfe (1971) arrived at and suggested an adjusted likelihood ratio test in which the statistic (not counting the mixing proportions):

$$(3.2) \qquad -2 \cdot \frac{1}{n}(n - 1 - p - \frac{k'}{2})\log_e\lambda \sim \chi^2_f \qquad \text{(chi-square)}$$

with degrees of freedom $f = 2p(k' - k)$, where

n = sample size,

p = number of variables,

k = number of component types in the null hypothesis,

k' = number of component types in the alternative hypothesis.

After performing a small scale simulation study, Wolfe (1971) on the basis of the results recommended using the modified likelihood ratio test given in (3.2) for k=1 against k=2, when under the alternative hypothesis the two components are assumed to have the same variance-covariance matrix. But, Wolfe's simulation results suggest that even in reasonably large sample sizes, the statistic in (3.2) does not appear to be asymptotically the usual chi-square. In Wolfe's simulation, some of the sample means and variances are quite different from those corresponding to the stipulated chi-square distributions. Also, the same results may not be true when under the alternative hypothesis the two components are assumed to have different variance-covariance matrices. Moreover, it is important to note that in the standard mixture problem, the likelihood function is a different function for different values of k, where k=1,2,...K. Therefore, in the context of the standard mixture model, the question that arises, and that remains largely unresolved still, is what the asymptotic null distribution is if the data actually contain k=1,2,...,K clusters.

Others in the statistical literature, have also cited the distributional problems of the likelihood ratio test statistic in the mixture problem. For example, Hartigan (1977) speculated that the log likelihood ratio lies between $\frac{1}{2}\chi^2_p$ and $\frac{1}{2}\chi^2_{p+1}$, where p is the number of variables. Binder (1978), on the other hand, argues that the likelihood ratio criterion given in (3.2) is not necessarily asymptotically chi-square distributed since

$$(3.3)\qquad \begin{aligned} &H_0\colon\ \Pi=1\\ &H_1\colon\ 0<\Pi<1. \end{aligned}$$

Here, under the null hypothesis, Π, the mixing proportion, is on the boundary of the parameter space, and the likelihood ratio criterion takes the value zero when $\hat{\Pi}$, the maximum likelihood estimate for Π, is 1 with probability $\frac{1}{2}$, and therefore, under the null hypothesis, the likelihood ratio criterion cannot be asymptotically χ^2.

Behboodian (1972), shows that as the component densities become closer and closer to each other, the information matrix approaches a singular matrix with some diagonal elements equal to zero. The same thing happens when the mixing parameter Π tends to one or zero. Consequently, Behboodian concludes that for estimating the parameters in a mixture where two component clusters are well separated, or which has a mixing proportion close to zero, very large samples may be needed. For example, for a fixed total sample size n, when we run the mixture algorithm for a very large hypothesized number of clusters K, the mixing proportion Π starts tending to zero. To put it in another way, as K, the desired total number of component clusters, gets larger and larger for a fixed sample size n, then the mixing proportion π tends to zero. Thus causing the information matrix to be singular. For this reason, we need very large samples to fit the standard normal mixture model to ensure that the component sample sizes are large enough so that the information matrix will not become singular.

This point raises another important question as to what the appropriate hypothesized number of component clusters K should be for a fixed sample size n to fit the mixture model. In the literature, this point has never been studied before, and certainly deserves more attention which will be a subject of futher study later.

A rule of thumb, however, is to use $K \sim (n/2)^{1/2}$ suggested by Mardia, et. al. (1979), where K is the total hypothesized number of component clusters, and n is the total sample size.

In reviewing the literature further, we see that some simulation results of Everitt (1981) show that the suggestion of Wolfe (1971) seems reasonable only in cases where n>10p. That is, the sample size n is of order 10p, where p is the number of variables, for testing one standard normal mixture model against only two standard normal mixture models when the two components are assumed to have the same variance-covariance matrix. According to Everitt's large scale simulation results, Hartigan's (1977) conjecture does not seem to be correct. However, at this point, Everitt's results cannot be extended to be true for testing two standard normal mixture models against three, three against four, four against five, and so forth, since there does not exist any reasonable Monte Carlo validation of the significance testing procedure given in (3.2).

Utilizing estabiisned results in the literature on the distribution of the log likelihood ratio test statistic when the true parameter is "near" the boundaries of the hypothesis regions, we can reflect the key distributional requirements of the model.

Following Feder (1968), we state that, when the data can be represented by n independent random variables with identical distributions depending on the parameters $(\Theta_1,\Theta_2,\ldots,\Theta_k)$ then the limiting distribution (as $n\to\infty$) of

$$(3.7) \qquad -2\log_e(\text{likelihood ratio})$$

is, under certain sequences of alternative hypotheses converging to the null hypothesis which appears to be the case in testing mixture models, a noncentral chi-squared distribution. This result is due also to Wald (1943).

According to this result, it seems that for the mixture problem the key distributional requirement for a test is

$$(3.8) \qquad -2C(n,p,K)\log_e\lambda \overset{a.d.}{\sim} \chi_f'^2(\delta) \qquad \text{(noncentral chi-square)}$$

where

f = number of degrees of freedom,

δ = noncentrality parameter, and

$C(n,p,K) = \frac{1}{n}(n - 1 - p - \frac{K}{2})$ = correction factor,

n = sample size,

p = number of variables,

K = total number of components hypothesized in the mixture model.

In the next section, that is, in Section 4, we shall introduce the two well known model selection criteria to be used to estimate k $(k=1,2,\ldots,K)$, the number of component clusters in the standard normal mixture model. First some general explanations on model-selection criteria will be appropriate.

4. Model-Selection Criteria

In the literature, model selection or identification problems continues to attract a great deal of interest among statisticians and other scientists. The major effort in this respect has been channeled towards simple criteria for choosing one of a set of competing models to describe a given data set. Much of this interest has been stimulated by the fundamental work of Akaike (1973) and by the appearance of an information criterion due to him, known as Akaike's Information Criterion (AIC). Therefore, one group of criteria we see in the current statistical literature are based on Boltzmann's (1877) entropy or the Kullback's (1959) information, such as Akaike's Information Criterion. The other main group of criteria are Bayesian. Among the Bayesian, in particular, here we shall consider only Schwarz' Criterion (SC).

Next, we give the formal definitions and some of the important characteristics of these two model-selection criteria.

4.1. Akaike's Information Criterion (AIC)

Suppose there are K alternative models M_k, k=1,2,...K, represented by the densities $f_1(\cdot|\underset{\sim}{\theta}_1)$, $f_2(\cdot|\underset{\sim}{\theta}_2),\ldots,f_K(\cdot|\underset{\sim}{\theta}_K)$ for the explanation of a random vector $\underset{\sim}{X}$ and given n observations. In 1971, Akaike first introduced an information criterion, which has become known as Akaike's Information Criterion (AIC) for the identification and comparison of statistical models among a class of competing models with different number of parameters. It is defined by

$$\text{(4.1.1)} \qquad AIC(k) = -2\ \ln[\max L(k)] + 2m(k),$$

or symbolically is defined by

$$\text{(4.1.2)} \qquad AIC = -2\ \ln(\text{maximized likelihood}) + 2\ (\text{number of parameters estimated within the model}).$$

In (4.1.1), L(k) is the likelihood when M_k is the model, max denotes its maximum over the parameters, and m(k) is the number of independent parameters when M_k is the model.

The statistic AIC(k), was obtained by Akaike (1973, 1974) with the aid of an information theoretic interpretation of the method of maximum likelihood. It is a natural estimate of minus twice the expected log likelihood of the model whose parameters are determined by the method of maximum likelihood. The minused expected log likelihood is, except for an additive constant, identical to the (generalized) entropy, or the "cross-entropy," which is a measure of goodness of fit or closeness of the estimated, fitted, or predictive model to the true model. From this point of view, when several competing models are being compared or fitted, AIC(k) 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 model chosen is the one which minimizes AIC and is called the minimum AIC procedure. The first term in (4.1.1) stands for the penalty of badness of fit when the maximum likelihood estimators of the parameters of the model is used. The first term is also known as the measure of inaccuracy [see, e.g., Stone (1982)]. The second term in the definition of AIC, on the other hand, is interpreted as representing a penalty that should be paid for increasing the number of parameters, or compensation for the bias or increased unreliability in the first term due to the increased number of parameters. The second term in AIC, is also known as the complexity of the selected model. If more parameters are used to describe the data, it is natural to get a larger likelihood, possibly without improving the goodness of fit. Thus, AIC avoids this spurious improvement of fit by penalizing the use of additional parameters. In this sense, the AIC may be regarded as an explicit formulation of principle of parsimony in model building. In the statistical literature, the interpretation of the second term in AIC as a measure of the complexity of the model M_k, corresponds to the principle known as Occam's Razor, which emphasizes the desirability of selecting accurate and parsimonious models of reality. This principle is also closely related to the principle in hypothesis testing which emphasizes the desirability of considering "substantive" significance as opposed to statistical significance. For more details on this, we refer the reader to Hodges and Lehmann (1954).

We now list some of the important characteristics of Akaike's Information Criterion (AIC) as follows:

(i) AIC is defined without specific reference to the true model $f(\cdot|\underset{\sim}{\theta}_k^*)$. Thus, for any finite number of parametric models, we may always consider an extended model that will play the roll of $f(\cdot|\underset{\sim}{\theta}_k)$. This suggests that AIC can be useful for the comparison of models which are nonnested, i.e., the situation where conventional log likelihood ratio test is not applicable as mentioned by Akaike (1982).

(ii) The value of AIC decreases quickly as the number of parameters being adjusted is increased and then increases almost linearly when too many redundant parameters are included in the model. For more on this, refer to Akaike (1978), Smith and Spiegelhalter (1980).

(iii) According to AIC, inclusion of an additional parameter is appropriate if ln[max L] increases by one unit or more, i.e., if max L increases by a factor of e or more.

(iv) AIC can have positive or negative values depending on the situation.

If we let $\hat{L}(k)$ = max L(k) when M_k is the model, with, say k, number of parameters, and L(k+1) = max L(k+1) when M_{k+1} is the model, with, say k+1, number of parameters, and if $\hat{L}(k+1)/\hat{L}(k) > e$, then AIC(k) is positive. If $\hat{L}(k+1)/\hat{L}(k) < e$, then AIC(k) is negative.

(v) AIC does not require level of significance or table look-up.

(vi) The relationship between the AIC and the conventional likelihood ratio test statistic can be written as

$$(-2)\ln \lambda(H_0;H_1) = \mathrm{AIC}(H_0) - \mathrm{AIC}(H_1) - 2k,$$

where the model H_1 contains the model H_0 as a restricted family of distributions of H_1 and k denotes the degrees of freedom of the chi-square distribution of the likelihood ratio test statistic.

4.2. Schwarz' Criterion (SC)

Schwarz (1978) proposed a model selection procedure which minimizes the criterion,

$$\text{(4.2.1)} \qquad SC(k) = -2 \ln[\max L(k)] + m(k)\ln(n),$$

where n is the number of independent observations. This criterion is obtained by analyzing the behavior of the posterior probability of the model M_k when n grows to infinity under the assumption of some arbitrary positive *a priori* probability distributions on the parameters. Therefore, this criterion is a Bayesian criterion. For this reason, we shall abbreviate it as SC, instead of SIC. One should note that, SC and AIC are qualitatively the same, but they are quantitatively different from one another only in that the number of estimated parameters is multiplied by ln(n), the natural logarithm of the sample size.

We now list some of the important characteristics of Schwarz' Criterion (SC) as follows:

(i) SC assumes a fixed penalty for guessing the wrong model.

(ii) For small sample sizes, SC favors lower-dimensional models as compared to AIC. However, depending on the nature of the priors on the parameters and the nature of the model fitted, Schwarz' approximation may fail in small samples. Nevertheless, for large sample sizes it has its own advantages.

(iii) According to Schwarz' Criterion (SC), an additional parameter will be included if it increases ln[max L] by an amount ln(n)/2, that is, if max L increases by a factor of $\sqrt{n}$ or more.

(iv) Like AIC, SC can also have positive or negative values depending on the situation. That is, if $\hat{L}(k+1)/\hat{L}(k) > \sqrt{n}$, then SC(k) is positive. On the other hand, if $\hat{L}(k+1)/\hat{L}(k) < \sqrt{n}$, then SC(k) is negative.

(v) Also SC does not require level of significance or table look-up.

5. The Forms of Model-Selection Criteria in Standard Normal Mixture Model

Despite the recent development of the use of statistical methodology and models in many disciplines, it seems that in many situations the difficulty of constructing an adequate model based on the available sample information is not fully recognized. Cluster analysis is a case in point.

Recall that k denotes the number of clusters or component clusters. Usually k is permitted to vary: k=1,2,...,K, say. Each choice of k corresponds to a different model for the data. One has to estimate the parameters, say ${}_k\underset{\sim}{\theta}$, of this model. Then one computes the likelihoods L(k), k=1,2,...,K and is faced with the problem of comparing them. That is, in classification and clustering we have the problems of identifying and discovering the number of clusters present in the standard mixture model, without any *a priori* information about the data.

Such problems of statistical model identification suggest the introduction and the application of practically useful and versatile, and yet theoretically sound criteria of "fit" of models such as the ones we discussed in Section 4.

We, next, give the forms of AIC and SC to be used in standard normal mixture model approach to clustering.

For the standard mixture model, we first, consider our conjecture in (3.8) and show the form of AIC under this conjecture by stating and proving the following theorem.

Theorem 5.1. If $-2\ln\lambda \overset{a.d.}{\sim} \chi'^2_f(\delta)$ (non-central chi-square) with f = 2(M-m) degrees of freedom, then

(5.1) $$AIC^*(k) = -2C\ln[\max L(k)] + 3m(k),$$

where $C = \frac{1}{n}(n - 1 - p - \frac{K}{2})$ = correction factor,

k=1,2,...,K = number of component clusters, or types,

$m = m(k)$,

$m(k) = kp + (k-1) + \frac{p(p+1)}{2} =$ number of parameters including the mixture proportions when covariances are equal,

$m(k) = kp + (k-1) + k\,\frac{p(p+1)}{2} =$ number of parameters including the mixture proportions when covariances are different between clusters, and

$M = m(K)$.

Proof. In general,

$$(5.2) \qquad -2nE[B(f;\hat{f})] = -2nE[\text{entropy}] \approx \delta + m.$$

where E denotes the expected value, $\delta = n||_k\underset{\sim}{\theta} - \underset{\sim}{\theta}_{true}||^2_{\underline{J}}$ is the noncentrality parameter, "$||\cdot||$" stands for the Euclidean norm with respect to $\underline{J} = (J_{ij})$, the (kxk) Fisher information matrix, and m denotes here, the number of parameters. We asserted in (3.8) that

$$(5.3) \qquad -2\ln \lambda \overset{a.d.}{\sim} \chi'^2_f(\delta),$$

where $f = 2(M-m)$ is the number of degrees of freedom, and δ, is the noncentrality parameter. As is well known,

$$(5.4) \qquad -2C\ln \lambda \approx E[-2C\ln \lambda] \approx E[\chi'^2_f(\delta)] = \delta + f = \delta + 2(M-m).$$

Hence, solving (5.4) for δ, the noncentrality parameter, we have

$$(5.5) \qquad \delta \approx -2C\ln \lambda - 2(M-m).$$

Now substituting (5.5) into (5.2), we obtain

$$(5.6) \qquad \begin{aligned} -2nE[B(f;\hat{f})] &\approx \delta + m \\ &\approx -2C\ln \lambda - 2(M-m) + m \\ &= -2C\ln \lambda - 2M + 3m. \end{aligned}$$

Since

$$(5.7)\qquad \ln \lambda = \ln \frac{\max L(k)}{\max L(K)} = \ln[\max L(k)] - \ln[\max L(K)],$$

and since AIC estimates the quantity $-2nE[B]$, then from (5.6), we have

$$(5.8)\qquad \mathrm{AIC} = -2C\ln[\max L(k)] + 3m - 2M + 2C\ln[\max L(K)]$$

For comparison purposes, it suffices to ignore the additive terms $-2M$ and $2C\ln[\max L(K)]$ in (5.8). Thus, for the standard mixture model AIC in (5.8) takes the simple form

$$(5.9)\qquad \mathrm{AIC}^*(k) = -2C\ln[\max L(k)] + 3m.$$

To make $\mathrm{AIC}^*(k)$ compatable with $SC(k)$, we can even drop C, the correction factor, and use

$$(5.10)\qquad \mathrm{AIC}^*(k) = -2\ln[\max L(k)] + 3m.$$

As we mentioned before, stimulated by the appearance of the Akaike's Information Criterion (AIC), Schwarz (1978) has recommended the model selection criterion,

$$(5.11)\qquad SC(k) = -2\ln[\max L(k)] + m(k)\ln(n),$$

where $k=1,2,\dots,K$ = number of component clusters, or types,

$m = m(k)$

$m(k) = kp + (k-1) + \frac{p(p+1)}{2}$ = number of parameters including the mixture proportions when covariances are equal,

$m(k) = kp + (k-1) + k\frac{p(p+1)}{2}$ = number of parameters including the mixture proportions when covariances are different between clusters, and

$M = m(K)$

for the standard mixture model.

Having defined these two well known model-selection criteria for the standard normal mixture model, in the next section, Section 6, we apply these two criteria to the famous Fisher iris data. In doing so, we shall attempt to improve Wolfe's and others' results without the worry of what the appropriate significance level α should be in testing the hypothesis of different component clusters in order to discover or identify and describe the clusters or types in the mixture model.

6. Application of Standard Normal Mixture Model to Fisher Iris Data

In this section we shall apply the standard normal mixture model to the well-known Fisher (1936) iris data. We shall give the numerical results from the mixture model by performing different analyses on the iris data by applying the model-selection criteria for differnt choices of k. We shall accomplish this by using the mixture algorithm under two assumptions: common covariance matrices between the component normals, and varying covariance matrices in determining the actual number of types or species in the Fisher iris data.

The iris data consist of four characteristics (p=4) for three species of iris; the species are *Iris setosa* (S), *Iris versicolor* (Ve), and *Iris virginica* (Vi), and the characteristics are sepal length, sepal width, petal length, and petal width. Each group is represented by 50 plants, and hence this data set is composed of 150 iris species in total.

This data set has been quite extensively studied in classification and cluster analysis since it was published by Fisher (1936), and still today, is being used to test the practical utility of various classification and clustering methods proposed by many investigators such as Friedman and Rubin (1967), Kendall (1966), Solomon (1971), Mezzich and Solomon (1980), and many others, including the present author.

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, for our purposes, if we were presented with the 150 irises in an unclassified manner (say, before the three species were established), then the mixture analysis using model-selection criteria attempts to discover and describe the types of irises without using any *a priori* classification information.

Using the *NORMIX programs* (i.e., normal mixture programs) of Wolfe (1967), which are modified and extended by this author, on the Fisher iris data, we ran *normal mixtures with different covariance matrices between the clusters (i.e., types)*, and *normal mixtures with common covariance matrices*. In both cases, we ran k=1,2,...,7 types and computed $AIC^*(k)$'s and SC(k)'s for identifying the best component cluster or clusters under the following situations:

1. When the mixture algorithm initially partitions the data into equal size groups;
2. When the data initially reordered to make the problem difficult for the mixture algorithm;

3. When the results from k-means algorithm are used to initialize the mixture algorithm to avoid the problem of local maxima of the likelihood function;
4. When a special initialization scheme is used to initialize the mixture algorithm which is proposed by this author; and finally
5. When a special initialization scheme is used on the reordered data to start the mixture algorithm, again to avoid the problem of local maxima of the likelihood function.

We present all our numerical results under each of the above situations respectively, as follows.

6.1. When Data Initially Partitioned into Equal Size Groups

When no special initialization is used, the mixture algorithm in the first step of iteration sets the belonging probabilities equal to one. That is, $P(k|\underset{\sim}{X}_i) = 1$ when the individual i is from component (or group) k and zero otherwise. This initialization is equivalent to partitioning the observations into equal size groups. Then the algorithm estimates the number of observations from the kth component in the second step. In the third and fourth steps, the algorithm estimates the cluster means and the within cluster variance-covariance matrices, respectively. In the fifth step, the determinants and inverses of the variance-covariance matrices are computed for each k and then the probability densities, the average densities, and the log likelihood function. This cycle is repeated until the maximum-likelihood estimates of the parameters converge, and until all the individuals or data units are assigned into their respective component clusters and no further reallocation occurs.

Under this situation, we ran k=1, k=2, k=3, k=4, k=5, k=6, and k=7 components or types and computed AIC*(k)'s and SC(k)'s for identifying and selecting the best component cluster or clusters. We obtained the following results.

TABLE 6.1.1. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE DIFFERENT BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m[b]	AIC*(k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.748
2	337.008	29	-587.016	-528.709*
3	371.177	44	-610.354*	-521.887**
4	385.842	59	-594.684**	-476.057
5	397.178	74	-572.356	-423.567
6	436.148	89	-605.296	-426.349
K=7	439.528	104	-567.056	-357.950

Where p=4 Variables, n=150 Observations, and

a. From Iterative Maximum Likelihood Estimates in Mixture Model After Convergence Took Place when 36 Iterations were used.

b. $m = kp+k-1+k\frac{p(p+1)}{2}$ = Number of Parameters.

c. $AIC^{*}(k) = -2\ln[\max L(k)] + 3m$.

d. $SC(k) = -2\ln[\max L(k)] + m\ln(n)$.

* First Minimum AIC* and SC.

** Second Minimum AIC* and SC.

TABLE 6.1.2. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE EQUAL BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m[b]	AIC*(k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.748
2	254.915	19	-452.830	-414.629
3	295.009	24	-518.018	-469.763
4	328.314	29	-569.628**	-511.321*
5	334.076	34	-566.152	-497.791**
6	339.142	39	-561.284	-482.870
K=7	355.353	44	-578.706*	-490.176

Where p=4 Variables, n=150 Observations, and

a. From Iterative Maximum Likelihood Estimates in Mixture Model After Convergence Took Place when 36 Iterations were used.

b. $m = kp+k-1+ \frac{p(p+1)}{2}$ = Number of Parameters.

c. $AIC^*(k) = -2\ln[\max L(k)] + 3m$.

d. $SC(k) = -2\ln[\max L(k)] + m\ln(n)$.

* First Minimum AIC* and SC.

** Second Minimum AIC* and SC.

Examining each table carefully, starting with Table 6.1.1 where the covariance matrices are different between clusters (or types), we see that the first minimum AIC* is when k=3 types, the second minimum AIC* is when k=4 types. That is, when k=3 types we have the best mixture submodel. This indicates that there are indeed three types of species in the iris data. On the other hand, the first minimum SC is when k=2 types, and the second minimum SC is when k=3 types. Thus, according SC k=2 types is the best mixture submodel indicating the fact that SC favors lower-dimensional models when compared with Akaike's AIC*. Nevertheless, the second minimum SC is when k=3 types where also AIC* achieves its first minimum. Hence, the mixture model has recovered the known structure among the 150 iris plants and we are capable of identifying it by using the minimum AIC* and the minimum SC procedures. For the three-types solution, by examining the confusion matrix of group membership, we see further that the I. setosa (Type or Cluster 1) were completely recovered, as I. virginica (Type or Cluster 3). However, five plants of I. versicolor (Type or Cluster 2) were classified with Type 3 and therefore these could be regarded as misclassified.

In Table 6.1.2 where the covariance matrices are considered to be equal between clusters (or types), we see that the first minimum AIC* is when k=7 types, the second minimum AIC* is when k=4 types. On the other hand, SC favors k=4 first, and then k=5 to be the second best mixture submodel. These results are not surprising since the population covariance matrices of the three types of irises are not equal to each other. Moreover, since mixture analysis attempts to find maximum-likelihood estimates of the parameters, the best solution for our purposes is the one with the greatest likelihood, or the greatest log likelihood. And hence, if we compare ln[max L(k)] of Table 6.1.1 and Table 6.1.2, respectively, we see that we have the greatest log likelihoods for each component clusters in Table 6.1.1, except when k=1 of course. Thus, this suggests that we should use the results of Table 6.1.1 where the covariance matrices are different for the iris data.

6.2. When Data Initially Reordered

In this case, we made the problem intentionally harder for the mixture algorithm through the reordering of the iris data sequentially. We chose first three species from each group and sequentially reordered the data until all the 150 flowers were scrambled completely. Such reordering of the data makes the algorithm start at different initial estimates of the parameters. The purpose of doing this is to obtain satisfactory initial estimates of the parameters which are essential if we need to avoid misleading solutions.

We ran again the NORMIX program assuming both different and equal covariance matrices between the clusters (or types) for k=1, k=2, k=3, k=4, k=5, k=6, and k=7 types. For each of the clustering alternatives, we computed AIC*(k)'s and SC(k)'s to be able to identify the best type and consequently determine the exact number of types. For these our results are shown in Tables 6.2.1 and 6.2.2.

TABLE 6.2.1. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE DIFFERENT BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m[b]	AIC*'k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.748
2	254.235	29	-427.470	-369.162
3	361.859	44	-591.718*	-503.251*
4	376.186	59	-575.372**	-456.745**
5	380.982	74	-539.964	-391.177
6	245.141#	89	-223.282#	- 44.337#
K=7	426.002	104	-540.004	-330.397

* First Minimum AIC* and SC.

** Second Minimum AIC* and SC.

# AIC* and SC Values During 5th Iteration. Mixture Algorithm Halted at 6th Iteration. Singular Variance-Covariance Matrix.

TABLE 6.2.2. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE EQUAL BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m^b	AIC*(k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.446
2	191.137	19	-325.274*	-287.072*
3	191.137	24	-310.274**	-262.018**
4	191.137	29	-295.274	-236.965
5	182.611	34	-263.222	-194.861
6	191.137	39	-265.274	-186.859
K=7	191.136	44	-250.272	-161.806

* First Minimum AIC* and SC.

** Second Minimum AIC* and SC.

# a, b, c, and d are as in Tables 6.1.1 and 6.1.2.

Now examining Tables 6.2.1 and 6.2.2, we see in Table 6.2.1 that the first minimum AIC* and SC occur at k=3 types, the second minimum AIC* and SC occur at k=4 types. Thus, both criteria choose k=3 types as the best mixture submodel.

In Table 6.2.2, however, we see completely the opposite of the results in Table 6.2.1. Here, the first minimum AIC* and SC both occur at k=2 types, and the second minimum AIC* and SC occur at k=3 types. We note, however, that, ln[max L(k)], except k=1, has converged to the same value for k=2,3,...,7 types even when we used 36 iterations. That is, ln[max L(k)] for k=2,...,7 are all stationary. Again, since mixture analysis attempts to find maximum-likelihood estimates of the parameters, the best solution for our purposes is the one with the greatest likelihood, or the greatest log likelihood. Therefore, comparing ln[max L(k)] of Table 6.2.1 and 6.2.2, we see that ln[max L(k)] are the greatest for each component clusters in Table 6.2.1, except when k=1. This suggests again that we should use the results of Table 6.2.1 where the covariance matrices are different for the iris data. However, one should not be puzzled with the noncovergence of ln[max L(k)] in Table 6.2.2, since we are not always guaranteed convergence in iterative procedures, nor are we guaranteed that the local optimum is always global. We show such a result to demonstrate that unexpected things also might happen.

6.3. When Data Initialized by K-Means Algorithm

It is a well known fact among the users of cluster analysis techniques that in the multivariate situation satisfactory or good initial estimates for the parameters are almost essential to start the iterative clustering algorithms to avoid misleading solutions. Specially, in the mixture analysis, there may be many different solutions of the maximum likelihood equations. Therefore, suitable initial values for the parameters are crucial when fitting mixtures of multivariate normal distributions to data to avoid the problem of local maxima of the likelihood function.

In the literature, Hartigan (1975, p. 124), Everitt (1981), and others, suggest "k-means" algorithm to be applied to data first, and then take the resulting cluster centroids (or means), etc., as starting values for component mean vectors, etc., in the maximum likelihood estimation algorithm. Following their suggestions, we ran "k-means" algorithm by using the BMDP K-MEANS PROCEDURE and asked for k=1,2,...,7 clusters on the 150 iris plants. We then took the resulting cluster centroids for each k and used them as starting values for component mean vectors in the mixture analysis for k=1,2,...,7. We obtained the following results.

TABLE 6.3.1. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE EQUAL BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m[b]	AIC*(k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.748
2	337.008	29	-587.016**	-528.709*
3	358.709	44	-585.418**	-496.950**
4	314.804#	59	-452.608#	-333.981#
5	412.012	74	-602.024*	-453.237
6	393.591#	89	-520.182#	-341.236#
K=7	391.616#	104	-471.232#	-252.125#

TABLE 6.3.2. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE EQUAL BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m[b]	AIC*(k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.748
2	254.915	19	-452.830	-414.629
3	295.001	24	-518.002	-469.763
4	328.314	29	-569.628**	-511.320*
5	334.065	34	-566.130	-497.768**
6	339.119	39	-561.238	-482.824
K=7	352.781	44	-573.562*	-485.095

* First Minimum AIC* and SC.

** Second Minimum AIC* and SC.

# AIC* and SC Values During 5th Iteration. Mixture Algorithm Halted at 5th Iteration. Singular Variance-Covariance Matrix.

a, b, c, and d are as in Tables 6.1.1 and 6.1.2.

Looking at Table 6.3.1 and 6.3.2, we see in Table 6.3.1 that the first minimum AIC* occurs at k=5 types and the first minimum SC occurs at k=2 types. The second minimum AIC* occurs at k=2 types and at k=3 since the values are significantly close to each other. Also, the second minimum SC occurs at k=3 types. For k=4, k=6, and k=7 types, the mixture algorithm halted at 6th iteration due to singular variance-covariance matrix.

In Table 6.3.2, we see that the first minimum AIC* occurs at k=7 types and the second minimum AIC* occurs at k=4 types. On the other hand, the first minimum SC occurs at k=4 types and the second minimum SC occurs at k=5 types. We further note here that these results are identical to those obtained in Table 6.1.2, when data initially partitioned into equal size groups by the algorithm.

Even though using "k-means" or other clustering techniques as a tool of initializing clusters appear to be the most obvious way to obtain suitable initial values for the parameters in the mixture analysis, but such an approach in general may not be the best as we shall see in the next two sections, that is, in Section 6.4 and 6.5, respectively.

6.4. When Data Initialized by Special Initialization Scheme

In Section 6.3, we gave the results of the mixture analysis when we initialized the mixture algorithm by using the results of "k-means" algorithm as our inputs or starting values for component mean vectors. As we mentioned, such an approach in general may not be the best and cheap. Therefore, in this section, we shall propose a simple and less expensive initialization scheme which has intuitive appeal and by-and-large philosophically is acceptable.

The proposed initialization scheme is as follows:

(i) We first compute the maximum and the minimum of the variables across all data. We denote this by $\underset{\sim}{X}_{max}$ and $\underset{\sim}{X}_{min}$. Let $R = \underset{\sim}{X}_{max} - \underset{\sim}{X}_{min}$ be the range of the data on the variable vector $\underset{\sim}{X}$.

(ii) Next, we compute the average of $\underset{\sim}{X}_{min}$ and $\underset{\sim}{X}_{max}$. We denote this by $\bar{\underset{\sim}{X}}_{11} = (\underset{\sim}{X}_{min} + \underset{\sim}{X}_{max})/2$. To initialize k=1 component mixture, we use $\bar{\underset{\sim}{X}}_{11}$ as the component mean vector in the mixture analysis.

(iii) To initialize k=2 component mixtures, we compute $\bar{\underset{\sim}{X}}_{21} = (\underset{\sim}{X}_{min} + \bar{\underset{\sim}{X}}_{11})/2$, and $\bar{\underset{\sim}{X}}_{22} = (\bar{\underset{\sim}{X}}_{11} + \underset{\sim}{X}_{max})/2$ to be entered as the component mean vectors in the mixture analysis.

(iv) To initialize k=3 component mixtures, we compute $\bar{\underset{\sim}{X}}_{31} = (\underset{\sim}{X}_{min} + \bar{\underset{\sim}{X}}_{21})/2$, $\bar{\underset{\sim}{X}}_{32} = (\bar{\underset{\sim}{X}}_{21} + \bar{\underset{\sim}{X}}_{22})/2$, and $\bar{X}_{33} = (\bar{\underset{\sim}{X}}_{22} + \underset{\sim}{X}_{max})/2$ to be entered as the component mean vectors in the mixture analysis, and so on.

Thus, we continue in this fashion until we generate all the initial mean vectors sequentially, and until we reach the larger hypothesized number of component clusters K. In doing this, we remain in the range of the data on the variable vector $\underset{\sim}{X}$. Such an initialization scheme sets up cluster centers regularly spaced at intervals on each variable which is less expensive and easy to program. Of course, we can also consider outer points (i.e., the points outside of the data range) and use the above initialization scheme to initialize the mixture and other clustering algorithms, which we did not pursue it here.

Our results obtained from this special initialization scheme are shown in Tables 6.4.1 and 6.4.2.

TABLE 6.4.1. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE DIFFERENT BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m[b]	AIC*(k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.748
2	337.008	29	-587.016**	-528.709*
3	371.177	44	-610.234*	-521.887**
4	381.395	59	-585.790	-467.163
5	405.493	74	-588.986	-440.200
6	426.428	89	-585.856	-406.911
K=7	433.193	104	-554.386	-345.279

TABLE 6.4.2. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE EQUAL BETWEEN CLUSTERS

No. of Types k	ln[max L(k)][a]	No. of Parameters m[b]	AIC*(k)[c]	SC(k)[d]
1	171.448	14	-300.896	-272.748
2	254.915	19	-452.830	-414.629
3	295.009	24	-518.018	-469.763
4	315.296	29	-543.592	-485.284
5	333.998	34	-565.996**	-497.635*
6	341.242	39	-565.448	-487.070
K=7	355.339	44	-578.678*	-490.210**

* First Minimum AIC* and SC

** Second Minimum AIC* and SC

a, b, c, and d are as in Tables 6.1.1 and 6.1.2.

Examining each table carefully, starting with Table 6.4.1 where the covariance matrices are different between clusters (or types), we see that the first minimum AIC* is when k=3 types, the second minimum AIC* is when k=2 types. That is, when k=3 types we have the best mixture submodel. On the other hand, the first minimum SC occurs at k=2 types, and the second minimum SC occurs at k=3 types. Thus, according to SC k=2 types is the best mixture submodel. Comparing these results with the results of mixture analysis obtained from initializing the mixture algorithm by using "k-means" results given in Table 6.3.1, we clearly see that our initialization scheme gives better results than what is suggested in the literature.

In Table 6.4.2 where the covariance matrices are considered to be equal between clusters (or types), we see that the first minimum AIC* occurs at k=7 types and the second minimum AIC* occurs at k=5 types. SC favors the same mixture submodels but in the reversed order as compared to AIC*. Again these results are not surprising since the population covariance matrices of the three types of irises are not equal to each other, and ln[max L(k)] values are greatest for each component cluster in Table 6.4.1 as compared to the ln[max L(k)] values given in Table 6.4.2, except when k=1.

6.5. When Special Initialization Scheme is Used on Reordered Data

Finally, when we use the special initialization scheme presented in Section 6.4 on the reordered data to start the mixture algorithm to avoid the problem of local maxima of the likelihood function, we obtained the following results.

TABLE 6.5.1. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE EQUAL BETWEEN CLUSTERS

No. of Types k	ln[max L(k)]^a	No. of Parameters m^b	AIC*(k)^c	SC(k)^d
1	171.448	14	-300.896	-272.748
2	257.235	29	-427.470	-369.162
3	358.219	44	-584.438*	-495.970*
4	374.422	59	-571.884**	-453.217**
5	220.659#	74	-219.318#	- 70.532#
6	218.458#	89	-169.916#	9.029#
K=7	226.395#	104	-140.790#	- 68.314#

TABLE 6.5.2. THE AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA WHEN COVARIANCE MATRICES ARE EQUAL BETWEEN CLUSTERS

No. of Types k	ln[max L(k)]^a	No. of Parameters m^b	AIC*(k)^c	SC(k)^d
1	171.448	14	-300.896	-272.748
2	191.135	19	-325.270	-287.068
3	295.009	24	-518.018*	-469.763*
4	287.889	29	-488.778**	-430.470**
5	171.531#	34	-241.062#	-172.701#
6	171.559#	39	-226.118#	-147.704#
K=7	171.576#	44	-211.152#	-122.685#

* First Minimum AIC* and SC.

** Second Minimum AIC* and SC.

# AIC* and SC Values During 5th Iteration. Mixture Algorithm Halted at 6th Iteration. Singular Variance-Covariance Matrix.

a, b, c, and d are as in Tables 6.1.1 and 6.1.2.

Looking at Tables 6.5.1 and 6.5.2, we see that under both different and equal covariance matrices between clusters (or types), the first minimum AIC* and SC occur at k=3 types. The second minimum AIC* and SC occur at k=4 types. Thus, in this case according to AIC* and SC k=3 types is the best mixture submodel. Comparing the values of AIC* and SC for k=2,3, and 4 types in Table 6.5.1 and 6.5.2, respectively, we can see that the AIC* and SC values in Table 6.5.2 are larger than the AIC* and SC values in Table 6.5.1, suggesting to us that when we are clustering iris data, and in general, we should use different covariance matrices rather than using equal covariance matrices. Thus, model-selection criteria can also be used to decide whether or not to assume a common covariance matrix.

From the results in Table 6.5.1 and 6.5.2, we further note that it suffices to fit K=5 hypothesized number of mixtures to Fisher iris data rather than fitting K=7 multivariate normal mixtures.

7. Conclusions and Discussion

From our numerical results in Section 6, we see that model-selection criteria can indeed be used to estimate k, the number of component clusters (or types) in the mixture model, when we do not know the group structure of the data *a priori*.

Summarizing the results on the number of times the minimum AIC* and SC selected each mixture submodel across all the tables given in Section 6, we obtain the following frequencies.

TABLE 7.1. SUMMARY OF THE RESULTS OF AIC*(k)'s AND SC(k)'s FOR STANDARD MIXTURE MODEL FOR THE IRIS DATA

No. of Types k	Number of Times AIC*(k) Selected	Number of Times SC(k) Selected
1	0	0
2	2	5
3	4	2
4	0	2
5	1	1
6	0	0
K=7	3	0

Looking at Table 7.1, we see that AIC* identifies the correct group structure (i.e., k=3 types) in the Fisher iris data four times as compared to SC which identifies the correct structure twice. AIC* chooses k=2 types twice, SC chooses k=2 types five times indicating that SC favors lower-dimensional models as compared to AIC*. The case where k=7 types was chosen three times by AIC* corresponds to the results where the covariance matrices between clusters were assumed to be equal instead of different. In these applications, however, these criteria often agree in identifying the correct model.

In the literature, objections have been raised that minimizing the AIC* does not produce an asymptotically consistent estimate of the model. For this, we shall refer the reader to Schwarz (1978), Bhansali and Downham (1977). But as also mentioned by Larimore (1983), no strong reasons have been offered for why such consistency would be desirable or would give sensible results generally, since in most applications such as the one we presented in this paper, we can vary the class of alternative models but not the number of observations. As

Akaike (1981) states: ". . . This inconsistency of order determination does not necessarily mean a serious problem, as expected deviation of the fitted model in terms of entropy decreases to its minimum possible value as the data length tends to infinity. This means that the procedure is inconsistent in terms of our basic criterion. If AIC is replaced by

$$\begin{array}{l}-2\ \ln(\text{maximized likelihood})\\ +f(n)(\text{number of free parameters}),\end{array}$$

where f(n) is a function which increases without bound, yet such that $f(n)/n \to 0$, as n tends to infinity, then the corresponding MAICE produces a consistent estimate of the order when this does exist."

Therefore, consistency for a given class of models within a fixed number of observations is not a problem for a good model-selection criterion. Specially in classification and clustering problems we do not have to worry about consistency or the order of a model.

For example, from Table 7.1, we see that Schwarz' Criterion (SC) which is a consistent modified version of AIC, does not necessarily pick up the correct group structure more often than AIC* in the Fisher iris data even when it is known _a priori_ that there are three types of species of irises. So the question is: "What kinds of penalty should the decision maker pay while trying to expect consistency for the model when indeed no consistency problem exists in a finite sample situation?"

Thus, it seems that to argue consistency when data contains a finite sample size is fruitless. The performances of these model-selection criteria most often depend strongly on the class of models, on the nature of the prior specification corresponding to which these criteria are derived, and of course, on the type of data sets they are applied.

Thus, in concluding, we see that our numerical results clearly demonstrate the potential of both Akaike's Information Criterion (AIC), and Schwarz' Criterion (SC) in identifying the best clustering alternative or alternatives, and estimating the number of component clusters present in the mixture model. These model-selection criteria are defined without any reference to a particular null hypothesis and are measures of the badness of the model which are free from the ambiguities inherent in the application of conventional procedures.

Acknowledgements

The author extends his appreciation to Mr. Ronald Ball for modifying the NORMIX programs to include the model-selection criteria. This research was supported by Army Research Office Contract DAAG29-82-K-0155, at the University of Illinois at Chicago.

REFERENCES

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

________ (1974). A New Look at the Statistical Model Identification. IEEE Trans. on Automatic Control, AC-19, 716-723.

________ (1978). Time Series Analysis and Control Through Parametric Models. In Applied Time Series Analysis (D. F. Findley, Ed.). New York: Academic Press, 1-23.

________ (1981). Modern Development of Statistical Methods. In Trends and Progress in System Identification (P. Eykhoff, Ed.). New York: Pergamon Press, 169-184.

________ (1982). Prediction and Entropy. MRC Technical Summary Report #2397, June 1982, University of Wisconsin-Madison, Mathematics Research Center.

Beale, E. M. L. (1969). Cluster Analysis. London: Scientific Control Systems.

Behboodian, J. (1972). Information Matrix for a Mixture of Two Normal Distributions. J. Statist. Comp. Simul., 1, 295-314.

Bhansali, R. J., and Downham, D. Y. (1977). Some Properties of the Order of an Autoregressive Model Selected by a Generalization of Akaike's FPE Criterion. Biometrika, 64, 547-551.

Binder, D. A. (1978). Bayesian Cluster Analysis. Biometrika, 65, 31-38.

Boltzmann, L. (1877). Uber die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Warmetheorie und der Wahrscheienlichkeitsrechnung respective den Satzen uber das Warmegleichgewicht. Wiener Berichte, 76, 373-435.

Calinski, T., and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. Communications in Statistics, 3, 1-27.

Day, N. E. (1969). Estimating the Components of a Mixture of Normal Distributions. Biometrika, 56, 463-474.

Dubes, R., and Jain, A. K. (1979). Validity Studies in Clustering Methodologies. Pattern Recognition, 11, 235-254.

Everitt, B. S. (1974). Cluster Analysis. London: Heinemann Educational Books.

___________ (1979). Unresolved Problems in Cluster Analysis. Biometrics, 35, 169-181.

Everitt, B. S. (1981). A Monte Carlo Investigation of the Likelihood Ratio Test for the Number of Components in a Mixture of Normal Distributions. Multivariate Behavioral Res., 16, 171-180.

Feder, P. (1968). On the Distribution of the Log Likelihood Ratio Test Statistic When the True Parameters are "Near" the Boundaries of the Hypothesis Regions. Ann. Mathematical Statistics, 39, 2044-2055.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Ann. of Eugenics, 7, 179-188.

Fleiss, J. L., and Zubin, J. (1969). On the Methods and Theory of Clustering. Multivariate Behavioral Res., 4, 235-50.

Friedman, H. P., and Rubin, J. (1967). On Some Invariant Criteria for Grouping Data. J. American Statistical Asso., 62, 1159-1178.

Gnanadesikan, R., and Wilk, M. .B. (1969). Data Analytic Methods in Multivariate Statistical Analysis. In Multivariate Analysis II (P. R. Krishnaiah, Ed.). New York: Academic Press, 593-638.

Hartigan, J. A. (1975). Clustering Algorithms. New York: John-Wiley and Sons, Inc.

____________ (1977). Distribution Problems in Clustering. In Classification and Clustering. (J. Van Ryzin, Ed.). New York: Academic Press, 45-71.

Hodges, J. L., and Lehmann, E. L. (1954). Testing the Approximate Validity of Statistical Hypotheses. J. Roy. Stat. Soc. Ser. B. 16, 261-268.

John, S. (1970). On Identifying the Population of Origin of Each Observation in a Mixture of Observations from Two Normal Populations. Technometrics, 12, 553-563.

Kendall, M. G. (1966). Discrimination and Classification. In Multivariate Analysis (P. R. Krishnaiah, Ed.). New York: Academic Press.

Kullback, S. (1959). Information Theory and Statistics. New York: John-Wiley and Sons, Inc.

Larimore, W. E. (1983). Predictive Inference, Sufficiency, Entropy and Asymptotic Likelihood Principle. Biometrika, 70, 175-181.

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

Maronna, R., and Jacovkis, P. M. (1974). Multivariate Clustering Procedures with Variable Metrics. Biometrics, 30, 499-505.

Marriott, F. H. C. (1971). Practical Problems in a Metnod of Cluster Analysis. Biometrics, 27, 501-514.

Matusita, K., and Noboru, O. (1980). A Criterion for Choosing the Number of Clusters in Cluster Analysis. In *Recent Developments in Statistical Inference and Data Analysis* (K. Matusita, Ed.) Amsterdam: North-Holland Publishing Co., 203-213.

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

Milligan, G. W. (1981). A Discussion of Procedures for Determining the Number of Clusters in a Data Set. Twelfth Annual Meeting, The Classification Society (North American Branch), May 31-June 2, 1981, Toronto, Canada.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* *6*, 461-464.

Sclove, S. L. (1977). Population Mixture Models and Clustering Algorithms. *Communications in Statistics-Theor. Meth.*, A6(*5*), 417-434.

__________ (1982). Application of the Conditional Population-Mixture Model to Image Segmentation Technical Report A82-1, Army Research Office Contract DAAG29-82-K-0155, University of Illinois at Chicago. To appear in *IEEE Trans. Pattern Analysis and Machine Intelligence*.

Scott, A. J., and Symons, M. J. (1971). Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics*, 27, 387-397.

Smith, A. F., and Spiegelhalter, D. J. (1980). Bayes Factors and Choice Criteria for Linear Models. *J. R. Statist. Soc.*, *B42*, 213-220.

Sokal, R. R. (1977). Clustering and Classification: Background and Current Directions. In *Classification and Clustering* (J. Van Ryzin, Ed.). New York: Academic Press, 1-15.

Solomon, H. (1971). *Numerical Taxonomy*. Mathematics in the Archaeological and Historical Sciences. Edinburgh: Edinburgh University Press.

Stone, C. J. (1982). Local Asymptotic Admissibility of a Generalization of Akaike's Model Selection Rule. *Ann. Inst. Statist. Math.*, *34A*, 123-133.

Symons, M. J. (1981). Clustering Criteria and Multivariate Normal Mixtures. *Biometrics*, *37*, 35-43.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Trans. of the American Math. Society*, *54*, 426-482.

Wolfe, J. H. (1967). NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. Research Memorandum, SRM 68-2. San Diego: U.S. Naval Personnel Research Activity.

__________ (1970). Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Res.*, *5*, 329-350.

__________ (1971). A Monte-Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distribution. Research Memorandum 72-2, San Diego, California: U.S. Naval Personnel and Training Research Laboratory.

LMED

8