SAS® Technical Report A-108

Cubic Clustering Criterion

SAS Institute Inc.

5903

SAS® Technical Report A-108
Cubic Clustering Criterion

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS® Technical Report A-108, Cubic Clustering Criterion*, Cary, NC: SAS Institute Inc., 1983. 56 pp.

SAS® Technical Report A-108, Cubic Clustering Criterion

ISBN 1-58025-185-4

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, November 1983
2nd printing, August 1998

Note that text corrections may have been made at each printing.

Table of Contents

Abstract. 1

Introduction. 1

Computation of the Cubic Clustering Criterion 4

Empirical Examination of the Performance of the CCC 8

Examples. 30

Conclusion. 49

References. 50

CUBIC CLUSTERING CRITERION

Abstract

The cubic clustering criterion (CCC) can be used to estimate the number of clusters using Ward's minimum variance method, k-means, or other methods based on minimizing the within-cluster sum of squares. The performance of the CCC is evaluated by Monte Carlo methods.

Introduction

The most widely used optimization criterion for disjoint clusters of observations is known as the within-cluster sum of squares, WSS, error sum of squares, ESS, residual sum of squares, least squares, (minimum) squared error, (minimum) variance, (sum of) squared (Euclidean) distances, trace(**W**), (proportion of) variance accounted for, or R^2 (see, for example, Anderberg 1973; Duran and Odell 1974; Everitt 1980). The following notation is used herein to define this criterion:

n = number of observations

n_k = number of observations in the k^{th} cluster

p = number of variables

q = number of clusters

X = n by p data matrix

$\bar{\mathbf{X}}$ = q by p matrix of cluster means

Z = cluster indicator matrix with element z_{ik} = 1 if the i^{th} observation belongs to the k^{th} cluster, 0 otherwise.

Assume that without loss of generality each variable has mean zero. Note that Z'Z is a diagonal matrix containing the n_ks and that

$$\bar{\mathbf{X}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \quad .$$

The total-sample sum-of-squares and crossproducts (SSCP) matrix is

$$\mathbf{T} = \mathbf{X}'\mathbf{X} \quad .$$

The between-cluster SSCP matrix is

$$\mathbf{B} = \boldsymbol{X}'\mathbf{Z}'\mathbf{Z}\boldsymbol{X} \quad .$$

The within-cluster SSCP matrix is

$$\begin{aligned}\mathbf{W} &= (\mathbf{X} - \mathbf{Z}\boldsymbol{X})'(\mathbf{X} - \mathbf{Z}\boldsymbol{X}) \\ &= \mathbf{X}'\mathbf{X} - \boldsymbol{X}'\mathbf{Z}'\mathbf{Z}\boldsymbol{X} \\ &= \mathbf{T} - \mathbf{B} \quad .\end{aligned}$$

The within-cluster sum of squares pooled over variables is thus trace(W). By changing the order of the summations, it can also be shown that trace(W) equals the sum of squared Euclidean distances from each observation to its cluster mean.

Since T is constant for a given sample, minimizing trace(W) is equivalent to maximizing

$$R^2 = 1 - \frac{\mathrm{trace}(\mathbf{W})}{\mathrm{trace}(\mathbf{T})} \quad ,$$

which has the usual interpretation of the proportion of variance accounted for by the clusters. R^2 can also be obtained by multiple regression if the columns of X are stacked on top of each other to form an np by 1 vector, and this vector is regressed on the Kronecker product of Z with an order p identity matrix.

Many algorithms have been proposed for maximizing R^2 or equivalent criteria (for example, Ward 1963; Edwards and Cavalli-Sforza 1965; MacQueen 1967; Gordon and Henderson 1977). This report concentrates on Ward's method as implemented in the CLUSTER procedure. Similar results should be obtained with other algorithms, such as the k-means method provided by FASTCLUS.

The most difficult problem in cluster analysis is how to determine the number of clusters. If you are using a goodness-of-fit criterion such as R^2, you would like to know the sampling distribution of the criterion to enable tests of cluster significance.

Ordinary significance tests, such as analysis of variance F tests, are not valid for testing differences between clusters. Since clustering methods attempt to maximize the separation between clusters, the assumptions of the usual significance tests, parametric or nonparametric, are drastically violated. For example, 25 samples of 100 observations from a single univariate normal distribution were each divided into two clusters by FASTCLUS. The median absolute t statistic testing the difference between the cluster means was 13.7, with a range from 10.9 to 15.7. For a nominal significance level of .0001 under the usual, but invalid, assumptions, the critical value is 3.4, yielding an actual type 1 error rate close to 1.

The first step in devising a valid significance test for clusters is to specify the null and alternative hypotheses. For clustering methods based on distance matrices, a popular null hypothesis is that all permutations of the values in the distance matrix are equally likely (Ling 1973; Hubert 1974). Using this null hypothesis you can do a permutation test or a rank test. The trouble with the permutation hypothesis is that, with any real data, the null hypothesis is totally implausible even if the data do not contain clusters. Rejecting the null hypothesis does not provide any useful information (Hubert and Baker 1977).

Another common null hypothesis is that the data are a random sample from a multivariate normal distribution (Wolfe 1970, 1978; Lee 1979). The multivariate normal null hypothesis is better than the permutation null hypothesis, but it is not satisfactory because there is typically a high probability of rejection if the data are sampled from a distribution with lower kurtosis than a normal distribution, such as a uniform distribution. The tables in Englemann and Hartigan (1969), for example, generally lead to rejection of the null hypothesis when the data are sampled from a uniform distribution. Hartigan (1978) and Arnold (1979) discuss both normal and uniform null hypotheses, and the uniform null hypothesis seems preferable for most practical purposes.

Hartigan (1978) has obtained asymptotic distributions for the within-cluster sum of squares criterion in one dimension for normal and uniform distributions. Hartigan's results require very large sample sizes, perhaps 100 times the number of clusters, and are, therefore, of limited practical use.

This report describes a rough approximation to the distribution of the R^2 criterion under the null hypothesis that the data have been sampled from a uniform distribution on a hyperbox (a p-dimensional right parallelepiped). This approximation is helpful in determining the best number of clusters for both univariate and multivariate data and with sample sizes down to 20 observations. The approximation to the expected value of R^2 is based on the assumption that the clusters are shaped roughly like hypercubes. In more than one dimension, this approximation tends to be conservative for a small number of clusters and slightly liberal for a very large number of clusters (about 25 or more in two dimensions). The cubic clustering criterion (CCC) is obtained by comparing the observed R^2 to the approximate expected R^2 using an approximate variance-stabilizing transformation. Positive values of the CCC mean that the obtained R^2 is greater than would be expected if sampling from a uniform distribution and therefore indicate the possible presence of clusters. Treating the CCC as a standard normal test statistic provides a crude test for the hypotheses:

H_0: the data have been sampled from a uniform distribution on a hyperbox.

H_a: the data have been sampled from a mixture of spherical multivariate normal distributions with equal variances and equal sampling probabilities.

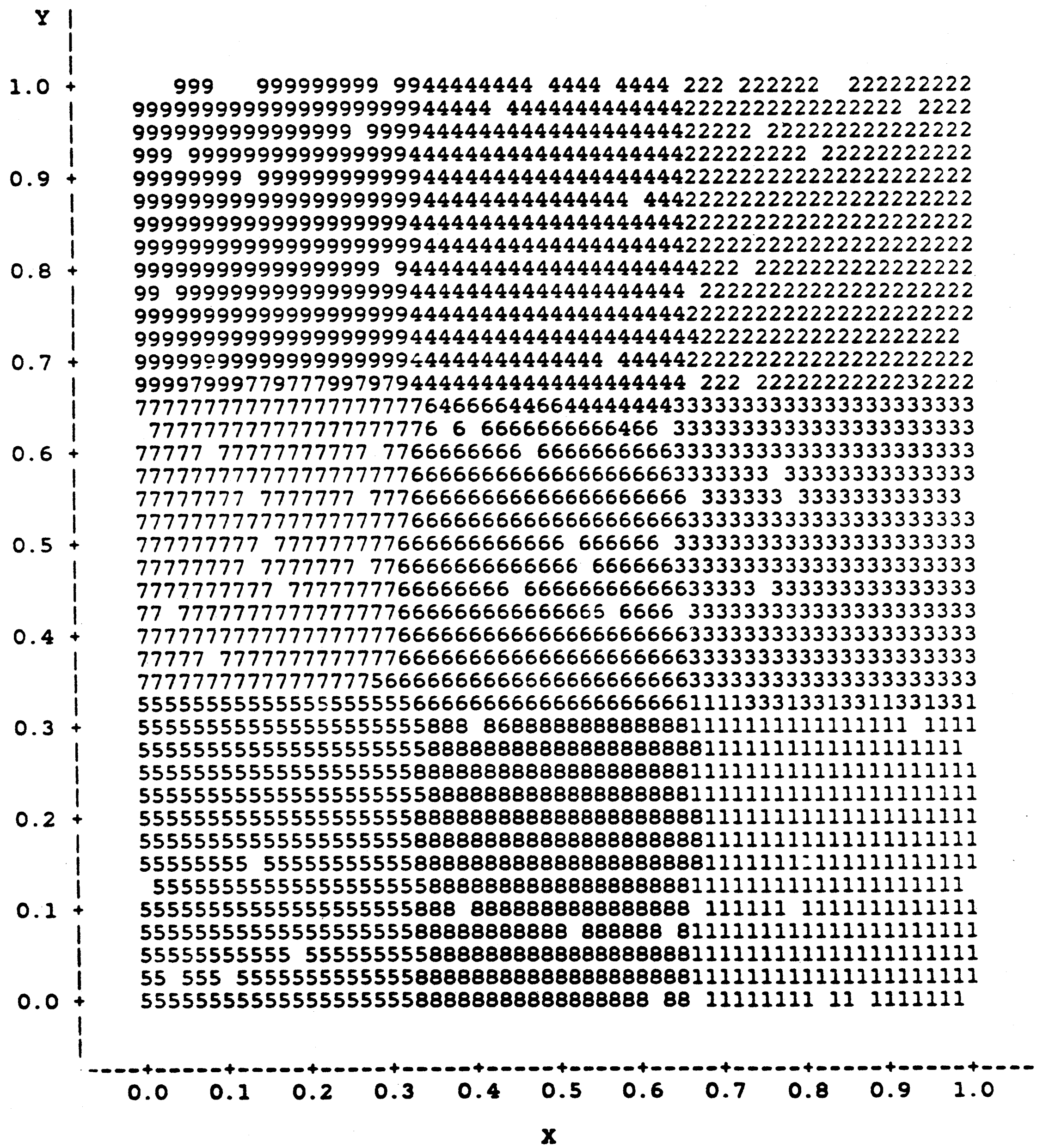
Under this alternative hypothesis, R^2 is equivalent to the maximum-likelihood criterion (Scott and Symons 1971).

Computation of the Cubic Clustering Criterion

The CCC is based on the assumption that clusters obtained from a uniform distribution on a hyperbox are hypercubes of the same size. The hypercube assumption is obviously false in most cases, but is generally conservative unless the number of clusters is very large in two or more dimensions. Wong (1982) has shown that, for many clusters in two dimensions from a uniform sample, the cluster shape tends to be hexagonal.

Figure 1 illustrates a case in which the hypercube (or square, since there are only two dimensions) assumption is correct. A sample of 10,000 points from a uniform distribution on the unit square was divided into nine clusters by FASTCLUS. Each cluster is nearly square with edge length 1/3.

FIGURE 1
NINE CLUSTERS FROM A UNIFORM DISTRIBUTION ON A UNIT SQUARE



NOTE: 7573 OBS HIDDEN

A first approximation to the value of R^2 for a population uniformly distributed on a hyperbox can be obtained as follows. Assume that the edges of the hyperbox are aligned with the coordinate axes. Let s_j be the edge length of the hyperbox along the j^{th} dimension. Assume further that the s_js are in decreasing order. The volume of the hyperbox is

$$v = \prod_{j=1}^{p} s_j \quad .$$

If the hyperbox is divided into q hypercubes with edge length c, then the volume of the hyperbox equals the total volume of the hypercubes, hence

$$c = (\frac{v}{q})^{\frac{1}{p}} \quad .$$

Let

$$u_j = \frac{s_j}{c}$$

be the number of hypercubes along the j^{th} dimension of the hyperbox. The total-sample variance along the j^{th} dimension is proportional to s_j^2, while the within-cluster variance along the j^{th} dimension is proportional to c^2. Thus

$$\begin{aligned} R^2 &\doteq 1 - \frac{\sum_{j=1}^{p} c^2}{\sum_{j=1}^{p} s_j^2} \\ &= 1 - \frac{p}{\sum_{j=1}^{p} u_j^2} \quad . \end{aligned}$$

In Figure 1, for example, $s_1=s_2=1$, $c=1/3$, and $u_1=u_2=3$, so the population R^2 is

$$R^2 \doteq 1 - \frac{2}{(3^2+3^2)}$$
$$= 0.88888\ldots ,$$

while the sample R^2 is 0.88967+.

The above approximation fails badly if the dimensionality of the between-cluster variation, say p^*, is less than p. Obviously, p^* must be less than the number of clusters, q. Also, $u_j<1$ implies $p^*<j$. For a better approximation, assume the clusters are hyperboxes with edge length c in the first p^* dimensions, and edge length s_j in the remaining dimensions. Let

$$v^* = \prod_{j=1}^{p^*} s_j \quad ,$$

$$c = \left(\frac{v^*}{q}\right)^{\frac{1}{p^*}} \quad ,$$

$$u_j = \frac{s_j}{c} \quad ,$$

where p^* is chosen to be the largest integer less than q such that u_{p^*} is not less than one. Then we have the following approximation to the population R^2:

$$R^2 \doteq 1 - \frac{p^* + \sum_{j=p^*+1}^{p} u_j^2}{\sum_{j=1}^{p} u_j^2} \quad .$$

In small samples from a uniform distribution on a hyperbox, the sample R^2s tend to exceed the population R^2 due to the phenomenon widely known as "capitalization on chance." Extensive simulations led to the following heuristic small-sample approximation for the expected value of R^2:

$$E(R^2) \doteq 1 - \left[\frac{\sum_{j=1}^{p^*} \frac{1}{n+u_j} + \sum_{j=p^*+1}^{p} \frac{u_j^2}{n+u_j}}{\sum_{j=1}^{p} u_j^2}\right]\left[\frac{(n-q)^2}{n}\right]\left[1 + \frac{4}{n}\right] .$$

Given a sample X, let s_j be the square root of the j^{th} eigenvalue of $T/(n-1)$, so that under the null hypothesis the length of hyperbox in the j^{th} dimension is proportional to the standard deviation of the j^{th} principal component of the data. The CCC is computed from the observed R^2 as

$$CCC = ln\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{(.001+E(R^2))^{1.2}} .$$

The above formula was derived empirically in an attempt to stabilize the variance across different numbers of observations, variables, and clusters.

Empirical Examination of the Performance of the CCC

A Monte Carlo study of the null distribution of the CCC was performed by clustering samples from uniform distributions on hypercubes by Ward's minimum variance method as implemented in the CLUSTER procedure. The design involved two factors:

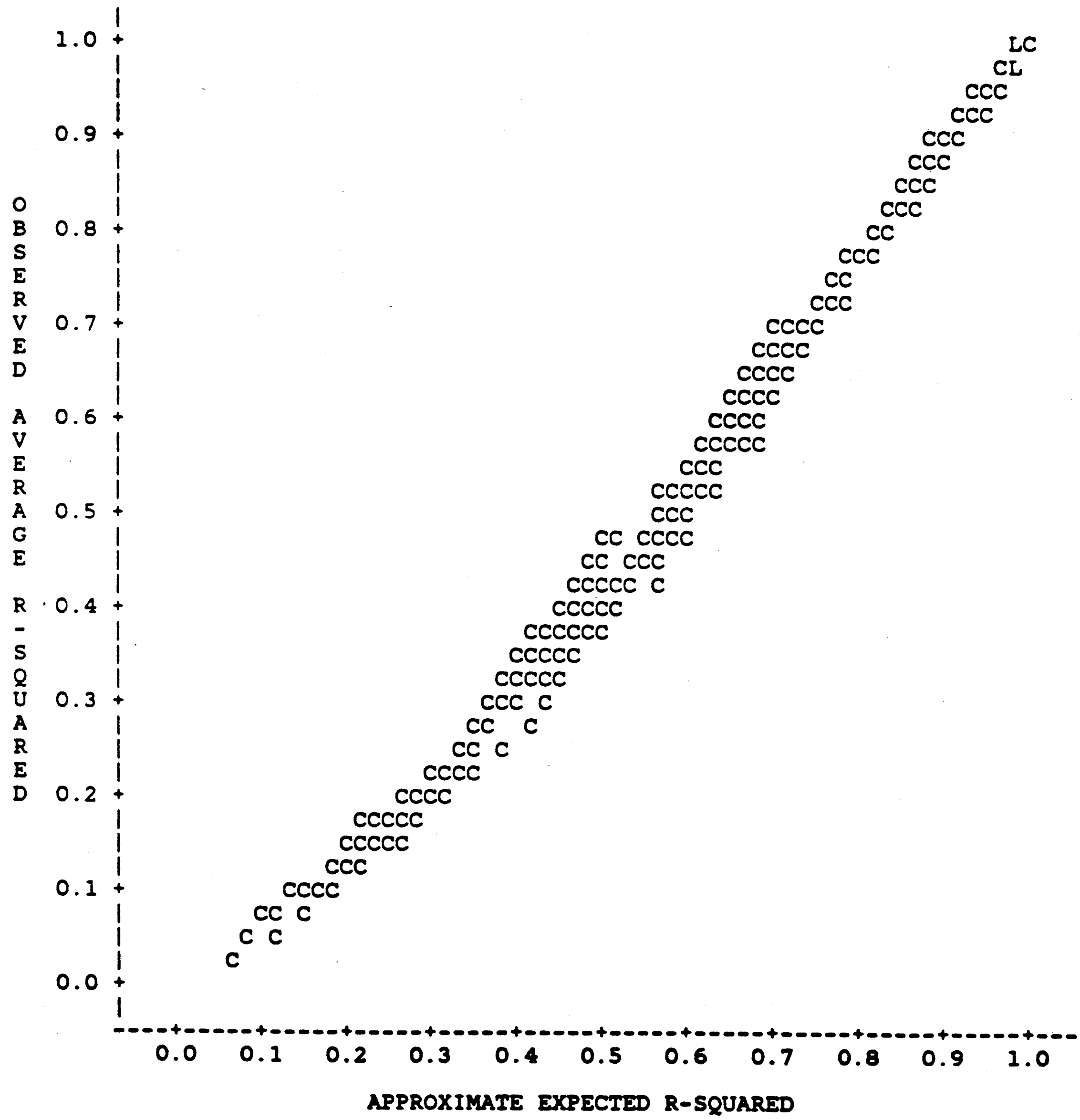
- The number of observations was 20, 40, 80, 160, 320, or 640.

- The number of variables was 1, 2, 4, 8, or 16.

For each combination of these factors, 50 samples were generated from a uniform distribution on a hypercube. Each sample was clustered and $E(R^2)$ and the CCC were computed with the number of clusters ranging from one to one-tenth the number of observations.

Figure 2 is a plot of the observed average R^2 against the theoretical approximate expected R^2. Each combination of number of observations, number of variables, and number of clusters is represented by a point giving the mean of 50 values of the observed R^2 and the approximate $E(R^2)$. Points for which the observed R^2 exceeds $E(R^2)$ are labeled "L" for liberal, while the remaining points are labeled "C" for conservative. If both liberal and conservative points fell at a given plotting position, "L" was printed. The vast majority of the points are mildly conservative.

FIGURE 2
PLOT OF OBSERVED AVERAGE R-SQUARED AGAINST
THEORETICAL APPROXIMATE EXPECTED R-SQUARED
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
POINT LABELS: C=CONSERVATIVE L=LIBERAL



NOTE: 453 OBS HIDDEN

Table 1 gives the mean and standard deviation of the CCC for each combination of number of observations, clusters, and variables. Nearly all the means are negative, showing that the CCC is generally conservative. The only exceptions are for 56 or more clusters with 640 observations and 2 variables. For a given number of observations and variables, the mean CCC reaches a minimum when the number of clusters is close to the number of variables plus one, in which case the assumption of hypercubical clusters is badly violated. The CCC becomes extremely conservative for 16 variables, especially with a large number of observations. The standard deviations are generally close to 1.0, but are larger for a small number of clusters, especially with the larger sample sizes for two clusters.

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 20

	VARIABLES									
	1		2		4		8		16	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS										
2	-0.6	1.2	-0.7	0.6	-1.2	0.4	-1.5	0.3	-1.8	0.2

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 40

	VARIABLES									
	1		2		4		8		16	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS										
2	-0.7	1.4	-1.0	0.7	-1.5	0.5	-2.0	0.4	-2.5	0.4
3	-0.5	1.1	-1.6	1.1	-2.2	0.6	-2.6	0.4	-3.3	0.4
4	-0.5	1.1	-1.0	1.1	-2.5	0.7	-3.0	0.4	-3.8	0.3

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 80

	VARIABLES									
	1		2		4		8		16	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS										
2	-0.7	1.6	-1.3	0.9	-2.3	0.8	-3.2	0.7	-3.8	0.5
3	-0.8	1.2	-2.8	1.2	-3.3	0.8	-4.1	0.7	-4.8	0.5
4	-0.7	1.2	-1.2	1.4	-4.0	0.9	-4.6	0.8	-5.4	0.5
5	-0.5	1.1	-1.1	1.2	-4.4	1.1	-5.0	0.8	-6.0	0.5
6	-0.6	1.1	-1.0	1.2	-3.6	1.1	-5.2	0.9	-6.4	0.5
7	-0.4	1.3	-0.8	1.1	-2.9	1.1	-5.3	1.0	-6.7	0.5
8	-0.4	1.1	-0.6	1.0	-2.4	1.1	-5.3	1.0	-6.9	0.6

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 160

	VARIABLES									
	1		2		4		8		16	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS										
2	-1.2	2.2	-1.9	1.2	-3.3	1.0	-5.1	0.7	-6.2	0.7
3	-1.6	1.4	-4.9	1.3	-5.0	0.9	-6.6	0.8	-7.8	0.7
4	-1.4	1.3	-2.7	1.6	-6.3	0.8	-7.5	0.8	-8.8	0.8
5	-1.2	1.2	-2.4	1.3	-7.4	0.9	-8.2	0.9	-9.5	0.8
6	-1.2	1.1	-2.0	1.1	-6.5	0.8	-8.8	1.0	-10.0	0.8
7	-1.1	1.0	-1.8	0.9	-5.5	1.0	-9.0	1.0	-10.4	0.8
8	-1.1	1.0	-1.5	0.9	-4.8	1.0	-9.1	1.1	-10.6	0.8
9	-1.1	1.0	-1.2	0.9	-4.1	1.1	-9.0	1.2	-10.8	0.8
10	-1.1	1.1	-1.1	0.9	-3.5	1.1	-8.4	1.2	-10.8	0.9
11	-1.1	1.0	-1.0	0.9	-3.0	1.1	-7.8	1.2	-10.8	0.9
12	-1.0	1.0	-0.8	0.9	-2.6	1.1	-7.2	1.2	-10.8	1.0
13	-0.9	1.1	-0.7	1.0	-2.3	1.1	-6.8	1.1	-10.6	1.0
14	-0.9	1.2	-0.6	1.0	-2.0	1.1	-6.4	1.1	-10.4	1.0
15	-0.9	1.2	-0.5	1.0	-1.7	1.1	-6.0	1.1	-10.0	1.0
16	-0.8	1.2	-0.4	1.1	-1.5	1.1	-5.6	1.1	-9.7	1.0

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 320

	VARIABLES									
	1		2		4		8		16	
CLUSTERS	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
2	-3.1	2.4	-2.6	1.4	-4.9	1.6	-8.4	0.8	-10.7	0.9
3	-1.9	2.0	-7.3	1.4	-7.7	1.3	-10.7	0.8	-13.4	0.9
4	-2.2	1.4	-4.5	1.6	-10.2	1.4	-12.4	0.9	-15.1	0.8
5	-2.2	1.5	-4.4	1.5	-12.4	1.3	-13.7	0.9	-16.4	0.8
6	-1.7	1.3	-4.1	1.3	-10.9	1.2	-14.8	0.9	-17.3	0.8
7	-1.7	1.1	-3.7	1.4	-9.5	1.2	-15.6	1.0	-18.1	0.7
8	-1.5	1.1	-3.2	1.3	-8.3	1.2	-16.2	1.1	-18.6	0.7
9	-1.5	1.0	-2.9	1.4	-7.5	1.1	-16.5	1.1	-19.0	0.7
10	-1.5	0.9	-2.6	1.3	-6.6	1.2	-15.4	1.0	-19.2	0.7
11	-1.4	1.0	-2.4	1.3	-5.9	1.2	-14.5	0.9	-19.4	0.8
12	-1.3	1.1	-2.2	1.2	-5.3	1.2	-13.7	0.9	-19.5	0.8
13	-1.3	1.1	-2.0	1.1	-4.8	1.1	-13.0	0.9	-19.4	0.8
14	-1.3	1.1	-1.8	1.0	-4.3	1.1	-12.3	0.8	-19.3	0.9
15	-1.3	1.0	-1.6	1.0	-4.0	1.1	-11.7	0.8	-19.2	0.9
16	-1.2	0.9	-1.5	1.0	-3.7	1.1	-11.1	0.8	-18.8	1.0
17	-1.2	0.9	-1.4	1.0	-3.5	1.0	-10.6	0.9	-18.1	1.0
18	-1.2	1.0	-1.3	1.0	-3.2	1.0	-10.1	0.9	-17.4	1.0
19	-1.2	1.0	-1.2	1.0	-3.0	1.0	-9.6	0.9	-16.8	1.0

(CONTINUED)

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 320

	VARIABLES									
	1		2		4		8		16	
CLUSTERS	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
20	-1.2	1.0	-1.1	1.0	-2.8	0.9	-9.1	1.0	-16.2	1.0
21	-1.1	1.0	-1.0	1.1	-2.6	0.9	-8.7	1.0	-15.8	1.0
22	-1.1	1.0	-0.9	1.1	-2.5	0.9	-8.2	1.0	-15.3	1.0
23	-1.0	1.0	-0.8	1.1	-2.3	1.0	-7.8	1.0	-14.8	1.0
24	-1.0	1.0	-0.7	1.1	-2.1	1.0	-7.5	1.0	-14.3	1.0
25	-1.0	1.0	-0.7	1.1	-1.9	1.0	-7.1	1.0	-13.9	1.0
26	-1.0	1.0	-0.6	1.1	-1.8	1.0	-6.8	1.0	-13.5	0.9
27	-1.0	0.9	-0.5	1.1	-1.6	1.0	-6.5	1.0	-13.1	0.9
28	-1.0	0.9	-0.4	1.1	-1.5	1.0	-6.2	1.0	-12.8	1.0
29	-1.0	0.9	-0.4	1.1	-1.4	1.0	-6.0	1.0	-12.4	1.0
30	-0.9	0.9	-0.3	1.1	-1.2	1.0	-5.7	1.0	-12.1	1.0
31	-0.9	1.0	-0.3	1.1	-1.1	1.0	-5.5	1.0	-11.8	0.9
32	-1.0	1.0	-0.2	1.1	-1.0	1.0	-5.2	0.9	-11.5	0.9

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	VARIABLES									
	1		2		4		8		16	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS										
2	-3.5	3.2	-2.8	1.4	-7.2	1.9	-12.8	1.1	-17.3	0.9
3	-3.1	2.0	-10.4	1.7	-11.0	1.6	-16.8	1.2	-21.6	0.9
4	-3.2	1.9	-5.4	2.4	-15.2	1.5	-19.7	1.4	-24.6	0.9
5	-2.8	1.7	-5.7	1.8	-19.2	1.4	-22.1	1.3	-26.9	0.9
6	-2.9	1.4	-5.6	1.5	-17.3	1.2	-24.1	1.3	-28.7	0.9
7	-2.5	1.4	-5.2	1.3	-15.6	1.1	-25.9	1.3	-30.2	1.0
8	-2.1	1.2	-4.9	1.2	-14.0	1.1	-27.4	1.2	-31.4	1.0
9	-1.9	1.2	-4.6	1.3	-12.7	1.3	-28.6	1.2	-32.4	1.0
10	-2.0	1.0	-4.3	1.4	-11.5	1.4	-27.2	1.2	-33.2	1.0
11	-2.1	0.9	-4.1	1.2	-10.5	1.4	-26.0	1.2	-33.9	1.0
12	-2.2	0.9	-3.9	1.1	-9.6	1.4	-24.8	1.1	-34.5	1.1
13	-2.2	1.0	-3.8	1.0	-8.8	1.4	-23.8	1.1	-34.9	1.1
14	-2.1	1.0	-3.6	0.9	-8.2	1.3	-22.9	1.1	-35.1	1.1
15	-2.0	1.1	-3.4	0.9	-7.7	1.3	-22.0	1.1	-35.3	1.2
16	-1.8	1.1	-3.2	0.9	-7.3	1.3	-21.2	1.1	-35.3	1.2
17	-1.8	1.1	-3.1	0.9	-7.0	1.3	-20.3	1.1	-35.3	1.2
18	-1.8	1.1	-2.9	0.9	-6.7	1.3	-19.6	1.1	-34.1	1.1
19	-1.8	1.0	-2.8	0.9	-6.5	1.2	-18.9	1.1	-33.0	1.1

(CONTINUED)

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	VARIABLES									
	1		2		4		8		16	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS										
20	-1.8	1.0	-2.7	0.9	-6.2	1.2	-18.2	1.1	-32.0	1.1
21	-1.8	1.0	-2.6	0.9	-6.0	1.2	-17.5	1.2	-31.0	1.1
22	-1.8	1.0	-2.5	0.9	-5.8	1.1	-16.9	1.2	-30.1	1.1
23	-1.8	0.9	-2.4	0.9	-5.6	1.1	-16.3	1.1	-29.3	1.1
24	-1.8	1.0	-2.3	0.9	-5.3	1.1	-15.7	1.1	-28.5	1.1
25	-1.8	1.0	-2.2	0.9	-5.1	1.1	-15.2	1.1	-27.7	1.1
26	-1.7	1.1	-2.1	0.9	-4.9	1.0	-14.7	1.1	-26.9	1.1
27	-1.7	1.1	-2.0	0.8	-4.7	1.0	-14.1	1.1	-26.2	1.1
28	-1.6	1.1	-1.9	0.9	-4.5	1.0	-13.6	1.1	-25.5	1.1
29	-1.5	1.1	-1.8	0.9	-4.3	1.0	-13.2	1.1	-24.9	1.1
30	-1.5	1.0	-1.8	0.9	-4.1	1.1	-12.8	1.1	-24.2	1.1
31	-1.4	1.0	-1.7	0.9	-3.9	1.1	-12.3	1.0	-23.6	1.0
32	-1.4	1.0	-1.6	0.9	-3.8	1.1	-12.0	1.0	-23.1	1.0
33	-1.5	1.0	-1.5	0.9	-3.6	1.1	-11.6	1.0	-22.5	1.0
34	-1.5	1.0	-1.4	0.9	-3.4	1.1	-11.2	1.0	-22.0	1.0
35	-1.5	1.0	-1.3	0.9	-3.3	1.1	-10.9	1.0	-21.5	1.0
36	-1.5	1.0	-1.2	0.9	-3.1	1.1	-10.5	1.0	-21.0	1.0
37	-1.5	1.0	-1.1	0.9	-3.0	1.1	-10.2	1.1	-20.5	1.0

(CONTINUED)

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	VARIABLES									
	1		2		4		8		16	
CLUSTERS	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
38	-1.6	1.0	-1.0	0.9	-2.9	1.1	-9.9	1.1	-20.1	1.0
39	-1.6	1.0	-0.9	0.9	-2.7	1.1	-9.6	1.1	-19.6	1.0
40	-1.6	1.0	-0.8	0.9	-2.6	1.0	-9.3	1.1	-19.2	1.0
41	-1.6	1.0	-0.8	0.9	-2.5	1.0	-9.0	1.1	-18.8	1.0
42	-1.6	1.0	-0.7	0.9	-2.3	1.0	-8.7	1.1	-18.4	1.0
43	-1.6	1.0	-0.6	0.9	-2.2	1.0	-8.5	1.1	-18.1	1.0
44	-1.6	1.0	-0.5	0.9	-2.1	1.0	-8.2	1.0	-17.7	1.0
45	-1.6	1.0	-0.5	0.9	-2.0	1.0	-7.9	1.0	-17.4	0.9
46	-1.5	1.0	-0.4	0.9	-1.9	1.0	-7.7	1.0	-17.0	0.9
47	-1.5	1.0	-0.4	0.9	-1.7	1.0	-7.4	1.0	-16.7	0.9
48	-1.5	1.0	-0.3	0.9	-1.6	1.0	-7.2	1.0	-16.4	0.9
49	-1.5	1.0	-0.3	0.9	-1.5	1.0	-7.0	1.0	-16.1	0.9
50	-1.4	1.0	-0.3	0.9	-1.4	1.0	-6.8	1.0	-15.8	0.9
51	-1.4	1.0	-0.2	0.9	-1.3	1.0	-6.5	1.0	-15.5	0.9
52	-1.4	1.0	-0.2	0.9	-1.2	1.0	-6.3	1.0	-15.2	0.9
53	-1.4	1.0	-0.1	0.9	-1.1	1.0	-6.1	1.0	-14.9	0.9
54	-1.4	1.0	-0.1	0.9	-1.1	1.0	-5.9	1.0	-14.7	0.9
55	-1.4	1.0	-0.0	0.9	-1.0	1.1	-5.8	1.0	-14.4	0.9

(CONTINUED)

TABLE 1
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	VARIABLES									
	1		2		4		8		16	
CLUSTERS	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
56	-1.4	1.0	0.0	0.9	-0.9	1.1	-5.6	1.0	-14.2	0.9
57	-1.4	1.0	0.1	0.9	-0.8	1.1	-5.4	1.0	-13.9	0.9
58	-1.4	1.0	0.1	0.9	-0.7	1.1	-5.2	1.0	-13.7	0.9
59	-1.4	1.0	0.2	0.9	-0.6	1.1	-5.1	1.0	-13.5	0.9
60	-1.3	1.0	0.2	0.8	-0.5	1.1	-4.9	0.9	-13.2	0.9
61	-1.3	1.0	0.3	0.8	-0.4	1.1	-4.8	0.9	-13.0	0.9
62	-1.3	0.9	0.3	0.8	-0.4	1.1	-4.6	0.9	-12.8	0.9
63	-1.3	0.9	0.4	0.8	-0.3	1.1	-4.5	0.9	-12.6	0.9
64	-1.3	0.9	0.5	0.8	-0.2	1.1	-4.3	0.9	-12.4	0.9

Figure 3 plots the probability of the CCC exceeding 2.0 for each combination of number of observations, clusters, and variables. All probabilities are less than .10 and most are less than .05. Table 2 shows the probability of the maximum CCC exceeding 2.0, where the maximum is taken over numbers of clusters, for each combination of number of observations and variables. All probabilities are less than .10. The maximum CCC exceeded 3.0 only once in the study, for 160 observations and 1 variable. Therefore, a CCC value exceeding 2 or 3 can be taken as evidence favoring rejection of the null hypothesis of a uniform distribution on a hyperbox, although a precise significance level cannot be specified.

FIGURE 3
PROBABILITY OF CCC EXCEEDING 2.0
PLOTTED AGAINST THE NUMBER OF CLUSTERS
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
PLOTTING SYMBOL IS FIRST DIGIT OF THE NUMBER OF OBSERVATIONS
NUMBER OF VARIABLES=1

```
    0.10 +
         |
    0.09 +
         |
    0.08 + 1
E        |
R   0.07 +
R        |
O   0.06 +
R        |
    0.05 +
R        |
A   0.04 + 8    88
T        |
E   0.03 +
         |
    0.02 + 2   1            3333
         |
    0.01 +
         |
    0.00 + 23448811111111113666633333333333366666666666666666666666666666666
         --+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+----
           1   5   9  13  17  21  25  29  33  37  41  45  49  53  57  61

                             NUMBER OF CLUSTERS

NOTE:     52 OBS HIDDEN
```

FIGURE 3
PROBABILITY OF CCC EXCEEDING 2.0
PLOTTED AGAINST THE NUMBER OF CLUSTERS
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
PLOTTING SYMBOL IS FIRST DIGIT OF THE NUMBER OF OBSERVATIONS
NUMBER OF VARIABLES=2

```
    0.10 +
         |
    0.09 +
         |
    0.08 +
E        |
R   0.07 +
R        |
O   0.06 +
R        |
    0.05 +
R        |
A   0.04 +
T        |
E   0.03 +
         |
    0.02 +   44 8                       333                    666666666666
         |
    0.01 +
         |
    0.00 + 2288818811111111333333333333333666666666666666666666
         --+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+----
           1   5   9  13  17  21  25  29  33  37  41  45  49  53  57  61

                              NUMBER OF CLUSTERS

NOTE:      56 OBS HIDDEN
```

FIGURE 3
PROBABILITY OF CCC EXCEEDING 2.0
PLOTTED AGAINST THE NUMBER OF CLUSTERS
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
PLOTTING SYMBOL IS FIRST DIGIT OF THE NUMBER OF OBSERVATIONS
NUMBER OF VARIABLES=4

```
    0.10 +
         |
    0.09 +
         |
    0.08 +
E        |
R   0.07 +
R        |
O   0.06 +
R        |
    0.05 +
R        |
A   0.04 +
T        |
E   0.03 +
         |
    0.02 +                                                             6666
         |
    0.01 +
         |
    0.00 + 224488881111111133333333333333336666666666666666666666666666
         --+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+----
           1   5   9  13  17  21  25  29  33  37  41  45  49  53  57  61

                              NUMBER OF CLUSTERS

NOTE:      62 OBS HIDDEN
```

FIGURE 3
PROBABILITY OF CCC EXCEEDING 2.0
PLOTTED AGAINST THE NUMBER OF CLUSTERS
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
PLOTTING SYMBOL IS FIRST DIGIT OF THE NUMBER OF OBSERVATIONS
NUMBER OF VARIABLES=8

```
     0.10 +
          |
     0.09 +
          |
     0.08 +
 E        |
 R   0.07 +
 R        |
 O   0.06 +
 R        |
     0.05 +
 R        |
 A   0.04 +
 T        |
 E   0.03 +
          |
     0.02 +
          |
     0.01 +
          |
     0.00 + 2244888811111111333333333333333366666666666666666666666666666666
          --+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+----
            1   5   9  13  17  21  25  29  33  37  41  45  49  53  57  61

                                 NUMBER OF CLUSTERS
```

NOTE: 62 OBS HIDDEN

FIGURE 3
PROBABILITY OF CCC EXCEEDING 2.0
PLOTTED AGAINST THE NUMBER OF CLUSTERS
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
PLOTTING SYMBOL IS FIRST DIGIT OF THE NUMBER OF OBSERVATIONS
NUMBER OF VARIABLES=16

```
     0.10 +
          |
     0.09 +
          |
     0.08 +
 E        |
 R   0.07 +
 R        |
 O   0.06 +
 R        |
     0.05 +
 R        |
 A   0.04 +
 T        |
 E   0.03 +
          |
     0.02 +
          |
     0.01 +
          |
     0.00 + 2244888811111111333333333333333366666666666666666666666666666666
          --+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+----
            1   5   9  13  17  21  25  29  33  37  41  45  49  53  57  61

                                 NUMBER OF CLUSTERS
```

NOTE: 62 OBS HIDDEN

TABLE 2
100 TIMES PROBABILITY OF MAXIMUM CCC EXCEEDING 2.0
CLASSIFIED BY NUMBER OF OBSERVATIONS AND VARIABLES
FOR UNIFORM HYPERCUBICAL DISTRIBUTIONS
EACH TABLE ENTRY IS BASED ON FIFTY SAMPLES

	VARIABLES				
	1	2	4	8	16
OBSERVATIONS					
20	2	0	0	0	0
40	2	2	0	0	0
80	8	2	0	0	0
160	8	2	0	0	0
320	2	2	0	0	0
640	0	2	2	0	0

The first Monte Carlo study examined hypercubical distributions. A second Monte Carlo was run to evaluate the CCC in uniform distributions on non-cubical hyperboxes. To keep computer time within reasonable limits the dimensionality was limited to four while the ranges were varied in three dimensions. The number of observations was 80, 160, 320, or 640. Fifty samples were generated in each cell. Table 3 gives the mean and standard deviation of the CCC for each combination of number of observations, clusters, and shape of hyperbox. The shapes are given as four numbers indicating the ranges in the four dimensions. Again the results are conservative. Error rates analogous to those in Table 2 were computed, and none exceeded the .02 level.

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 80

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
2	-2.4	0.7	-1.1	0.8	-1.5	0.8	-1.9	0.9	-0.8	1.0	-1.1	1.0	-1.2	1.0
3	-3.2	0.8	-2.1	0.7	-2.7	0.8	-2.9	0.8	-0.6	0.6	-1.9	0.9	-1.7	0.8
4	-4.0	0.8	-2.7	0.7	-1.9	1.1	-3.7	0.9	-1.0	0.6	-1.9	0.9	-2.5	0.8
5	-4.3	0.9	-3.2	0.8	-1.8	1.0	-2.9	0.9	-1.3	0.6	-1.6	0.8	-2.4	0.8
6	-3.6	1.0	-2.9	0.8	-1.8	0.9	-2.4	0.8	-1.5	0.6	-1.4	0.8	-2.2	0.9
7	-3.1	1.0	-2.7	0.9	-1.8	0.9	-1.9	0.9	-1.7	0.7	-1.3	0.9	-1.9	0.9
8	-2.6	0.9	-2.4	0.8	-1.8	0.8	-1.6	1.0	-1.8	0.7	-1.3	0.8	-1.7	0.9

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 80

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS						
2	-1.3	1.1	-1.5	0.8	-2.0	0.8
3	-2.7	1.2	-2.8	1.0	-2.8	0.9
4	-1.4	1.2	-2.0	1.1	-3.7	1.1
5	-1.4	0.9	-2.1	0.9	-2.7	1.2
6	-1.2	0.9	-2.1	0.8	-2.0	1.2
7	-1.1	0.8	-2.0	0.8	-1.5	1.1
8	-1.0	0.7	-1.9	0.8	-1.3	1.0

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 160

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
2	-3.5	1.0	-1.7	1.1	-2.0	1.1	-2.7	1.2	-1.7	1.5	-1.3	1.2	-1.1	1.0
3	-5.0	0.8	-2.9	1.0	-4.4	1.0	-4.5	0.7	-1.3	0.7	-2.5	1.0	-2.4	0.8
4	-6.2	0.9	-4.2	0.9	-3.1	1.4	-6.3	0.9	-1.6	0.8	-2.8	0.8	-3.8	0.8
5	-7.2	0.9	-5.1	1.0	-3.5	1.2	-4.8	1.0	-2.1	0.7	-2.4	0.8	-3.7	0.9
6	-6.0	1.0	-4.8	0.9	-3.6	1.2	-3.7	1.1	-2.6	0.7	-2.0	0.7	-3.4	1.0
7	-5.1	1.0	-4.5	0.9	-3.5	1.1	-3.1	1.1	-2.8	0.7	-1.8	0.8	-3.1	0.9
8	-4.4	1.0	-4.2	0.8	-3.4	1.0	-2.8	1.2	-2.9	0.6	-2.0	0.8	-2.8	0.9
9	-3.7	1.0	-3.9	0.8	-3.3	1.0	-2.6	1.1	-2.9	0.6	-2.0	0.9	-2.5	0.8
10	-3.2	1.0	-3.7	0.8	-3.1	0.9	-2.5	1.1	-2.9	0.6	-2.0	0.8	-2.2	0.8
11	-2.7	1.0	-3.4	0.9	-3.0	0.9	-2.3	1.2	-2.8	0.6	-2.1	0.8	-2.0	0.8
12	-2.3	1.0	-3.1	0.9	-2.9	1.0	-2.2	1.2	-2.7	0.7	-2.1	0.8	-1.8	0.8
13	-2.0	1.0	-2.8	0.9	-2.7	1.0	-2.1	1.1	-2.6	0.7	-2.1	0.8	-1.6	0.8
14	-1.8	1.0	-2.6	0.9	-2.5	1.0	-1.9	1.1	-2.5	0.7	-2.0	0.8	-1.5	0.9
15	-1.5	1.0	-2.3	0.9	-2.3	1.0	-1.8	1.0	-2.4	0.7	-2.0	0.8	-1.3	0.9
16	-1.4	1.0	-2.1	0.9	-2.2	1.0	-1.7	1.0	-2.3	0.7	-1.9	0.8	-1.2	0.9

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 160

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS						
2	-1.9	1.1	-1.6	1.0	-2.5	1.1
3	-4.6	1.1	-4.3	1.1	-4.4	0.9
4	-2.4	1.4	-2.5	1.1	-5.9	1.2
5	-2.4	1.1	-2.7	1.1	-4.7	1.1
6	-2.1	0.9	-2.7	1.0	-3.7	1.3
7	-1.9	0.8	-2.6	0.9	-3.0	1.3
8	-1.6	0.8	-2.5	0.9	-2.4	1.2
9	-1.5	0.8	-2.3	0.9	-2.1	1.1
10	-1.4	0.8	-2.2	0.9	-1.8	1.0
11	-1.3	0.7	-2.0	0.9	-1.6	1.0
12	-1.2	0.7	-1.9	0.8	-1.5	1.0
13	-1.2	0.7	-1.7	0.8	-1.3	0.9
14	-1.2	0.8	-1.6	0.8	-1.2	0.9
15	-1.2	0.7	-1.4	0.8	-1.1	0.9
16	-1.3	0.7	-1.3	0.8	-1.0	0.9

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 320

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
2	-5.2	1.3	-2.6	1.2	-2.9	1.2	-4.4	1.5	-2.0	1.6	-1.8	1.5	-2.3	1.4
3	-7.6	1.0	-4.1	1.0	-6.8	1.2	-6.8	1.3	-1.7	0.9	-3.9	1.0	-3.7	1.1
4	-10.0	1.1	-6.3	1.0	-4.7	1.6	-9.8	1.2	-2.1	1.0	-4.3	0.8	-6.1	1.1
5	-12.5	1.1	-8.2	1.0	-5.4	1.5	-7.9	1.1	-2.9	0.9	-3.8	0.9	-6.1	0.9
6	-11.0	1.2	-8.2	1.0	-5.5	1.2	-6.4	1.2	-3.7	0.8	-3.2	1.0	-5.9	0.9
7	-9.8	1.2	-7.9	1.0	-5.6	1.1	-5.3	1.2	-4.2	0.8	-2.8	1.1	-5.6	0.9
8	-8.7	1.3	-7.6	1.1	-5.6	1.0	-4.9	1.4	-4.6	0.8	-2.9	1.2	-5.2	0.9
9	-7.7	1.3	-7.2	1.1	-5.5	0.9	-4.9	1.2	-4.8	0.8	-3.1	1.2	-4.7	0.9
10	-6.8	1.4	-6.8	1.1	-5.4	0.9	-4.7	1.1	-4.9	0.8	-3.2	1.2	-4.3	0.9
11	-6.1	1.4	-6.3	1.1	-5.3	0.8	-4.6	1.0	-4.9	0.8	-3.3	1.1	-3.9	0.9
12	-5.5	1.4	-5.9	1.1	-5.2	0.7	-4.4	1.0	-4.9	0.8	-3.3	1.0	-3.6	0.9
13	-5.0	1.3	-5.5	1.1	-5.0	0.7	-4.3	0.9	-4.9	0.8	-3.3	1.0	-3.3	0.9
14	-4.5	1.3	-5.1	1.1	-4.8	0.7	-4.1	0.9	-4.8	0.8	-3.3	0.9	-3.0	0.9
15	-4.2	1.3	-4.8	1.1	-4.6	0.7	-4.0	0.9	-4.6	0.7	-3.3	0.9	-2.9	0.9
16	-3.8	1.2	-4.4	1.1	-4.4	0.7	-3.8	0.9	-4.5	0.7	-3.3	0.9	-2.8	1.0
17	-3.5	1.2	-4.1	1.0	-4.1	0.7	-3.7	0.9	-4.3	0.7	-3.3	0.8	-2.8	1.0
18	-3.3	1.2	-3.8	1.0	-3.9	0.7	-3.6	0.9	-4.1	0.7	-3.2	0.8	-2.7	0.9
19	-3.1	1.1	-3.5	1.0	-3.7	0.7	-3.4	0.9	-4.0	0.7	-3.1	0.8	-2.6	0.9

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 320

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
20	-2.8	1.1	-3.3	1.0	-3.5	0.8	-3.3	0.9	-3.8	0.7	-3.0	0.8	-2.5	0.9
21	-2.6	1.1	-3.0	1.0	-3.3	0.8	-3.1	0.9	-3.6	0.7	-2.9	0.8	-2.4	0.9
22	-2.5	1.1	-2.8	1.0	-3.1	0.8	-3.0	0.9	-3.4	0.7	-2.9	0.8	-2.3	0.9
23	-2.3	1.1	-2.6	1.0	-2.9	0.8	-2.9	0.9	-3.2	0.7	-2.8	0.8	-2.2	0.8
24	-2.1	1.1	-2.4	1.1	-2.7	0.8	-2.7	0.9	-3.1	0.7	-2.7	0.8	-2.1	0.8
25	-2.0	1.1	-2.2	1.1	-2.5	0.8	-2.6	0.9	-2.9	0.7	-2.5	0.8	-2.0	0.8
26	-1.8	1.1	-2.1	1.1	-2.3	0.8	-2.5	0.9	-2.8	0.7	-2.4	0.8	-1.9	0.8
27	-1.7	1.1	-1.9	1.1	-2.2	0.8	-2.4	0.9	-2.6	0.7	-2.3	0.8	-1.8	0.8
28	-1.5	1.1	-1.8	1.0	-2.0	0.8	-2.2	0.9	-2.5	0.7	-2.2	0.8	-1.7	0.8
29	-1.4	1.1	-1.7	1.0	-1.9	0.8	-2.1	0.9	-2.3	0.7	-2.1	0.8	-1.6	0.8
30	-1.3	1.1	-1.5	1.0	-1.7	0.8	-2.0	0.9	-2.2	0.7	-2.0	0.8	-1.5	0.8
31	-1.2	1.1	-1.4	1.0	-1.6	0.8	-1.8	0.9	-2.0	0.7	-1.9	0.8	-1.4	0.7
32	-1.1	1.1	-1.3	1.0	-1.4	0.8	-1.7	0.9	-1.9	0.7	-1.8	0.8	-1.3	0.7

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 320

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS						
2	-2.6	1.5	-2.9	1.5	-3.7	1.6
3	-6.8	1.3	-6.8	1.2	-6.6	1.5
4	-3.9	1.8	-4.1	1.6	-9.6	1.3
5	-3.6	1.3	-4.4	1.3	-8.0	1.2
6	-3.3	1.0	-4.5	1.1	-6.5	1.3
7	-3.0	0.9	-4.4	1.0	-5.3	1.4
8	-2.7	0.8	-4.3	1.0	-4.4	1.2
9	-2.5	0.8	-4.2	1.0	-4.0	1.2
10	-2.4	0.8	-4.0	1.0	-3.8	1.2
11	-2.3	0.7	-3.8	1.0	-3.6	1.1
12	-2.2	0.7	-3.6	1.1	-3.5	1.0
13	-2.1	0.7	-3.4	1.1	-3.2	0.9
14	-2.1	0.7	-3.2	1.0	-3.0	0.9
15	-2.2	0.8	-2.9	1.0	-2.8	0.9
16	-2.2	0.8	-2.7	1.0	-2.6	0.9
17	-2.3	0.8	-2.5	1.0	-2.4	0.9
18	-2.3	0.8	-2.4	1.0	-2.2	0.9
19	-2.3	0.8	-2.2	1.0	-2.1	1.0

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 320

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS						
20	-2.3	0.8	-2.1	1.0	-1.9	0.9
21	-2.3	0.8	-1.9	1.0	-1.8	0.9
22	-2.3	0.7	-1.8	1.0	-1.7	0.9
23	-2.2	0.7	-1.7	1.0	-1.6	0.9
24	-2.2	0.7	-1.6	1.0	-1.5	0.8
25	-2.2	0.7	-1.5	1.0	-1.4	0.8
26	-2.1	0.7	-1.4	1.0	-1.4	0.8
27	-2.1	0.7	-1.4	1.0	-1.3	0.8
28	-2.0	0.7	-1.3	1.0	-1.2	0.8
29	-2.0	0.7	-1.2	1.0	-1.2	0.8
30	-1.9	0.7	-1.2	1.0	-1.1	0.8
31	-1.9	0.7	-1.1	1.0	-1.0	0.8
32	-1.8	0.7	-1.1	0.9	-1.0	0.8

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
2	-7.5	1.6	-3.3	1.5	-4.4	1.8	-5.3	2.3	-2.7	2.7	-3.5	2.2	-3.5	1.9
3	-11.3	1.1	-6.1	1.1	-9.8	1.5	-10.1	1.5	-2.9	0.9	-5.9	1.3	-5.7	1.2
4	-15.2	1.2	-9.4	1.1	-7.1	2.1	-15.0	1.4	-3.3	1.1	-6.7	1.0	-9.3	1.1
5	-19.4	1.3	-12.4	1.3	-8.5	1.6	-12.5	1.4	-4.7	1.1	-6.1	0.9	-9.7	1.1
6	-17.3	1.3	-12.6	1.2	-8.5	1.4	-10.2	1.4	-5.9	0.9	-5.2	1.1	-9.4	1.2
7	-15.4	1.3	-12.3	1.1	-8.6	1.3	-8.5	1.5	-6.8	0.9	-4.5	0.9	-8.8	1.1
8	-13.8	1.4	-12.0	1.1	-8.8	1.2	-7.9	1.8	-7.4	0.9	-4.8	1.2	-8.2	1.0
9	-12.5	1.3	-11.6	1.2	-8.9	1.2	-7.9	1.5	-7.8	0.8	-5.3	1.2	-7.7	0.9
10	-11.3	1.3	-11.1	1.2	-9.0	1.1	-7.8	1.4	-8.1	0.8	-5.6	1.0	-7.1	1.0
11	-10.3	1.3	-10.6	1.2	-9.0	1.1	-7.6	1.3	-8.2	0.8	-5.7	0.9	-6.6	1.0
12	-9.5	1.3	-10.1	1.2	-8.9	1.1	-7.5	1.3	-8.2	0.8	-5.8	0.8	-6.1	1.1
13	-8.7	1.4	-9.6	1.2	-8.7	1.1	-7.4	1.3	-8.2	0.8	-5.9	0.8	-5.8	1.0
14	-8.1	1.5	-9.1	1.2	-8.6	1.0	-7.2	1.2	-8.1	0.8	-5.9	0.8	-5.4	1.0
15	-7.6	1.4	-8.6	1.2	-8.4	1.0	-7.0	1.2	-7.9	0.8	-6.0	0.8	-5.2	1.1
16	-7.2	1.3	-8.2	1.2	-8.2	1.0	-6.8	1.2	-7.8	0.8	-6.0	0.8	-5.2	1.2
17	-6.9	1.2	-7.8	1.2	-7.9	1.0	-6.7	1.1	-7.6	0.8	-6.0	0.8	-5.3	1.2
18	-6.6	1.2	-7.4	1.2	-7.7	1.0	-6.6	1.1	-7.5	0.8	-6.0	0.8	-5.2	1.1
19	-6.4	1.1	-6.9	1.2	-7.4	1.0	-6.4	1.1	-7.3	0.8	-6.0	0.8	-5.1	1.0

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
20	-6.1	1.1	-6.5	1.2	-7.1	0.9	-6.2	1.1	-7.1	0.7	-5.9	0.8	-5.0	1.0
21	-5.9	1.1	-6.2	1.2	-6.8	0.9	-6.1	1.1	-6.9	0.7	-5.9	0.8	-4.9	1.0
22	-5.7	1.1	-5.9	1.2	-6.6	0.9	-5.9	1.0	-6.7	0.7	-5.9	0.8	-4.8	1.0
23	-5.5	1.1	-5.6	1.2	-6.3	0.9	-5.7	1.0	-6.5	0.7	-5.8	0.8	-4.7	1.0
24	-5.3	1.1	-5.4	1.2	-6.1	0.9	-5.5	1.0	-6.3	0.7	-5.7	0.8	-4.6	1.0
25	-5.1	1.1	-5.1	1.2	-5.8	0.9	-5.3	1.0	-6.0	0.7	-5.6	0.8	-4.5	1.0
26	-4.9	1.1	-4.9	1.2	-5.6	0.9	-5.2	1.0	-5.8	0.7	-5.5	0.8	-4.5	1.0
27	-4.7	1.1	-4.7	1.2	-5.4	0.8	-5.0	1.0	-5.6	0.7	-5.4	0.9	-4.4	0.9
28	-4.6	1.1	-4.5	1.2	-5.2	0.8	-4.8	1.0	-5.4	0.7	-5.3	0.9	-4.3	0.9
29	-4.4	1.0	-4.3	1.2	-5.0	0.8	-4.6	1.0	-5.2	0.7	-5.2	0.8	-4.2	0.9
30	-4.2	1.0	-4.1	1.2	-4.8	0.8	-4.5	0.9	-5.0	0.7	-5.1	0.8	-4.1	0.9
31	-4.1	1.0	-3.9	1.2	-4.6	0.8	-4.3	0.9	-4.8	0.7	-5.0	0.8	-4.0	0.9
32	-3.9	1.0	-3.8	1.2	-4.4	0.8	-4.1	0.9	-4.6	0.7	-4.9	0.9	-3.9	0.9
33	-3.8	0.9	-3.6	1.2	-4.2	0.8	-3.9	0.9	-4.4	0.7	-4.8	0.9	-3.8	0.9
34	-3.6	0.9	-3.5	1.1	-4.1	0.8	-3.8	0.9	-4.3	0.7	-4.6	0.9	-3.7	0.9
35	-3.5	0.9	-3.4	1.1	-3.9	0.7	-3.6	0.9	-4.1	0.8	-4.5	0.8	-3.6	0.9
36	-3.3	0.9	-3.3	1.1	-3.7	0.7	-3.5	0.9	-4.0	0.8	-4.4	0.9	-3.5	0.9
37	-3.2	0.9	-3.1	1.1	-3.6	0.7	-3.4	0.9	-3.8	0.8	-4.3	0.9	-3.5	0.9

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
38	-3.0	0.9	-3.0	1.1	-3.4	0.7	-3.2	0.9	-3.7	0.8	-4.1	0.9	-3.4	0.8
39	-2.9	0.9	-2.9	1.1	-3.3	0.8	-3.1	0.9	-3.6	0.8	-4.0	0.9	-3.3	0.8
40	-2.8	0.9	-2.8	1.1	-3.2	0.8	-3.0	0.9	-3.4	0.8	-3.9	0.9	-3.2	0.8
41	-2.6	0.9	-2.7	1.1	-3.0	0.8	-2.8	0.9	-3.3	0.8	-3.7	0.9	-3.1	0.8
42	-2.5	0.9	-2.6	1.1	-2.9	0.8	-2.7	0.9	-3.2	0.8	-3.6	0.9	-3.0	0.8
43	-2.4	0.9	-2.5	1.1	-2.8	0.8	-2.6	0.9	-3.1	0.8	-3.5	0.9	-2.9	0.8
44	-2.2	0.9	-2.4	1.1	-2.7	0.8	-2.5	0.9	-2.9	0.8	-3.4	0.9	-2.8	0.8
45	-2.1	0.9	-2.3	1.1	-2.5	0.8	-2.3	0.9	-2.8	0.8	-3.2	0.9	-2.8	0.8
46	-2.0	0.9	-2.2	1.1	-2.4	0.8	-2.2	0.9	-2.7	0.8	-3.1	0.9	-2.7	0.8
47	-1.9	0.9	-2.1	1.1	-2.3	0.8	-2.1	0.9	-2.6	0.8	-3.0	0.9	-2.6	0.8
48	-1.8	0.9	-2.0	1.0	-2.2	0.8	-2.0	0.9	-2.5	0.8	-2.9	0.9	-2.5	0.8
49	-1.7	0.9	-1.9	1.0	-2.0	0.8	-1.9	0.9	-2.4	0.8	-2.8	0.9	-2.4	0.8
50	-1.6	0.9	-1.8	1.0	-1.9	0.8	-1.8	0.9	-2.3	0.8	-2.6	0.9	-2.3	0.8
51	-1.4	0.9	-1.7	1.0	-1.8	0.8	-1.7	0.9	-2.2	0.8	-2.5	0.9	-2.2	0.8
52	-1.3	0.9	-1.6	1.0	-1.7	0.8	-1.6	0.9	-2.1	0.8	-2.4	0.9	-2.2	0.8
53	-1.2	0.9	-1.5	1.0	-1.6	0.8	-1.5	0.9	-2.0	0.8	-2.3	0.9	-2.1	0.8
54	-1.2	0.9	-1.5	1.0	-1.5	0.8	-1.4	0.9	-1.9	0.8	-2.2	0.9	-2.0	0.8
55	-1.1	0.9	-1.4	1.0	-1.4	0.8	-1.3	0.9	-1.8	0.8	-2.1	0.9	-1.9	0.8

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE													
	1 1 1 1		2 1 1 1		2 2 1 1		2 2 2 1		4 1 1 1		4 2 1 1		4 2 2 1	
	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS														
56	-1.0	0.9	-1.3	1.0	-1.3	0.8	-1.2	0.9	-1.7	0.8	-2.0	0.9	-1.8	0.8
57	-0.9	0.9	-1.2	1.0	-1.2	0.8	-1.1	0.9	-1.6	0.8	-1.9	0.9	-1.7	0.8
58	-0.8	0.9	-1.1	1.0	-1.1	0.8	-1.0	0.9	-1.5	0.8	-1.8	0.9	-1.6	0.8
59	-0.7	0.9	-1.1	1.0	-1.0	0.8	-0.9	0.9	-1.5	0.8	-1.8	0.9	-1.6	0.8
60	-0.6	0.9	-1.0	0.9	-0.9	0.8	-0.9	0.9	-1.4	0.8	-1.7	0.9	-1.5	0.8
61	-0.5	0.9	-0.9	0.9	-0.8	0.8	-0.8	0.9	-1.3	0.8	-1.6	0.9	-1.4	0.8
62	-0.4	0.9	-0.9	0.9	-0.8	0.8	-0.7	0.9	-1.2	0.8	-1.5	0.9	-1.3	0.8
63	-0.4	0.9	-0.8	0.9	-0.7	0.9	-0.6	0.9	-1.1	0.8	-1.4	0.9	-1.3	0.8
64	-0.3	0.9	-0.7	0.9	-0.6	0.9	-0.5	0.9	-1.1	0.8	-1.4	0.9	-1.2	0.8

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS						
2	-3.4	1.6	-3.8	1.4	-4.9	2.2
3	-10.3	1.6	-9.9	1.6	-9.5	1.6
4	-5.8	1.9	-6.5	1.8	-14.8	1.7
5	-5.7	1.5	-7.1	1.5	-12.4	1.4
6	-5.3	1.1	-7.0	1.1	-10.2	1.4
7	-5.0	1.0	-7.2	0.9	-8.4	1.4
8	-4.6	1.0	-7.2	0.8	-7.2	1.2
9	-4.3	0.9	-7.2	0.8	-6.7	1.1
10	-4.1	0.8	-7.1	0.8	-6.4	1.1
11	-4.0	0.8	-7.0	0.8	-6.2	1.1
12	-3.8	0.8	-6.8	0.8	-6.0	1.1
13	-3.6	0.8	-6.5	0.8	-5.7	1.1
14	-3.5	0.7	-6.3	0.8	-5.4	1.1
15	-3.5	0.9	-6.0	0.8	-5.2	1.1
16	-3.9	0.9	-5.7	0.8	-5.0	1.1
17	-4.0	0.9	-5.4	0.8	-4.8	1.0
18	-4.1	0.9	-5.1	0.8	-4.6	1.0
19	-4.2	0.9	-4.9	0.9	-4.4	1.0

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
	MEAN	STD	MEAN	STD	MEAN	STD
CLUSTERS						
20	-4.1	0.8	-4.7	0.9	-4.2	0.9
21	-4.1	0.8	-4.5	0.9	-4.0	0.9
22	-4.1	0.8	-4.3	0.9	-3.9	0.9
23	-4.1	0.8	-4.1	0.9	-3.7	0.9
24	-4.1	0.8	-4.0	0.9	-3.6	0.9
25	-4.1	0.8	-3.8	0.9	-3.5	0.9
26	-4.1	0.8	-3.7	0.9	-3.4	0.9
27	-4.1	0.8	-3.6	0.9	-3.3	0.9
28	-4.0	0.8	-3.5	0.9	-3.1	0.9
29	-4.0	0.8	-3.4	0.9	-3.0	0.9
30	-4.0	0.8	-3.3	1.0	-2.9	0.9
31	-3.9	0.8	-3.3	1.0	-2.8	0.9
32	-3.9	0.8	-3.2	1.0	-2.7	0.9
33	-3.9	0.8	-3.2	1.0	-2.6	0.9
34	-3.8	0.8	-3.2	1.0	-2.5	0.9
35	-3.8	0.8	-3.1	0.9	-2.4	0.9
36	-3.8	0.8	-3.0	0.9	-2.3	0.9
37	-3.7	0.9	-2.9	0.9	-2.2	0.9

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
CLUSTERS	MEAN	STD	MEAN	STD	MEAN	STD
38	-3.7	0.9	-2.8	0.9	-2.2	0.9
39	-3.6	0.9	-2.8	0.9	-2.1	0.9
40	-3.5	0.9	-2.7	0.9	-2.0	0.9
41	-3.5	0.9	-2.6	0.9	-2.0	0.9
42	-3.4	0.9	-2.5	1.0	-1.9	0.9
43	-3.4	0.9	-2.4	0.9	-1.8	0.9
44	-3.3	0.9	-2.4	0.9	-1.8	0.8
45	-3.2	0.9	-2.3	1.0	-1.7	0.8
46	-3.2	0.9	-2.2	1.0	-1.6	0.8
47	-3.1	0.9	-2.2	1.0	-1.6	0.8
48	-3.0	0.9	-2.1	0.9	-1.5	0.8
49	-3.0	0.9	-2.0	0.9	-1.4	0.8
50	-2.9	0.9	-2.0	0.9	-1.4	0.8
51	-2.8	0.9	-1.9	0.9	-1.3	0.8
52	-2.8	0.9	-1.8	0.9	-1.3	0.8
53	-2.7	0.9	-1.8	1.0	-1.2	0.8
54	-2.6	0.9	-1.7	1.0	-1.2	0.8
55	-2.6	0.9	-1.6	1.0	-1.1	0.8

(CONTINUED)

TABLE 3
MEANS AND STANDARD DEVIATIONS OF THE CUBIC CLUSTERING CRITERION
CLASSIFIED BY NUMBER OF OBSERVATIONS, CLUSTERS, AND HYPERBOX SHAPE
FOR UNIFORM DISTRIBUTIONS ON HYPERBOXES IN FOUR DIMENSIONS
EACH MEAN AND STANDARD DEVIATION IS BASED ON FIFTY SAMPLES

NUMBER OF OBSERVATIONS = 640

	SHAPE					
	4 4 1 1		4 4 2 1		4 4 4 1	
CLUSTERS	MEAN	STD	MEAN	STD	MEAN	STD
56	-2.5	0.9	-1.6	1.0	-1.1	0.8
57	-2.4	0.9	-1.5	1.0	-1.0	0.8
58	-2.3	0.9	-1.5	1.0	-1.0	0.8
59	-2.3	0.9	-1.4	1.0	-0.9	0.8
60	-2.2	0.9	-1.3	1.0	-0.9	0.8
61	-2.1	0.9	-1.3	1.0	-0.8	0.8
62	-2.1	0.9	-1.2	1.0	-0.8	0.8
63	-2.0	0.9	-1.2	1.0	-0.8	0.8
64	-1.9	0.9	-1.1	1.0	-0.7	0.8

Table 4 provides an indication of the power of the CCC for detecting a mixture of two spherical normal distributions with unit variance and equal sampling probabilities. Ten samples of 100 observations were generated from each of 15 populations with 1, 2, 4, 8, or 16 variables and a distance between component means of 4, 5, or 6 standard deviations. Table 4 shows the frequency with which the CCC exceeded 2. The power decreases as the dimensionality increases, as expected. With 1 variable, a separation of 4 or 5 standard deviations is required for good power, while 16 variables require a separation of 6 or more standard deviations.

TABLE 4
FREQUENCY OF THE CCC EXCEEDING 2.0
IN TEN SAMPLES OF 100 OBSERVATIONS FROM A MIXTURE
OF TWO SPHERICAL MULTIVARIATE NORMAL DISTRIBUTIONS
WITH UNIT VARIANCE AND EQUAL SAMPLING PROBABILITIES

	DISTANCE BETWEEN CENTROIDS		
	4	5	6
VARIABLES			
1	5	10	10
2	3	10	10
4	0	8	10
8	0	4	10
16	0	0	7

Milligan and Cooper (1983) performed a Monte Carlo comparison of 30 criteria for the number of clusters, including the CCC. In the overall evaluation, the CCC ranked sixth best, correctly identifying the number of clusters 321 times of 432 attempts. The CCC tended to overestimate the number of clusters, probably because some of the clusters were elliptical rather than spherical.

Examples

Figures 4 through 6 show CCC plots for samples of 100 observations from various normal distributions clustered by Ward's method. In each case the CCC values are negative and generally decreasing as the number of clusters increases. Figure 4 is based on a univariate normal distribution. Figure 5 comes from a spherical multivariate normal distribution in 16 dimensions. Figure 6 illustrates an elliptical normal distribution in 16 dimensions, for which the standard deviation in the j^{th} dimension is j.

FIGURE 4
CCC PLOT
100 OBS FROM A NORMAL DISTRIBUTION

PLOT OF _CCC_*_NCL_   SYMBOL IS VALUE OF _NCL_

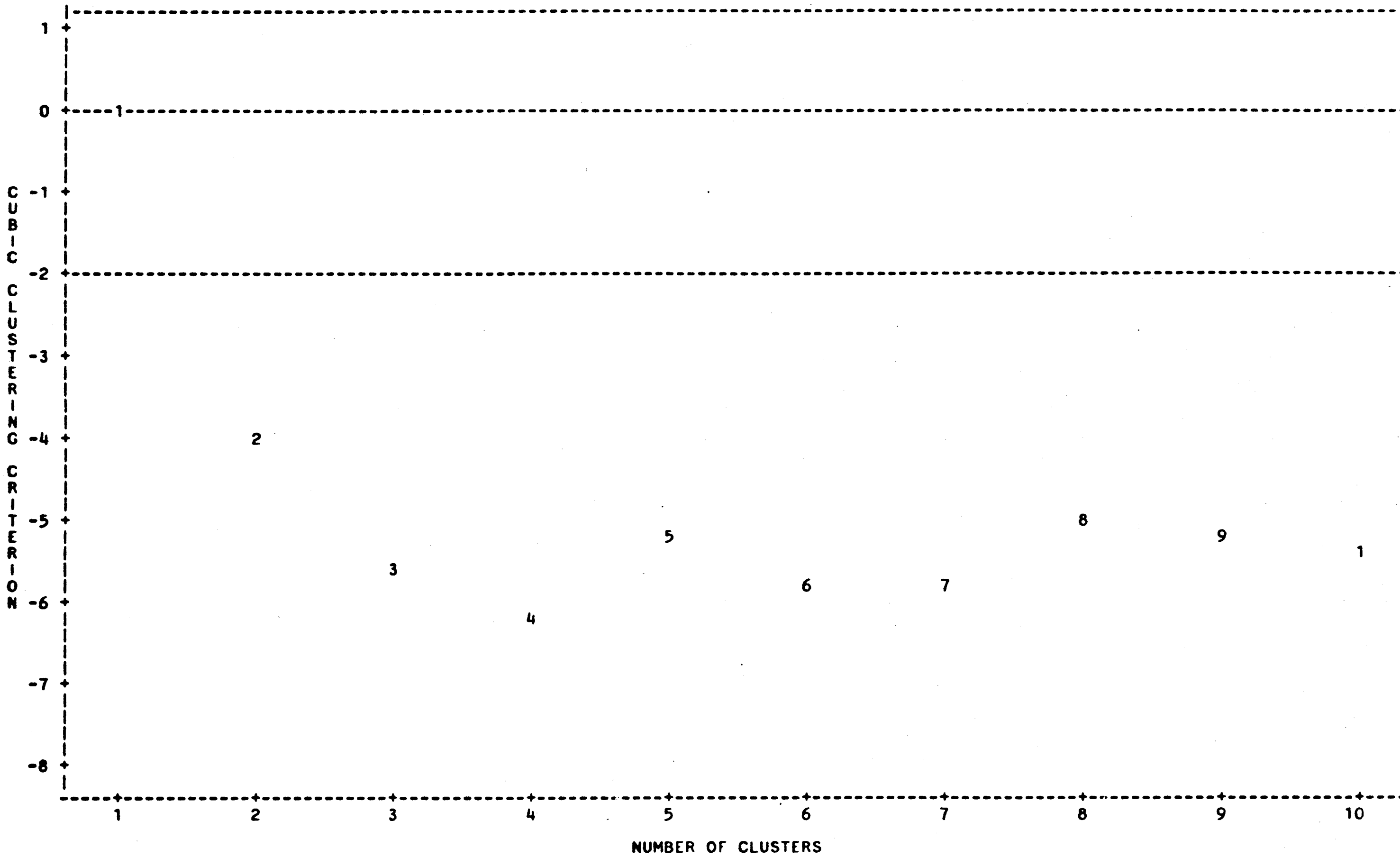


FIGURE 5
CCC PLOT
100 OBS FROM A SPHERICAL MULTIVARIATE NORMAL DISTRIBUTION
IN 16 DIMENSIONS

PLOT OF _CCC_*_NCL_   SYMBOL IS VALUE OF _NCL_

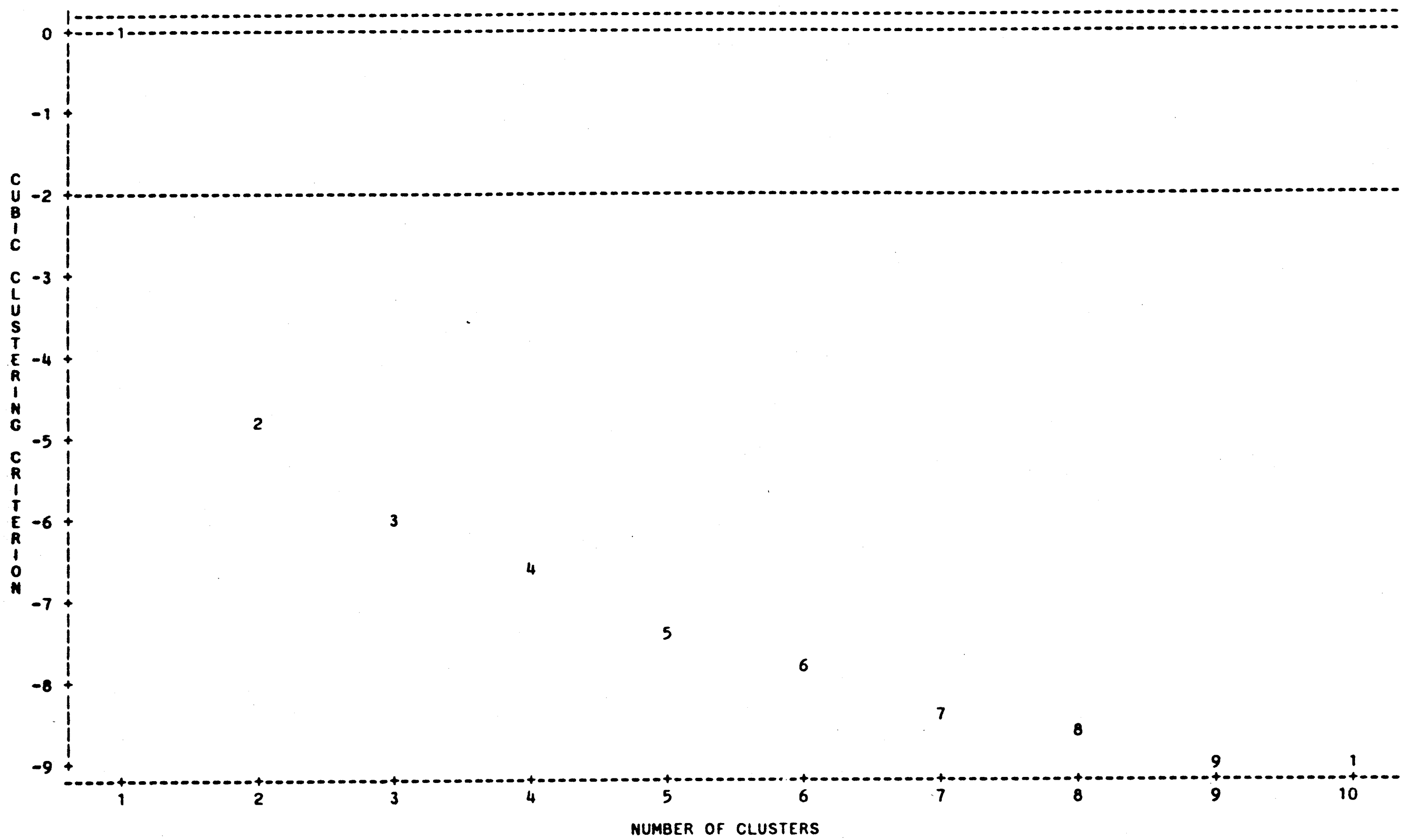


FIGURE 6
CCC PLOT
100 OBS FROM AN ELLIPTICAL MULTIVARIATE NORMAL DISTRIBUTION
IN 16 DIMENSIONS

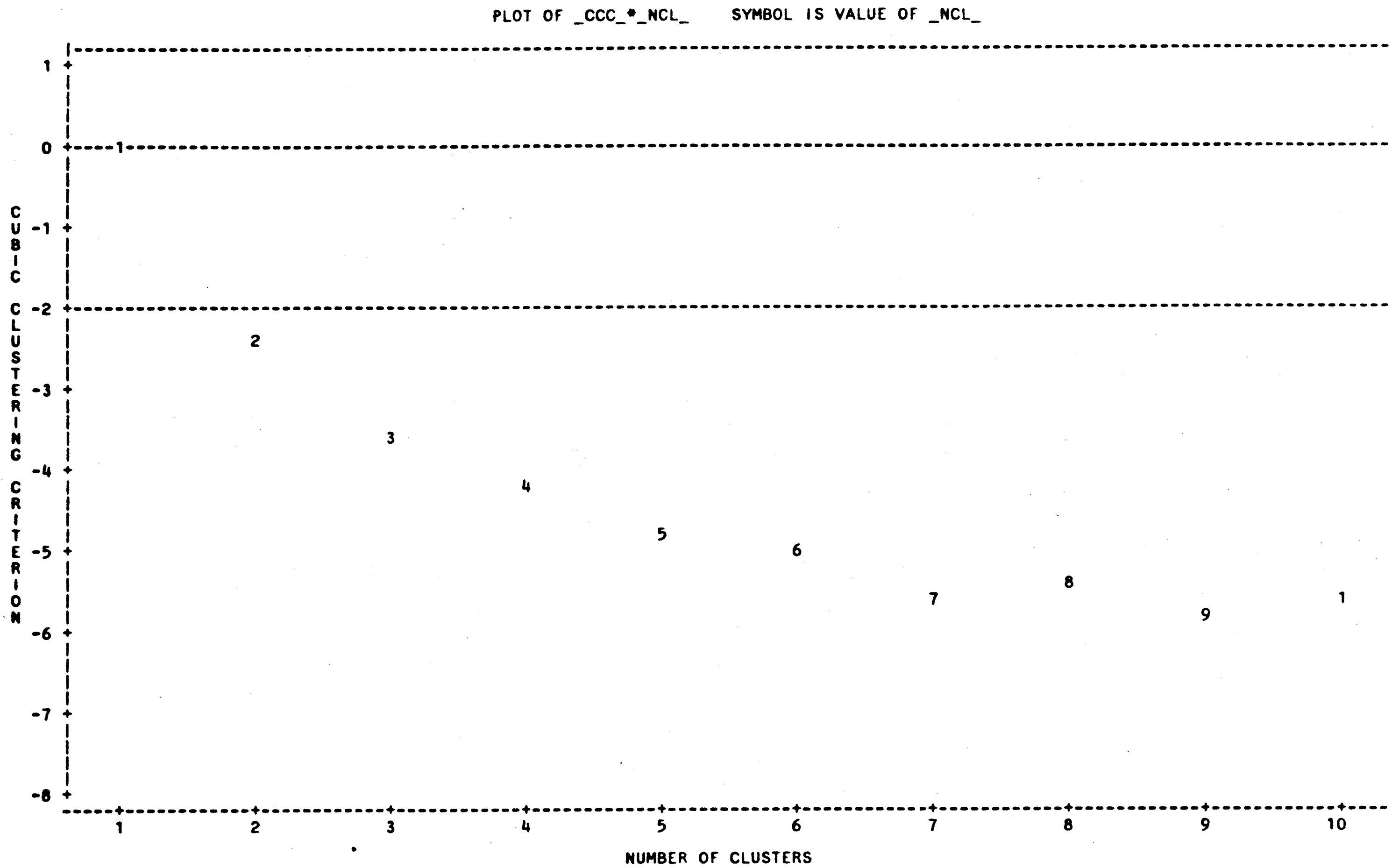
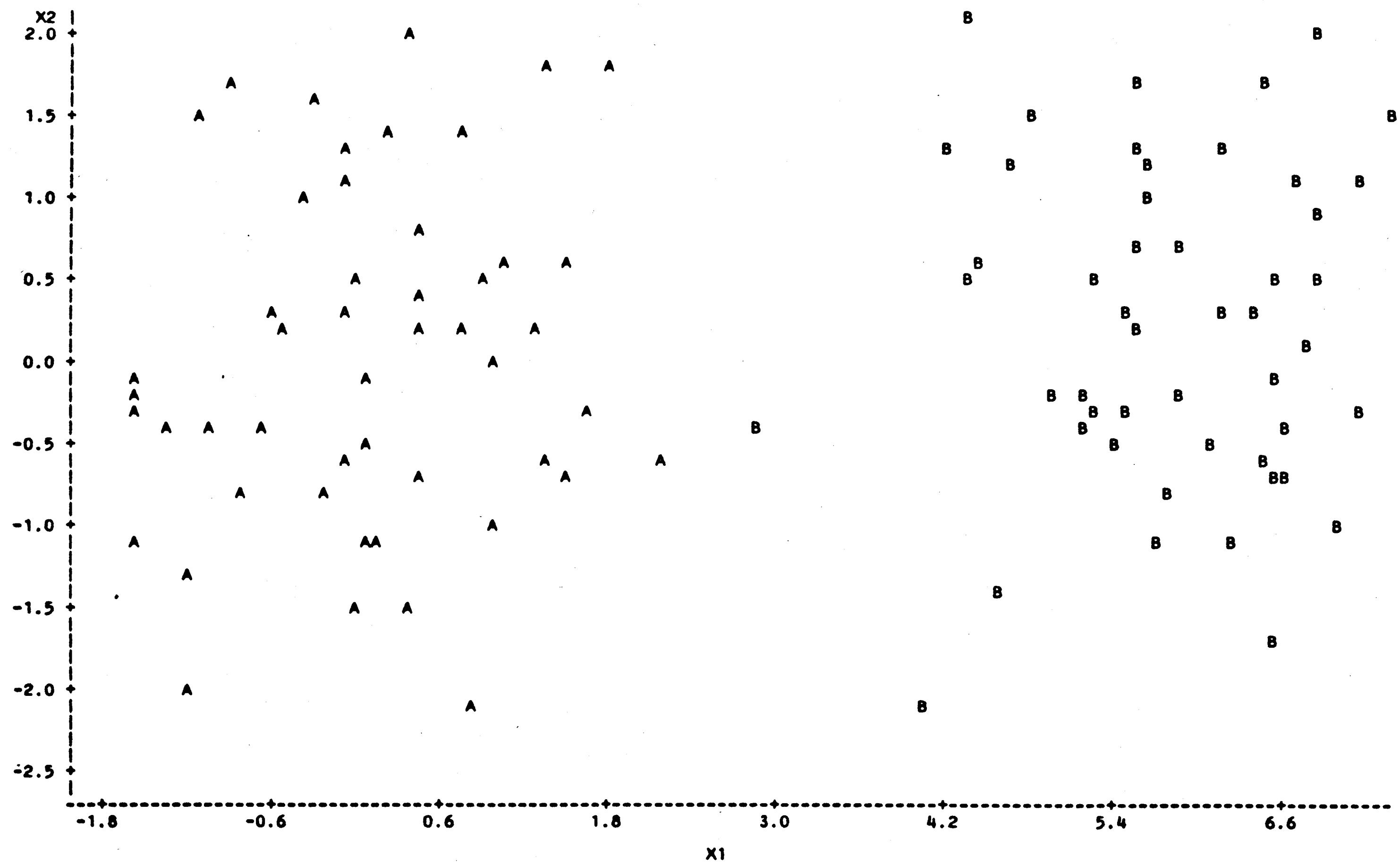


Figure 7 is a scatterplot of 100 observations from a mixture of two circular normal distributions separated by 6 standard deviations. Figure 7A shows the corresponding CCC plot with a sharp peak clearly indicating two clusters. Figure 7B presents an analysis of the data in Figure 7 standardized to unit standard deviations. Standardization causes the clusters to become highly elliptical in violation of the alternative hypothesis on which the CCC is based. The resulting plot suggests the possibility of four or nine clusters. This example illustrates the danger of indiscriminate standardization.

FIGURE 7
PLOT OF 50 OBS FROM EACH OF TWO CIRCULAR NORMAL DISTRIBUTIONS
SEPARATED BY 6 STANDARD DEVIATIONS

PLOT OF X2*X1   SYMBOL IS VALUE OF NAME



NOTE: 2 OBS HIDDEN

FIGURE 7A
CCC PLOT
FOR RAW DATA IN FIGURE 7

PLOT OF _CCC_*_NCL_   SYMBOL IS VALUE OF _NCL_

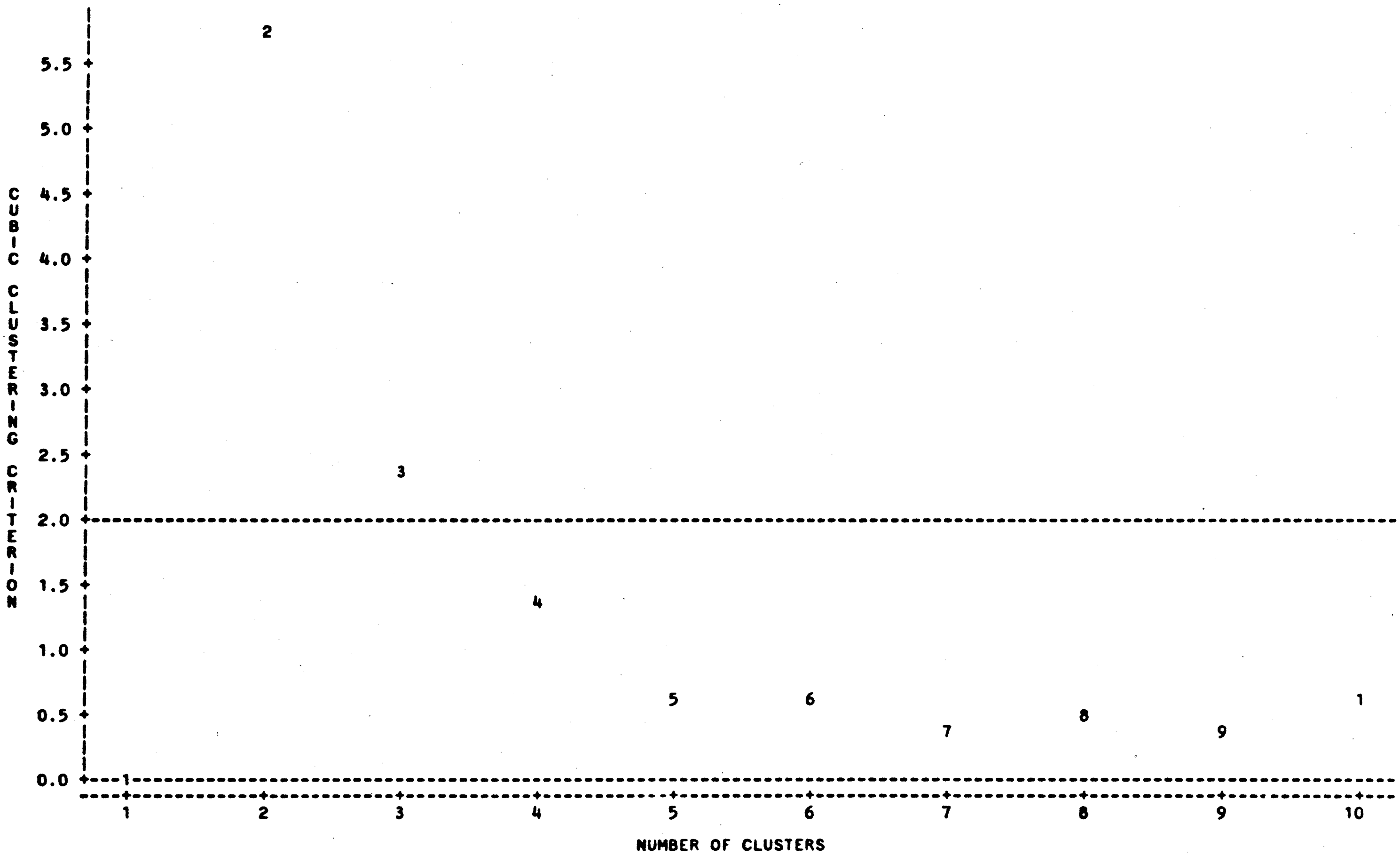


FIGURE 7B
CCC PLOT
FOR STANDARDIZED DATA IN FIGURE 7

PLOT OF _CCC_*_NCL_ SYMBOL IS VALUE OF _NCL_

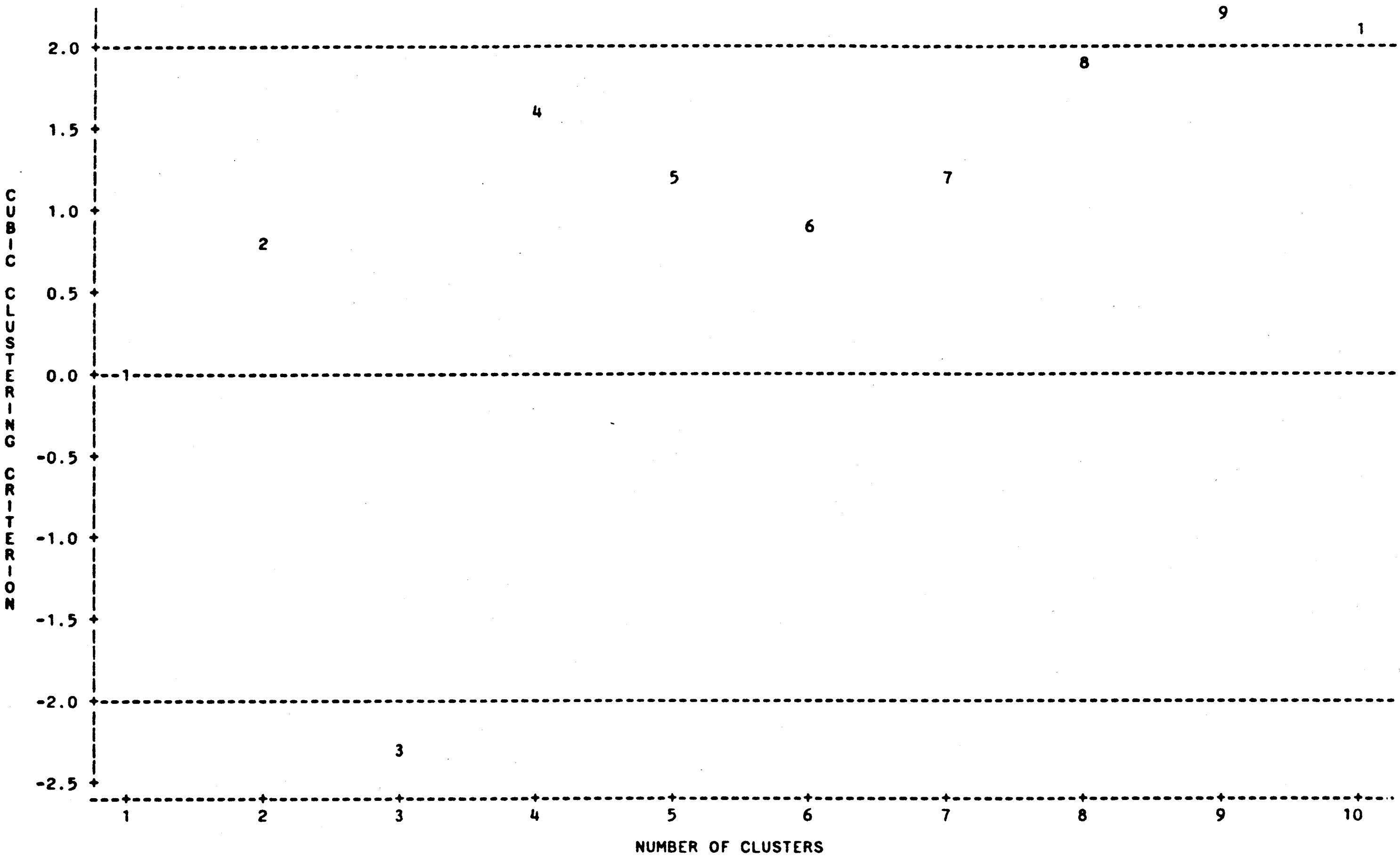


Table 5 gives the means of six clusters in two dimensions that are used to generate data in Figures 8 and 9. Figure 8 shows 120 observations from six circular normal distributions with the given means and standard deviations of 1.0 unit. By eye it is apparent that there are at least four clusters, but the clusters labeled A, B, and C cannot be easily distinguished. The CCC plot in Figure 8A has a peak at five clusters, but the peak is rather blunt, indicating that two of the five clusters are not well separated, or perhaps that there are only four clusters, one of which may be elliptical. In the scatterplot in Figure 9, the standard deviation of each cluster is reduced to .25 units. The CCC plot in Figure 9A has a blunt peak at six clusters, suggesting either six circular clusters or five clusters of which one may be elliptical.

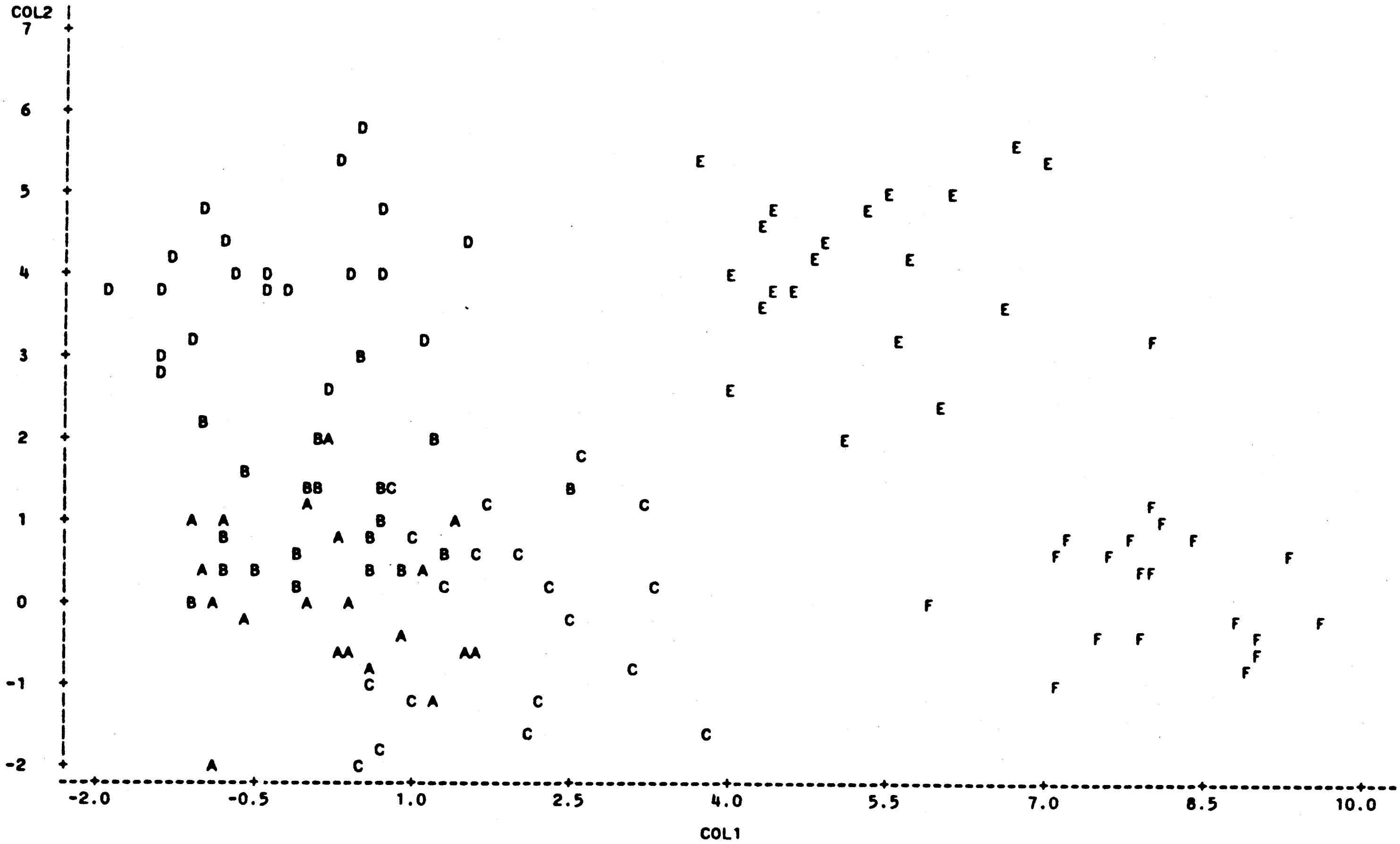
TABLE 5
CLUSTER MEANS FOR FIGURES 8 AND 9

CLUSTER MEANS

MEAN	COL1	COL2
ROW1	0	0
ROW2	0	1
ROW3	2	0
ROW4	0	4
ROW5	5	4
ROW6	8	0

FIGURE 8
PLOT OF 120 OBSERVATIONS
FROM A MIXTURE OF 6 CIRCULAR MULTIVARIATE NORMAL DISTRIBUTIONS
WITH STANDARD DEVIATION 1.0

PLOT OF COL2*COL1 SYMBOL IS VALUE OF ROW



'E: 1 OBS HIDDEN

FIGURE 8A
CCC PLOT
FOR RAW DATA IN FIGURE 8

PLOT OF _CCC_*_NCL_   SYMBOL IS VALUE OF _NCL_

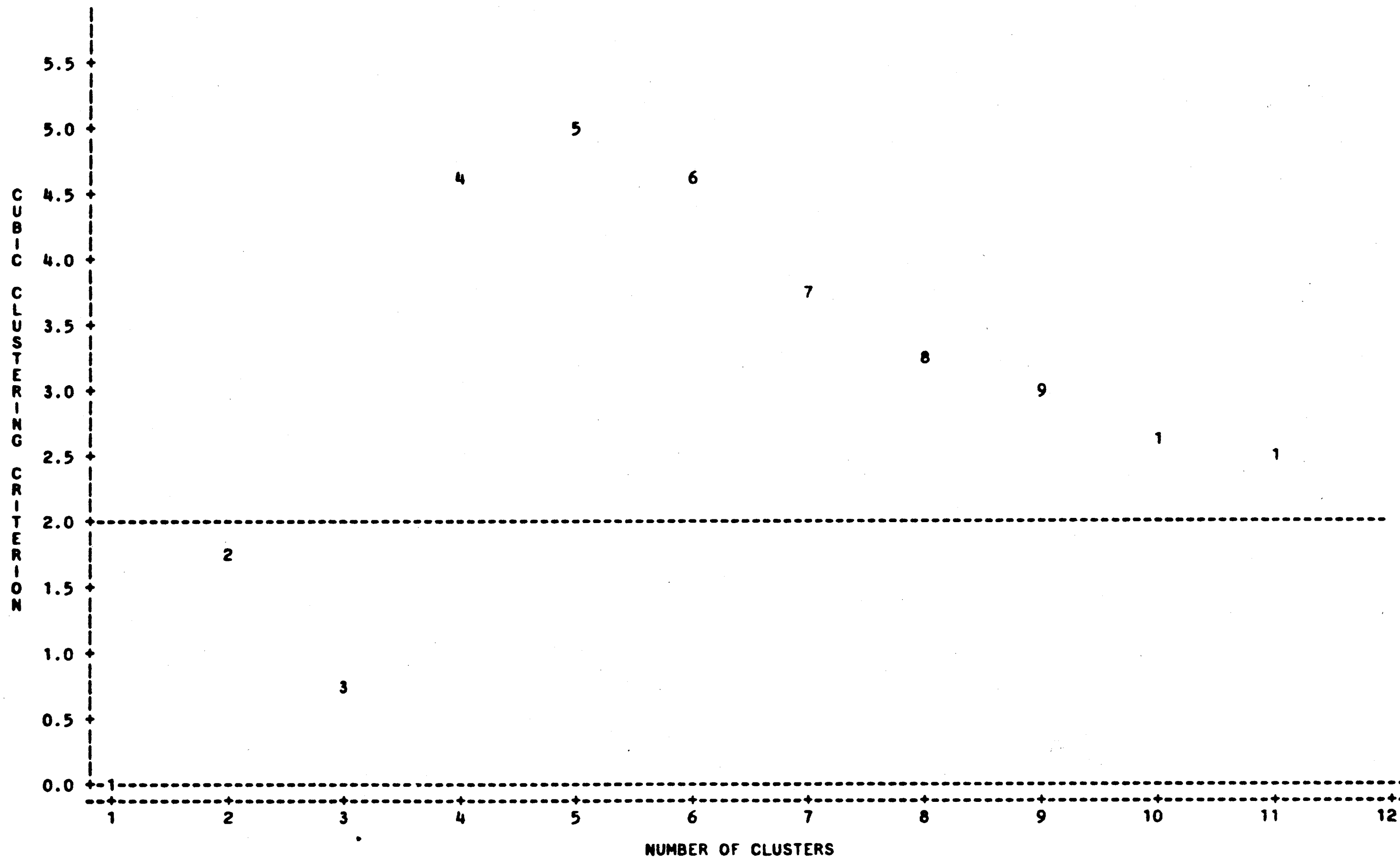
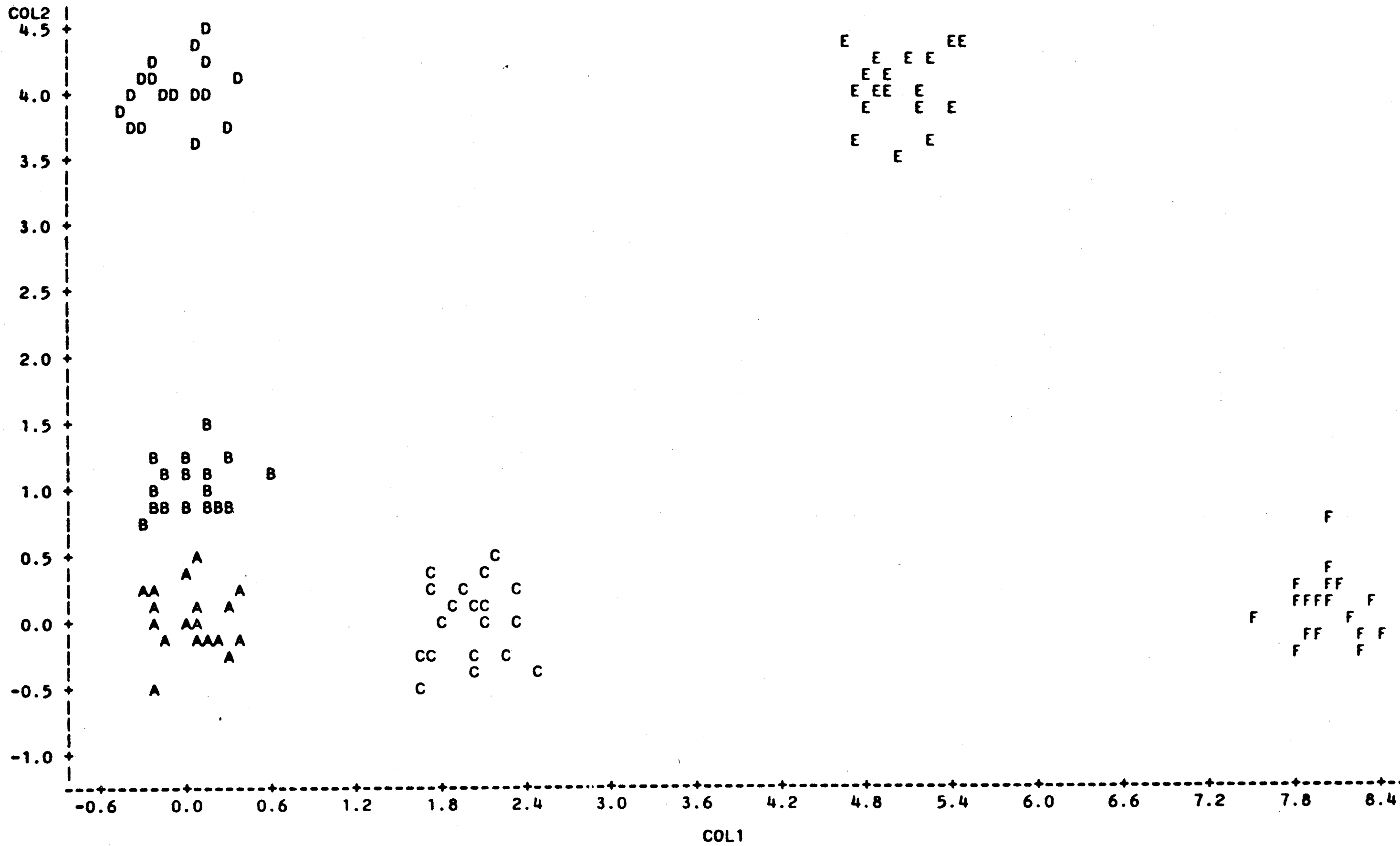


FIGURE 9
PLOT OF 120 OBSERVATIONS
FROM A MIXTURE OF 6 CIRCULAR MULTIVARIATE NORMAL DISTRIBUTIONS
WITH STANDARD DEVIATION .25

PLOT OF COL2*COL1   SYMBOL IS VALUE OF ROW



NOTE: 13 OBS HIDDEN

FIGURE 9A
CCC PLOT
FOR RAW DATA IN FIGURE 9

PLOT OF _CCC_*_NCL_   SYMBOL IS VALUE OF _NCL_

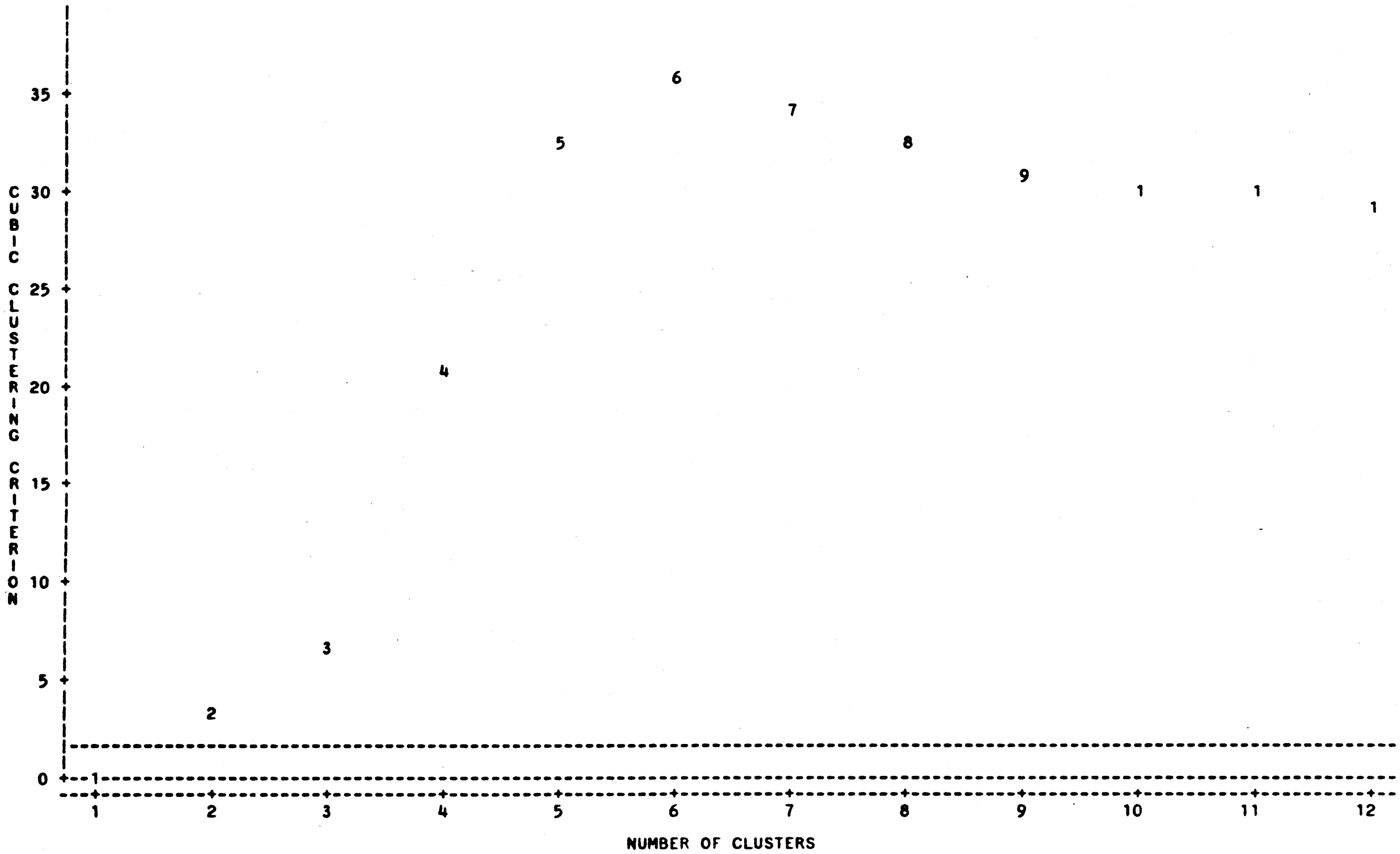


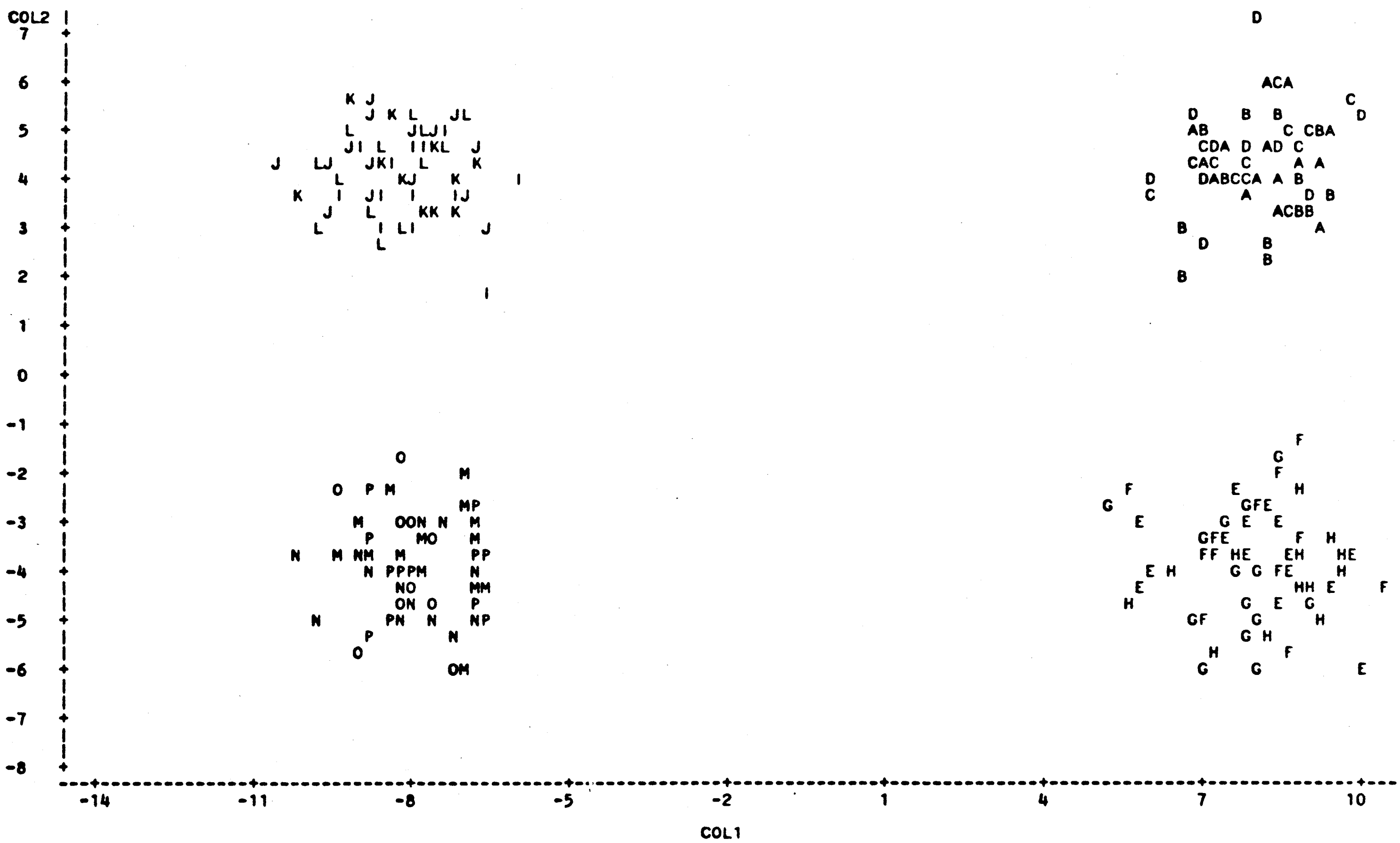
Table 6 contains the means of 16 clusters in a hierarchical arrangement used to generate data in Figures 10 and 11. Figure 10 plots the first two dimensions, showing clusters with standard deviations of 1.0 unit. There are four apparent clusters, each one of which actually is four clusters separated in the two dimensions not shown on the plot. The view through the other two dimensions would be similar but the apparent separation among the clusters would be reduced by a factor of four. The levels of interest in the hierarchy are 2, 4, 8, or 16 clusters. The first three of these levels can be seen in the CCC plot in Figure 10A as large jumps or local peaks in the CCC. In Figure 11 the standard deviations are reduced to .25 units, and the corresponding CCC plot in Figure 11A shows all four levels of the hierarchy.

TABLE 6
CLUSTER MEANS FOR FIGURES 10 AND 11

	CLUSTER MEANS			
MEAN	COL1	COL2	COL3	COL4
ROW1	8	4	2	1
ROW2	8	4	2	-1
ROW3	8	4	-2	1
ROW4	8	4	-2	-1
ROW5	8	-4	2	1
ROW6	8	-4	2	-1
ROW7	8	-4	-2	1
ROW8	8	-4	-2	-1
ROW9	-8	4	2	1
ROW10	-8	4	2	-1
ROW11	-8	4	-2	1
ROW12	-8	4	-2	-1
ROW13	-8	-4	2	1
ROW14	-8	-4	2	-1
ROW15	-8	-4	-2	1
ROW16	-8	-4	-2	-1

FIGURE 10
PLOT OF 240 OBSERVATIONS IN THE FIRST TWO DIMENSIONS
OF A MIXTURE OF 16 SPHERICAL MULTIVARIATE NORMAL DISTRIBUTIONS
WITH STANDARD DEVIATION 1.0

PLOT OF COL2*COL1   SYMBOL IS VALUE OF ROW



NOTE: 34 OBS HIDDEN

FIGURE 10A
CCC PLOT
FOR RAW DATA IN FIGURE 10

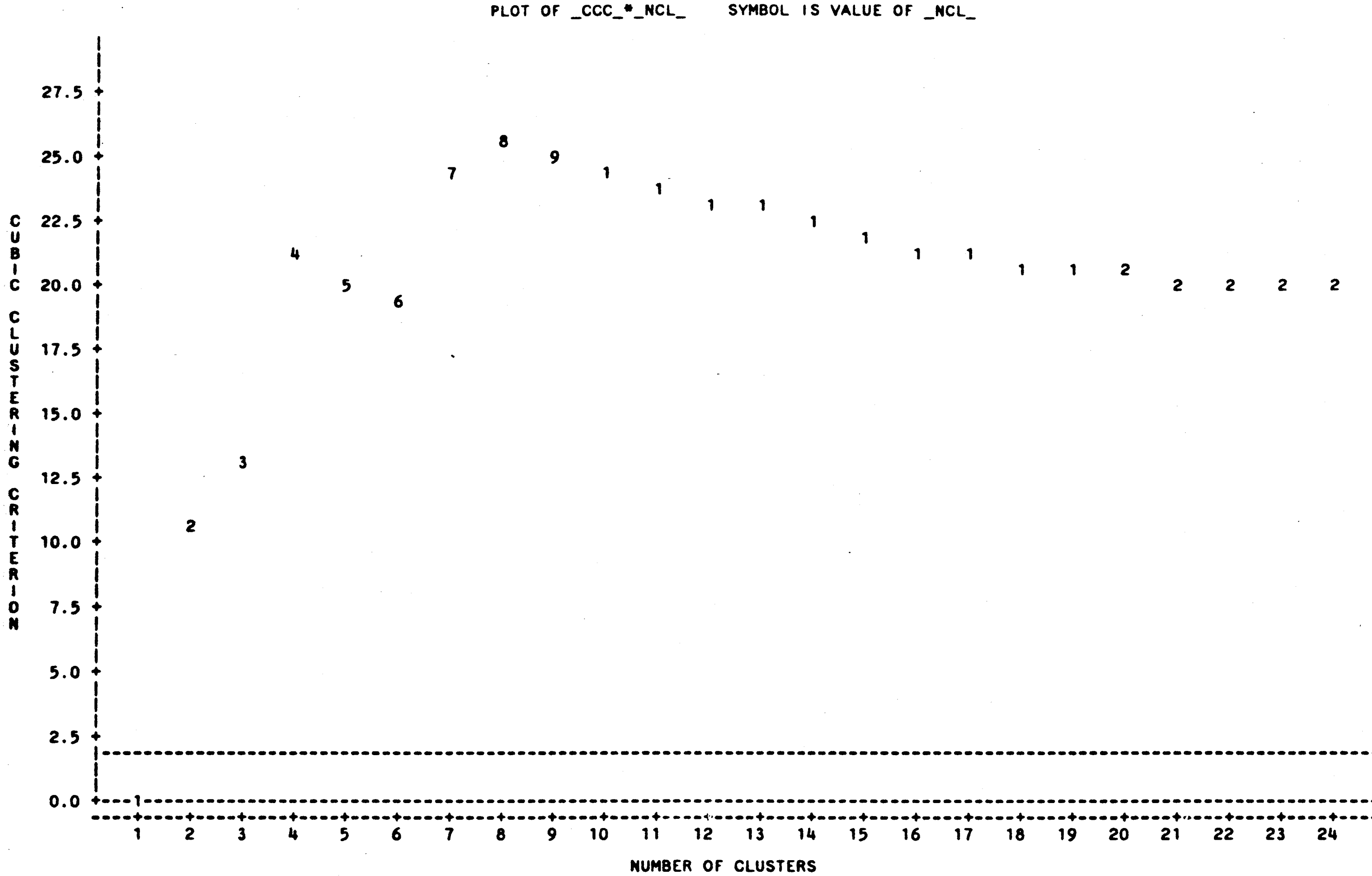
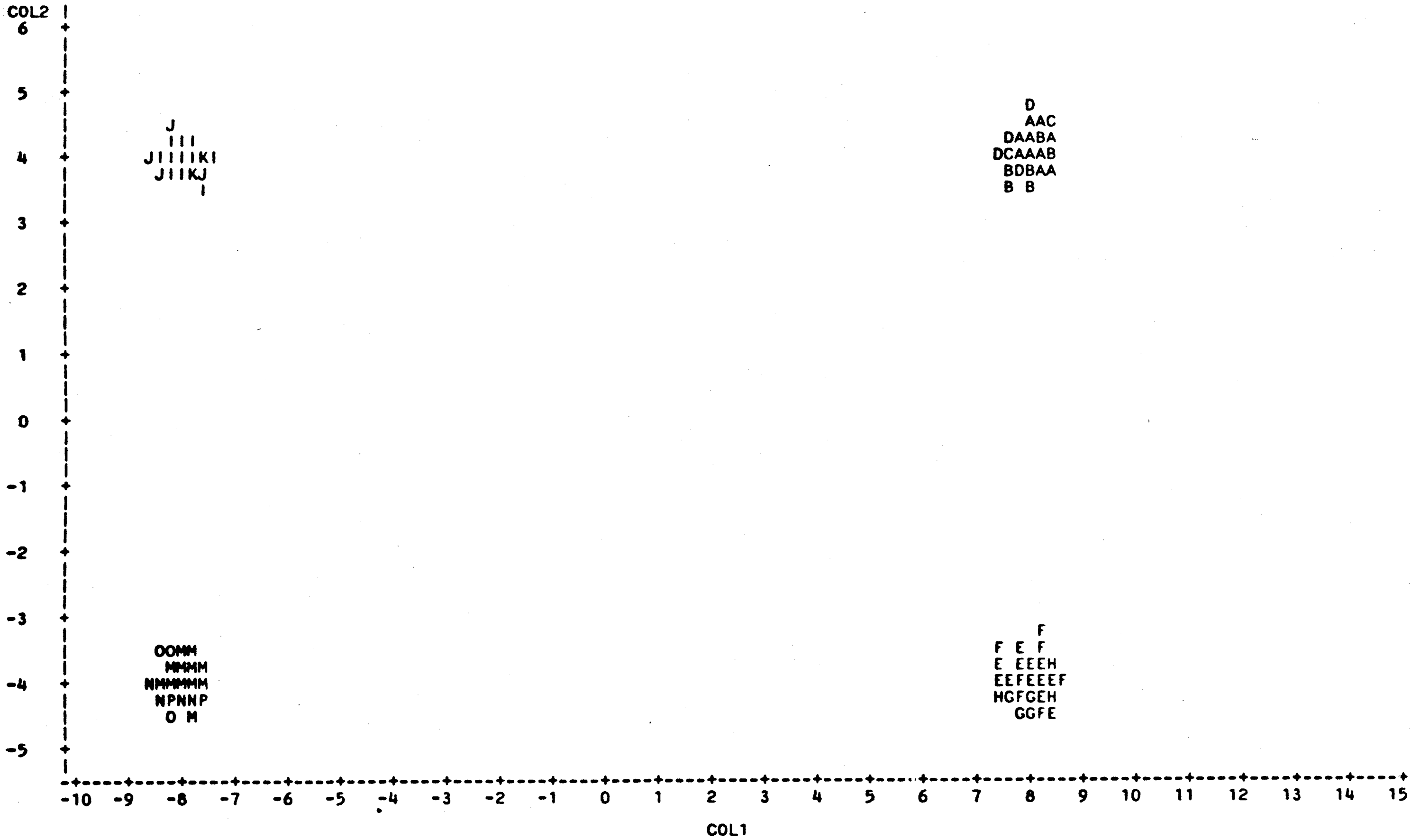


FIGURE 11
PLOT OF 240 OBSERVATIONS IN THE FIRST TWO DIMENSIONS
OF A MIXTURE OF 16 SPHERICAL MULTIVARIATE NORMAL DISTRIBUTIONS
WITH STANDARD DEVIATION .25

PLOT OF COL2*COL1   SYMBOL IS VALUE OF ROW



NOTE: 154 OBS HIDDEN

FIGURE 11A
CCC PLOT
FOR RAW DATA IN FIGURE 11

PLOT OF _CCC_*_NCL_   SYMBOL IS VALUE OF _NCL_

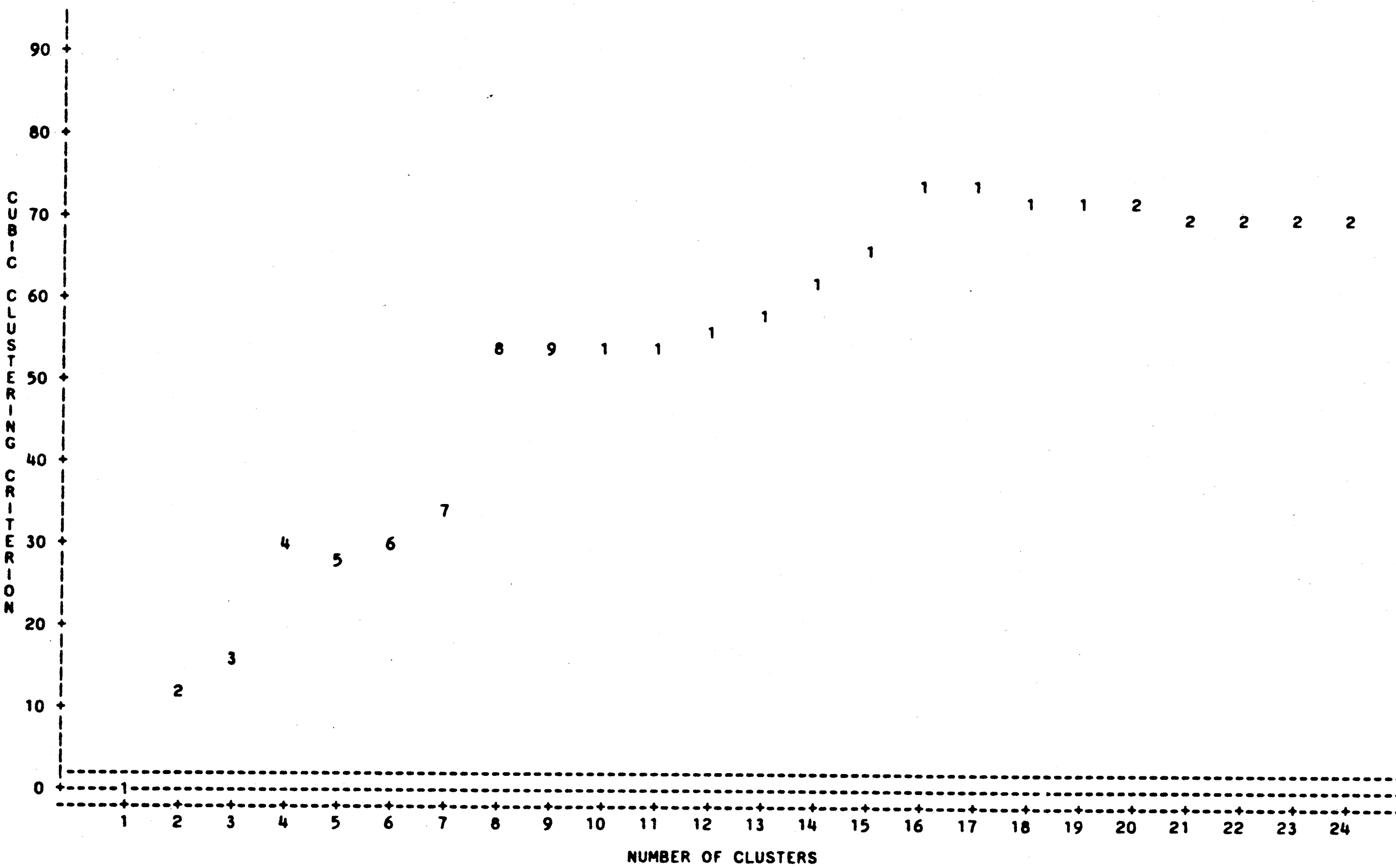


Figure 12 shows a CCC plot for the raw iris data from Fisher (1936). There is a local peak at three clusters, the correct value, but also a much higher peak at five or six clusters due to the elliptical nature of the clusters. If the data are standardized, the three cluster solution becomes apparent as shown in Figure 13.

FIGURE 12
CCC PLOT FOR RAW IRIS DATA

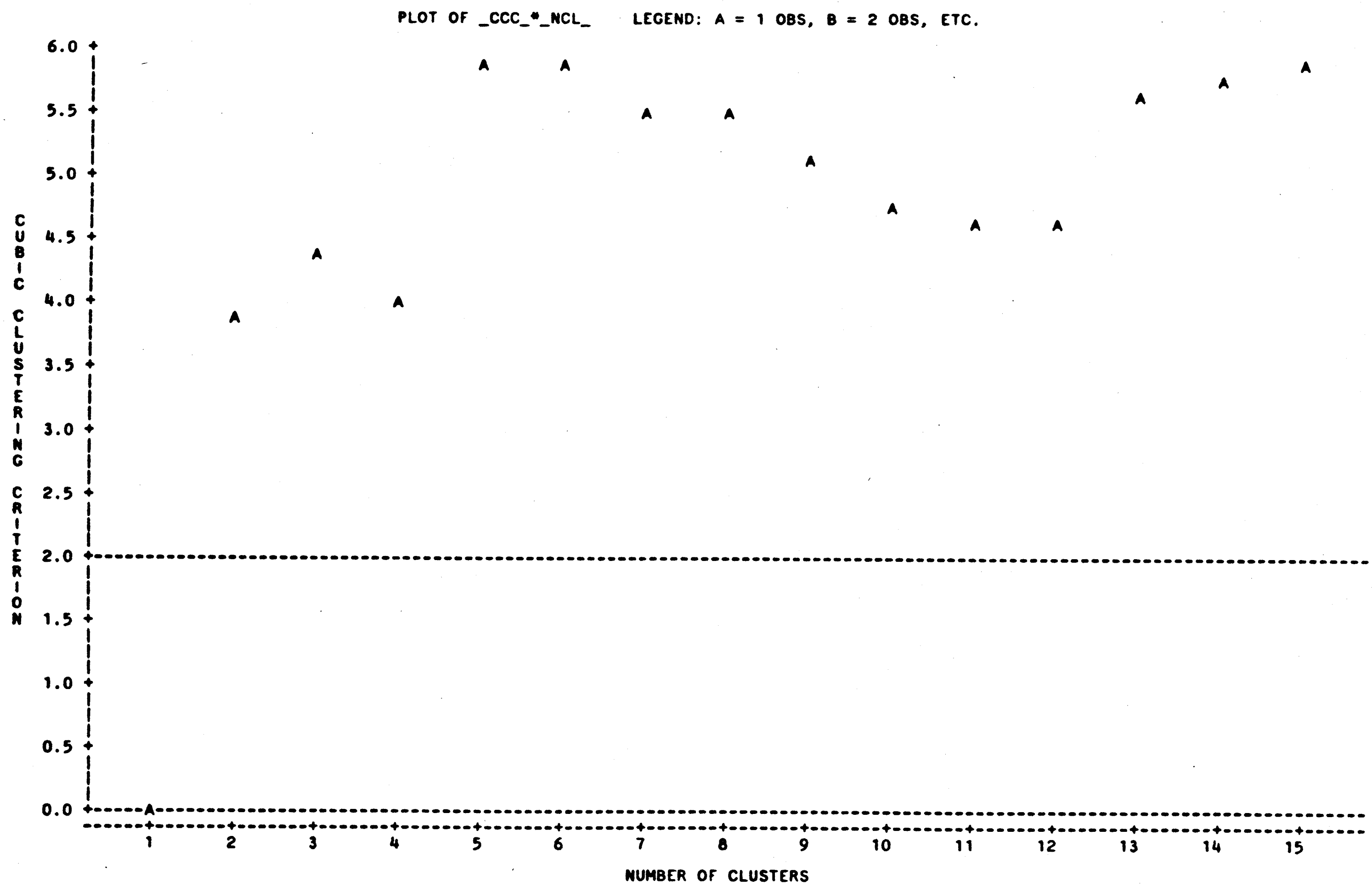
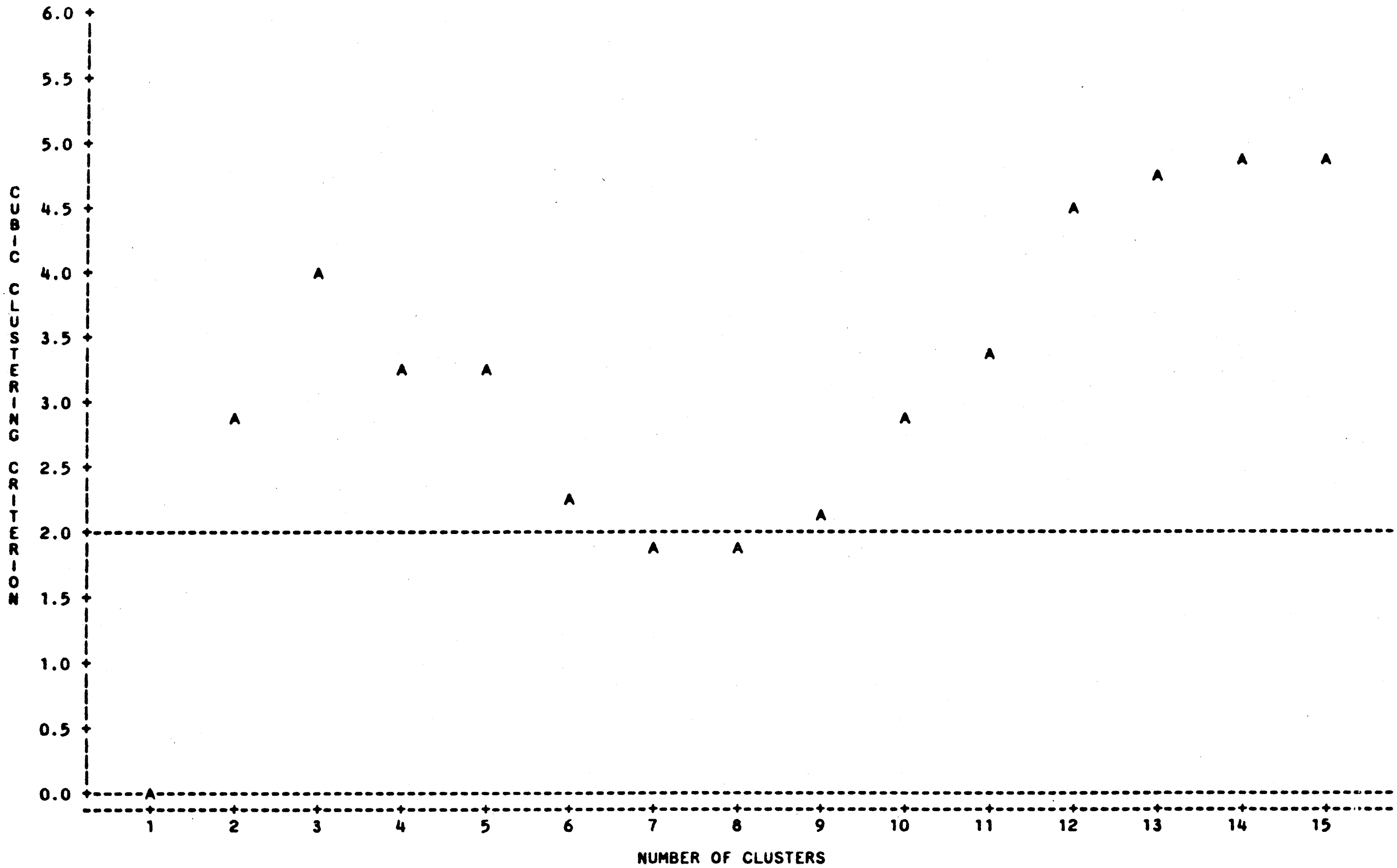


FIGURE 13
CCC PLOT FOR STANDARDIZED IRIS DATA

PLOT OF _CCC_*_NCL_   LEGEND: A = 1 OBS, B = 2 OBS, ETC.



Conclusion

The best way to use the CCC is to plot its value against the number of clusters, ranging from one cluster up to about one-tenth the number of observations. The CCC may not be well-behaved if the average number of observations per cluster is less than ten. The following guidelines should be used for interpreting the CCC:

- Peaks on the plot with the CCC greater than 2 or 3 indicate good clusterings.

- Peaks with the CCC between 0 and 2 indicate possible clusters but should be interpreted cautiously.

- There may be several peaks if the data have a hierarchical structure.

- Very distinct non-hierarchical spherical clusters usually show a sharp rise before the peak followed by a gradual decline.

- Very distinct non-hierarchical elliptical clusters often show a sharp rise to the correct number of clusters followed by a further gradual increase and eventually a gradual decline.

- If all values of the CCC are negative and decreasing for two or more clusters, the distribution is probably unimodal or long-tailed.

- Very negative values of the CCC, say -30, may be due to outliers. Outliers generally should be removed before clustering.

- If the CCC increases continually as the number of clusters increases, the distribution may be grainy or the data may have been excessively rounded or recorded with just a few digits.

A final and very important warning: neither the CCC nor R^2 is an appropriate criterion for clusters that are highly elongated or irregularly shaped. If you do not have prior substantive reasons for expecting compact clusters, use a nonparametric clustering method such as Wong and Lane's (1983), rather than Ward's method or k-means.

References

Anderberg, M.R. (1973), Cluster Analysis for Applications, New York: Academic Press.

Duran, B.S. and Odell, P.L. (1974), Cluster Analysis, New York: Springer-Verlag.

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965), "A method for cluster analysis," Biometrics, 21, 362-375.

Englemann, L. and Hartigan, J.A. (1969), "Percentage Points of a Test for Clusters," Journal of the American Statistical Association, 64, 1647-1648.

Everitt, B.S. (1980), Cluster Analysis, 2nd ed., London: Heineman Educational Books Ltd.

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

Gordon, A.D. and Henderson, J.T. (1977), "An Algorithm for Euclidean Sum of Squares Classification," Biometrics, 33, 355-362.

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

Hartigan, J.A. (1978), "Asymptotic Distributions for Clustering Criteria," Annals of Statistics, 6, 117-131.

Hubert, L. (1974), "Approximate Evaluation Techniques for the Single-link and Complete-link Hierarchical Clustering Procedures," Journal of the American Statistical Association, 69, 698-704.

Hubert, L.J. and Baker, F.B. (1977), "An Empirical Comparison of Baseline Models for Goodness-of-Fit in r-Diameter Hierarchical Clustering," in Classification and Clustering, ed. J. Van Ryzin, New York: Academic Press.

Ling, R.F (1973), "A Probability Theory of Cluster Analysis," Journal of the American Statistical Association, 68, 159-169.

MacQueen, J.B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

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

Milligan, G.W. and Cooper, M.C. (1983), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," College of Administrative Science Working Paper Series 83-51, Columbus: The Ohio State University.

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

Ward, J.H. (1963), "Hierarchical grouping to optimize an objective function," Journal of the American Statistical Association, 58, 236-244.

Wolfe, J.H. (1970), "Pattern Clustering by Multivariate Mixture Analysis," Multivariate Behavioral Research, 5, 329-350.

Wolfe, J.H. (1978), "Comparative Cluster Analysis of Patterns of Vocational Interest," Multivariate Behavioral Research, 13, 33-44.

Wong, M. A. (1982), "Asymptotic Properties of Bivariate k-means Clusters," Communications in Statistics, Theory and Methods, 11, 1155-1171.

Wong, M. A. and Lane, T. (1983) "A k^{th} Nearest Neighbor Clustering Procedure," Journal of the Royal Statistical Society , Series B, in press.

Index

C

cubic clustering criterion 1
  computation 4
  empirical examination of performance 8
  examples 30

SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513

ISBN 1-58025-185-4

9 781580 251853