Journal of the Korean
Statistical Society
Vol. 20, No. 2, 1991

A Method to Predict the Number of Clusters[†]

Seong-San Chae[*] and William D. Warde[**]

ABSTRACT

The problem of determining the number of clusters, K, is the main objective of this study. Attention is focused on the use of Rand(1971)'s C_k statistic with some agglomerative clustering algorithms(ACA) defined in the (β, π) plane in predicting the number of clusters within the given set of data. The (k, C_k) plots for $k=1, 2, \cdots, N$ are explored by a Monte Carlo study. Based on its performance, the use of C_k with the pair of ACA, $(-.5, .75)$ and $(-.25, .0)$, is recommended for predicting the number of clusters present within a set of data.

1. Introduction

In partitioning N individuals to be clustered into k groups for a set of p-dimensional multivariate data, one may wish to find the best procedure to predict the number of distinct groups, K. If a large body of data can be reduced to a relatively compact description, it may become the basis for further statistical research.

Fowlkes and Mallows(1983) suggest useful and interpretable methods for exploring the number of groups and comparing the results of clustering algorithms by using a similarity measure, B_k, under some assumptions. They indicate that in comparing the original clustering of mixture data with the clustering of perturbed data, the (k, B_k) plots tend to peak at the k which is equal to the true number of clusters, where $k=1, 2, \cdots, K, \cdots, N$. This stimulates the consideration of applying Rand's (1971) C_k for predicting the number of clusters present in a given set of data.

Two similarity measures, B_k and C_k, are somewhat similar in construction and have the following properties :

1. They depend on the matching martix, $[n_{ij}]$, where $i, j=1, 2, 3, \cdots, k$, and $k=1, 2, \cdots, N$;

[†] This research is supported by Korea research Foundation, 1990.

[*] Department of Statistics, Taejon University, Taejon, Korea.

[**] Department of Statistics, Oklahoma State Univ., Stillwater, OK, USA

2. They lie between 0.0 and 1.0 ;
3. They are 1.0 if the k clusters within each clustering correspond completely(except at k =N) ;
4. They are 0.0 if every pair of objects that appear in the same cluster in one clustering is assigned to different clusters in another clusterings.

Hence, the behavior of the measure C_k for every k in some situations is examined to predict the number of clusters for the given set of data. This will provide useful information on the properties of different agglomerative clustering procedures.

Some notations which is useful for understanding a cluster, a clustering, an hierarchy and an agglomerative clustering methods(ACM) can be found in DuBien and Warde(1987).

2. ACA and A Comparative Statistic

The application of an ACM requires that a measure of distance, d, be imposed on data points. The measure of similarity or dissimilarity explicates "close", initially ; and the ACM reevaluates the "closeness" of clusters after each join. For the purpose of this study, the squared Euclidean distance, which is only a semi-metric measure of distance, is considered since it is not as important in determining the resultant clusterings as the algorithm of ACM is (DuBien and Warde, 1987).

Letting d_{ij} denote the joining distance between cluster Y_i and cluster Y_j, where Y_i, $Y_j \in Y^K$, K=1, 2, ···, N. Then $Y_{(ij)}=Y_i \cup Y_j$ will denote the new cluster within clustering Y^{K-1}. It should be noted that the joining distance, d_{ij}, is always the smallest distance remaining in the set of all distances between clusters in clustering Y^K.

For any clustering Y^K in the hierarchy, if the distances d_{ij}, d_{ik}, and d_{jk} between pairs of clusters Y_i, Y_j and Y_k are obtained recursively from clustering Y^{K+1}, K<N, then the distance between the new cluster $Y_{(ij)}$ and any other cluster $Y_k \in Y^K$ can be computed from the following formula originally presented by Lance and Williams(1966, 1967) :

$$d_{(ij)k}=\alpha_i\ d_{ik}+\alpha_j\ d_{jk}+\beta\ d_{ij}+\pi\ |\ d_{ik}-d_{jk}\ | , \qquad (2.1)$$

where d_{ij} denotes the distance between the clusters Y_i and Y_j with n_i and n_j elements, respectively, and α_i, α_j, β, and π are specified parameters defining the particular member of the family of ACA.

Further, DuBien and Warde(1979) have explored the properties of the sequence of distances, $d_{(ij)k}$, by placing a suitable set of constraints on the parameters given in equation (2.1) and derviving a two parameter family of ACA. Then equation (2.1) becomes

$$d_{(ij)k}=\frac{1-\beta+2\pi}{2}\ d_{jk}+\ \frac{1-\beta-2\pi}{2}\ d_{ik}+\beta\ d_{ij} \qquad (2.2)$$

where $d_{ij}<d_{ik}<d_{jk}$.

For more details on $(\beta,\ \pi)$ family of ACA, refer to DuBien and Warde(1987).

For the present study, only nine ACA are chosen. The $(\beta,\ \pi)$ values which define these nine ACA are conveniently delineated in three groups of three algorithms as follows :

(1) $\beta=0.0$ with $\pi=-0.5,\ 0.0,\ 0.5$;
(2) $\beta=-0.25$ with $\pi=-0.25,\ 0.0,\ 0.5$;

(3) $\beta=0.5$ with $\pi=-0.0$, 0.25, 0.75.

In (β, π) family, (.0, −.5) is known as single linkage ; (.0, .0) as average linkage ; (.0, .5) as complete linkage ; (−.25, .0) or (−.5, .0) as flexible strategy.

It is known that two distinct clustering methods often produce two quite different clusterings from the same set of data, depending on the structure within the data. However, if the results of several different clustering procedures agree closely, then one may have more confidence in the reality of any group structure which is indicated by several clustering procedures as mentioned by Gordon(1981).

Rand's (1971) C statistic measures the similarity between two clusterings derived from any source. Further, a computational form for the C derived from an incidence matrix is given. If the clusters within each clustering are arbitrarily numbered and n_{ij} represents the number of data points simultaneously in the i-th cluster of Y and the j-th cluster of Y′, then

$$C(Y, Y')=\frac{\binom{N}{2}-\frac{1}{2}\left[\sum_i\left(\sum_j n_{ij}\right)^2+\sum_j\left(\sum_i n_{ij}\right)^2\right]+\sum_{i,j} n_{ij}^2}{\binom{N}{2}} \tag{2.3}$$

In this formulation, if two different clustering algorithms are applied to the same set of data and the clusters within each clustering are similar, the values of C(Y, Y') might be close to 1. Also, C(Y, Y')=0 when the two clusterings have no similarities.

In this study, the examination of the behavior of C for changing k is of interest in some situations. Thus, C will be represented as $C_k(Y, Y')$, which is the similarity measure between one clustering Y and another clustering Y' having the same number of clusters, k, resulting from two different ACA applied to the same set of N data points, where k=1, 2, 3, ⋯, N.

Then three observations concerning the C_k statistic will suffice for the purpose of this study :

1. The closer C_k is to 1.0, the more similar are the two clusterings ;
2. If $C_k(Y, Y')>C_k(Y, Y'')$, then Y and Y' are more similar than Y and Y'' ;
3. If $C_k(Y, Y')\geq C_{k-1}(Y, Y')$ and $C_k(Y, Y')>C_{k+1}(Y, Y')$, then C_k is the local maxium for given k for the two clusterings.

3. Monte Carlo Experiments

3.1 Design of a Comparative Study

A clustering method is purported to be a functional mechanism for finding or retrieving the "natural" structure within data. Hence, the degree to which a clustering method "retrieves" the structure within generated data is an important characteristic of the clustering method. Moreover, if two different ACM are applied to the same set of data, the degree to which the two retrieved structures correspond to each other through their resultant clusterings is another characteristic to be considered. This characteristic could be thought of as the "agreement" between two ACM for any specific number of clusters for given set of data.

Let Y represent the "true" structure of the data. Let Y' and Y" denote the two different clusterings which result from applying two different ACM to the same N data points. Then C_k(Y,

$Y^{\prime}$) is a measure of the "retrieval" ability of the ACM to the true structure generated, while $C_k(Y^{\prime}, Y^{\prime\prime})$ is a measure of the "agreement" between the two ACM through their resultant clusterings for $k=2, 3, \cdots, K, \cdots, N-1$.

Some of the possible structural parameters considered in this comparative study are defined as follows :

1. N, the number of data point in X ;
2. p, the number of variables defining each data points ; i.e., the dimensionality of the Euclidean p-space in which X is embedded ;
3. The noise(i.e., ρ for MVN, or θ for MVLN) within set of data ;
4. K, the number of populations from which the data points are generated ;
5. The types of population or the probability distribution from which each of the K populations of data points are generated ;
6. The split or n_k, $k=1, 2, \cdots, K$, the size of cluster generated from each population of data points ;
7. The distance, δ_k. between mean vectors for MVN, or median vectors for MVLN.

For the purpose of this study, the probability distribution for each of the K populations generated is fixed to be multivariate normal(MVN) and lognormal(MVLN). The subroutine GGNSM from the IMSL(International Mathematical and Statistical Library) is used to generate data. Generations of the MVN and the MVLN populations will be discussed in detail.

3.2 MVN case

For the convenience, $N=60$, $p=2$, and $K=3$ in this study. Then a brief summary of data structure may be outlined as follows :

$X_i \sim BVN(m_k, \Sigma)$

where : $i=1, 2, \cdots, 60$ with split into the $K=3$ populations of either $20-20-20$ or $30-20-10$;

: m_k, $k=1, 2, 3$, is constrained by an equilateral triangle spatial configuration ;

: $\delta_k=\delta=4.0, 6.0$, is the distance between mean vectors ;

: $\Sigma_k=\Sigma=\begin{pmatrix} 1.0 & \rho \\ \rho & 1.0 \end{pmatrix}$, $\rho=0.0, 0.4,$ and 0.8.

3.3 MVLN case

As it is well-known, the application of techniques developed on multivariate normal distrbution is often limited. Hence, the investigation on the use of Rand's C_k to determine the number of clusters by applying the ACM is extended to a skewed distribution, the multivariate lognormal (MVLN).

Since an ACM is used to find the natural structure present in data, the data structure generated should be reasonably well suited. The desire is to have MVLN data that has similar structure to that constructed for MVN data.

Let X_i be a random vector that follows $N_p(0, \Sigma)$ where

$X_i=[X_{i1}, X_{i2}, \cdots, X_{ip}]^{\prime}$ and set

$Z_{ip}=[Z_{i1}, Z_{i2}, \cdots, Z_{ip}]^{\prime}$.

The transformation

$$Z_{ip}=m_i \ \exp(X_{ip}), \tag{3.1}$$

is applied to obtain a lognormal variate Z_{ip} having

$E(Z_{ip})=\xi_i=m_i \ \exp(\sigma_i^2/2)$,

$VAR(Z_{ip})=\lambda_i^2=m_i \ \exp(\sigma_i^2)(\exp(\sigma_i^2)-1)$,

where m_i, $m_i>0$, is the median.

Then the correlation ρ_{ij}^* between Z_i and Z_j with respect to the correlation ρ_{ij} in the $N_p(0, \Sigma)$ distribution is given by

$$\rho_{ij}^*=\frac{\exp(\rho_{ij}\sigma_i\sigma_j)-1}{[\exp(\sigma_i^2)-1]^{1/2}[\exp(\sigma_j^2)-1]^{1/2}}$$

Thus to cbtain a specified correlation ρ_{ij}^* between Z_i and Z_j, the corresponding correlation ρ_{ij} is

$$\rho_{ij}=\frac{1}{\sigma_i\sigma_j}\ln\{1+\rho_{ij}^* \ [\exp(\sigma_i^2)-1]^{1/2} \ [\exp(\sigma_j^2)-1]^{1/2}\}.$$

It is possible that particular ρ_{ij}'s violate $|\rho_{ij}|\leq 1$ or that the ρ_{ij}'s give a martix Σ that is not positive definite(Johnson, 1987). In this study, the correlation ρ_{ij}^* is set to 0.0 to provide for any general σ_i and σ_j in many data sets. Instead of investigating the effect of correlation (or, noise) between the two variables, the angle, θ, used to set the spatial configuration of data points for each of the population median vectors was varied. Difference in angle by rotating the equilateral triangle would be interpreted in terms of noise(or, perturbation) in the data structure generated from MVLN distribution since the shape of the data structure generated depends on the median vectors which are also dependent on the degree of rotation.

Since a similar data structure which was used for the MVN data is desired, N, p, and K are fixed to be the same as in the MVN study. Thus this study is limited to bivariate lognormal distribution (BVLN) which could be extended to MVLN distribution.

It should be mentioned that the mean vector, ξ, was considered to set the data points for each population with fixed median vector, m. However, a large number of the data points overlapped within the area below the fixed median vectors with skewed-right and long positive tail data regardless of ξ_i, where $\xi_i>m_i>0$. Intuitively, the application of a clustering method was not reasonable even for large differences among the mean vectors. However, the use of the median vector to locate the data points for each population did not suffer from this problem.

Moreover, the variance depends on the median when σ^2 is fixed. The variance of Z_i increases rapidly as the median increases. A large portion of the data points which were generated with a large median always overlapped with another population generated with a small median because of the large difference in the variances. Even if the distance among the median vectors set for the different populations was large, the same type of data structure was obtained. At this point, a reasonable data structure for an application of clustering methods could not be obtained without controlling the variance. The variance for a BVLN random variate Z_{ip} is

$$\lambda_i^2=m_i^2 \ \exp(\sigma_i^2)(\exp(\sigma_i^2)-1).$$

Let λ_i be 1.0 where the median m_i is specified for each population of data points. By solving the equation,

$$\sigma_i^2+\ln[\exp(\sigma_i^2)-1]+2\ \ln(m_i)=0.0, \qquad (3.2)$$

σ_i^2 was obtained to generate BVN .with specified variance and hence a BVLN with variance 1.0 with specified median. Thus ξ decreases rapidly as the median increases. In addition, the shape of data structure generated for BVLN is close to normal (Johnson and Kotz, 1970) for any specified median if σ_i^2 is small, which in this study is a consequence of the choice of a large value for the median. Since the shapes of the distribution of the data points for each population differ from each other as a function of the median vectors, the size of the cluster (split) might effect the "retrieval" ability and "agreement" for unequal sized cluster.

Hence BVLN vectors for each population are generated by solving the equation (4.2) for fixed constant values of the median vectors. And the transformation (4.1) is applied to BVN vectors obtained from a population having a mean vector of zero with specified variance-covariance matrix. Since the number of data points in each population effect the retrieval ability of clustering algorithms, the number of data points is designated for each population generated at the median vectors as:

(1) n_1 at$(1, 1)$,
(2) n_2 at$(1+\delta\cos(\theta), 1+\delta\sin(\theta))$,
(3) n_3 at$(1+\delta\cos(\theta+60), 1+\delta\sin(\theta+60))$.

The data structure for the comparative study may be outlined as follows:

$Z_i \sim \text{BVLN}(m_k, \psi)$,
where $Z_i=[Z_{i1}, Z_{i2}, \cdots, Z_{ip}]$, $i=1, 2, \cdots, 60$, with split into the $K=3$ populations of $n_1-n_2-n_3$ (i.e., $20-20-20$, $30-20-10$, $30-10-20$, $\cdots$, $10-20-30$);
: m_k, $k=1, 2, 3$, is the median vector of each population;
: $\delta_k=\delta=4.0, 6.0$, is the distance between median vectors;
: $\psi_k=\psi=\begin{pmatrix}1.0 & 0.0\\ 0.0 & 1.0\end{pmatrix}$, is the variance-covariance structure;
: $\theta=15°, 30°$, is the angle to set the spatial configuration between median vectors.

4. Analyses and Results

Observations and discussions from the comparative study on the use of C_k were made with respect to the ACA defined by (β, π) and the settings of the structural parameters $(\rho, \delta, \text{split})$ for MVN and $(\theta, \delta, \text{split})$ for MVLN.

For each setting of the structural parameters, a value $C_k(Y, Y')$ is computed for each algorithm, and $C_k(Y', Y'')$ is computed for each pair of the 36 possible pairs of ACA in 100 replications for all $k=2, 3, \cdots, K, \cdots, N-1$. Based on the 100 replications, AC_k, the sample mean, and $STDC_k$, the sample standard deviation of AC_k values, $k=2, 3, \cdots, N-1$, are obtained. Further, the percent(%) of the replications, which is the number of times that C_k satisfy the conditions,

$$C_{k-1}\leq C_k \text{ and } C_{k+1}\leq C_k,$$

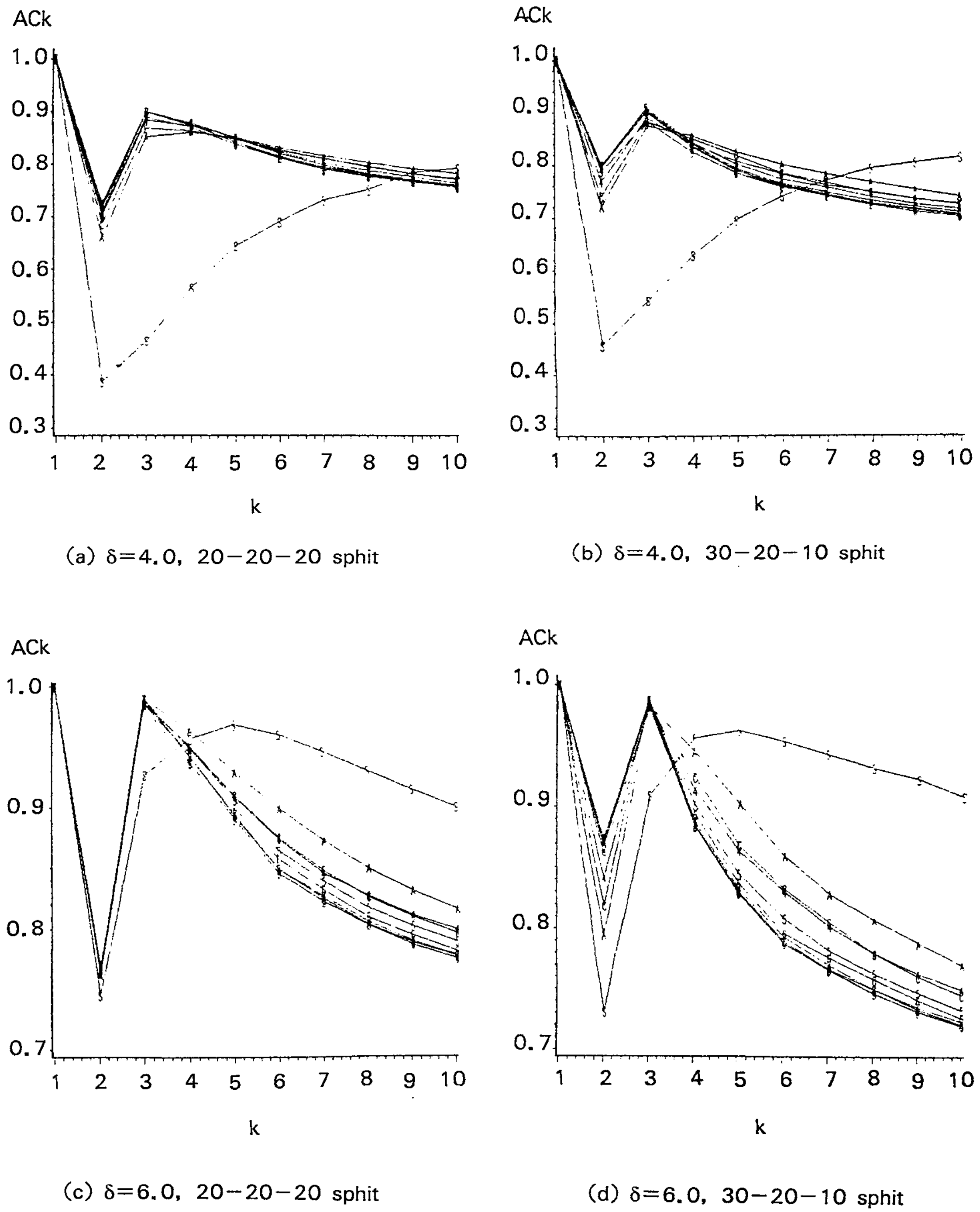
for a known number of clusters, K, where $k=2, 3, \cdots, K, \cdots, N-1$, is obtained for nine ACA and all possible pairs of them. Then, the % obtained by $C_k(Y, Y')$ for each of the nine ACA quantifies how well an ACA "retrieves" the known structure. The % calculated by $C_k(Y', Y'')$ for possible pairs of ACA quantifies how well two ACA in each pair agree to each other through their resultant clusterings giving a local maximum at the specified number k. And the

Table 1. Percent Retrieval of true Population for All Algorithms on MVN

(β, π)	split	20−20−20			30−20−10			AVG%
	δ\ρ	.0	.4	.8	.0	.4	.8	
(.0, −.5)	4.0	13	9	36	19	14	62	25.5
	6.0	67	74	73	64	71	69	69.7
(.0, .0)	4.0	63	68	59	72	65	50	62.8
	6.0	87	87	86	83	88	87	86.3
(.0, .5)	4.0	72	73	56	76	73	50	66.7
	6.0	93	92	91	91	94	90	91.8
(−.25, −.25)	4.0	77	74	73	81	77	72	75.7
	6.0	96	94	92	95	95	88	93.3
(−.25, .0)	4.0	81	81	81	90	89	75	82.8
	6.0	98	99	98	96	96	93	96.7
(−.25, .5)	4.0	83	88	89	85	81	84	85.0
	6.0	93	94	98	96	97	95	95.5
(−.5, .0)	4.0	85	84	91	84	86	69	83.2
	6.0	100	98	97	96	97	95	97.2
(−.5, .25)	4.0	88	87	89	83	83	78	84.7
	6.0	100	100	100	96	99	97	98.7
(−.5, .75)	4.0	79	78	88	78	83	72	83.0
	6.0	99	100	99	95	95	95	97.2

Table 2. Precent retrieval of True Population for All Algorithms on MVLN

(β, π)	split	20-20-20		30-20-10		20-10-30		20-30-10		30-10-20		10-20-30		20-30-20		AVG%
	δ/δ	15°	30°	15°	30°	15°	30°	15°	30°	15°	30°	15°	30°	15°	30°	
(.0,	4.0	18	21	17	21	17	20	25	24	15	22	23	22	27	34	21.9
−.5)	6.0	48	50	43	43	47	41	53	46	45	46	54	44	53	52	47.4
(.0,	4.0	47	53	50	49	53	45	53	47	56	35	53	57	58	57	50.9
.0)	6.0	76	72	75	75	77	72	77	72	67	72	85	76	82	83	75.8
(.0,	4.0	55	55	62	60	54	59	58	62	64	52	71	67	59	66	60.8
.5)	6.0	79	78	84	84	80	71	83	86	82	79	79	86	83	91	81.8
(−.25,	4.0	73	63	65	68	66	68	71	69	61	59	58	58	55	73	67.0
−.25)	6.0	84	82	80	83	80	77	83	80	83	88	88	88	92	89	83.5
(−.25,	4.0	76	73	65	66	74	74	78	77	73	61	78	82	74	82	73.8
−.0)	6.0	93	89	80	82	87	78	84	90	87	79	94	86	94	93	86.9
(−.25,	4.0	81	79	74	76	76	74	75	82	67	62	77	83	80	88	76.7
−.5)	6.0	89	87	86	89	89	84	87	92	86	82	96	91	95	95	89.1
(−.5 ,	4.0	82	84	72	69	86	83	88	89	77	71	84	87	85	89	81.9
−.0)	6.0	83	84	86	90	90	89	90	98	92	82	95	94	95	97	91.8
(−.5 ,	4.0	82	80	76	71	86	85	81	88	75	68	83	86	83	91	81.0
−.25)	6.0	94	98	88	91	95	92	90	95	88	88	98	93	95	95	92.9
(−.5 ,	4.0	79	80	78	79	79	83	80	79	75	68	83	86	83	90	81.0
−.75)	6.0	96	91	86	81	93	90	94	95	82	87	93	94	97	92	91.0



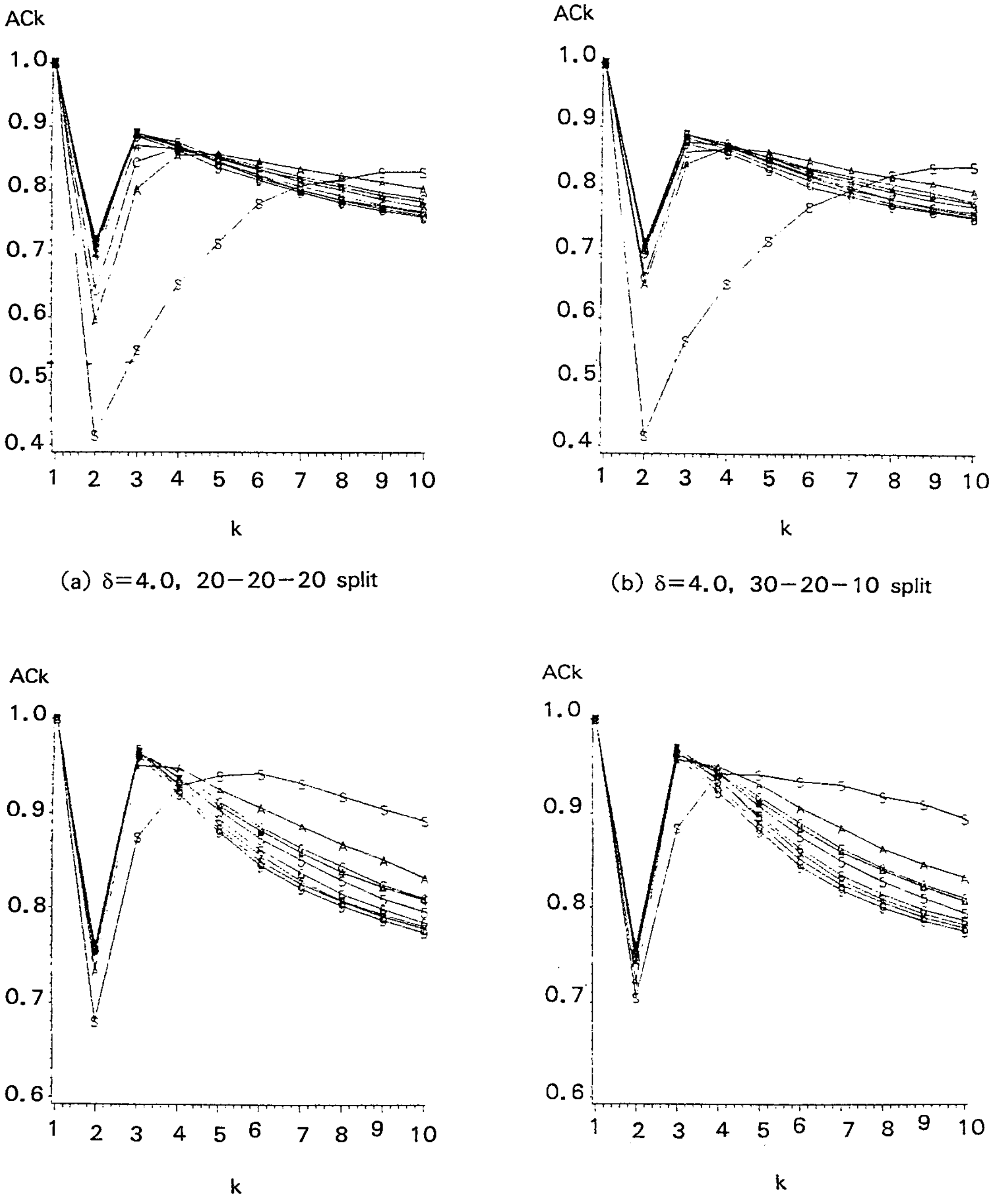
(a) δ=4.0, 20−20−20 sphit

(b) δ=4.0, 30−20−10 sphit

(c) δ=6.0, 20−20−20 sphit

(d) δ=6.0, 30−20−10 sphit

Fig. 1. Retrieval results of the nine ACA with ρ=.0 on MVN



(a) δ=4.0, 20−20−20 split

(b) δ=4.0, 30−20−10 split

(c) δ=6.0, 20−20−20 split

(d) δ=6.0, 30−20−10 split

Fig. 2. Retrieval results of the nine ACA with θ=15 on MVLN

% calculated by $C_k(Y', Y'')$, which is the number of times that two ACA "estimates" the number of clusters correctly, will be defined as $\%_s$. Finally, AVG% and AVG$\%_s$, the sample means of the % and $\%_s$ across all settings of the structural parameters are obtained for nine ACA and possible pairs of them. In addition, STD$\%_s$, the standard deviation of AVG$\%_s$ is calculated for all possible pairs of nine ACA.

Hence, AVG% and AVG$\%_s$ provide informations on how well the C_k "retrieves" the true structure and "estimates" the specified number of clusters, respectively, across all settings of the structural parameters.

Using the results given in tables 1−2 and figures 1−2, at first, the following conclusions may be made for MVN and MVLN :

1) The single linkage algorithm at (.0, −.5) is different from all of the other ACA : i.e., the single linkage is the worst algorithm, in general, however the only good algorithm for high noise ;
2) The average linkage at (.0, .0) and the complete linkage at (.0, .5) perform worse when ρ is close to 1.0 than when ρ is close to 0.0, regardless of the size of cluster (split) for fixed δ with MVN ;
3) For any other ACA defined by $\beta \leq -0.25$ and $\pi \geq 0.0$ in the (β, π) plane, the number of clusters for the the population structure generated is well predicted by C_k for all settings of the structural prameters (ρ, δ, split) with MVN ;
4) For any ACA defined by $\beta \leq -0.5$ and $\pi \geq 0.0$ in the (β, π) plane, the number of clusters are well predicted for MVLN.

At this point, investigation on the general use of C_k with clustering algorithms when any prior information is unknown for given set of data was our objective. It was necessary to choose several pairs of clustering algorithms that cooperate with the comparative statistic, C_k, indicating the number of clusters $k=3$ across all settings of the structural parameters. If the clusterings produced by the nine ACA agree closely, we may have more confidence in prdicting the number of clusters by observing the comparative statistics C_k. The number of local maxima at $k=3$ with respect to $\%_s$ was used to determine the performance of C_k in conjunction with the possible pairs of ACA for the settings of the structural parameters for MVN and MVLN. In fact, the $\%_s$ is the "agreement" between two ACA consisting of a pair. Hence in predicting the number of clusters by using C_k for the settings of the structural parameters, the pair of ACA, A and B, that agree more closely with respect to $\%_s$ in terms of the clusterings than the pair of ACA, A' and B', were chosen for further study iff

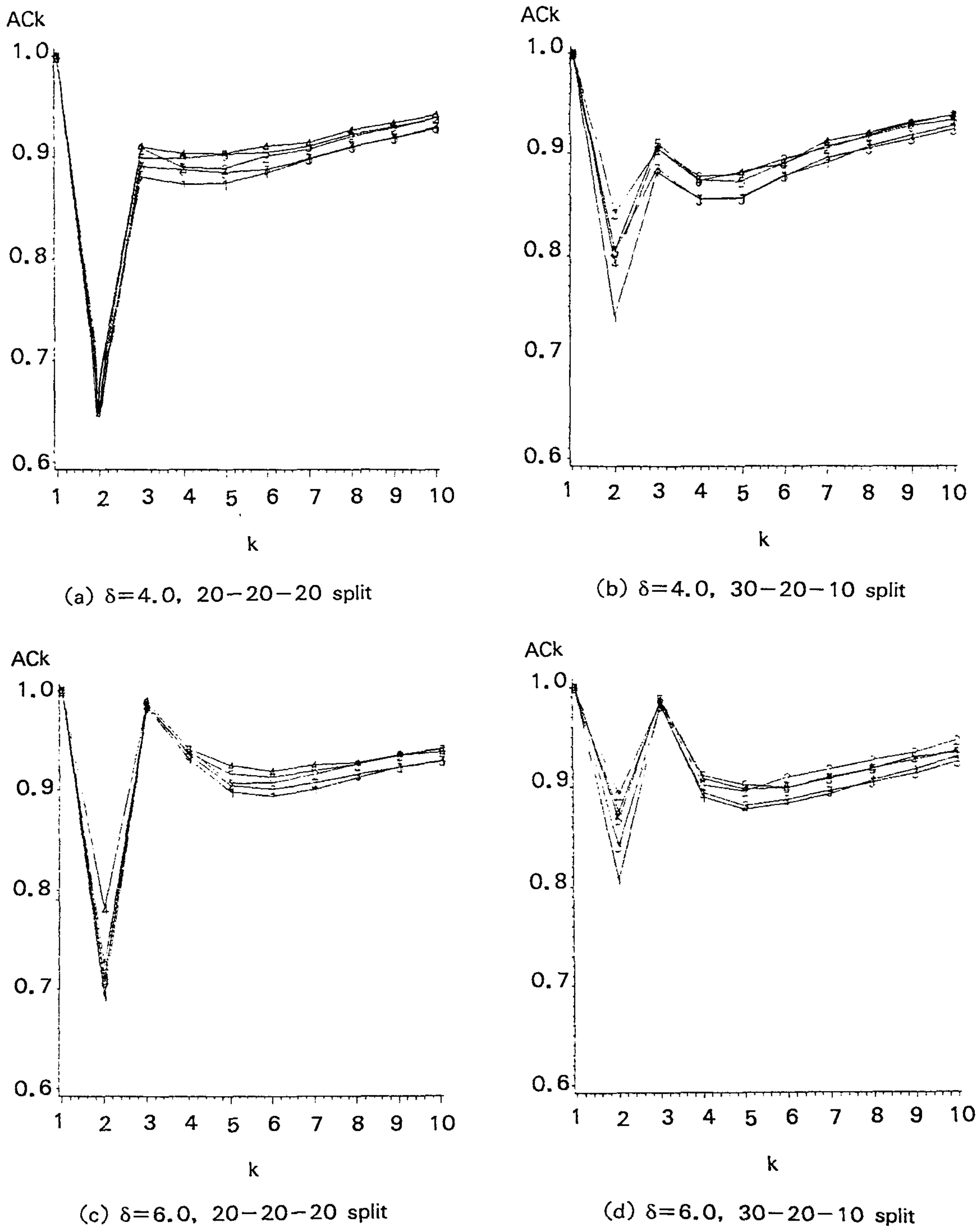
1) $\%_s[A, B] > \%_s[A', B']$, where $\%_s[A, B]$ is the percent of local maxima obtained for paired algorithms A and B ;
2) the %'s, the "percent retrievals" of A and B algorithms, are considered large for the settings of the structural parameters.

In this way, a few general observations with respect to the settings (ρ, δ, split) for MVN and (θ, δ, split) for MVLN was made as follows :

1) ρ does not greatly affect the agreement between the ACA with respect to $\%_s$ for the two splits with the effect becoming less for increasing δ, whenever the pairs with single linkage algorithm are not considered for MVN ;
2) θ has little effect on the agreement between the ACA for the several splits with MVLN ;
3) The different splits with respect to $\%_s$ have little effect on the prediction of the number of clusters for MVN and MVLN ;

Table 3. AVG%$_s$ and STD$_s$ of Correct Prediction by using All Possible Pairs of the nine Algorithms on MVN and MVLN

	POPULATION	MVN				MVLN			
(β, π)	δ	4.0		6.0		4.0		6.0	
	(β, π)	AVG%$_s$	STD%$_s$	AVG%$_s$	STD%$_s$	AVG%$_s$	STD%$_s$	AVG%$_s$	STD%$_s$
(.0 , −.5 ,	(.0 , .0)	24.5	4.45	55.3	2.44	18.5	1.16	40.2	1.60
	(, .5)	28.2	3.73	71.3	1.61	23.3	1.66	50.4	1.91
	(−.25, −.25)	28.2	3.91	65.0	1.59	25.1	1.74	50.1	1.68
	(, .0)	35.5	6.40	72.2	1.62	39.9	1.98	57.4	2.03
	(, .5)	38.0	7.69	75.5	1.61	32.4	2.39	60.3	2.16
	(−.5 , .0)	37.0	6.28	74.0	2.07	34.1	2.27	62.6	1.96
	(, .25)	38.5	7.66	77.0	1.83	33.9	2.44	63.9	1.80
	(, .75)	35.2	6.09	74.0	1.10	31.4	2.00	63.4	1.88
(.0 , .0)	(.0 , .5)	37.3	2.12	59.2	2.17	29.6	1.08	48.2	1.47
	(−.25, −.25)	33.7	2.80	56.3	3.29	31.3	1.18	48.6	1.19
	(, .0)	43.3	3.53	67.0	3.76	36.0	1.36	59.1	1.64
	(, .5)	47.8	2.01	76.2	2.23	41.8	1.40	72.6	1.31
	(−.5 , .0)	46.0	2.68	73.2	3.80	42.0	2.01	71.7	1.85
	(, .25)	49.0	2.02	76.8	3.08	43.8	1.68	74.7	1.40
	(, .75)	47.5	1.45	80.3	3.17	43.6	1.33	77.9	1.69
(.0 , .5)	(−.25, −.25)	39.8	2.52	61.2	2.57	36.4	0.98	55.7	1.50
	(, .0)	42.2	3.90	60.8	3.37	41.2	1.53	56.8	1.80
	(, .5)	48.0	3.62	71.7	1.78	48.0	1.28	68.9	1.51
	(−.5 , .0)	46.7	2.87	69.2	2.27	47.8	1.68	70.3	1.78
	(, .25)	47.0	2.96	74.3	2.89	51.1	1.61	74.5	1.40
	(, .75)	48.2	4.25	78.7	1.43	51.3	1.26	77.6	0.75
(−.25, −.25,	(−.25, −.0)	38.5	1.95	47.0	2.46	32.0	1.58	41.7	1.57
	(, .5)	50.5	0.99	69.0	2.38	47.6	0.98	67.2	1.58
	(−.5 , .0)	46.3	0.99	58.7	3.93	43.6	1.24	59.6	1.60
	(, .25)	51.3	1.43	69.8	2.59	48.4	1.36	67.6	1.33
	(, .75)	53.2	0.91	75.7	1.69	50.4	0.56	74.9	1.11
(−.25, .0)	(−.25, −.5)	47.8	1.14	63.9	2.52	47.1	1.62	60.9	0.98
	(−.5 , .0)	36.7	1.38	50.0	1.91	42.9	1.82	51.4	2.14
	(, .25,	45.7	1.20	61.7	1.74	47.4	1.52	61.7	1.84
	(, .75)	51.7	1.28	72.8	1.99	53.0	1.28	71.9	1.36
(−.25, .5 ,	(−.5 , .0)	44.7	2.20	58.0	2.52	47.4	1.19	59.4	1.62
	(, .25)	42.2	1.08	51.2	1.51	46.1	1.45	53.8	1.51
	(, .75)	46.0	2.94	64.7	2.01	52.4	1.65	65.2	1.18
(−.5 , .0)	(−.5 , .25)	33.8	1.90	40.0	2.25	38.4	1.75	41.9	1.56
	(, .75)	46.0	0.93	67.2	2.54	51.1	1.28	66.3	1.50
(−.5 , .25)	(−.5 , .75)	45.2	1.22	57.8	2.29	47.4	1.22	60.6	1.91



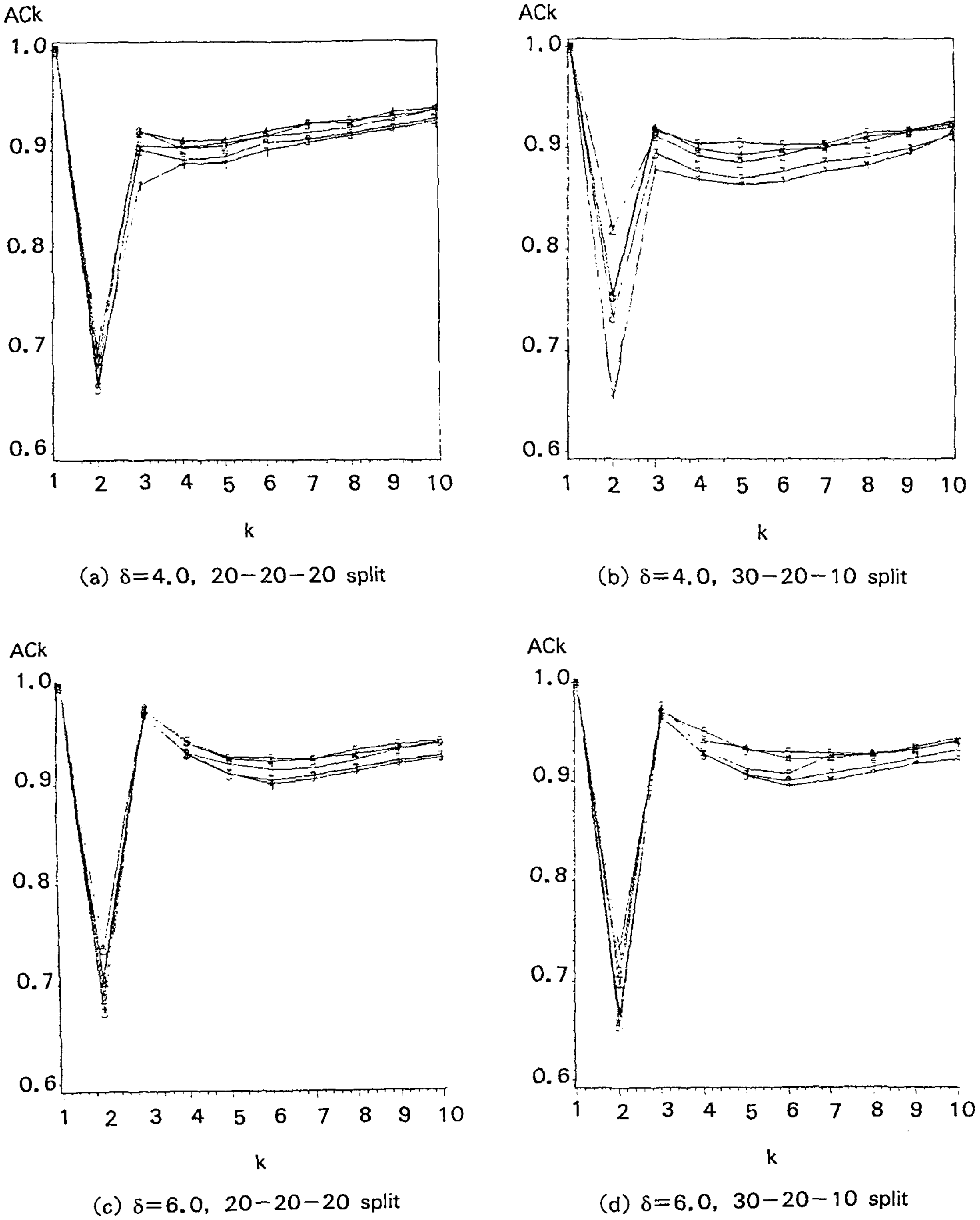
(a) δ=4.0, 20−20−20 split

(b) δ=4.0, 30−20−10 split

(c) δ=6.0, 20−20−20 split

(d) δ=6.0, 30−20−10 split

Fig. 3. Retrieval results of the 5 pairs of ACA with ρ=.0 on MVN



(a) δ=4.0, 20−20−20 split

(b) δ=4.0, 30−20−10 split

(c) δ=6.0, 20−20−20 split

(d) δ=6.0, 30−20−10 split

Fig. 4. Retrieval results of the 5 pairs of ACA with θ=15 on MVLN

4) Increasing δ from 4.0 to 6.0 causes an increase in $\%_s$ for all settings (ρ, split) for MVN and (θ, δ) for MVLN.

Overall, it is not necessary to consider the all structural settings(i.e., ρ or θ, split and δ), since the structures in many data sets are usually unknown. Then the summary on the $\%_s$ for all possible pairs of nine ACA is given in table 3.

Based on the AVG$\%_s$ and STD$\%_s$ from table 3, the pairs with (−.5, .75) in the (β, π) plane perform better with respect to C_k than the other pairs of clustering algorithms for both MVN and MVLN. Some pairs of algorithms indicate the number of clusters better than the others for specific settings of the structural parameters. In addition, the behaviors of C_k through AC_k, k=1, 2, ⋯, 10, for subjectively chosen five pairs among other pairs of algorithms with MVN and MVLN are represented in figures 3−4, respectively. Five pairs are,

1) (−.5, .75) vs .(.0, .5),
2) (−.5, .75) vs .(−.25, .0),
3) (−.5, .75) vs .(−.25, −.25),
4) (−.5, .25) vs .(−.25, −.25),
5) (−.25, .5) vs .(−.25, −.25).

Moreover, the % retrieval of the true population generated with the specific structural parameter for each clustering algorithm was considered from tables 1−2 for MVN and MVLN, respectively. If both algorithms combined as a pair have high retrieval abilities for the true population, we will consider the pair to be the best among five pairs of algorithms for both MVN and MVLN data. In this way the structure of clusterings produced by the pair of clustering algorithms is also similar to the data structure generated.

From the results of the comparative study, it is concluded that the use of C_k with the pair of algorithms, (−.5, .75) vs.(−.25, .0), defined in the (β, π) plane is recommended in predicting the number of clusters, regardless of the characteristics of the given set of data.

This confirms that the flexible strategy at (−.25, .0) recommended by DuBien and Warde (1987) is at least one algorithm for finding the unknown structure present in many data sets. Moreover, the pair of algorithms (−.5, .75) vs. (−.25, 0) generally performs better than any combinations of single, complete, and average linkages, regardless of the degree of noise(ρ, or θ) and the relative sizes(splits) of the clusters present in the data.

5. Concluding Remarks

A great of flexiblity in a limited extension of the comparative study could be achieved by applying Rand's C_k and choosing different agglomerative clustering algorithms to pair with the (−.5, .75) algorithm defined in the (β, π) plane. Since the use of C_k with the pairs of (−.5, .75) with other clustering algorithms predicted the number of clusters fairly well.

In conclusion, it appears from the all evidence on its performance that the pair of agglomerative clustering algorithms, (−.5, .75) vs. (−.25, .0), with C_k statistic is a useful method on determining the number of clusters present in the set of data for MVN and MVLN. Also, this could be extended to the other types of data sets. However, the performance of C_k is dependent on the characteristics of the data, the choices of agglomerative clustering algorithms and distance measures. Therefore, the results on the use of C_k should be examined critically to make sure they are meaningful.

Acknowledgement

We should like to thank to the editor and the referees for their helpful comments and suggestions in improving this paper.

References

1. DuBien, J.L. and Warde, W.D.(1979). A Mathematical Comparison of the Members of an Infinite Family of Agglomerative Clustering Algorithms. *Canadian Journal of Statistics*, 7. 29-38.
2. DuBien, J.L. and Warde, W.D.(1987). A Comparison of Agglomerative Clustering Methods with respect to Noise. *Communication in Statistics : Theory and Method*, 16, 1433-1460.
3. Fowlkes, E.B. and Mallows, C.L.(1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of American Statistical Association*, 78, 553-584.
4. Gordon, A.D.(1981). *Classification* : Methods for the Exploratory Analysis of Multivariate Data, New York : Champman and Hall.
5. Johnson, Mark E.(1987), *Multivariate Statistical Simulation*. New York : John Wiley & Sons.
6. Johnson, N.L. and Kotz, S.(1970). *Distribution in Statistics* : Continuous Univariate Distributions 1, New York : John Wiley & Sons.
7. Lance, G.N. and Williams, W.T.(1966). A Generalized Sorting Strategy for Computer Classification, *Nature*, 212, 218.
8. Lance, G.N. and Williams, W.T.(1967). A General Theory of Classificatory Sorting Strategies, 1. Hierachical Systems. *The Computer Journal*, 9, 373-380.
9. Rand, William Medden.(1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of American Statistical Association*, 66, 846-850.