A Comparison of Four Clustering Methods Using MMPI Monte Carlo Data

Roger K. Blashfield and Leslie C. Morey
University of Florida

Monte carlo procedures were used to generate data sets that resembled MMPI psychotic (8-6), neurotic (1-3-2), and personality disorder (4-9) patterns. Lorr's clumping method, inverse factor analysis, average linkage, and Ward's method were the clustering methods compared. The solutions were found to vary in terms of misclassifications and coverage. The clustering solutions also varied as a function of different optional parameters associated with each method.

Cluster analysis is a generic term encompassing a large and diverse family of statistical classification procedures. The use of cluster analysis in the study of the classification of psychopathology has increased considerably in the past 15 years. Surprisingly, research using cluster analysis has not had the impact that might be expected in a field that so needs methodological improvement (Kendell, 1975). This lack of impact exists for various reasons. First, the existence of over 100 different clustering algorithms, coupled with the disturbing fact that these algorithms can result in different solutions to the same problem (Bartko, Strauss, & Carpenter, 1971), has led to considerable confusion about the efficacy of cluster analysis. Also, many clustering procedures have been developed to address problems of dubious relevance to psychopathology, such as differentiating forest soils or classifying the penile morphology of New Guinea rodents. These problems are compounded by the conflicting results of the few studies that have attempted to decide which clustering method is most useful in applied research (Everitt, 1979).

During the 1970s a small set of research studies that used monte carlo techniques were developed. By using these techniques it was possible to create data sets in which the true classificatory structures were known precisely. That is, artificial data were created in such a way that the exact classificatory structure of the data could be specified before any cluster analysis was performed. The use of monte carlo techniques thus permitted the comparison of cluster analysis solutions with the known classificatory structure of the data sets.

A serious problem with existing monte carlo studies is that the similarity between the monte carlo generated data and actual data used in clinical research is unknown. For example, Gross (1972) generated data sets in which there were three clusters. Twenty-five entities were sampled to represent each cluster along five measurement variables. Each variable was independently sampled and was assumed to be normally distributed. All variables had the same variance, but different clusters had different

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 1 Winter 1980 pp. 57–64

mean vectors. However, in most clinical research, variables are correlated and have different variances. Thus, it is not known how well the results found by Gross (1972) would generalize to actual clinical research problems. Future studies that compare cluster analysis methods might best attempt to use monte carlo data sets that are modeled after data structures actually encountered in research on the classification of mental disorders. This study attempts such an analysis.

The monte carlo data sets in this study were modeled after a classificatory structure that has been suggested to underlie that of the MMPI. Recent research by Skinner and Jackson (1978) has pointed to the existence of three basic categories underlying this instrument: (1) a psychotic category, whose profile resembles an 8-6 code type; (2) a neurotic category, whose profile is similar to a 1-3-2 code type; and (3) a personality disorder category whose example code type is a 4-9.

The major purpose of this study was to examine how different cluster analysis methods performed when applied to monte carlo data sets. The methods analyzed in this paper are among the most popular of the cluster analysis methods in previous research on psychopathology. The resulting classifications were analyzed in detail to determine the weaknesses and strengths of the various methods.

Method

Subjects

A mixture model was used to create the data set (Wolfe, 1970). Each parent sample was created using a routine that created multivariate random normal deviates having a specified covariance matrix. The parameters needed for creating each parent sample were a vector of means (the centroid profile), a vector of standard deviations, and a matrix of MMPI scale intercorrelations. The mean vectors for the three parent samples were the values specified for the Gilberstadt and Duker (1965) 1-3-2 profile, the Marks and Seeman (1963) 8-6 profile, and the Gilberstadt and Duker (1965) 4-9 profile. The correlation matrix for each parent sample was the matrix of MMPI scale intercorrelations reported by Swenson, Pearson, and Osborne (1973) from a sample in excess of 25,000 medical patients. Eleven MMPI scales were used, with Mf and Si not included from the standard 13 scales.

No published vectors of standard deviations were found in the MMPI literature for the three code type profiles. Hence, the standard deviations were estimated by using the code book rules to assign MMPI profiles from a large data set (Goldberg, 1965) to the 1-3-2, 8-6, and 4-9 code types. The standard deviation vectors were multiplied by a constant, which led to a moderate amount of overlap among the three clusters.

To represent each parent sample, 30 MMPI profiles were created. The 3 groups of 30 profiles were then mixed together to form data sets (mixtures) with 90 profiles and a known classificatory structure of 3 categories. After the first mixture was created, a second mixture was formed in order to permit a replication of results noted with the first mixture.

Cluster Analysis Methods

The cluster analysis methods selected for inclusion on the study were (1) average linkage (Sneath & Sokal, 1973); (2) Ward's (1963) method; (3) inverse factor analysis; and (4) Lorr's (1966) clumping method. These methods were selected because they have all been used in studies of psychiatric classification. A brief description of each of these methods will be provided in the following section.

Misclassification and Coverage

In order to provide a comparison of the four clustering methods and how well they solved the MMPI mixtures, two criteria were initially examined: the number of misclassifications and the coverage rates. A misclassification was defined as a profile that was assigned to a cluster in which the dominant membership was of profiles from a different parent sample. For example, in a cluster that had four members—three profiles from the 1-3 parent sample and one from the 4-9 sample—this last profile was considered misclassified.

The other criterion was coverage, or the number of profiles in a data set that were not assigned to any cluster at all. It should be noted that as standardly used, the two hierarchical agglomerative methods always assign profiles to some cluster. However, inverse factor analysis and Lorr's method permit less than complete coverage.

Results

Lorr's Clumping Procedure

Lorr's method creates clusters one at a time. This method uses cutoff criteria for inclusion into and exclusion from cluster membership. The similarity measure used by this method is Pearson correlation. Lorr suggested that the profile correlation value significant at the .05 level be used as the inclusion cutoff. He also recommended the use of the .01 level to determine the exclusion cutoff. Thus, for 11 variables the .05 value (inclusion cutoff) was $r = .60$, and the exclusion was $r = .74$. Lorr's program, BUILDUP, was used to perform this method.

In the first mixture, six clusters were found. The first cluster was the largest, with 26 members, 15 of which came from the 8-6 sample and 11 of which came from the 4-9 sample. The next five clusters averaged 10 members per cluster (range of 6 to 14) and were generally pure representations of different samples. For example, the second and fourth cluster contained only members of the 1-3 parent sample. In total, there were 12 profiles assigned to clusters in which the predominant membership was from a different sample (i.e., 12 misclassifications). Of the 12 misclassifications 11 occurred in the first cluster. Additionally, 16 of the 90 profiles were not assigned to any cluster in the data set (82% coverage). For the replication data set, 7 clusters were found with steadily decreasing membership size. There were 5 misclassifications and 84% coverage.

Table 1 compares the results from the original data set with the replication data set. It can be noted that the first three clusters in the original data also appeared in the replication. However, the remaining three clusters were not replicated in the second data set. Thus, the number of clusters was overestimated by this clumping procedure, and some clusters failed to be substantiated on replication. If only the replicated clusters from the original data set were accepted (a procedural step recommended by Lorr), then the coverage for the original data set would fall to 54% and there would be 12 misclassifications.

Table 1
Solutions to the Original and Replication Data Sets Using Lorr's Method with .60 Inclusion and .74 Exclusion Parameters

Original		Replication	
Number	Mean Profile	Number	Mean Profile
26	8*764F9"2'	23	86F*72"4013'
14	1"3'	14	13"2
9	4"9'	16	49"
11	1*23"78'	8	4"96873'
8	6*8"470'	4	18*237"60'
6	4"K9'	5	469'
		6	3'

Note These profiles are represented using the Welsh code.

The three clusters that did replicate represented the actual cluster structure of the data.

In order to provide further information about the characteristics of Lorr's clumping procedure, two additional analyses were performed on the data. The two modifications involved the alteration of the inclusion and exclusion parameters in order to analyze the effect of these parameters on the cluster solutions. In the first analysis, the inclusion parameter was reduced to $r = .40$, and the exclusion parameter was set at $r = .60$. The resulting solution contained four clusters, with the first two clusters being quite large. The first cluster had 34 members, 19 from 8-6 sample and 15 from the 4-9 sample. The second cluster had 25 members, all but one of which were from the 1-3-2 sample. Of the 90 profiles 16 were not classified, yielding an 82% coverage rate, which is the same as the 82% coverage rate when the standard inclusion and exclusion cutoffs were used.

The second additional analysis increased the inclusion and exclusion cutoffs to .70 and .80, respectively. The results were that Lorr's clumping procedure located 8 clusters that were relatively small (membership range was 15 to 4). There was only one profile misclassified. However, the coverage rate was only 73% (24 profiles not assigned).

Summary. Lorr's clumping procedure overestimated the number of clusters. Also, the solutions tended toward finding a relatively large first cluster that was a mix of profiles from two samples. The conclusion from the analyses with varied cutoffs is that if the cutoffs are set fairly low, a better estimate of the number of clusters will be made and the coverage of the solution may be higher. However, the resulting cluster solution is less likely to represent the actual structure of the data. If relatively high cutoffs are used, the cluster solution will overestimate the number of clusters, but the clusters that are found will not contain many misclassifications. In other words, high cutoff solutions will tend to find clusters, representing "local densities" in the data space, that are unlikely to be replicated on another sample. In addition, high cutoff solutions will increase the percentage of entities not assigned to any cluster.

Inverse or Q Factor Analysis

Inverse factor analysis forms clusters by factoring a matrix of correlations among subjects (instead of correlations among variables). Principal components factor analysis, followed by a varimax rotation, was performed using the SPSS program.

Two judgmental decisions had to be made in order to use inverse factor analysis. The first concerned the number of factors. A popular technique for this decision is to choose the number of factors that have eigenvalues greater than 1.0. A subjective analysis of the eigenvalues suggested that the eigenvalue-greater-than-one rule was inappropriate. For example, the eigenvalues from an inverse factor analysis on one data set were 34.11, 19.69, 12.61, 6.43, 5.26, 3.93, 3.34, 2.00, 1.44, and 1.14. The eigenvalue-greater-than-one rule would suggest the presence of 10 factors. However, common judgmental rules, such as the scree test (Cattell, 1966), would indicate that three factors were present. In both data sets, the scree test suggested the presence of three factors; hence, three-factor solutions were chosen for the rotations.

How large a factor loading is necessary for a profile to be assigned to a factor? Initially, each profile was assigned to the factor on which it had a loading of .6 or greater. If a profile had loadings of .6 or greater on more than one factor, it was not assigned to any factor. The reason for choosing the .6 loading was that this was the inclusion cutoff in the Lorr clumping procedure, and any correlation less than .6 would not be significantly (at the .05 level) different from zero. Three factors accounted for 74% of the variance (75% on replication). Eight profiles were misclassified (12 on replication); 2 were assigned to more than 1 factor (on replication, 5); and 16 profiles were not assigned at all (9 in the replication data set). Thus, the coverage rate was 80%. Most of the misclassified profiles had a high negative loading on a factor associated with a different cluster. For example, one profile that came from the 1-3-2 sample had a high negative loading on the 4-9 factor, suggesting that this profile could be described as the polar opposite of the 4-9 profiles.

In order to give some idea of the effect of using a different cutoff for assigning profiles to a factor, the varimax solution was analyzed twice more. When the decision was made to accept any loading greater than .4 as sufficient for membership on a factor, there were 5 misclassifications; 39 profiles were assigned to more than 1 factor; and every entity had a loading of .4 or greater with some factor (100% coverage). Assigning each profile to the factor on which it had the highest loading (regardless of the absolute value of that loading) resulted in 19 misclassifications; no multiple assignments to factors occurred, and the coverage was 100%.

Summary. Inverse factor analysis required two major decisions: choosing the number of factors and deciding on the size of the factor loading sufficient for cluster membership. The first decision generally was not difficult with these data sets. However, the choice of factor-loading cutoff had the effect of altering the relative tradeoff between coverage and misclassifications. The most conservative rule, using a .6 cutoff, kept down misclassifications; but coverage was poor. On the other hand, the decision rule to use the largest loading made the coverage complete but markedly increased the number of misclassifications.

Average Linkage

The third method of cluster analysis was average linkage, also called UPGMA, or unweighted pairwise group mean average clustering, by Sneath and Sokal (1973). The similarity measure was the Pearson product-moment correlation, the same as was used in inverse factor analysis, and Lorr's clumping procedure. CLUSTAN was the program used to perform the analyses.

Hierarchical agglomerative methods (of which average linkage is one) do not determine the number of clusters in the data set. As a result, the decision was made to apply Mojena's (1977) first rule for determining the number of clusters implied by a hierarchical agglomerative method. Mojena's rule indicated that three clusters were present. In contrast, a visual analysis of the UPGMA dendrogram did not suggest an obvious three-cluster solution.

For the three-cluster solution, average linkage had 26 misclassifications (12 on replication). Most misclassifications occurred because many of the profiles from the 4-9 sample were assigned to a cluster containing all the 8-6 profiles. Examination of the misclassified 4-9 profiles showed that these profiles tended to have 3 to 6 scales with T scores greater than 70.

Another important feature of the average linkage cluster solution was that if the five-cluster solution was adopted instead of the three-cluster solution, then relatively few misclassifications would be made. The five-cluster solution contained three relatively large clusters (26 to 28 members) and two small clusters (2 to 6 members). If this five-cluster solution was made, there were three misclassifications.

This example exhibited a general result noted in all hierarchical clustering solutions. The further up the hierarchical tree the number of clusters were selected, the fewer were the number of misclassifications. In other words, when a solution with a relatively large number of clusters was adopted, the clusters were relatively "tight" and contained profiles from the same parent sample. However, in such solutions the number of clusters was strikingly overestimated.

Edelbrock (1979) has made a useful suggestion about such solutions. He pointed out that it is not necessary for a hierarchical solution to be interpreted in such a way that the coverage needs to be 100%. Suppose the decision had been to look at the seven-cluster solution from the average linkage result but to delete clusters that contained less than 10% of the total sample as "outliers." Using this procedure proposed by Edelbrock, the number of misclassifications would only be 3, and the coverage would be reduced to 91%. For the replication sample, the number of misclassifications was 12 and coverage was 100%.

Ward's Method

Ward's (1963) method, like average linkage, is a hierarchical agglomerative method. However, this method uses Euclidean distance measures and attempts to minimize the within-cluster variance. To perform Ward's method, CLUSTAN was used.

As with the average linkage solutions, Mojena's rule indicated the presence of three clusters in the Ward's solutions. Thus, misclassification rates were computed for the three-cluster solutions. Ward's method had 22 misclassifications (12 misclassifications on replication). Edelbrock's modification did not work with Ward's method because it tends to generate clusters of nearly equal membership size. For example, the four-cluster solution for the original data using Ward's method had 29, 29, 18, and 14 members.

Examination of the misclassified profiles indicated that Ward's method appeared to be most sensitive to profile elevation. For example, the major number of misclassifications occurred between the 8-6 sample and the 4-9 sample. The misclassified 8-6 profiles had profiles with relatively low elevations. On the other hand, the misclassified 4-9 profiles had elevations that were relatively high.

This tendency of Ward's method to be sensitive to profile elevation was particularly evident in the solution to another data set in which cluster overlap was large. The three-cluster solutions of these data contained (1) a group of highly elevated profiles in which the cluster centroid had only three scales with T-score means below 70; (2) a group of moderately elevated profiles whose cluster centroid was not greatly different from the grand mean of the total data set; and (3) a group of relatively low profiles that matched the 1-3-2 sample fairly well. In the face of noisy data, Ward's method tended to form clusters that represent a categorical scaling on an "elevation continuum."

Summary of the Results

The optimum solutions obtained by the clustering methods are presented in Table 2. From the table the variability in the cluster solutions across clustering methods and across data sets is apparent. With the original MMPI monte carlo data, average linkage using Edelbrock's modification had the least number of misclassifications, relatively high coverage, and a correct estimate of the number of clusters. On the replica-

Table 2
Comparison of Solutions to the Original (and Replication) Data Using Modifications to the Clustering Methods

Method	Number of Clusters	Percent Coverage	Number of Misclass-ifications
Lorr (standard)	6 (7)	82 (84)	12 (5)
Inverse Factor Analysis (.6 Loading) with scree test-no double assignments	3 (3)	80 (84)	8 (12)
Average Linkage (Edelbrock's modification)	3 (3)	91 (100)	3 (12)
Ward's Method	3 (3)	100 (100)	22 (12)

tion data, Lorr's method had the fewest number of misclassifications, but at the cost of relatively low coverage and an overestimate of the number of clusters. Additional runs using these four cluster methods showed that the methods were sensitive to optional control parameters.

Discussion

What do the results of this study indicate for the use of cluster analysis with psychological test data? First, it should be noted that when using Lorr's method and inverse factor analysis, there was a general tradeoff between coverage and misclassification rates. It seems that these two methods will yield solutions representing relative pure modal "types," even in the face of highly variable data. However, the methods accomplish this goal by ignoring the information provided by a large number of entities.

On the other hand, the hierarchical techniques utilize all of the information provided in the sample, and as such, these methods are more susceptible to error variance. For instance, Ward's method relies primarily on scale elevation to separate MMPI groups. When presented with highly variable data, this is not necessarily a bad procedure, but it yields little information about the category structure of data. Average linkage also has problems when cluster variance becomes large. When Edelbrock's procedure for defining outliners was used, this method could be relatively accurate but with some cost in terms of coverage.

The newness of the procedures suggested by Mojena (1977) and Edelbrock (1979) have not allowed much time for systematic research. However, the results of this study seem to indicate that they have much promise. Also Cattell's (1966) scree test was quite accurate as a test for the number of clusters when inverse factor analysis was used. More research on these procedures needs to be done, as they may be helpful in eliminating some of the subjective decisions made when performing cluster analysis.

The objectivity that cluster analysis promises to lend to the science of psychiatric classification is long overdue. In studying human decision making, it is apparent that humans, even skilled ones, are often overwhelmed by the complexity of data with which they are faced and tend to overuse traditional preconceptions about the data set rather than being sensitive to whatever de facto structure may exist. The Kraepelinian system of classification has so pervaded the thinking of contemporary clinicians that it is difficult to get a fresh perspective using human classifiers. Cluster analysis, as an empirical classification methodology, has the promise of providing an objective and unbiased look at the actual structure of descriptive data on psychiatric patients.

References

Bartko, J. J., Strauss, J. S., & Carpenter, W. T. An evaluation of taxometric techniques for psychiatric data. *Classification Society Bulletin*, 1971, *2*, 2-28.

Cattell, R. B. The scree test for the number of factors. *Multivariate Behavioral Research*, 1966, *1*, 245-276.

Edelbrock, C. Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. *Multivariate Behavioral Research*, 1979, *14*, 367-384.

Everitt, B. S. Unresolved problems in cluster analysis. *Biometrics*, 1979, *35*, 169-181.

Gilberstadt, H., & Duker, J. *A handbook for clinical actuarial MMPI interpretation*. Philadelphia: W. B. Saunders Company, 1965.

Goldberg, L. R. Diagnosticians vs diagnostic signs: The diagnosis of psychosis vs neurosis from the MMPI. *Psychological Monographs*, 1965, *79*, (Whole No. 602).

Gross, A. L. A Monte Carlo study of the accuracy of the hierarchical grouping procedure. *Multivariate Behavioral Research*, 1972, *7*, 379-389.

Kendell, R. E. *The role of diagnosis in psychiatry*. Oxford: Blackwell, 1975.

Lorr, M. M. *Explorations in typing psychotics*. New York: Pergamon, 1966.

Marks, P. A., & Seeman, W. *The actuarial description of abnormal personality*. Baltimore: Williams & Wilkins, 1963.

Mojena, R. Hierarchical grouping methods and stopping rules—An evaluation. *Computer Journal*, 1977, *20*, 359-363.

Skinner, H. A., & Jackson, D. N. A model of psychopathology based on an integration of MMPI actuarial systems. *Journal of Consulting and Clinical Psychology,* 1978, *46,* 231-238.

Sneath, P. H. A., & Sokal, R. R. *Numerical taxonomy.* San Francisco: W. H. Freeman, 1963.

Swenson, W. M., Pearson, J. S., & Osborne, D. *An MMPI source book: Basic item, scale, and pattern data on 50,000 medical patients.* Minneapolis: University of Minnesota Press, 1973.

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

Wolfe, J. H. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research,* 1970, *5,* 329-350.

Acknowledgments

The computer analysis discussed in this paper was supported by the Department of Psychiatry at the University of Florida.

Author's Address

Send requests for reprints or further information to Roger Blashfield, Box J-256, JHMHC, Gainesville, FL 32610.