SIMULTANEOUS CLUSTERING OF CASES AND VARIABLES
-A Conceptual Framework and Some Examples-

by

Yasuo Ohashi

University Hospital Computer Center
University of Tokyo Hospital
7-3-1, Hongo, Bunkyo-ku, Tokyo, JAPAN

Institute of Statistics Mimeo Series No. 1827

July 1987

SIMULTANEOUS CLUSTERING OF CASES AND VARIABLES
-A Conceptual Framework and Some Examples-

Yasuo Ohashi

University Hospital Computer Center
University of Tokyo Hospital
7-3-1, Hongo, Bunkyo-ku, Tokyo, JAPAN

Abstract : An approach for simultaneous clustering of cases and variables is proposed and examples of strategies of data analysis are discussed using three types of data.

In this paper, it is assumed that data is given in the form of a cases×variables data matrix where measurements are taken on continuous scale. Existing models for approximating a rectangular data matrix are summarized and (simultaneous) clustering is characterized as a method of increasing the goodness of fit of each model or decreasing the number of parameters of a well-fitted model through two operations named "localization" and "merging", respectively.

Properties of two operations and estimation procedures are summarized and relationship to some existing clustering techniques is mentioned.

Keywords : Simultaneous clustering, Block clustering, two-way table, Additive model, Multiplicative model

This research was done while the author stayed at the Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, from October, 1986 to July, 1987.

1. INTRODUCTION

It is our usual experience that the results of clustering of cases depend, often largely, on the variables used for analysis and, *vice verse*, those of variables depend on the cases used, whether those results are obtained by the application of automated classification techniques such as *k*-means method or by the visual inspection of the output of ordination techniques such as principal component analysis. So, it is often recommended to reanalize the data by deleting or subsetting cases and/or variables in order to verify the "stability" and "generality" of the results of analysis.

One possible remedy for the problems in selecting cases and/or variables to be used in classification is to classify all cases and all variables simultaneously, as Hartigan[9] points out. Although a number of block clustering methods(Hartigan[8],[9]), one of which is implemented in BMDP as P3M[3], are proposed for the simultaneous clustering, it seems to the author that they are *ad hoc* and an approach which has some correspondence with well-established data analysis techniques is worth developing.

In this paper, it is assumed that data is given in the form of a cases×variables data matrix where measurements are taken on continuous scale. In Section 2, existing models for approximating a rectangular data matrix are summarized and (simultaneous) clustering is characterized as a method of increasing the goodness of fit of each model or decreasing the number of parameters of a well-fitted model through two operations named "localization" and "merging", respectively. In section 3, strategies of simultaneous clustering based on

the proposed method are discussed with numerical examples. Several future problems are briefly discussed in Section 4.

The author should confess that this paper discusses only a conceptual framework and a small number of examples. Some important practical problems including development of efficient algorithms and softwares, determination of the number of clusters, empirical comparison with other approaches and interpretability of the results are little discussed in this paper, and, at present, the object of application of the proposed approach is limited to a relatively small data because of its heavy requirements to computer resources. Prototype programs used for analyzing data in Section 3 were developed as SAS/IML(Interactive Matrix Language)[20] modules and can be used interactively in Proc IML of SAS.(These lists are available upon direct request to the author.) Supplementary analysis including usual clustering techniques are also carried out by using SAS[19]. A design of an integrated software based on the proposed approach is under study.

2. MODELS AND THEIR FITTING

2.1. Approximation of a rectangular data matrix

Let $\mathbf{X}=(x_{\alpha i})$ be a data matrix where $\alpha=1,...,A$ and $i=1,...,I$ refer to cases and variables, respectively. A variety of models are proposed for approximating $\mathbf{X}$ by $\hat{\mathbf{X}}=(\hat{x}_{\alpha i})$ as follows:

Constant model:	$\hat{x}_{\alpha i} = c$	(M0)
Case(row) effect model:	$\hat{x}_{\alpha i} = t_{\alpha}$	(MC)

Variable(column) effect model: $\hat{x}_{\alpha i} = u_i$ (MV)

Additive model: $\hat{x}_{\alpha i} = t_\alpha + u_i$ (MA)

$(\Sigma t_\alpha = 0 \text{ or } \Sigma u_i = 0)$

Multiplicative model: $\hat{x}_{\alpha i} = dp_\alpha q_i$ (MM)

$(\Sigma p_\alpha^2 = 1, \Sigma q_i^2 = 1)$

Mixture model: $\hat{x}_{\alpha i} = t_\alpha + dp_\alpha q_i$ (MXC)

$(\Sigma p_\alpha^2 = 1, \Sigma q_i^2 = 1, \Sigma q_i = 0)$

$\hat{x}_{\alpha i} = u_i + dp_\alpha q_i$ (MXV)

$(\Sigma p_\alpha^2 = 1, \Sigma q_i^2 = 1, \Sigma p_i = 0)$

$\hat{x}_{\alpha i} = t_\alpha + u_i + dp_\alpha q_i$ (MXA)

$(\Sigma t_\alpha = 0 \text{ or } \Sigma u_i = 0, \Sigma p_\alpha^2 = 1, \Sigma q_i^2 = 1, \Sigma p_i = 0, \Sigma q_i = 0)$

Saturated model: $\hat{x}_{\alpha i} = v_{\alpha i}$ (MS)

where restricted conditions in parentheses are necessary for identifiability of each parameter value.

We adopt the following squared norm as a measure of goodness of fit of each model:

$$||\mathbf{X}-\hat{\mathbf{X}}||^2_{\mathbf{W},\mathbf{V}} = \Sigma_\alpha \Sigma_\beta \Sigma_i \Sigma_j w^{\alpha\beta} v^{ij} (x_{\alpha i} - \hat{x}_{\alpha i})(x_{\beta j} - \hat{x}_{\beta j}) \quad (1)$$

where $\mathbf{W}=(w^{\alpha\beta})$ and $\mathbf{V}=(v^{ij})$ are metric matrices which accommodate intra-cluster correlation structure or give weights to cases and variables, respectively. If each case is independent and considered to have to give the same weight to analysis, $w^{\alpha\beta}$ should be equal to

$$\delta^{\alpha\beta} = 1 \text{ (when } \alpha=\beta), 0 \text{ (otherwise).} \quad (2)$$

It is well known that *k*-means method assumes implicitly a metric matrix of (2) for $\mathbf{V}$ as well as $\mathbf{W}$ and the resulting clusters tend to be spherically shaped and "natural" but long-shaped clusters often fail to be detected(Everitt[5]). Although there are some attempts for estimating intra-cluster

variable correlation (Ace et al.[1] and ACECLUS in SAS[19]), in the following we adopt the usual squared Euclidean norm(the residual sum of squares;RSS):

$$|| \mathbf{X} - \hat{\mathbf{X}} ||^2_{I,I} = \Sigma_\alpha \Sigma_i (x_{\alpha i} - \hat{x}_{\alpha i}) \qquad (1)'$$

as a measure of goodness of fit.

The models (M0)~(MA) are classical ANOVA models. The model (MM) was first introduced by Fisher and Mackenzie[6] instead of an additive model (MA) and applied to the data of manurial response of potato varieties. A generalization of (MM):

$$\hat{x}_{\alpha i} = \sum_{k=1}^{K} d_k p_{k\alpha} q_{ki} \qquad (MSV)$$

was studied by Eckart and Young[4] and Householder and Young[10] and is known as the Eckart-Young decomposition in quantitative psychology. It is well known that the least squares(LS) solution which minimizes RSS is given by the largest K singular values and the corresponding eigenvectors of the raw data matrix $\mathbf{X}$. The ordination technique based on the solution of $K=2$ is called biplot.

Mandel[14] and Gollob[7], independently and possibly influenced by the work of Fisher and Mackenzie, developed a generalization of (MXA):

$$\hat{x}_{\alpha i} = t_\alpha + u_i + \sum_{k=1}^{K} d_k p_{k\alpha} q_{ki}. \qquad (MMA)$$

Generalizations of (MXC) and (MXV) are also possible. Statistical inference theory based on (MMA) or (MXA) under the normality assumption is summarized in Krishnaiah and Yochmowitz[13].

For each model, the LS solution of an additive part is easily obtained as the usual ANOVA solution for a two-way table without replication, and that of a multiplicative part is calculated from the singular value decomposition of the residual matrix whose element is given by $(x_{\alpha i} - (\hat{x}_{\alpha i}$ *by the additive part*$))$.

Okamoto[17] called the singular value decomposition(principal component analysis) based on (MXA), (MSV), (MXC) and (MXV) as *M* method, *N* method, *R* method and *Q* method, respectively.

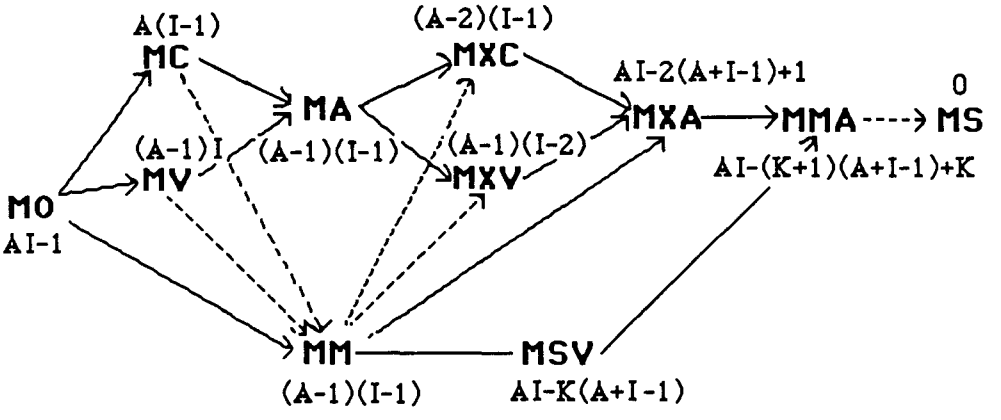

Fig.1 Hierarchy of models and degrees of freedom of RSS

When enough goodness of fit is not attained with a simpler model, we usually proceed to a more complex one with more parameters and, *vice verse*, when enough fit is attained, we proceed to a more parsimonious model for the easiness of interpretation. But it is not always easy to interpret complicated models such as (MXA) and (MMA) and the difference of degrees of freedom between models are large when *A* or *I* is large. (See Fig.1 for the degrees of freedom of RSS of each model.)

In the next section, simultaneous clustering is characterized giving an intermediate continuum, through two operations defined as follows, between (M0) and (MXA) (or (MS)), which would be more interpretable than the complex mixture model with more than one multiplicative term.

2.2. Two operations generating simultaneous clustering

In this paper, we confine ourselves to a relatively simple simultaneous clustering where clustering structure of cases does not depend on variables as shown in Fig.2a and exclude a "mosaic" block structure as shown in Fig.2b.

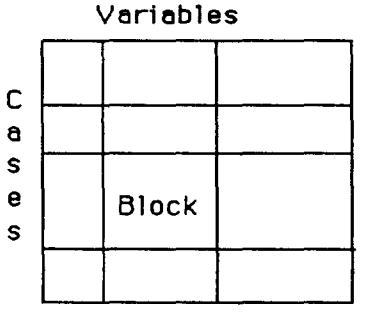

Fig.2a "Simple"structure

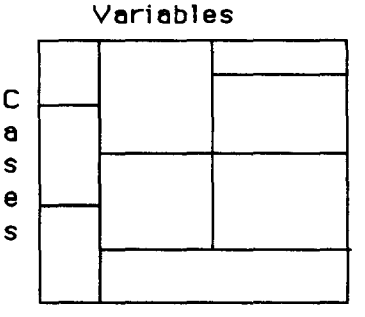

Fig.2b Complicated"mosaic"structure

Let $C(\cdot)$ be a function which gives the label of cluster to which each case or variable belongs and define $\alpha^*=C(\alpha)$ and $i^*=C(i)$. The label α^* varies from 1 to L and i^* varies from 1 to M where L and M are the number of clusters of cases and variables, respectively. We denote the parameters c, t_α, u_i, d, p_α, q_i and $v_{\alpha i}$ in (M0)~(MS) representatively by z. We define that z is global as to cases or variables when z does not include suffix α or i, respectively and define z is local otherwise.

Localization This operation is defined as "localizing" a

global parameters so that the value varies among clusters. We denote the localized parameters as to cases, variables and both by $z|_{\alpha^*}$, $z|_{i^*}$ and $z|_{\alpha^* i^*}$, respectively. For example, models $c|_{\alpha^*}$, $c|_{i^*}$ and $c|_{\alpha^* i^*}$ have L, M and LM parameters, respectively.

Merging This operation is defined as equating the value of local parameters within each cluster. We denote the merging operator and the merged parameters as to cases, variables and both by

$$z_{\alpha^*} = (z_\alpha)_{\cdot \alpha^*}, \quad z_{i^*} = (z_i)_{\cdot i^*}, \quad z_{\alpha^* i^*} = (z_{\alpha i})_{\cdot \alpha^* i^*},$$

respectively.

It is easy to show the following rules hold:

$$(z)|_{\alpha^* i^*} = (z|_{\alpha^*})|_{i^*} = (z|_{i^*})|_{\alpha^*}, \quad (z_{\alpha i})_{\cdot \alpha^* i^*} = (z_{\alpha^* i})_{\cdot i^*} = (z_{\alpha i^*})_{\cdot \alpha^*} \tag{3}$$

$$(z_1 + z_2)|_{X^*} = z_1|_{X^*} + z_2|_{X^*}, \quad (z_1 + z_2)_{\cdot X^*} = z_{1 \cdot X^*} + z_{2 \cdot X^*} \tag{4}$$

$$(z_1 \cdot z_2)|_{X^*} = z_1|_{X^*} \cdot z_2|_{X^*}, \quad (z_1 \cdot z_2)_{\cdot X^*} = z_{1 \cdot X^*} \cdot z_{2 \cdot X^*} \tag{5}$$

$$z_X|_{X^*} = z_X, \; z_{\cdot Y^*} = z \text{ if } z \text{ does not include suffix } Y \tag{6}$$

where X^* is α^*, i^* or $\alpha^* i^*$ and Y is α or i.

Restrictions to parameters for identifiability are modified as follows:

$$\text{localization:} \quad \Sigma z_Y = \text{const.} \rightarrow \Sigma z_Y|_{X^*} = \text{const.} \tag{7}$$

$$\text{merging:} \quad \Sigma z_X = \text{const.} \rightarrow \Sigma N(X^*) z_{X^*} = \text{const.} \tag{8}$$

where $N(\cdot)$ is a function which gives the size of each cluster.

Estimation of parameters under an localized model is easily carried out by treating each block (where both rows(cases) and columns(variables) belong to the same cluster) as a whole matrix and applying the LS calculation as described before independently to other blocks. For a

merged model, estimation can be carried out as follows:

•For (M0)~(MV) and the additive parts of (MXC) and (MXV), each parameter is estimated as the corresponding mean of merged cases or variables.

•For (MA) and the additive part of (MXA), t_{α^*} and u_{i^*} are estimated by the application of the ANOVA method for a two-way table with different replications.

•For the multiplicative part, a whole matrix is replaced by a $L \times M$ matrix whose element is given

$$(N(\alpha^*))^{1/2}(N(i^*))^{1/2}(\tilde{x}_{\alpha^* i^*} - (\hat{x}_{\alpha^* i^*} \textit{ by the additive part }))$$

where $\tilde{x}_{\alpha^* i^*}$ is a mean of the block, that is

$$\tilde{x}_{\alpha^* i^*} = \sum_{c(\alpha)=\alpha^*} \sum_{c(i)=i^*} x_{\alpha i} / (N(\alpha^*)N(i^*)).$$

Singular value decomposition is applied to the resulting matrix to yield the parameters d, $\tilde{p}_{\alpha^*}$ and $\tilde{q}_{i^*}$ and, finally, p_{α^*} and q_{i^*} are obtained by

$$p_{\alpha^*} = \tilde{p}_{\alpha^*} / N(\alpha^*)^{1/2}, \quad q_{i^*} = \tilde{q}_{i^*} / N(i^*)^{1/2}.$$

2.3. Model generation and relationship to existing methods

Intermediate models are generated by applying a localization operation to parsimonious models and, *vice verse*, applying a merging operation to non-parsimonious models, taking account of the rules (3)~(6) . One model can be derived by more than one different paths, for example,

$$v_{\alpha^* i^*} = c_{|\alpha^* i^*}, \qquad \text{(MHA)}$$
$$v_{\alpha^* i} = u_{i|\alpha^*}, \qquad \text{(MKM)}$$
$$v_{\alpha i^*} = t_{\alpha|i^*}.$$

It is worth noting some models which are bases of existing (simultaneous) clustering methods are derived by applying these two operations to a simple model. Hartigan's block clustering[8] and BMDP3M[3] can be interpreted as fitting the above (MHA) model. Sarle[18] assumes the same model and develops an iterative computer program for finding a non-hierarchical optimal solution. SAS VARCLUS[19] implicitly assumes the model

$$\hat{x}_{\alpha i} = u_i + d_{|i^*} p_{\alpha|i^*} q_i \qquad \text{(MVC)}$$

which is derived from (MXV) and a localization as to variables.

3. STRATEGIES OF DATA ANALYSIS

As usual clustering techniques of cases or variables based on an optimizing criterion, two strategies are possible for simultaneous clustering proposed here; that is, a hierarchical approach and non-hierarchical one. And, moreover, the former is dichotomized into an agglomerative one and a divisive one. As regards an non-hierarchical approach, technical difficulties such as determination of the number of clusters, remedies of avoiding local optimum solutions and development of rapid algorithms are multiplied in simultaneous clustering. At present, it seems to the author that the hierarchical approach with an interactive computer program is feasible and practical solution for exploring a data matrix which is possibly cluster-structured unless we have certain information on the number of clusters.

Merging operations which start from (MS), (MA) or (MM) are appropriate for an agglomerative approach which yields more parsimonious models, and localization operations are appropriate for an divisive one which seeks homogeneous blocks in which a simple model such as (M0), (MC) or (MV) fits well locally. The starting models for an agglomerative approach and the local models for an divisive approach should be selected, problem by problem, based on physical interpretability and goodness of fit of global models.

In the following, examples of strategies are briefly discussed using three data types.

Applicants×ratings Kendall, Stuart and Ord gives an example ([11],p.358) where 48 applicants are evaluated from score 0 to 10 on 15 items(see Appendix 1 for data values and names of items). The objective of analysis is grouping applicants together whose scoring patterns are similar and, at the same time, grouping items together which take similar values on each applicant. The model $v_{\alpha^* i^*} = c_{|\alpha^* i^*}$ is appropriate for this objective. Starting from the saturated model $v_{\alpha i}$, merging of two cases or two variables which yields the minimum increase of RSS can be carried out successively just as Ward's method(for example, Everitt[5]). Several modules are programmed for initialization, merging of two clusters, searching a specified number of pairs of clusters whose merging yield the smallest increases of RSS and printing of intermediate results. Mean square defined as

(*Increase of RSS/ difference of degrees of freedom*)

was used in order to determine whether cases or variables should be merged at some stage of agglomeration.

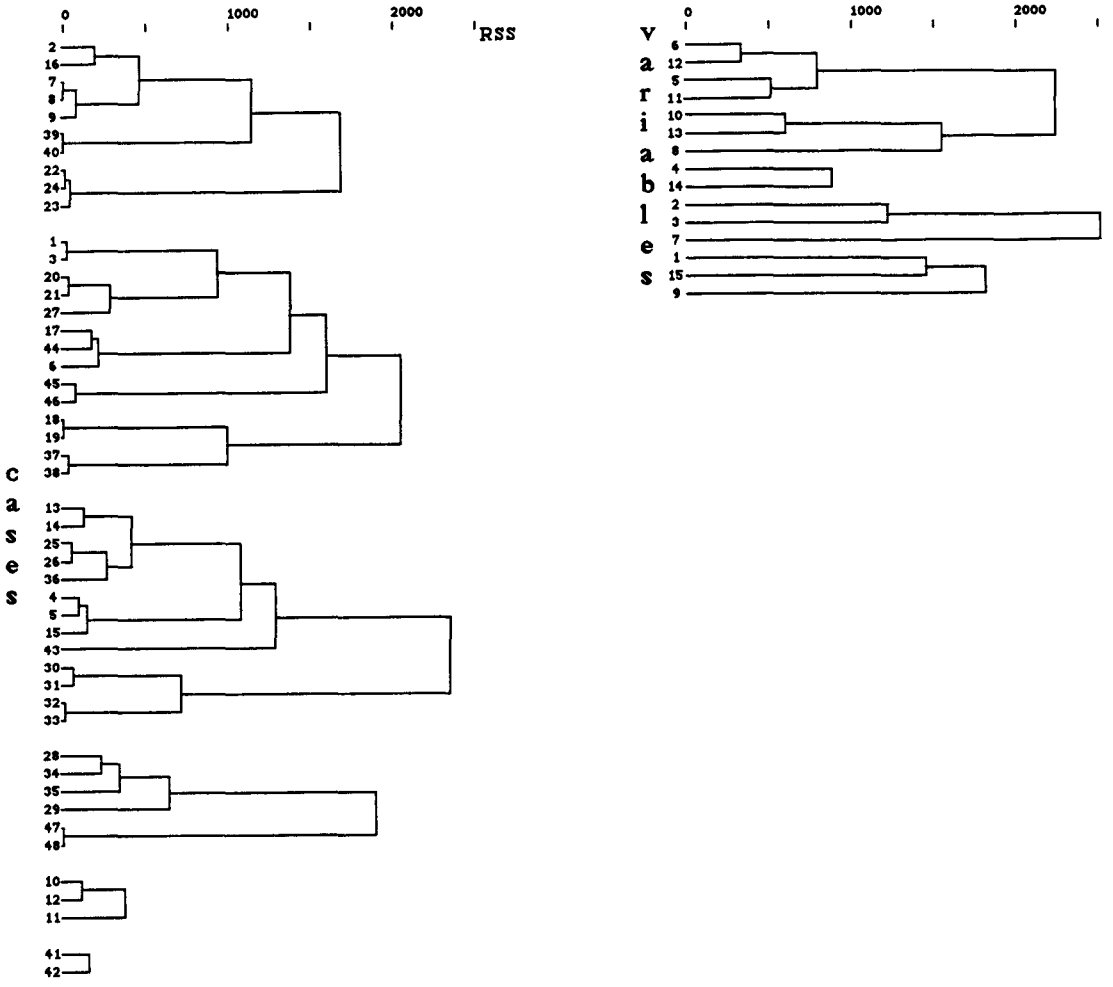

Fig.3 Dendrogram of simultaneous clustering (Applicants data)

A part of the dendrogram generated is shown in Fig.3 where the value of horizontal axis corresponds to RSS of the model at each stage. Kendall([12], pp.34-36) classifies variables by the simple inspection of the correlation matrix and forms the following clusters:

(X5,X6,X8,X10,X11,X12,X13 and X2 might perhaps be included)

(X1,X9,X15) (X4,X7,X14) (X3).

A main difference of Kendall's results and ours is that in our dendrogram X2(Appearance of applicants) and X3(Academic ability of applicants) merge at an early stage of agglomeration. A scatter plot of case-cluster means of X2 and X3 (at the stage just before the merging of X2 and X3) is shown in Fig.4. Except a rather outlying cluster consisting of cases 1 and 3 (applicants with good appearance and low academic ability) and case 43(with poor appearance and high academic ability), the means are close to each other although the correlation coefficient between X2 and X3 (which is calculated based on the deviation from the grand mean) is almost zero, which explains the above discrepancy.

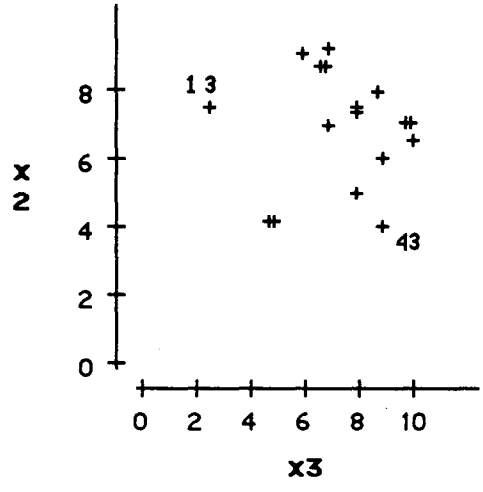

Fig.4 Scatter plot of X2 and X3 (Applicants data)

Table 1. Block means (Applicants data)

Variables	6 12 5 11 10 13 8	1 15 9	4 14	2 3 7
Cases				
1 3 20 21 27 17 44 6 45 46 18 19 37 38	6.9	5.1	6.1	7.7
13 14 25 26 36 4 5 15 43 30 31 32 33	4.5	4.7	6.4	7.4
2 16 7 8 9 39 40 22 24 23	8.9	8.4	8.7	8.3
28 34 35 29 47 48	1.7	1.4	3.1	6.2
10 12 11	9.4	4.9	2.3	7.6
41 42	0.7	10	0.0	4.7

Mean values of 24(=6×4) blocks which correspond to the final stage of the merging in Fig.3 is shown in Table 1. Cases are classified into 2 "middle-score" clusters, one "high-score" cluster, one "low-score" cluster and two outlying groups with a few cases. Variables are classified into 4 clusters: the first, one main cluster with 7 variables (Kendall named it an "ability to project am outlooking personality"); the second, a cluster of "experience" factors; the third, likability and keenness to join; the fourth, a mixture of appearance, academic ability and honesty. the application of usual Ward's method generated almost the same case-clustering as ours except that cases 1, 3, 6, 17 and 44 are classified into the "high-score" group and cases 30, 31, 32 and 33 are classified into the first "middle-score group". Ward's method applied to the transformed matrix generated

essential the same variable-clustering as Fig.3. Although it may be possible to derive almost the same interpretation as Table 1, combining the 2 "dual" results of Ward's method, our approach is more straightforward and more simple to carry out.

Individuals×measurements of sizes of various parts The famous Fisher's Iris data(for example, Kendall, Stuart and Ord[11], p.375) is this example. Here, individuals come from a mixture of several "natural" clusters and, at least conceptually, we can take as many measurements as we want from each individual, however, we can assume the existence of hypothetical factors (in Iris data, "size" of petal and "size" of sepal or "length" and "width") and each variable has high correlation with at least one of these factors. Therefore, adding measurements(variables) does not always contribute to the separation of "natural" clusters, and to make matters worse, adding measurements which have high correlation with one factor comes to neglect other factors possibly with high discriminating power.

One approach for finding natural clusters while taking account of such structure of variables is to estimate intra-cluster correlation structure of variables, as is mentioned in the Introduction. Another possible approach from the point of simultaneous clustering is to fit a model:

$$\hat{x}_{\alpha i} = u_{i|\alpha^*} + d_{|\alpha^* i^*} p_{\alpha|i^*} q_{i|\alpha^*} .$$

This model is a localization of (MVC) as to cases and assumes that cases in each cluster have one dimensional structure in each cluster of variables. After fitting the above model and finding out the intra-cluster correlation structure of

variables (and if it is similar among the clusters), it may be possible to calculate principal components from the pooled sum of squares matrix and use them for case-clustering.

When the numbers of clusters of cases and variables are fixed, the LS solution for the above model can be sought by alternating the minimization step as to cases (while fixing the cluster structure of variables) and that as to variables (while fixing the cluster structure of cases). The algorithm for the former step is developed by modifying the algorithm of k-means method and the latter step is carried out by modifying the VARCLUS algorithm. In this example, because the number of variables is small, a module was programed for seeking the LS solution for every possible clustering of variables. In this module, the increase (decrease) of RSS due to the deletion (inclusion) of one case from (to) the case-cluster is calculated by using the approximation of Appendix 3.

Table 2 is the results of the case-clustering where each variable is standardized to having unit variance and the number of case-clusters is fixed to 3. If both RSS and within-cluster sum of squares take small values, the resulting case-clustering would be well-separated in Euclidean norm and the resulting variable-clustering would be a parsimonious presentation of data, which is unfortunately not the case of ours. The results of simultaneous clustering, however, give us some insights. Table 3 is the cross-classification of original iris species and the result of the usual k-means case-clustering as well as the "best" solution for 2 variable-clusters ("1212" pattern; that is, variable 1 and 3 form one cluster and variable 2 and 4 form the other). The "misclassification" rate is much smaller in the "best" solution and the simultaneous clustering appears

to have succeeded rather well in estimating intra-(case-) cluster correlation structure as shown in Table 4.

Table 2 RSS for variable-clustering pattern (Iris data)

number of var-clusters	cluster pattern					RSS	within sum of squares
4	1	2	3	4	*1	0	138.9
3	1	2	3	3	*2	1.3	368.9
3	1	2	3	2		4.8	190.1
3	1	2	3	1		5.9	218.8
3	1	2	2	3		5.3	177.7
3	1	2	1	3		2.1	193.6
3	1	1	2	3		15.1	177.8
2	1	1	2	2		18.9	188.0
2	1	2	2	2		11.6	182.8
2	1	1	1	2		22.6	212.1
2	1	1	2	1		29.1	200.1
2	1	2	1	1		9.1	208.7
2	1	2	1	2		8.4	185.7
2	1	2	2	1		14.9	165.7
1	1	1	1	1		33.3	211.9

*1: Results of *k*-means clustering
*2: Variable 1(sepal length) forms the first cluster, variable 2(Sepal width) forms the second one and variables 3 and 4(petal length and petal width) form the third one

Table 3 Misclassification of cases (Iris data)

	k-means			"1212" clustering		
cluster number	1	2	3	1	2	3
Setosa	50	0	0	50	0	0
Versicolor	0	11	39	0	3	47
Versinica	0	36	14	0	45	5

Table 4 Intra-cluster correlation matrix (Iris data)

Setosa (cluster 1 of "1212")

1	.743	.267	.278
	1	.178	.233
		1	.332

Versicolor

1	.526	.754	.546
	1	.561	.664
		1	.787

Cluster 3 of "1212"

1	.718	.870	.732
	1	.760	.850
		1	.871

Versinica

1	.457	.864	.281
	1	.401	.538
		1	.322

Cluster 2 of "1212"

1	.320	.829	.386
	1	.339	.713
		1	.442

In passing, if a special type of factor analysis models in Appendix 4 and the normality assumption can be postulated, it is possible to measure the deviation of data from the model and do some inferences on the number of variable-clusters. In the above "1212" clustering, χ^2's of one common factor model (df=2) in each cluster are 4.08, 27.2 and 22.9, respectively; χ^2's of two common factor model (df=1) in the latter two clusters are 0.2 and 5.4, which shows two factor (2 variable-clusters)model fits rather well to each case-cluster, although we cannot interpret χ^2 in a usual way because the optimization is carried out in allocating cases.

Varieties×manurial treatments The data of Fisher and Mackenzie[6] mentioned earlier, to which the multiplicative model (MM) fits better than both the additive model (MA) and the additive model after log-transformation(Ohashi[16]) and, at the same time, gives a physically meaningful interpretation, is an example of this type of data (see Appendix 2 for data values). Merging operations starting

from the well-fitted multiplicative model give us more parsimonious interpretation of response pattern and necessary calculation steps are straightforward if we confine ourselves to the hierarchical clustering. In this case, the degrees of freedom due to merging of case-clusters and variable-clusters are the same (that is, 1). Some sort of "stopping rules" may be invented by comparing the increase of RSS and RSS of the initial model.

Several modules which have the same functions as those in the first example are programmed and a part of dendrogram generated is shown in Fig.5 where the value of the horizontal axis corresponds to RSS of each step.

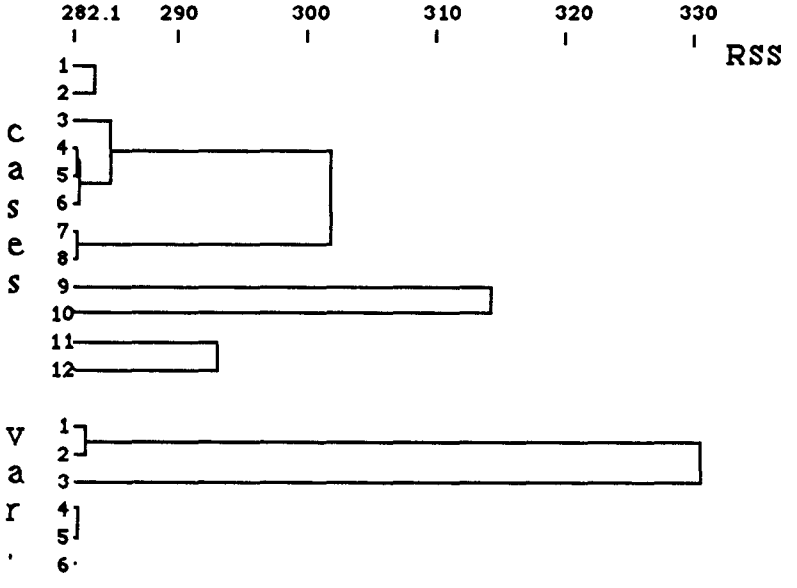

Fig.5 Dendrogram of simultaneous clustering (Potato data)

The mean square of the initial model is 282.117/55=5.129, and an approximate F value for merging is calculated by dividing the increase of RSS by this mean square. If we

stop merging when the F value exceeds 10% point, the resulting case-clusters are

(1 2) (3 4 5 6 7 8) (9 10) (11 12),

and variable-clusters are

(1 2) (3) (4 5) (6);

that is, there is no "significant" difference between the sulphate row and the chloride row. Predicted values and residuals from the merged model (standardized by dividing the standard error $\sqrt{5.129}=2.26$) are shown in Table 5. There seems neither peculiar patterns nor outliers.

Table 5. Predicted values and residuals (Potato data)

Predicted value

variables(manurial tr.) cases(potato)	1 2	3	4 5	6
1 2	26.5	24.7	20.3	6.2
3 4 5 6 7 8	21.5	20.0	16.5	5.0
9 10	16.6	15.4	12.7	3.9
11 12	12.5	11.6	9.6	2.9

Standardized residuals

variables(manurial tr.) cases(potato)	1	2	3	4	5	6
1	-0.5	-0.2	0.8	1.2	-0.8	1.5
2	0.7	0.2	-0.4	0.0	-1.5	0.1
3	0.8	1.3	-2.6	0.8	1.9	-0.0
4	-0.7	-1.1	0.0	1.6	0.7	1.2
5	0.6	-0.4	0.0	-0.3	0.4	-0.3
6	-0.3	1.3	0.8	-0.3	-0.9	-1.2
7	0.4	-2.1	0.8	-1.7	1.4	-0.4
8	0.2	-0.3	0.3	-1.6	-1.2	0.7
9	0.8	1.6	0.3	-0.4	0.1	-1.0
10	-0.8	-0.4	-0.5	-0.1	-0.3	-0.8
11	0.6	-0.7	-0.2	1.3	1.4	-0.3
12	-1.1	-0.3	0.8	-0.6	-0.6	-0.6

4. NOTES AND FUTURE PROBLEMS

The following are notes and future problems of the proposed simultaneous clustering procedure.

(1)Efficient algorithms

In the hierarchical simultaneous merging operation illustrated in the first example of the previous section, a main computational difficulty is that the "distances" between case-clusters (variable-clusters) should be recalculated from the original data matrix after every merging of variable-clusters (case-clusters), while in the usual case-clustering by Ward's method we need not refer to the original data once we have calculated the distance of every pair of cases. It is not clear to the author whether some inventions may be possible for reducing this computational burden.

(2)Fuzzy clustering

The fuzzy k-means clustering (Bezdek[2]) is sometimes a useful pattern recognition technique for finding out an overlapping cluster structure. A generalization for the simultaneous clustering is rather straight forward, although it is difficult to justify it theoretically. A "solution" of the fuzzy variable-clustering is formally comparable to those of classical principal component analysis and factor analysis. The comparison of both solution may be interesting.

(3)Determination of the number of clusters

For case-clustering, many criteria have been proposed, based on the normal-theory likelihood (for example, Milligan and cooper[15]). These criteria seem to work well, if the data follow the assumption postulated, especially when there is no outlier. There seems no criterion proposed for variable-clustering even for this ideal situation. A model

in Appendix 4 is an attempt for evaluating the number of variable-clusters, although an exact distribution theory for the likelihood is formidable because of the optimizing procedure involved in the clustering.

REFERENCES

[1]Art, D., Gnanadesikan, R. and Kettenring, J.R.(1982). "Data-based metrics for cluster analysis", Utilitas Mathematica, 21A, 75-79.

[2]Bezdek, J.C.(1981). Pattern Recognition with Fuzzy Objective Algorithms, Plenum Press.

[3]Dixon, W.J. et al. eds.(1985). BMDP Statistical Software Manual, University of California Press.

[4]Eckart, C. and Young, G.(1939). "A principal axis transformation for non-Hermitian matrices", Bull.Amer.Math.Soc., 45,118-121.

[5]Everitt, B.(1979). "Unresolved problems in cluster analysis", Biometrics, 35, 169-181.

[6]Fisher, R.A. and Mackenzie, W.A.(1923). "Studies in crop variation II: The manurial response of different potato varieties", J.Agricultural Sci., 13, 311-320.

[7]Gollob, H.F.(1968). "A statistical model which combines features of factor analytic and analysis of variance techniques", Psychometrika, 33, 73-115.

[8]Hartigan, J.A.(1975). Clustering Algorithms, Wiley.

[9]Hartigan, J.A.(1982). "Classification" IN:Encyclopedia of Statistical Sciences (Kotz, S. and Johnson, N.L. eds.), vol.2,1-10, Wiley.

[10]Householder, A.S. and Young, G.(1938). "Matrix approximation and latent roots", Amer.Math.Monthly, 45, 165-171.

[11]Kendall, M.G., Stuart, A. and Ord, J.K.(1983). The Advanced Theory of Statistics, vol.3, fourth ed., Griffin.

[12]Kendall, M.G.(1981). Multivariate Analysis, second ed.,Griffin.

[13]Krishnaiah, P.R. and Yochmowith, M.G.(1980). "Inference on the structure of interaction in two-way classification model",IN:Handbook of Statistics (Krishnaiah, P.R.ed.), vol.1, 973-994, North-Holland.

[14]Mandel, J.(1969). "The partitioning of interaction in analysis of variance", J.Res.National Bureau of Stand., B, 73B, 309-328.

[15]Milligan, G.W. and Cooper, M.C.(1985). "An examination of procedures for determining the number of clusters in a data set", Psychometrica, 50, 159-179.

[16]Ohashi, Y.(1982). (in Japanese) "A multiplicative model", Kokyuroku, 526, 13-45.
[17]Okamoto, M.(1972). "Four techniques of principal component analysis", J.Japan Statist.Soc., 2, 63-69.
[18]Sarle, W. (1982). "Cluster analysis of least squares" Proc.7th SAS User's Group International, 651-653.
[19]SAS Institute Inc.(1985) SAS User's Guide: Statistics, Version 5 edition, SAS Inc.
[20]SAS Institute Inc.(1985) SAS/IML User's Guide Version 5 edition, SAS Inc.

Appendix 1 Applicants data

15 ratings(variables)

X1: Form of letter of application
X2: Appearance
X3: academic ability
X4: Likeability
X5: Self-confidence
X6: Lucidity
X7: Honesty
X8: Salesmanship
X9: Experience
X10:Drive
X11:ambition
X12:Grasp
X13:Potential
X14:Keenness to join
X15:Suitability

48 applicants(cases)

CASE_ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
1	6	7	2	5	8	7	8	8	3	8	9	7	5	7	10
2	9	10	5	8	10	9	9	10	5	9	9	8	8	8	10
3	7	8	3	6	9	8	9	7	4	9	9	8	6	8	10
4	5	6	8	5	6	5	9	2	8	4	5	8	7	6	5
5	6	8	8	8	4	4	9	2	8	5	5	8	8	7	7
6	7	7	7	6	8	7	10	5	9	6	5	8	6	6	6
7	9	9	8	8	8	8	8	8	10	8	10	8	9	8	10
8	9	9	9	8	9	9	8	8	10	9	10	9	9	9	10
9	9	9	7	8	8	8	8	5	9	8	9	8	8	8	10
10	4	7	10	2	10	10	7	10	3	10	10	10	9	3	10
11	4	7	10	0	10	8	3	9	5	9	10	8	10	2	5
12	4	7	10	4	10	10	7	8	2	8	8	10	10	3	7
13	6	9	8	10	5	4	9	4	4	4	5	4	7	6	8
14	8	9	8	9	6	3	8	2	5	2	6	6	7	5	6
15	4	8	8	7	5	4	10	2	7	5	3	6	6	4	6
16	6	9	6	7	8	9	8	9	8	8	7	6	8	6	10
17	8	7	7	7	9	5	8	6	6	7	8	6	6	7	8
18	6	8	8	4	8	8	6	4	3	3	6	7	2	6	4
19	6	7	8	4	7	8	5	4	4	2	6	8	3	5	4
20	4	8	7	8	8	9	10	5	2	6	7	9	8	8	9
21	3	8	6	8	8	8	10	5	3	6	7	8	8	5	8
22	9	8	7	8	9	10	10	10	3	10	8	10	8	10	8
23	7	10	7	9	9	9	10	10	3	9	9	10	9	10	8
24	9	8	7	10	8	10	10	10	2	9	7	9	9	10	8
25	6	9	7	7	4	5	9	3	2	4	4	4	4	5	4
26	7	8	7	8	5	4	8	2	3	4	5	6	5	5	6
27	2	10	7	9	8	9	10	5	3	5	6	7	6	4	5
28	6	3	5	3	5	3	5	0	0	3	3	0	0	5	0
29	4	3	4	3	3	0	0	0	0	4	4	0	0	5	0
30	4	6	5	6	9	4	10	3	1	3	3	2	2	7	3
31	5	5	4	7	8	4	10	3	2	5	5	3	4	8	3
32	3	3	5	7	7	9	10	3	2	5	3	7	5	5	2
33	2	3	5	7	7	9	10	3	2	2	3	6	4	5	2
34	3	4	6	4	3	3	8	1	1	3	3	3	2	5	2
35	6	7	4	3	3	0	9	0	1	0	2	3	1	5	3
36	9	8	5	5	6	6	8	2	2	2	4	5	6	6	3
37	4	9	6	4	10	8	8	9	1	3	9	7	5	3	2
38	4	9	6	6	9	9	7	9	1	2	10	8	5	5	2
39	10	6	9	10	9	10	10	10	10	10	8	10	10	10	10
40	10	6	9	10	9	10	10	10	10	10	10	10	10	10	10
41	10	7	8	0	2	1	2	0	10	2	0	3	0	0	10
42	10	3	8	0	1	1	0	0	10	0	0	0	0	0	10
43	3	4	9	8	2	4	5	3	6	2	1	3	3	3	8
44	7	7	7	6	9	8	8	6	8	8	10	8	8	6	5
45	9	6	10	9	7	7	10	2	1	5	5	7	8	4	5
46	9	8	10	10	7	9	10	3	1	5	7	9	9	4	4
47	0	7	10	3	5	0	10	0	0	2	2	0	0	0	0
48	0	6	10	1	5	0	10	0	0	2	2	0	0	0	0

Appendix 2 Potato data

6 manurial treatments(variables)

T1: Dunged series: sulphate row
T2: : chloride row
T3 : basal row
T4: undunged series: sulphate row
T5: : chloride row
T6: : basal row

12 potato varieties(cases)

VARIETY	T1	T2	T3	T4	T5	T6
1	25.3	26.0	26.5	23.0	18.5	9.5
2	28.0	27.0	23.8	20.4	17.0	6.5
3	23.3	24.4	14.2	18.2	20.8	4.9
4	20.0	19.0	20.0	20.2	18.1	7.7
5	22.9	20.6	20.1	15.8	17.5	4.4
6	20.8	24.4	21.8	15.8	14.4	2.3
7	22.3	16.8	21.7	12.7	19.6	4.2
8	21.9	20.9	20.6	12.8	13.7	6.6
9	18.3	20.3	16.0	11.8	13.0	1.6
10	14.7	15.6	14.3	12.5	12.0	2.2
11	13.8	11.0	11.1	12.5	12.7	2.2
12	10.0	11.8	13.3	8.2	8.3	1.6

Appendix 3 Approximation of the change of RSS of a localized (MVC) model

Let the sum of squares matrix within a case-cluster (measured from the cluster-mean values) and the cluster-mean vector be $\mathbf{S}$ and $\mathbf{x}$, respectively. The contribution of each block (case-variable-cluster) to RSS of the localized version of the model (MVC) is given by

$$\text{trace}(\mathbf{S}) - (\text{the first eigen value of } \mathbf{S}). \tag{A1}$$

After a simple algebra, the change of $\mathbf{S}$ due to the deletion of one case with the measurement vector $\mathbf{x}$ is shown to be

$$\Delta_{-}\mathbf{S} = -n/(n-1)(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})', \tag{A2}$$

where n is the number of the cases in the case-cluster (before the deletion). If we denote the first eigenvalue and the corresponding eigenvector (standardized to norm 1) by λ and $\mathbf{v}$, respectively, and denote their changes due to the deletion by $\Delta_{-}\lambda$ and $\Delta_{-}\mathbf{v}$, respectively,

$$\mathbf{S}\mathbf{v} = \lambda\mathbf{v}, \tag{A3}$$

$$(\mathbf{S}+\Delta_{-}\mathbf{S})(\mathbf{v}+\Delta_{-}\mathbf{v}) = (\lambda+\Delta_{-}\lambda)(\mathbf{v}+\Delta_{-}\mathbf{v}) \tag{A3}'$$

by definition. By subtracting (A3) from (A3)', multiplying $\mathbf{v}'$ from the left and ignoring the second order terms (and noting $\mathbf{v}'(\Delta_{-}\mathbf{v})=0$), it is shown that

$$\Delta_{-}\lambda \cong \mathbf{v}'(\Delta_{-}\mathbf{S})\mathbf{v}. \tag{A4}$$

In the same way, the change of the first eigenvalue ($\Delta_{+}\lambda$) due to the inclusion of a case to a cluster is approximated by the change of $\mathbf{S}$ ($\Delta_{+}\mathbf{S}$) as follows:

$$\Delta_+S = n'/(n'+1)(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})' , \qquad (A2)'$$
$$\Delta_+\lambda \cong \mathbf{v}'(\Delta_+S)\mathbf{v}, \qquad (A4)'$$

where n' is the number of cases before the inclusion.

Appendix 4 A factor analysis model for (MVC)

Let $\mathbf{x}$ be a $p\times 1$ measurement vector (rearranged according to variable-clusters) and $\mathbf{f}$ be a hypothetical $m\times 1$ common factor vector where m is a number of variable-clusters. As a basis for the (MVC) (or the VARCLUS procedure), we can assume the following structural model:

$$\mathbf{x} = \mathbf{c} + \Lambda\mathbf{f} + \boldsymbol{\epsilon}. \qquad (A5)$$

In the above model, $\mathbf{c}$ is a constant vector, $\boldsymbol{\epsilon}$ is a $p\times 1$ mutually uncorrelated error vector (and also uncorrelated with $\mathbf{f}$), and Λ is a $p\times m$ matrix of factor loadings with 0's in the off-diagonal part, that is,

$$\Lambda = \begin{pmatrix} \Lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Lambda_m \end{pmatrix} . \qquad (A6)$$

where Λ_i is a $p_i\times 1$ vector $(\Sigma p_i=p)$.

In the usual factor analysis model, each element of $\mathbf{f}$ is assumed to be uncorrelated to one another for the identifiability of the model, but in the model (MVC) (the VARCLUS procedure) each element may correlate with one another. In the following we assume

$$E(\mathbf{f}\mathbf{f}') = \Phi \qquad (A7)$$

without any restriction on off-diagonal elements if Φ is positive definite. (Without loss of generality, we can assume the mean and the variance of each element of $\mathbf{f}$ is 0 and 1, respectively.) From (A6) and (A7), the variance structure of $\mathbf{x}$ is represented by

$$\Sigma \equiv V(\mathbf{x}) = \Lambda\Phi\Lambda' + \Psi, \tag{A8}$$

where Ψ is a diagonal matrix of the variances of the elements of ϵ. If $\mathbf{f}$ and ϵ follow the normal distribution, the maximum likelihood estimates of the parameters Λ, Φ and Ψ which maximize the log-likelihood based on n cases

$$\text{log.lik.} = n/2 \cdot \log|\Sigma^{-1}S| - n/2 \cdot \text{trace}(\Sigma^{-1}S) + pn/2 \tag{A9}$$

are calculated numerically and the goodness of fit of the model can be assessed by comparing the value of the maximized log-likelihood with χ^2 distribution. The degrees of freedom of χ^2 is

$$(p^2 - 3p - m^2 + m)/2. \tag{A10}$$

A numerical optimization may be possible via Newton-Raphson method, but the EM-algorithm was used in the example of the Section 3 because the calculation of the second derivative of (A9) by the parameters is cumbersome. (A computer program for this EM-algorithm is available upon the request to the author.)

The results of a small scale of simulation study using artificial data show that the EM-algorithm is stable and the χ^2 approximation of the maximized log-likelihood is good when sample size is not so small, say over 50 (although the speed of convergence is slow) if the data follow the model and the specification of the positions of 0's in Λ is correct.