Bulletin
of
The Japanese Classification Society

Volume 3
December 1989
JAPAN

Contents

Conversational Selection of Variables in Logistic Regression Analysis

Masakazu Ando and Toshiro Haga
Science University of Tokyo
1-3, Kagurazaka, Shinjyuku-ku
Tokyo Japan

1. Introduction

These days logistic regression analysis is widely used principally in biostatistics. Many program packages for this method have been developed, but there are few program packages that have the function of selection of variables in spite of selection of variables being common in ordinal regression analysis and discriminant analysis. For example, SAS (Statistical Analysis System) has three procedures, PROBIT, LOGIST, CATMOD for logistic regression analysis, but it is only LOGIST PROCEDURE that has this function. In LOGIST PROCEDURE, variables are selected automatically, according to an assigned α level. However it is necessary to take technical judgments for selection of variables. CDA (Conversational Data Analysis) has a function of conversational selection of variables in regression analysis and discriminant analysis.

Following such a notion of CDA, we tried to include a program of logistic regression with the function of conversational selection of variables. To verify this program, the output of this program is compared with that of LOGIST, and PROBIT PROCEDURE of SAS.

2. Criterion Index for Selection of Variables and Items

In ordinal regression analysis an F-value that evaluates the amount of change of residual sum of squares is used as the criterion index for selection of variables. Corresponding to the F-value, log-likelihood ratio statistic is used in logistic regression analysis. To compute log-likelihood ratio statistic, it is necessary to create two models and compute their log-likelihood, therefore a large amount of computation is required. Score statistic and Wald's statistic are approximate values of log-likelihood ratio statistic, and the former is used for forward selection of variables, the latter is used for backward elimination. These two statistics need less computation than log-likelihood ratio statistic does. Actually these two statistics are used for the criterion index of selection of variables in LOGIST and we use these two statistics in CDA.

3. Index for evaluation of model

In ordinal regression analysis, the coefficient of determination (R^2) or that adjusted for the degree of freedom (R^{*2}) is used for index of evaluation of fitness of the model. R^2 is a ratio of regression sum of squares to total sum of squares. Let the ratio of residual sum of squares to total sum of squares be the error ratio, coefficient of determination is given by (l - error ratio). In logistic regression, let log-likelihood of the model only intercept be $\ln L_0$ as total sum of squares S_T and log-likelihood of current model be lnL* as residual sum of squares Se. The coefficient of determination(R_L^2) can be computed by log-likelihood statistic. It is shown in Equation (1).

$$R_L^2 = 1 - \frac{\ln L^*}{\ln L_0} = \frac{-2\,(\ln L_0 - \ln L^*)}{-1 \ln L_0} = \frac{\text{log-likelihood ratio}}{-2 \ln L_0} \quad (1)$$

LOGIST gives the statistic of Equation (1) after subtracting two times of number of variables in the model from the numerator with the view of AIC. This statistic corresponds to coefficient of determination doubly adjusted for the degree of freedom. Then CDA gives the statistic of Equation (1) and coefficient of determination doubly adjusted for the degree of freedom for the index; of fitness of model. As the fitness of the model can't be evaluated from only coefficient of determination, we are trying to include the function of such as regression diagnostic.

4. Selection of Categorical Variables

If the independent variable is categorical variable(item), logistic regression can be applied after conversion of the categorical variable to a set of dummy variables in a similar way to ordinal regression analysis(the first method of Hayashi's quantification theory). In CDA, dummy variables are generated by the program.

5.Example of output

(1) data

An example of the manual of LOGIST PROCEDURE is used to explain the process of conversational selection of variables in logistic regression analysis.

(2) Specifying the number of times of iteration

By specifying "A" at initial menu, logistic regression analysis starts.

>List, Fund.-stat, Graph, Monit, Analysis, or End

A

The fundamental policy of this program is conversational analysis, so the speed of response must be quick. The amount of computation for addition or elimination of variable in ordinary regression analysis is order of $n*p^2$(p is the number of independent variables), but in logistic regression analysis it is order of $n*p^2$(n is sample size) because it involves a computation of weighted sum of squares and crossproducts matrix. If the number of iteration times is k, then the order of computation becomes$n*p^2*k$. Then, by specifying the number of iteration times in the renewal of the model, the approximate value can be obtained in a short time. For example specifying A 3, the maximum of the iteration times is 3. If it is not specified then default value is used.

(3) Initial display. The initial display is shown in Fig.1

```
> List, Fund.-stat, Graph, .Monit, Analysis, or End
A
Var.          L.R.      D (L.R.)      D          b
0 CON       –2.883      –2.883      0.090      –0.693
1    x1      1.889       1.889      0.169
2    x2      1.074       1.074      0.300
3    x3      1.882       1.882      0.170
4    x4      7.931       7.931      0.005
5    x5      3.526       3.526      0.060
6    x6      0.659       0.659      0.417
```

Fig.1 Initial display

L.R. is the log-likelihood ratio statistic after including a variable or dummy variables generated from an item. D(L.R.) is the difference of L.R. before and after including a variable or an item and it is Score statistic or Wald's statistic. The column P is upper probability of chi-square distribution. B is the regression coefficient.

(4) Selection of variables

In Fig.1, p-value (0.005) of x_4 is the smallest, so x_4 is added to the model by specifying S4.Likelihood ratio statistic, two coefficients of determination and regression coefficient are given.

```
R> Sel., eXact., Res., E
S4
Enter        L.R.      R^2          R**^2
4    x4      8.089       0.235      0.177
Var.         L.R.    D (L.R.)       D          b
0 CON       –0.721     –8.813       0.003     –3.724
1    xl      9.049       0.960      0.327
2    x2      8.207       0.118      0.731
3    x3      8.587       0.498      0.480
4    x4      1.332      –6.757      0.009      2.859
5    x5      8.171       0.082      0.774
6    x6      9.124       1.034      0.309
```

The value of log-likelihood ratio statistic was expected to be 7.931 in Fig.1, but actually it is 8.089. This can be attributed to the fact that Wald's statistic is an approximation. If a p-value c selection is close to 0.3, the exact solution can be obtained by more iteration specifying X.

```
R> Sel., eXact., Res., E
X
Exact        L.R.      R^2          R**^2
             8.299       0.241      0.183
Var.         L.R.    D (L.R.)       p          b
0 CON        0.792     –7.506       0.006     –3.777
1    xl      9.417       1.118      0.290
2    x2      8.436       0.137      0.711
3    x3      8.870       0.571      0.450
4    x4      2.339      –5.959      0.015      2.897
5    x5      8.392       0.093      0.760
6    x6      9.558       1.259      0.262
```

The p-value of x_6 is less than 0.3, x_6 is added to the model.

```
R> Sel., eXact., Res., E
S6
Enter        L.R.      R^2          R**^2
6    x6      9.623       0.280      0.164
Var.         L.R.    D (L.R.)       p          b
O CON        8.437     –1.187       0.276      46.676
1    x1     10.969       1.346      0.246
2    x2      9.778       0.155      0.691
3    x3     10.372       0.749      0.387
4    x4      2.703      –6.920      0.009      3.263
5    x5     10.610       0.986      0.321
6    x6      8.250     –1.373       0.241     –51.195
```

```
R> Sel., eXact., Res., E
S1
Enter          L.R.        R^2          R**^2
1     x1      11.525       0.335        0.161
Var.           L.R.      D (L.R.)       p            b
O CON          9.978      –1.547        0.214       58.098
1     x1       9.326      –2.200        0.138        6.777
2     x2      11.566       0.041        0.840
3     x3      11.578       0.053        0.818
4     x4       5.162      –6.363        0.012        3.431
5     x5      11.584       0.059        0.809
6     x6       9.507      –2.018        0.155      –69.196
```

It seems that there are no significant variables, and if X is specified then exact regression coefficient is computed.

(5) Selection of items

The selection of items is similar to that of variables.

6. Conclusion

To apply this program for practical use, functions of regression diagnosis and its graphical outputs are required.

References

(1) Lawless,J.F. and Singhal,K.(1978), "Efficient Screening Nonnormal Regression Models", Biometrics 34, 318-327.
(2) SAS Institute Inc.(1985), SAS User's Guide: Supplementary, Version 5 Edition, Cary, North Carolina.
(3) Haga,T.(1984), "Conversational data analysis system", Japanese Journal of Applied Statistics vol.13 No.3, 125-138.

Evaluation Procedure for Space-Distorting Properties in Agglomerative Hierarchical Clustering Algorithms

Nagatomo Nakamura
Department of Construction
Junior College
Nihon University
1-24-7, Narashino-dai, Funabashi-shi
Chiba Japan

Noboru Ohsumi
Department of Statistical Methodology
The Institute of Statistical Mathematics
7-6-4, Minami-Azabu, Minato-ku
Tokyo Japan

The results of hierarchical classification often displays as dendrogram. However the distances (dissimilarities) on the dendrogram is not in accord with the original distances (dissimilarities). The differences between these two distortion measures were suggested by many researchers (for a review, see Gordon [1987] and Cormack [1971]), but these distortion measures do not indicate degree of space contraction or space dilation.
The objective of this report is to define the measure that represents the degree of space distortion (space contraction, space conservation, and space dilation), and to order some method containing the Lance and Williams combinatorial method using the distortion measure.

The combinatorial method was proposed by Lance and Williams (1967). At the same time, they introduced the concepts of space distortion occurring in the clustering process. However, their discussion was an experimental trial based on data and was not sufficiently proved by mathematical examination. The mathematical conditions for space distortion were clarified by Ohsumi and Nakamura (1989) using between cluster-distances related to the merging process. Moreover, they evaluated a space distortion conditions for some method containing a combinatorial method. Special attentions to the conditions, the boundary conditions of space contraction and dilation equivalent to the single linkage and complete linkage methods respectively.

Definition 1 Distortion measure

Suppose that d_{ij} is distance (dissimilarity) matrix and h_{ij} is distance matrix on the dendrogram, then the distortion measure is

$$M_{\mathrm{n}} = \frac{\sum h_{ij}}{\sum d_{ij}} \quad (i>j).$$

where the symbol “n” means abbreviated method name.

The greater (smaller) value of M_{n} tend to space dilation (contraction respectively).

Definition 2 Degree of distortion (standardized index)

If M_{n} was defined on the above, then the distortion degree is

$$DD_{\mathrm{n}} = \frac{M_{\mathrm{n}} - M_{SL}}{M_{CL} - M_{SL}}$$

where *SL* and *CL* indicate single linkage and complete linkage methods respectively.

Standardization of this distortion measure is indicated the following boundary conditions:

$0 \geq DD_{\mathrm{n}}$: space contraction
$1 > DD_{\mathrm{n}} > 0$: space conservation
$DD_{\mathrm{n}} \geq 1$: space dilation

Based on the above definitions, we performed a numerical examination. The results of the method order can be expressed as follows:

Single linkage method << Centroid method << Group average method<<
Median method << Weighted average method <<
Complete linkage method << Flexible method << Ward's method.

Where the symbol "<<" indicates the ordering of distortion (from contraction to dilation)

References

(1) Cormack, R. M. (1971), "A Review of Classification," Journal of the Royal Statistical Society, Series A, 134, 321-367.
(2)Gordon, A. D. (1987), "A Review of Hierarchical Classification," Journal of the Royal Statistical Society, Series A, 150, 119-137.
(3)Lance, G.N. and Williams, W. T. (1967), "A General Theory of Classificatory Sorting Strategies, I. Hierarchical Systems," The Computer Journal, 9, 373-380.
(4) Ohsumi, N. and Nakamura, N. (1989), "Space-Distorting Properties in Agglomerative Hierarchical Clustering Algorithms." in Data Analysis -Learning Symbolic and Numeric Knowledge- , 47th ISI Satellite Meeting Session.

Analysis of Eating Habit and Eating Disorder of Female by The Third Method of Quantification

Yoko Yoshinaga[1], Koujirou Shukutani[2], Shinobu Tatsunami[3], Nagasumi Yago[3] and Yukio Yamamura[1]

[1]Department of Public Health, [2]Department of Neuropsychiatry and [3]Radioisotope Research Institute, St.Marianna University School of Medicine

In an attempt to define possible risk factors for eating disorders such as anorexia nervosa and bulimia nervosa in young female adults, a questionnaire survey was performed on a volunteer group of a total 171 female students (age ranged 18-25 years old with the mean of 20.3 + 1.2 years old) from three nursing schools in Kawasaki City. The questionnaire consisted of a total 54 questions on their eating habit as well as on their physical images about their own bodies. By transforming the answers of "yes" or "no" into "1" or "2", respectively, a data file was constructed in a FACOM M130F computer, and was analyzed by the third method of quantification.

1) Sixty % of the students whose body weights were with in the age- and sex-matched Japanese standard +/– 10 %, considered themselves as being fat. Thirty-four % of those whose body weights were within the standard –10%, also considered themselves as being fat.
2) Fifty four % of the students maintained the standard eating habit, i.e. three meals a day. Those who usually have night snacks or extra foods amounted to 87% of the total.
3) The third method of quantification revealed the presence of four distinct groups of students according to their eating habits.
4) Between the four distinct groups, there was a directional order from "normal, healthy" to "eating disorders", and that was termed provisionally the ordered structure of eating habit.
5) The major factors for Axis I were tendency of bulimia and regret for overeating, while those for Axis II tendency of self-restriction of meals and vomiting.

These results clearly indicated that the most influential risk factors for eating disorders should include impulse to overeating, actual overeating, regret for it, vomiting and self-restriction of meals.

The distorted physical image of young female adults was discussed to result in their eating habit which finally lead to eating disorders.

Numerical Classification of EEG Spectra

Takenobu Tasaki, Masami Zaizen, and Masashi Goto

Shionogi Kaiseki Center, 1-22-41, Izumicho
Suita City, Osaka 564, Japan

For study of possible effects of drugs on electroencephalograms(EEG's) it is common to compare two long (for example, six hours) EEG traces recorded when a drug under the study and its vehicle are administrated to a same animal. Researchers of EEG usually look at a number of tracing papers and classify each of them into one of several categories based on his professional knowledge and skill of pattern recognition. The categories may be awaking, resting, slow-wave-light-sleeping, slow-wave-deep-sleeping and fast-wave-sleeping (or paradoxical sleeping). Then the researchers draw a sleep-awake-cycle diagram for thesequence of classified categories. They look for differences of the two such diagrams for the drug andits vehicle visually, and also compare the categorical compositions in a particular period numerically.

We introduce two clustering methods into these common processes of the EEG experts. As data units of analysis spectra are used in the methods. Note that one spectrum corresponds to one tracing paper (typically, 20 seconds segment) and their ordinates are sampled on discrete frequencies (in our following example, 0.5Hz equi-interval points from 0.5Hz to 20Hz). One method here is a RELOC (Anderberg,1973), which is used to classify each tracing paper numerically. The other is a two-dimensional EPP (Friedman,1987), which is used to explore the underlying structures of a collection of spectra.

We have applied the two clustering algorithms to the set of data obtained from a study of pharmacological effects of anti-depressants in cats. Principal results from these analyses can be illustrated in Figure 1 and 2. In Figure 1 right graphs are results for an anti-depressant, namely amitriptyline and left ones are those for a vehicle control. The first row is a sleep-awake-cycle diagram by an EEG researcher and the others are the diagrams by the RELOC clustering of the vehicle spectra from four recording leads (namely. amygdala, hippocampus, motor cortex and visual cortex). For each of the leads the clustering-based diagram is very similar to the human diagram. Note that the two slow-wave-sleeping categories in the human recognition have more categories in the clustering-based classification.

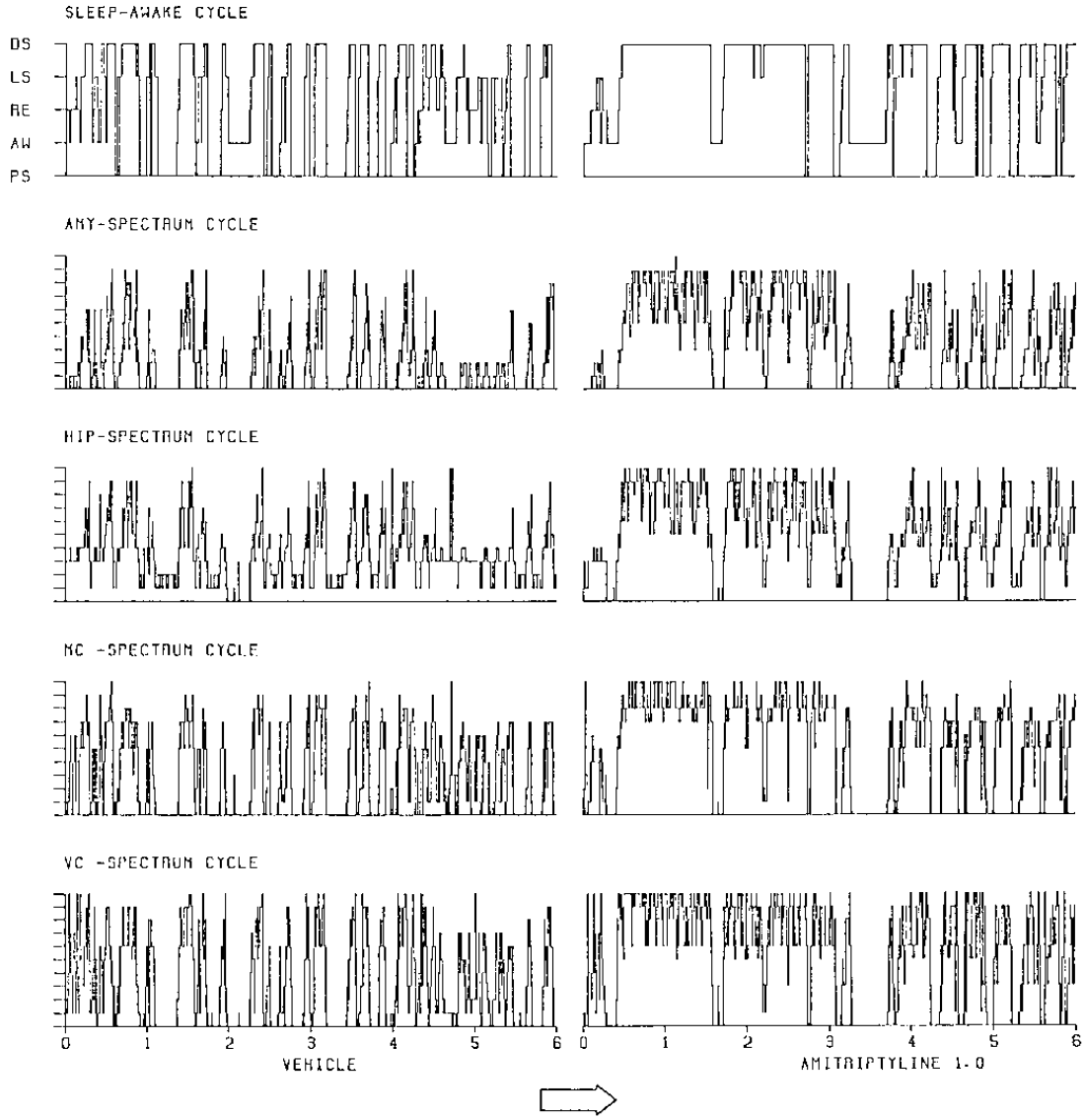

Figure 1. Sleep-swake-cycle diagrams based on a human pattern recognition and clusterings of spectra from four leads.

Left four graphs in Figure 2 show boomerang-like structures found by applying the EPP to four sets of vehicle spectra of four anti-depressants from a lead, hippocampus. Right four graphs plot respective scores of the drug spectra mapped onto the EPP solutions. Note that the boomerang-like structure is conserved for desipramine and maprotyline, but it hardly is left for imipramine and amitriptyline.

From these tentative applications it is suggested that the clustering methods here provide useful tools in understanding possible effects of drugs on the EEG.

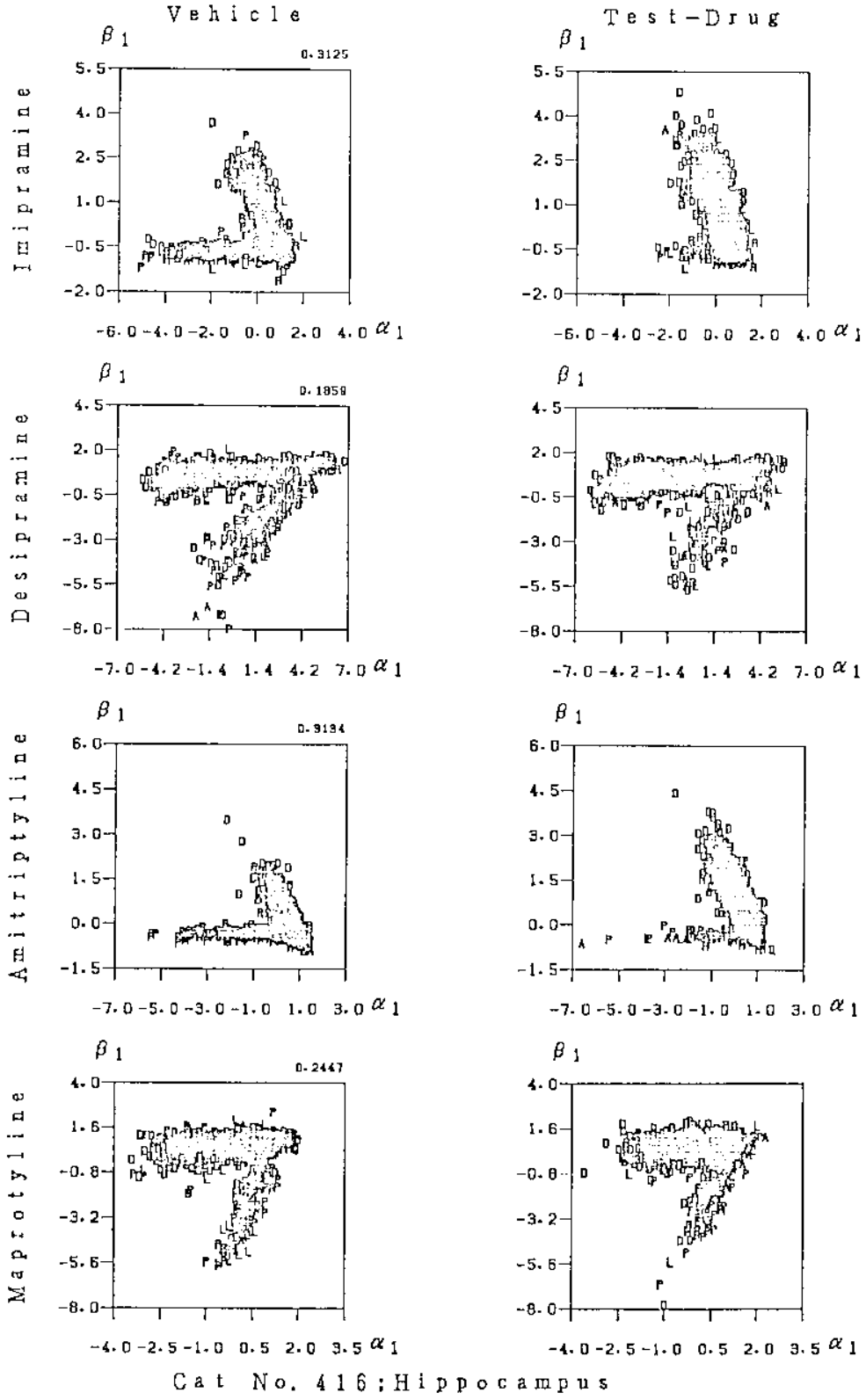

Figure 2. Project scores by a two-dimensional EPP for four anti-depressants.

References

(1)Michael R. Anderberg (1973). *Cluster Analysis for Applications.* New York and London, Academic Press.
(2)Jerome H. Friedman (1987). Exploratory projection pursuit. *J.Amer.Statist.Assoc.*, 82, 249-266.

Identification Problem of Drugs Prescribed for Patients

Institute of Japanese Union of Scientists and Engineers
Kiyoshi Katayama
Chang Yuan-Tsung

Based on the clinical inspection data a drug identification problem was studied and following procedure has been recommended. As a first step the principal component analysis would be applied to inspection data to extract serious variables which behave similarly with regard to the same drug utilizing factor loadings scatter diagrams. Next to the variables selection hierarchical classification procedure is adopted utilizing those selected variables. Choosing a certain number of clusters which is supposed to be a number of drugs. a new variable equals to a cluster number is attached to the original data sets. Then we apply the discrimination analysis using this membership variables, getting more efficient discrimination function.

Statistical software packages JUSE-QCAS, JUSE-QCAS/MA1, JUSE-CLUSTER have come to use throughout the microcomputer.

Theory and Practice in Classification
— In Case of Mineralogy —

Science University of Tokyo
Faculty of Engineering
Keiji Yajima

Although theoretical classification should never be contrary to practical classification, there is a significant difference between them. In other words we need to consider a number of problems applying the theoretical identification in practice. When we treat the material that exists very widely for example we need to get fast conclusion. The complicated operation should come later in classification work and also a compound notion which is to be used in identification stages might be arranged in a later step. If we establish a double checking procedure an optimal position of verification work with regard to efficiency will arise. Those problems are considered taking the classification procedure in mineralogy by means of polarized light microscope as a case study.

Classification of Time-Series Data with Generalized Logistic Growth Model

Shoichi Ueda
Faculty of Economics
Ryukoku University
Fushimi-ku, Kyoto, Japan

In this paper a procedure of classification is presented for time series data X_{lT} (l=1,2,..k; T=1,2,..N) according to their pattern of change over time.

Data matrix is presented as k points in N dimensional space. But, any distance defined in this space would distort the data structure with suffix running over time.

Data are plotted alternatively as K lines in X–T plain. We can see and classify lines for different item by this graphical presentation, but difficulties remain in defining distance by which classification is performed. Further essential difficulties would follow in the step of interpretation to tell in what means the data are classified.

Therefore, a preliminary step is required to find characteristics of pattern of change over time. The following step of classification is to be performed in the parameter space thus defined. Strictly speaking, the definition of parameters in the preliminary step is essential to derive appropriate classification.

We need model well generalized to describe various type of time series data.
The level-rate plot is one of widely applicable tool to find and specify such a model. In this plot, data X_{IT} are divided into two parts, namely

LEVEL X_{IT}	as horizontal scale X and
RATE $X_{IT}-X_{IT-l}$	as Vertical scale DX

Many kind of time-series data exhibit a quadratic curve

$$DX = -ß\,(X-L)(X+K) \qquad (1)$$

which is integrated to derive explicit formula (2) for X(T).

$$X = \frac{L - K\exp[-ß\,(L+K)(T-T_0)]}{1 + \exp[-ß\,(L+K)(T-T_0)]} \qquad (2)$$

This is “generalized Logistic curve” for standard case with K=0 and L=l. This curve generally startup from level X=K and grows up to level X=L. The generalized curve have lower level K other than 0 and upper level L other than 1.

The case with K>0 suggests existence of initial level on which additional changes are observed over time.

The case with K<0 suggests existence of threshold level and a kind of auto-adjusting mechanism works to suppress changes bellow that level.

Parameters $ß$ and T_0 are corresponding to speed of change and shift of time scale respectively. Thus, the problem comes to the classification in three dimensional space (K,L,$ß$) putting parameter T_0 outside of consideration.

An illustrative example is shown to classify time series data on prevalence rates of different durable goods by 47 geographic sectors.