Evaluation of Multivariate Graphs

(1995年3月31日)

Mamoru Fukumori

Key words: Radar chart, Face graph, Letter graph

Abstract

This paper discusses to what extent graphical methods convey the information of multivariate data. Three kinds of graphical representation,i.e., radar chart, letter graph and face graph, are selected and compared from the cognitive aspect. For this purpose three sets of 30 multivariate data, which had three clusters, were generated for the cases of three variables, five variables and seven variables, and the graphs of those data sets were shown to college students and were classified into three groups based on their subjective impression. The results show that those three graphical methods have different properties such that it is easier to classify the letter graphs when the number of variables are small while the case of five variables is the easiest for the other two kinds of graphical representation.

1 Introduction

Today, graphical representation is an important part of statistical data analysis. Graphical methods make it possible to find features in the data which can not be found by numerical methods. As Cox(1978) points out, there is a major need for theory of graphical methods. As Kruskal(1975) points out, graphical methods for data analysis are largely unscientific. It seems that graphical methods need scientific foundation.

Until now, most researches into graphical methods have focused on the technical aspect of data representation. Though the development of techniques is very important, it deals with a small piece of the whole process of graphical methods.

This study is based on human recognition. The purpose of this study is to evaluate the graphical methods through the human recognition by means of an experiment in which examinees are asked to classify some graphs into groups with their subjective impression.

Wakimoto et al(1979) discusses the relationship between multivariate statistical methods and graphical methods. They classify graphical methods into groups based on their relations with multivariate methods. For example, one group which is associated with regression analysis includes: scatter plot, linked vector graph and probability plot; and the second group which is associated with PCA includes: scatter plot, radar chart, face graph, biplot, triangle polynomial graph and statistical ellipse. This research chooses the techniques which highlight the nature of the data.

According to Wakimoto et al(1979) the graphical methods concerned with cluster analysis contain scatter plot, radar chart, face graph, body graph, constellation graph, vector separation graph, biplot, non-linear mapping plot and dendrogram. Further work by Goto et al(1986) discusses graphical representation in accordance with usage of the statistical method. In their study, the following graphical methods can be used for the purpose of cluster analysis or discriminant: scatter plot, histogram, constellation graph, face graph, dendrogram, SHADE, AID, tree and castle graph.

2 Method

2.1 Selection of multivariate graphs

Generally, graphical methods can be divided broadly into descriptive and analytical types as mentioned in the preface. Of these two types one includes methods such as radar chart, face graph, constellation graph, tree graph and linked vector graph, and the other includes methods such as biplot, non linear mapping, triangle polynomial graph, probability plot and statistical ellipse.

In this study, the graphs of the descriptive type were chosen because this purpose is to establish to what extent the human eye can cognize data depending on the type of graphical representation stimulus it receives.

Wakimoto et al(1979) discussed the relationship between multivariate statistical analysis and graphical methods. They arranged the graphical methods into groups in accordance with multivariate methods. For example, one group which is associated with regression analysis includes: scatter plot, linked vector graph and probability plot; and the second group which is associated with PCA includes: scatter plot, radar chart, face graph, biplot, triangle polynomial

graph and statistical ellipse.

In this study, the graphical techniques concerned with cluster analysis were chosen.

According to Wakimoto et al(1979) the graphical methods concerned with cluster analysis can be summarized as scatter plot, radar chart, face graph, body graph, constellation graph, vector separation graph, biplot, non-linear mapping plot and dendrogram. Further work by Goto et al(1986) arranged graphical representation in accordance with usage of the statistical method. In their study, the following graphical methods can be used for the purpose of cluster analysis or discriminant: scatter plot, histogram, constellation graph, face graph, dendrogram, SHADE, AID, tree and castle graph.

Having established this, it is possible to choose from the radar chart, face graph and body graph. However, since the face graph and body graph are fundamentally similar displaying data of resembling qualities, in this study the more popular face graph is chosen to avoid data analysis problems.

Furthermore the letter graph was additionally selected to join the face graph and radar chart since it can clearly and succinctly show the characteristics of the data. The letter graph is made possible by the letter size describing the value of the variable. These three graphical techniques were thus chosen, below is an outline of each graph type.

Of all the graphical methods the radar chart is originally the most popular. It makes for easy composite comparisons of the data since its appearance is readily understood. Thus in previous research the radar chart has been utilized frequently. However it's validity is not entirely proved as a representative medium.

Like the radar chart, the face graph(Chernoff,1973) is a popular means of graphical technique and there are many studies concerned with the face graph. It displays the data via facial expression on a human face representation. All the data is shown on one face. Thus differing quantities are spotted with ease. It can also deal with distinctive features in the data well. However, the problem of which variable to assign to which facial feature and the complicating factor of too many variables are among it's disadvantages.

According to Hirai et al(1990) the letter graph also shows data and it's peculiarities well. It can utilize one letter to represent two variables. In this way it permits interrelation and qualitative analysis of the graph.

The letter graph is produced as shown below:

1. The first step is to prepare the data which consists of n observations on $2 \times p$ variables. According to the situation of data, appropriate transformations performed on $x_{ij}, y_{ij} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, p)$ such as

$$x'_{ij} = \frac{x_{ij} - \overline{x_j}}{S_{xj}} + b, y'_{ij} = \frac{y_{ij} - \overline{y_j}}{S_{yj}} + b \tag{1}$$

where, b is a positive constance, $\overline{x_j}, \overline{y_j}$ is the sample mean of X_j, Y_j respectively and S_{xj}, S_{yj} is the sample standard deviation of each variable.

In the case of examination marks at school, it is suitable to transform the data into deviation values as

$$x'_{ij} = 10\left(\frac{x_{ij} - \overline{x_j}}{S_{xj}}\right) + 50, y'_{ij} = 10\left(\frac{y_{ij} - \overline{y_j}}{S_{yj}}\right) + 50. \tag{2}$$

2. The next step is to decide the letter. As 2 variable can be represented by a letter, p types of letters must be prepared in the case of $2 \times p$ variables.

3. For step 3, the length and width of the letters is assigned to the values of x_i, y_i. In observation i, length and width of the first letter are transformed as

$$l\left(\frac{x'_{i1}}{b}\right), l\left(\frac{y'_{i1}}{b}\right) \tag{3}$$

or

$$l\left(\frac{x'_{i1}}{50}\right), l\left(\frac{y'_{i1}}{50}\right) \tag{4}$$

respectively, where l is a length of a letter.

Similarly, the p-th letter is drawn by the values of x'_{ip}, y'_{ip}.

4. For the final step, letters are drawn and arranged from left to right.

2.2 Data generation

3,5 and 7 dimensional observations were generated based on models $N(\mu_k, I)$ of three groups,*i.e.*,k=1,2 and 3, where the distances between respective set as 3,4 and 5 regardless of their dimensions. For each variable, samples of 12,10 and 8 were drawn at random. These three respective groups make a sample total of 30. Table 1.1 $\sim$ 1.3 shows each variable average for each group.

The distance between respective groups was set at 5 for between group 1 and group 2, 4 for between group 1 and group 3 and at 3 for between group 2 and 3.

Table 1.1 Center of each group(3 variables)

	variable 1	variable 2	variable 3
Group 1	1.565	3.332	1.565
Group 2	-1.767	1.660	-1.767
Group 3	0.000	0.000	0.000
Mean	-0.067	1.664	-0.067
SD	1.853	1.850	1.853

Table 1.2 Center of each group(5 variables)

	variable 1	variable 2	variable 3	variable 4	variable 5
Group 1	0.990	0.990	0.990	2.555	2.555
Group 2	-1.565	-1.565	-1.565	0.910	0.910
Group 3	0.000	0.000	0.000	0.000	0.000
Mean	-0.192	-0.192	-0.192	1.155	1.155
SD	1.106	1.106	1.106	1.118	1.118

Table 1.3 Center of each group(7 variables)

	variable 1	variable 2	variable 3	variable 4	variable 5	variable 6	variable 7
Group 1	0.750	0.750	0.750	0.750	2.141	2.141	2.141
Group 2	-1.390	-1.390	-1.390	-1.390	0.650	0.650	0.650
Group 3	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Mean	-0.213	-0.213	-0.213	-0.213	0.930	0.930	0.930
SD	0.787	0.787	0.787	0.787	0.804	0.804	0.804

2.3 Graphical representation and experiment

From the random data a radar chart, face graph and letter graph was drawn for each of the 3,5 and 7 variable cases. Thus a total of 9 kinds of graphs were produced and 30 kinds of each graph classification were established. The face graph was drawn utilizing the following 7 components(as shown in table 2); face length, curvature of the mouth, slant of the eyes, ellipticity of the eyes, slant of the eyebrows, distance from the centre to the edge of the face and the length of the nose.

Table 2 Assignment of variables in face graph

Variables	part of face
Variable 1	length of face
Variable 2	curvature of the mouth
Variable 3	ellipticity of the eyes
Variable 4	slant of the eyes
Variable 5	slant of the eyebrows
Variable 6	distance from the center to the edge of the face
Variable 7	the length of the nose

In accordance with table 2, the case of 3 variables utilized variables 1-3 inclusive, 5 variables, 1-5 inclusive and 7 variables used all 7 of the parameters listed. The decision of which variables to use was problematic but with due consideration to previous research, by Matsubara(1977), Chernoff(1975), it was possible.

For example Honda et al recommended that eyebrow shape, distance between the eyebrow and eye, eye shape, distance between the eyes, nose shape and mouth shape were important factors in describing the face graph's features.

Other research by Ellis et al discovered, that when using composite(montage) photos, only the forehead, eyes and mouth could be accurately represented.

If we consider these studies we can see that only the eye, eyebrow, facial outline and mouth can produce a strong recognitive quotient in subjects.

As a pre-experiment, I investigated the facial expression idea loading at 18 parts of the face. I examined the extent of changes and transformations in each part over a 10 stage process. In my experiment involving 87 subjects I found that the largest changes occurred in 5 features;

face length, slant of the eyes, slant of the eyebrows, curvative of the mouth and ellipticity of the eyes. With Chernoff's and other previous investigations taken into account we finally selected those features shown in table 2 as the most suitable variables in our experiment.

Finally the letter graph employs the alphabet to display its findings where one letter represents one variable. Representation of each graph conformed to Seto/B(Tarumi and Hayashi,1990). The face graph which is used in this study linked the slant of the eye and slant of the eyebrow giving it a slightly different appearance from other face graphs.

For each graph type, I presented 30 graphs to the subjects and instructed to place the similar looking graphs into 3 separate classifications. Moreover, since the focus of this experiment was to observe the way the graphs are looked at, the subjects were 120 Junior College students(aged 20-21) who had little or no knowledge of statistics or graphical analysis. In this way the investigation erased any preconceived ideas of the subject and enabled bias to be reduced to a minimum. The subjects were shown 9 graphs(3 each of face graph, radar chart and letter graph) and 3 of variables. Having been put into 3 groups, 40 subjects were asked to asses each group in the way mentioned above.

2.4 Statistical analysis

1)Correct response

First, in order to evaluate the 'correctness' of the classification of the 30 graphs I calculated the correct response by using the data which did not have a missing value. In this experiment the definition of correct response is two-fold. So, if samples of the same group are classified by subjects into the same group or if samples of differing groups are classified in different groups, the response is 'correct'. From this definition, all those samples assigned to the same groups were given a value of 1 and those assigned to differing groups were given 0 value in the data matrix for each subject. Such a matrix was drawn for each subject. Then each matrix was compared to the correct matrix and the samples which accord with the correct matrix were given a value of 1. The other samples were given a value of 0.

Table 3 shows the standard deviation and average of the 'correct' rate. Since the sample number was 40 in each graph, the sample number including missing values is 40-n.

Table 3 Mean and standard deviation of correct rate in each subject

	Radar chart			Face graph			Letter graph		
	Ratio	SD	N	Ratio	SD	N	Ratio	SD	N
3 dimension	0.6791	0.0589	28	0.6634	0.0484	25	0.7392	0.0643	32
5 dimension	0.7769	0.1011	30	0.8093	0.1008	31	0.7253	0.1029	23
7 dimension	0.7237	0.0796	27	0.6206	0.0430	26	0.6659	0.0717	31

As table 3 shows, we can see that comparing the face graph and radar chart yields extremely high correct response value in the case of 5 variables in spite of the difficulty of cognition of variable numbers. The high correct response value maybe connected with the problem of variable assignment. Additionally the number of the variables in the face graph may affect the judgement of the subject adversely. Table 3 may also indicate that there is an optimum variable number for the face graph.

The correct rate of the radar chart has a high score in the case of 5 variables. In the case of 3 variables the effect of one unusually large or small value can over-influence the viewer's perception. Furthermore, with many variables the graph becomes chaotic so as to render it difficult to comprehend. So the case of 5 variables is the simplest and clearest for interpretation.

Finally the results of the letter graph indicate that as the number of variables increased the correct response value decreased. In this experiment it is best to regard the letter graph as a bar-chart since with many variables the appraisal of the schematic is complicated. The letter graph shows the highest correct ratio for the 3 variables case of any of the graph types.

2)Relation between the distance between the groups and misclassified data

In order to investigate the relationship between misclassified data and the distance between the groups, an accumulated response matrix was producted for each graph type. Table 4 shows an accumulated matrix of the face graph for the case of 5 variables. From table 4, the relation between each group interval can be expressed as shown in figure 1.

Table 4 Accumulated response matrix of the face graph for the case of 5 variables.

31																														
17	31																													
15	9	31																												
15	10	15	31																											
14	14	12	17	31																										
17	15	13	10	13	31																									
15	11	10	8	11	16	31																								
12	11	8	11	12	15	12	31																							
15	10	10	13	16	10	16	6	31																						
19	17	15	15	15	16	14	11	14	31																					
11	8	12	10	10	13	12	11	9	12	31																				
16	13	8	15	9	13	8	12	11	16	9	31																			
1	6	6	7	5	8	9	14	6	0	17	6	31																		
17	14	9	14	9	11	11	13	9	15	7	11	9	31																	
11	12	7	14	10	12	9	10	10	12	11	14	16	11	31																
0	6	6	8	6	8	9	15	7	0	16	6	30	8	17	31															
20	14	9	8	12	16	12	12	9	13	10	11	7	15	10	6	31														
9	9	11	12	9	8	13	2	11	6	10	9	16	14	13	16	9	31													
13	10	15	10	11	12	16	9	10	17	12	6	10	11	8	10	11	12	31												
2	6	7	9	7	9	8	16	6	0	14	7	28	9	17	29	8	16	9	31											
12	11	9	10	12	14	8	9	11	12	9	14	8	16	13	8	10	10	12	9	31										
7	6	8	12	12	10	6	12	12	4	2	9	16	13	13	17	12	12	11	18	13	31									
12	12	11	8	9	12	16	10	9	13	12	13	9	11	12	9	16	14	11	9	9	6	31								
10	11	9	8	13	8	8	7	13	10	7	10	3	7	10	4	12	7	4	6	10	6	10	31							
12	14	10	12	14	14	12	3	15	10	2	12	0	12	3	0	13	13	9	1	17	10	15	17	31						
7	8	11	6	7	10	11	12	10	4	12	7	16	11	1	15	9	8	13	14	10	9	10	8	11	31					
12	11	11	8	12	9	7	8	11	11	7	12	2	7	10	3	13	5	5	5	12	7	9	29	16	8	31				
12	12	9	8	13	7	8	8	13	11	6	12	2	8	10	3	14	6	5	5	11	8	11	29	16	8	29	31			
16	17	10	15	12	11	11	10	14	17	12	11	5	14	13	6	11	9	14	4	9	7	16	9	12	7	9	9	31		
8	10	10	12	14	9	8	4	15	15	4	12	1	10	6	1	10	11	7	2	12	7	11	25	22	7	23	23	11	31	

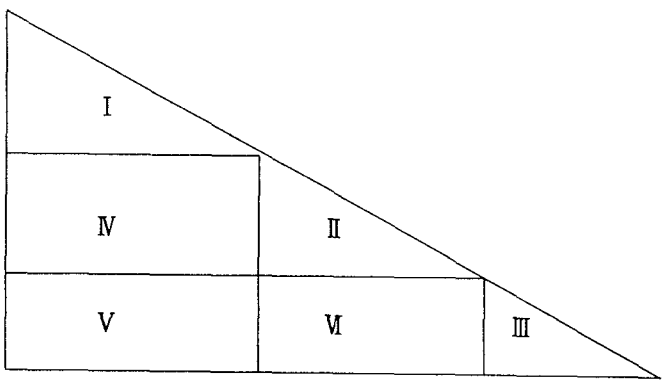

Figure 1 Territory among three groups

In figure 1, each territory, 1, 2 and 3 represents their respective classified groups 1,2 and 3. Furthermore, territory 4 shows the classification of groups 1 and 2. Similarly territory 5 shows that of groups 1 and 3 and territory 6 of groups 2 and 3. Thus in order to observe the relationship of the distance between the groups and the data classification, territory 4, 5 and 6 need to be examined.

Table 5 shows the mean response rate of correct classification for the 6 territories for the 5 variable cases. The distance between the groups is 5 for territory 4, 4 for territory 5 and 3 for territory 6.

Table 5 The mean response rate of correct classification for each territory.

	territory 1	territory 2	territory 3	territory 4	territory 5	territory 6
Radar chart	0.3849	0.3763	0.3345	0.3347	0.3139	0.3142
Face graph	0.4047	0.4201	0.4724	0.3124	0.3364	0.2698
Letter graph	0.3484	0.3411	0.3649	0.3489	0.3211	0.3326

From table 5 we can see that in territory 4 the response rate is higher than for territories 5 and 6 in both the letter graph and radar chart. This is caused by the slight influence of the distance between the groups. However in the face graph, territory 5 has a higher value than territory 6. Thus we can suggest that the distance between the groups may have little effect on the cognition of the graph. In addition we can see that the values for territories 5 and 6 are almost identical in the case of the radar and letter graph, whereas the face graph shows a clear numerical difference between the two territories. From this result we can extract two facts: One, that there is little difference between the three types of graphs and two, that it is difficult to classify the data by looking at the graphs only.

3)Quantification 4 analysis.

To examine the degree of reproduction of the original data, a comparison was made between the result which applied PCA to the low random data and the result which applied quantification method to the accumulated response matrix.

In quantification 4, similarity matrix which transformed the diagonal component of the accumulated response matrix to 0 was used.

Figure 2 shows the scatter plot of coordinate based on PCA by the random data for the 5 variable case of the face graph. Figure 3 shows the scatter plot of coordinate based on

quantification 4 by the accumulated response matrix data.

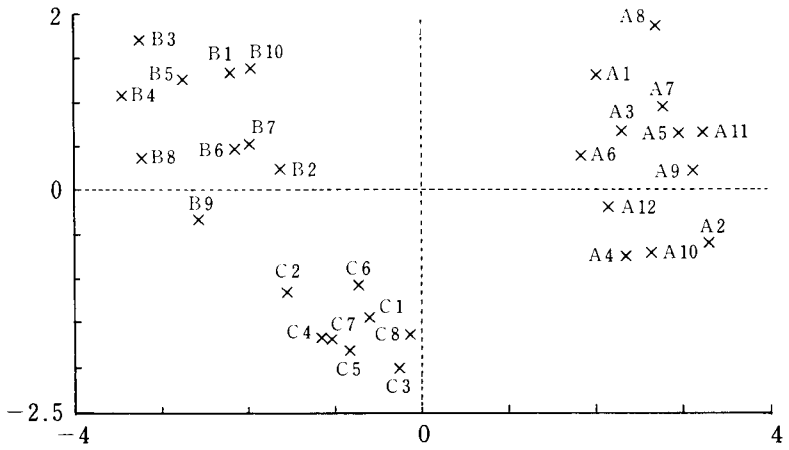

Figure 2 Scatter plot based on PCA

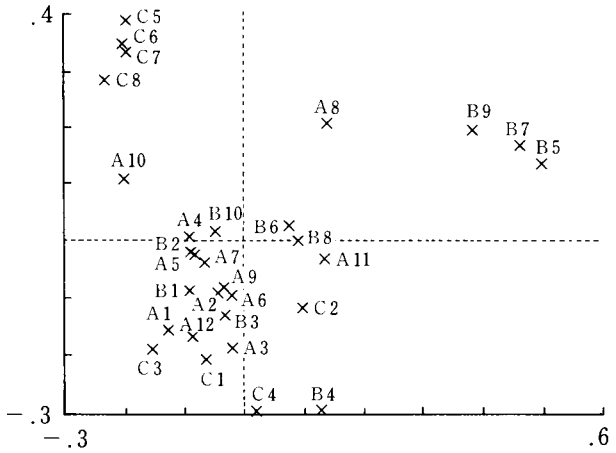

Figure 3 Scatter plot based on Quantification 4

In figure 2, three groups can be clearly distinguished and the relationship of distance between groups can be represented. However in figure 3 all samples with the exception of(C5, C6, C7, C8) and (B5, B7, B9) can not be classified easily. Thus I can say from this result that data cannot be classified by only human recognitive abilities as well as numerical methods can.

3 Concluding remarks

In this experiment, to discuss the classification of multivariate data through human recognition ability, an experiment of graph classification was carried out by using 3 types of graphs: radar chart, letter graph and face graph. The results are shown below.

1. The highest correct response rate was in the case of both the face graph and radar chart with 5 variables. The difficulty of classification increased as the number of variables increased for all graph types. Additionally the letter graph gave an extremely high correct response rate with few variables. This suggests that there may be an optimum variable number for classification of data in terms of human recognition.

2. From the results of the analysis on the incorrectly classified data I can say that the distance between the graphs has little effect on the mean response rate of each territory. It is perhaps inappropriate to classify the graphs in this way in this study.

3. Reproducing the original data by application of the quantification 4 was difficult with the exception of one part of the sample. In future experiments evaluation by means of non-metric MDS would be a possible advantage.

From the results obtained I can see that it is not always best to classify data through human recognition. Thus, though a detailed classification by graphical techniques is difficult, an approximate classification is possible.

However, in an attempt to produce a detailed view, limitations were found with the classification. These limitations were a result of each individual's recognitive ability of the graphical representations.

At the outset I considered that the graphs could be cognized by humans well, however in this experiment it was shown that the capability of the graph to differentiate data was limited by the

cognitive ability of the subject. If we were to continue the investigation, clearly we would need to develop a graph method with higher evaluative qualities for use in statistical experiments.

Finally, the problems which could not be examined in this study included: the way in which the variables were assigned for the face graph, the analysis of correlating data and of individual differences between subjects. Each problem should be examined independently in detail to further validate this type of experiment.

References

[1] Andrews,D.F.(1972). Plots of high-dimensional data. *Biometrics*, 28, 125-136.

[2] Chernoff,H.(1973). The use of faces to represent points in k-dimensional space graphically. *J. Amer. Statist. Assoc.*, 68, 361-368.

[3] Chernoff,H. and Rizvi,M.H.(1975). Effect on classification error of random permutations of features in representing multivariate data by faces. *J. Amer. Statist. Assoc.*, 70, 548-554.

[4] Cox,D.R.(1978). Some remarks on the role in statistics of graphical methods. *Applied Statistics, 27, 4-9.*

[5] Everitt, B.(1978). Graphical techniques for multivariate data. *Gower Publishing Ltd.*, London.

[6] Fukumori,M. and Tanaka,Y.(1994). Evaluation of multivariate graphs from the aspect of human recognition. *Bulletin of The Computational Statistics of Japan*, Vol.7, No.1, 37-45.

[7] Fukumori,M. and Tanaka,Y.(1994). Modified non-linear mapping and its application to the problem of detecting influential subsets. *Proceedings of Japan and China symposium.*

[8] Hirai,Y., Fukumori,M. and Wakimoto,K.(1988). Kanji graph representation for multivariate data and its application to cluster analysis. *Bulletin of The Computational Statistics*, Vol.1, No.1, 11-21.

[9] Honda,N.(1981). A search algorithm of variables assignment in face method considering psychometrical characteristics of facial expressions. *J. Behaviormetrics*, 8, 53-65 (in Japanese).

[10] Honda,N. and Aida,S.(1982). Analysis of multivariate medical data by face method. *Pattern recognition*, 15, 231-242.

[11] Honda,N. and Sugimoto,F.(1987). Multivariate data classification by face pattern considering psychometrical distances of facial expressions. *Behaviormetrika*, No.21, 29-43.

[12] Kruskal, W.H.(1975). Visions of maps and graphs, in *Auto-Carto II, Proceedings of the International Symposium on Computer Assisted Cartography (ed. J.Kavaliunas), Washington,D.C., U.S. Bureau of the Census and American Congress on Survey and Mapping, 27-36.*

[13] Wakimoto,K. and Taguri,M.(1978). Constellation graphical method for representing multidimensional data. *Ann. Inst. Statist. Math.*, 30, Part A, 77-84.