Developments in Data Analysis
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 12, Ljubljana: FDV, 1996

Analysis of "Don't Know" Answers by Cluster Analysis with Optimal Scaling Features

Herbert Matschinger[1]

Abstract

The aim of this paper is to approach several basic problems that occur when evaluating the effect of "don't know" answers on the dimensional structure of a set of items which are considered to measure a latent dimension that can be represented by a low-dimensional Euclidean space. Within a study on attitudes and beliefs about mental disorders a set of 15 dichotomous items was employed to identify groups of respondents displaying similar patterns of attitudes toward the psychiatric clinic, simultaneously taking into account the amount of knowledge about such institutions. We employ a particular combination of multiple correspondence analysis and k- means cluster analysis (GROUPALS) for both exploring the dimensionality of the 15 item instrument and the location of the "don't know" answers within the reduced space. Results show that denying the more positive aspects of a mental hospital or accepting the negative aspects cannot be separated from "don't know" answers, because these categories turned out to be very similar. In order to check the stability of the results, a balanced bootstrap analysis is performed.

1. Introduction

The aim of this paper is to approach several basic problems that occur when evaluating the effect of "don't know" answers on the dimensional structure of a set of items which are considered to measure a latent dimension that can be represented by a low-dimensional Euclidean space. We assume that "don't know" answers depend both on the underlying dimension and certain exogenous characteristics of the respondents. These "don't know" answers are treated as a particular response and not as is frequently done as missing values. Therefore the problem to be tackled is not to impute these values but rather to evaluate and interpret these responses with respect to a low- dimensional representation of the data matrix.

[1] Department of Psychiatry, University of Leipzig, Johannisallee 20, D-04317 Leipzig

Within a study on attitudes and beliefs about mental disorders a set of 15 items (see Appendix) was employed to identify groups of respondents displaying similar patterns of attitudes toward the psychiatric clinic, simultaneously taking into account the amount of knowledge about such institutions. It seems reasonable to assume that the location of respondents in a low-dimensional latent space is not independent of their knowledge about psychiatric clinics in general. Furthermore we expected the probability of a "don't know" answer for a particular item to depend on the latent dimension under consideration. Since these dependencies are certainly different for each individual item, the evaluation of the "don't know" answers is directly connected with the evaluation of the meaning of each item for each dimension of the reduced space. The analysis known under the name "GROUPALS" takes two main steps:

1. First a parsimonious Euclidean space has to be generated by principal component analysis. Since all the variables can only be treated as nominal variables due to the inclusion of the "don't know" response as an independent category, this has to be done by so called "non-linear principal component analysis", also called multiple correspondence analysis with rank one restriction (Gifi, 1990; Jolliffe, 1986:203; Jackson, 1991:224; Nishisato, 1980; Van Rijckevorsel and de Leeuw, 1988). The optimal scaling features of the singular value decomposition of the scaled indicator matrix provide both the quantification of the item categories and the objects (subjects) of the data.
2. As a second step we want to find a partition of the observations into a fixed number of mutually exclusive groups which is optimal with respect to internal cohesiveness and the external isolation of these groups. Internal cohesiveness is maximized by minimizing the pooled- within- group variance. This criterion ensures that the clusters are as tight as possible. The external isolation is maximized by maximizing the between- cluster variance (see Van Buuren, 1986:1/1). Since all the variables are nominal we do not want to use one of the many, often arbitrary, (dis)similarity measures proposed by Gower (1971) or by Gordon (1981) but rather transform these variables into numerical ones, so that the use of Euclidean metric becomes possible. Using the object score of non-linear PCA (see point 1) as linear combinations of the original variables provides the property of variance- maximization which suffices for the criterion of external isolation.

It is one of the important features of the approach described in the following that the quantification of the variables and the clustering problem are done simultaneously, since both minimization problems are incorporated into **one** loss-function.

2. Method

Let each variable h_j (j = 1,....,m) be coded into a (n x k) indicator matrix G_j. Define X as a (n x p) matrix of object scores and m (k x p) matrices Yj of category quantifications. The ordinary Homogeneity analysis minimizes a loss function very similar to the loss functions of principle component analysis (Bekker and de Leeuw, 1988:6)

$$L(X; Y_1,...., Y_m)=\frac{1}{m}\sum_{j=1}^{m} tr(X-G_jY_j)'(X-G_jY_j) \qquad (1)$$

In the framework of reciprocal averaging (see Bekker and De Leeuw, 1988:10) partial derivatives with repect to X or Y yield $X \div m^{-1}Gy$ and $y \div D^{-1}$ G'x respectively . D is the diagonal of G'G, representing the marginal totals of all item categories. Now the observations (usually called objects) are in the centre of "their" categories and the category quantifications are in the centre of the objects who choose the particular category. The two equations necessarely lead to $X \div GD^{-1}$ G'X. X is a latent vector of this expression. If we normalize X'X = n1 or y'Dy=1 then

$$X\Psi^2 = GY \qquad \text{and} \qquad Y = D^{-1}\, G'X$$

By means of singular value decomposition of $GD^{-1/2}=V\Psi W'$ we obtain X = V and $Y=D^{-1/2}W\varphi$ which demonstrates the basic identity between homogeneity analysis and multiple correspondence analysis.

If we now introduce Y (k x p) as a matrix of cluster points we can replace the i-th row of X by a corresponding clusterpoint y, imposing the restriction $X = G_cY$. If v is a vector of the first k integers, then $c = G_cv$. Using $X = G_cY$ means that we also scale the clusters in a p- dimensional space (Van Buuren and Heiser, 1989:700). Now we can write equation (1)

$$L(X;Y_1,...,Y_m) = m^{-1}\sum tr(G_cY-G_jY_j)'(G_cY-G_jY_j) \qquad (2)$$

Let $Z=\frac{1}{m}\sum G_jY_j$ and inserting $G_cY = Z-(Z-G_cY)$ into (2) yields a loss function with two components, where the first component is constant for fixed Y_1, ..., Y_m.

$$L(X; Y_1,...., Y_m)=\frac{1}{m}\sum_{j=1}^{m} tr(Z-G_jY_j)'(Z-G_jY_j)+tr(Z-G_cY)'(Z-G_cY) \qquad (3)$$

Therefore only the second part must by minimized over G_c and Y. This problem is known as the sum of squared distances clustering (SSQD). If this criterion is minimized over Y (cluster points) setting : $Y := (G_c'G_c)^{-1}G_c'Z$ the clusterpoints are set equal to the cluster centroids in terms of Z (Van Buuren and Heiser, 1989:701). For minimizing the SSQD criterion the k- means algorithm (Hartigan, 1974) is employed. As a final step X (object scores) are set to G_cY. In order to avoid trivial solution several normalization conditions have to be met, which will not be discussed here.

In order to get a single quantification of the item categories we impose rank-one restriction of the form: $Y_j = y_jb'_j$. The columns of Y now are all linearly related by b (the correlation of the item with the object scores of one dimension. Y is of rank one).

3. Data

The method has been applied to a set of 15 items from a survey on attitudes and beliefs about mental disorders conducted in the new *Länder* of Germany in the spring of 1993. Everyone of German nationality of at least 18 years of age living in private households was included in the target population. The survey was carried out in cooperation with the ZUMA (Zentrum für Umfragen, Methoden und Analysen e.V.) in Mannheim. The field work was entrusted to the GFM-GETAS (Gesellschaft für Marketing, Kommunikations und Sozialforschung mbH) in Hamburg. The respondents were asked to decide wether or not each of the 15 entities could be found in a psychiatric clinic. The third possible reaction was to say : "don't know" .

4. Results

The category quantifications of both the homogeneity analysis and non-linear PCA are prone to what is called "rare pattern" . Unique and simultaneously rare patterns are located far out in the space thus dominating the solution and causing degenerated solutions. The solution capitalizes on peripheral effects in the data (Van de Geer, 1985:38; 1993; Bekker and De Leeuw, 1988:10). Therefore it was necessary to run not only one cluster analysis for the entire sample, but to reduce the sample with respect to the maximum number of acceptable "don't know" answers in stepwise manner, and to repeat the analysis for each subpopulation. In the following we will present results for a maximum of 3 and 6 "don't know" answers as well as for the entire sample (max. 15 "don't know" answers). All estimations were carried out using the program GROUPALS (Van Buuren, 1986)

The interpretation of the results will be presented in two steps:

1. First of all the dimensions of the space need to be interpreted according to the coordinates of the item categories.
2. Secondly the location of the categories within this space yield the essential information for the interpretation of the partitions of observations (cluster) within the same space.

In order to do this we exploit the principle of reciprocal averages, which means that the object points are plotted as if they were located within the category centroids, although their optimal position are the cluster means. In this way we can inspect a low- dimensional **continuous** representation closest to the optimal cluster solution (Van Buuren and Heiser, 1989:704) . The precision of the k- means solutions is evaluated by means of the so called silhouette width (Kaufman and Rousseeuw, 1990:87) for each cluster and the average silhouette width (ASW) for the entire solution. Surprisingly a 3 clusters solution in a two-dimensional space was sufficient for all the subpopulations. Therefore we will only present this solutions.

In a first step the total population of N=1517 allowing for a total of 15 "don't know" answers for each respondent is analyzed. The population is partitioned into 3 clusters of very different size, thus exhibiting a very simple degenerated structure . Cluster 1 contains those interviewees who respond either yes or no; cluster 2 and 3 are „don't know" clusters. The 3rd cluster is characterized by the "don't know" answers of item 5,6 and 7 , a result which is hardly interpretable (for the coordinates see Figure 1). It seems to be the very artifact described earlier in this paper, as there are much less "don't know" answers compared with the other items.

Table 1: Maximal 15 "don't know" answers (n=1517)

	λ = 217	059	
ASW	1.	2.	N
0.83	0.15	-0.05	851
0.54	0.13	0.10	596
0.78	-0.69	-0.28	70

ASW = average silhouette width λ = latent roots (each column represents one dimension)

Although the average silhoutte width seemed to be sufficient, the latent roots are fairly low indicating that only a little amount of dispersion is explained by the two dimensions.

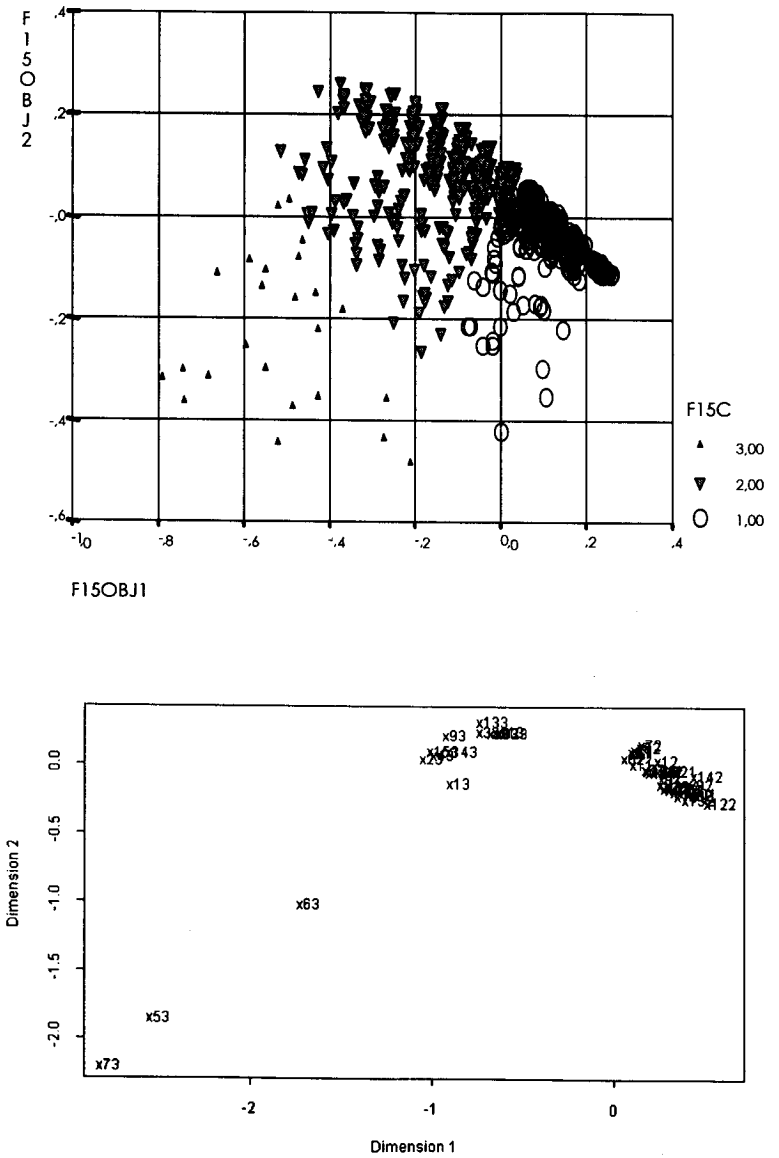


Figure 1 : Scatterplot of objects and coordinates for Max. 15 "don't know" answers

The pattern reported earlier changes when we only allow for a maximum of 6 "don't know" answers. Although a solution with 3 clusters in 2 dimensions is still sufficient , these clusters are characterized differently by the category - coordinates. The typical „don't know" cluster has vanished because these reactions become very similar to those categories which indicates a more "negative" or critical attitude towards the mental hospital. The cluster located mainly in the upper left quadrant of the plot is characterized by „yes" answers for the majority of the items and "don't know" answers for the more "negative" items. One cluster contains respondents saying „no" to the more negative items addressing the custodial aspect of the clinic (right lower quadrant of the plot), whereas respondents in the cluster located in the left lower quadrant either deny the more „positive" and liberal aspects or say "don't know" (see Figure 2). The population was reduced to 1215 observations. It is not surprising that all the "yes" answers coincide in one single cluster, since most of the questions are affirmed much more often than denied, so the "no" answers dominate the solution. Furthermore, also "don't know" answers are less frequently observed than "yes" answers but only those for the more "positive" characteristics of a mental hospital can be found in the same region as the "no" answers.

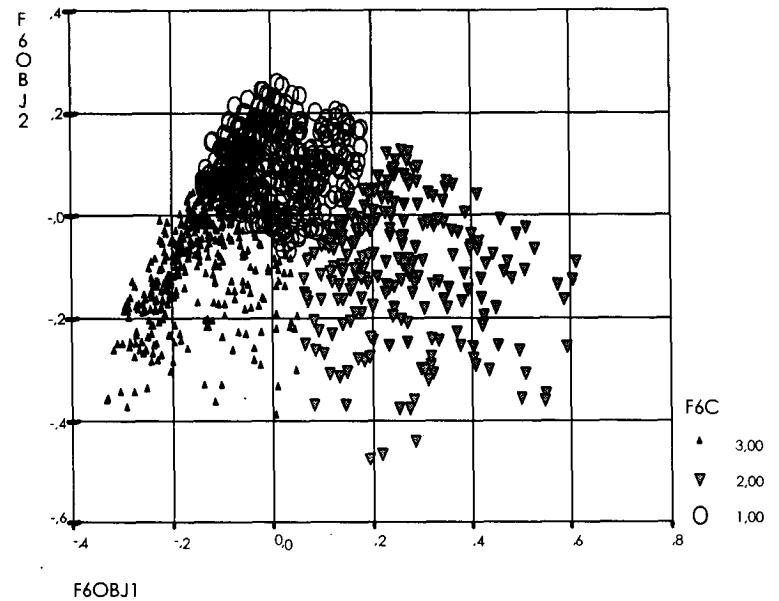
The eigenvalues for this solution are low but the average silhoutte widths are fairly sufficient allowing an interpretation of the 3 clusters solution. Furthermore the size of the clusters are more similar indicating that the solution does less capitalizing on rare pattern.

Table 2: Maximal 6 "don't know" answers (n=1215)

	λ = 119	064	
ASW	1.	2.	N
0.72	-0.01	0.11	583
0.47	0.26	-0.09	249
0.60	-0.16	-0.11	383

ASW = average silhouette width λ = latent roots (each column represents one dimension)

For all the further reduced subpopulations this basic pattern holds true, although some minor changes can be observed. Therefore a solution for max. 3 "don't know" answers (reducing the population to 808 observations) may serve as an example for all the solutions in-between.



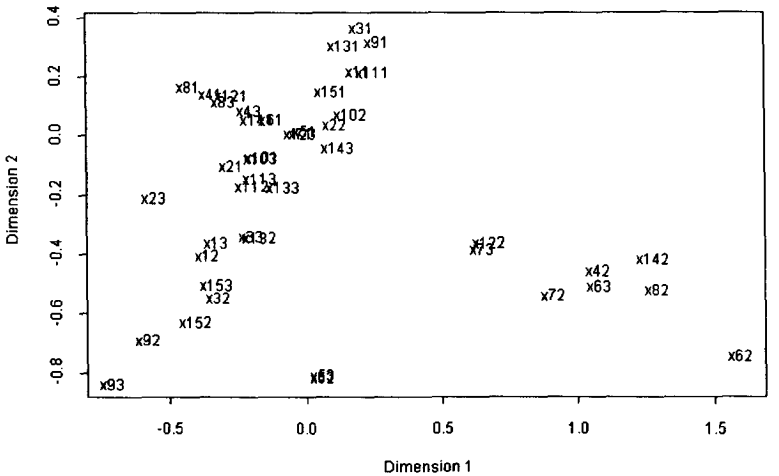


Figure 2: Scatterplot of objects and coordinates for Max. 6 "don't know" answers

We can easily see that the first and smallest cluster is again characterized by the denial of the more negative aspects of the clinic, such as: „restraining patients", „electroshock equipment"etc. The biggest cluster is located in the lower right quadrant of the plot showing the same artifact of either answering „yes" regardless of the content of the items or "don't know" for what we called before a "negative" characteristic. The third cluster, containing 230 persons, is mainly characterized by the response „no" to the more positive aspects of a clinic. Surprisingly the „no" answers regarding the different forms of treatment (client- centered therapy and pharmacological therapy) are contained in cluster 1 and not in different clusters, as might be expected from solutions presented earlier. In most of the solutions allowing for more then 3 „don't know" answers the „no" for client-centered therapy is located in the more"negative" cluster, whereas the „no" for medical treatment is located in the cluster indicating a more positive attitude toward the psychiatric clinic.

Table 3: Maximal 3 "don't know" answers (n=808)

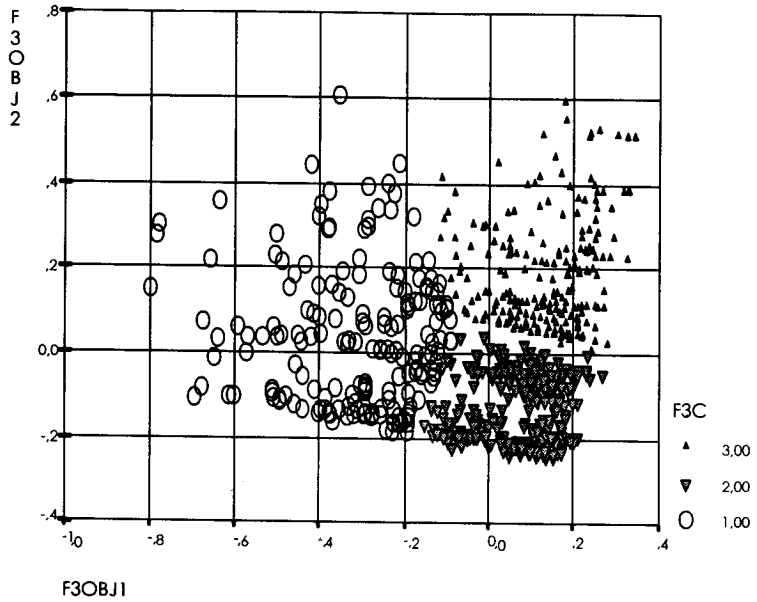
	λ = 114	072	
ASW	1.	2.	N
0.41	0.31	0.04	193
0.78	0.07	-0.14	385
0.52	0.14	0.20	230

ASW = average silhouette width λ = latent roots (each column represents one dimension)

The solution just presented exhibits sifficient silhouette-widths for each cluster but fairly low eigenvalues for the dimensions, indicating a low fit similar to solution allowing for more than 3 "don't know" answers .

5. Stability of results

Since all interpretations rely on the similarity of quantifications with respect to their projections on the two axes it is necessary to evaluate the stability of the results. The extraction of two dimensions seems sufficient at least with respect to the size of the latent roots. Investigating the similarity of coordinates by contructing stability regions sheds light on at least some of the systematic properties of the two dimensional solutions presented earlier. The result of the cluster analysis leads to the supposition that the two numerical "variables" representing the two dimensions stand for only three main groups of observations which can be described in a very parsimonious manner. Those groups have been characterized by those categories which are located within the boundaries of a particular cluster, as the object scores are proportional to the mean of the quantifications of those categories a particular object has chosen.



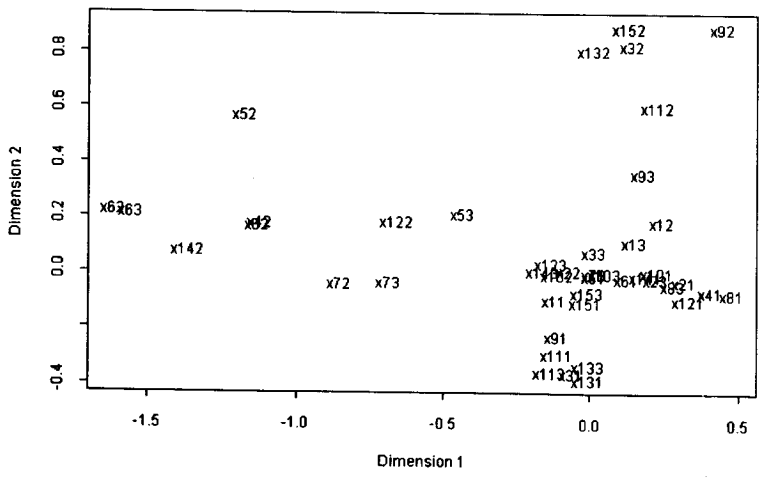


Figure 3: Scatterplot of the objects and coordinates for Max. 3 "don't know" answers

Hence it is of interest to investigate the stability in a sense that tiny changes in data should result in only tiny changes of the parameter estimations (Gifi, 1990; Markus, 1994). We do not focus on the calculations of confidence intervals or regions in order to make inferences from the sample to the population but rather look on the two-dimensional scatterplots of a series of solutions generated from bootstrap resampling (Efron and Tibshirani, 1993) . Contrary to the results recently published by Markus (1994) we calculated only 100 (and not 1000) **balanced** bootstrap samples (Hinkley, 1988) By this procedure each of the n observations are sampled equally often, so that each observation appears exactly B times in B bootstrap samples. This sampling procedure keeps the marginal totals for all categories proportional to the marginals of the original sample. Since we are less interested in making inferences a B=100 seems sufficient. For all calculations the program BOJA (Boomsma, 1990) has been used. This program applies the algorithm BB3 as described by Gleason (1988) which is supposed to be efficient when nB is "large" (larger than 50000, which is always the case in the applications presented here).

The distribution of the 100 replications of category quantification may serve as an indicator for the stability of the solutions. If the clouds for different category quantifications do not overlap the solutions may be considered quite stable. It will be shown that it is unnecessary to draw peeled convex hulls because the scatterplots can be evaluated easily without the polygon around the peeled convex hull.

As an example for the structure of stability the bootstrap results for a solution which allows for a maximum of 6 "don't know" answers are presented (Figures 4, 5 and 6). For the sake of saving space only the results for the first 6 items are presented, because these scatterplots will demonstrate the systematics of similarities between "don't know" answers and the other reactions which hold true for all the other items.

The matrixplots show distinctly that the denying of the more positiv aspects of a mental hospital or saying "yes" with regard to the negative characteristics cannot be separated from the "don't know" answers, because these categories are located within the same region of the plot. This very systematic pattern encourages the conclusion that a more critical appraisal of the mental hospital may not only result in particular patterns of "yes" and "no" answers for negative and positive characteristics respectively but also increase the probability for "don't know" answers. The results of the stability analysis facilitate the interpretation of the two dimensions and the three clusters of observations. They show that the observed systematics rather depend on the dimensions of the solution than on the restriction of the objects to three clusters. A mere inspection of the average silhoutte widths of the 100 bootstrap solution makes clear that also the 3-cluster solution is acceptable for all the solutions.

We should not forget that there is little knowledge about the bootstrap with this form of a non-linear analysis. Some validation studies were perfomed for non-linear canonical analysis (Van der Burg and De Leeuw, 1988) showing that results for the category quantification may be inaccurate and less precise, because they depend on the square root of the eigenvalue. Only little work has been done yet to clarify these problems.

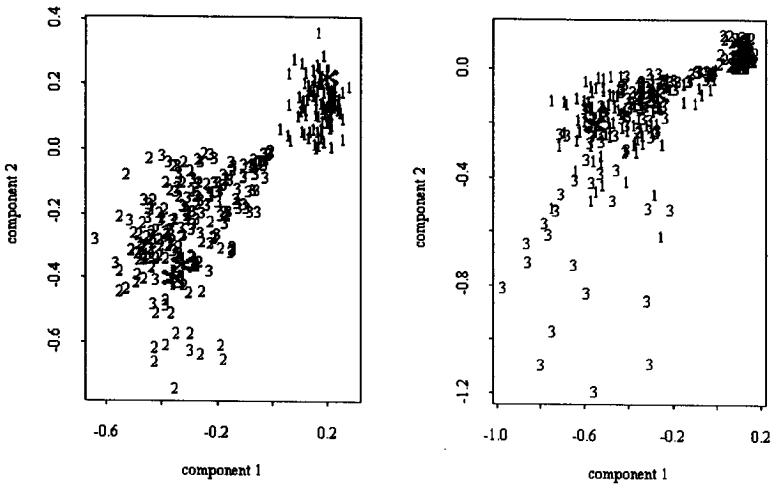


Figure 4: Matrixplot for variable 1 and variable 2

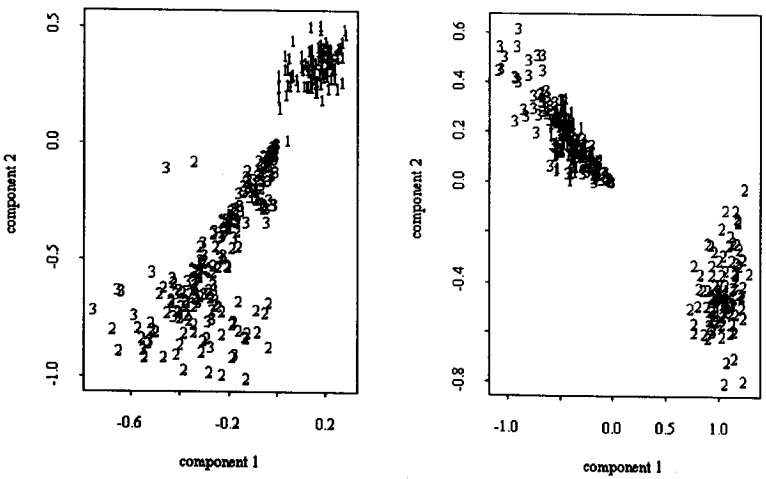


Figure 5: Matrixplot for variable 3 and variable 4

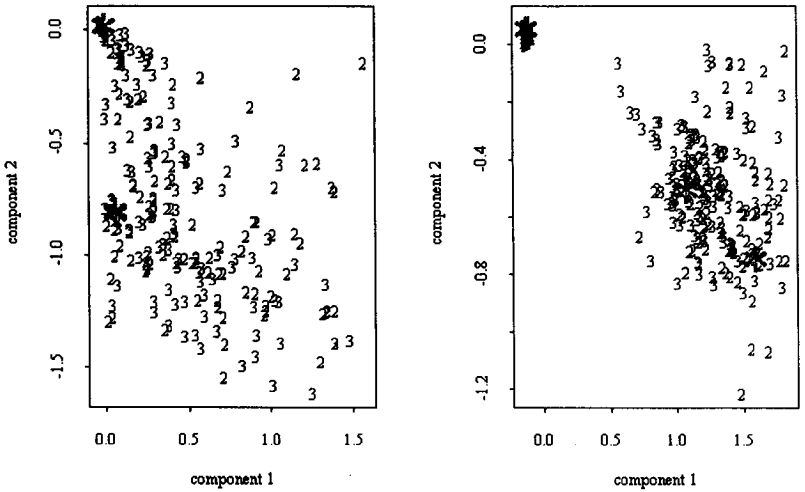


Figure 6: Matrixplot for variable 5 and variable 6

6. Discussion

Comparing the results for the different subpopulations we observe a relative stability with respect to the number of dimensions of the NLPCA, the number of cluster in the euclidean space and the meaning of the clusters. More complex solution show much smaller silhouette widths or result in well isolated but very small and hardly interpretable clusters.

It seems rather convincing to interpret two of the clusters as clusters of observations with either positive or negative attitudes toward the clinic, but we have to keep in mind that there is a strong asymmetry between "no" and "yes" answers, as the latter occurs much more frequently and therefor has much less power to discriminate between the observations. Hence these two clusters are mainly constituted by "no" answers. Critical attitudes toward the mental hospital increases the probability of "don't know" answers. The third cluster contains respondents exhibiting a mere „yeah-saying" artifact which is strongly connected with "don't know" answers for the "negative" characteristics. These clusters always contain about half of the sample under consideration. We have to assume that there exist only very rough imaginations about the domain and that there is no latent attitude or belief about the psychiatric clinic which determines to a certain extent the probability of

showing one of the three possible reactions. An inspection of the frequencies of the items for each subpopulation shows that there are more „yes" answers for all the items. For items 5, 6 and 7 only a very few „don't know answers are recorded. Usually people who either know nothing or only a little bit about a particular subject tend to adopt the convictions of the majority. The analysis presented supports the hypothesis that respondents having a certain attitude toward the mental hospital are forced to express it by answering "no" to particular questions. If this attitude deviates from the attitude of the majority as the mental hospital is seen more negatively, it turned out to be more likely to give a "don't know" answers.

References

[1] Bekker P. and De Leeuw J. (1988): Relations between variants of non-linear principal component analysis. In Van Rijckevorsel J.L.A. and De Leeuw J. (Eds.) *Component and Correspondence analysis*. Chichester Wiley.

[2] Boomsma A. (1991): BOJA - A program for bootstrap and jackknife analysis. Groningen iec ProGAMMA.

[3] Efron B. and Tibshirani R.J. (1993): *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[4] Jackson J.E. (1991): *A Users's Guide to Principal Components*. New York: Wiley.

[5] Jolliffe I.T. (1986): *Principal Component Analysis*. NewYork: Springer.

[6] Gifi A. (1990): *Nonlinear Multivariate Analysis*. New York: Wiley.

[7] Gleason J.R. (1988): Algorithms for balanced bootstrap simulations. *The American Statistician*, **42**, 263-266.

[8] Gordon A.D. (1981): *Classification*. London: Chapman & Hall.

[9] Gower J.C. (1971): A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857-874.

[10] Hartigan J.A. (1974): *Clustering Algorithms*. New York: Wiley.

[11] Hinkley D.V. (1988): Bootstrap methods (with discussion). *Journal of the Royal Statistical Society*, B, **50**, 321-337.

[12] Kaufmann L. and Rousseeuw P.J. (1990): *Finding Groups in Data*. New York: Wiley.

[13] Markus M. (1994): *Bootstrap Confidence Regions in Nonlinear Multivariate Analysis*. Leiden: DSWO-Press.

[14] Nishisato S. (1980): *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.

[15] Van Buuren S. and Heiser W. (1989): Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, **54**, 699-706.

[16] Van de Geer J.P. (1985): *HOMALS User Manual*. Leiden Departement of Data Theory (UG-85-2).

[17] Van Buuren S. (1986): *GROUPALS: A method to cluster objects for variables with mixed measurement levels*. RR-86-10, Leiden: Departement of Data Theory.

[18] Van der Burg E. and de Leeuw J. (1988): Use of the multinomial jackknife and bootstrap in generalized non-linear canonical correlation analysis. *Applied Stochastic Models and Data Analysis*, **4**, 159-172.

Appendix

What do you think is found in this hospital or in this mental hospital? To each point I am now going to list for you please tell me weather you believe that this is or isn't found in this hospital:

	yes	no	don't know	positiv negativ
1) Free passage for all patients				positiv
2) Patient uniform				negativ
3) Cafeteria				positiv
4) Straight jacket				negativ
5) Client- centered therapy				positiv
6) Closed wards				negativ
7) Pharmacological therapy				?
8) Rubber room				negativ
9) Recreation rooms for men and women				positiv
10) Pharmacological experiments				negativ
11) Comfortable double- rooms				positiv
12) Electroshock equipment				negativ
13) Patient hair-dresser				positiv
14) Restraining of patients				negativ
15) Exercise activities for patients				positiv