USE OF DISCRIMINANT ANALYSIS TO CLASSIFY PEOPLE WITH MENTAL DISABILITIES

Eric R. Langlet, Statistics Canada
R.H.Coats building, 15N, Tunney's Pasture, Ottawa, Canada, K1A 0T6

KEY WORDS: Canonical Discriminant Analysis, Nearest Neighbour Approach, Activity Limitations

1. Overview of the problem

The household component of the 1991 Health and Activity Limitation Survey (HALS) collects information on the nature and severity of disabilities, and information on the barriers which disabled persons face in the conduct of their daily activities. Two questionnaires are used, one for the adults aged 15 and over and one for the children aged 14 and under. This study is only using the adult part of the household component.

HALS sample was determined through a pre-selection process that took place as people filled their 1991 census form. Census questions 18 and 19, known as the activity limitation questions, were the basis for the selection of the sample. A sample of people who answered "YES" to these questions was taken with a high sampling fraction and a sample of people who answered "NO" was taken with a low sampling fraction to form the HALS 91 sample.

The HALS adult questionnaire first identifies in Section A the disabled population through a series of screening questions related to physical and mental disabilities. Screened-in individuals have to complete the follow-up portion of the questionnaire. The questionnaire identifies clearly the different types of physical disabilities but not the different types of mental disabilities. Questions on mental disabilities are also more prone to subjectivity than questions on physical disabilities. Theoretically, mental disabilities can be grouped in three categories. The first one is the "Mental Handicap" category (MH). A common example of this category would be someone with a certain degree of mental retardation. The second group is the "Learning Disability" category (LD). This is generally someone having difficulty in one or more specific learning areas. Dyslexia is an example of this. The third group is called the "Psychiatric Disability" category (P). This group concerns people having psychiatric conditions such as depression, schizophrenia, etc. Theoretically, there cannot be any overlap between the MH and LD group but there could be overlap between the MH and P group as well as between the LD and P group.

It is extremely difficult to obtain, from answers to the questionnaire, a deterministic rule to separate the three groups. We start with a universe of people with potential indication of mental disability. This is defined by screened-in individuals who have at least a "yes" to one of the mental disability questions in Section A of the questionnaire. Using suggestions given by different associations of mental disabilities, we were able to define deterministicly a core group of individuals in each category. We are fairly confident that each individual belonging to one of the three core groups is correctly classified. Unfortunately, these deterministic rules enable us to classify only a small portion of the individuals in the universe. Although not everybody in the universe should belong to one of the three groups, it is felt that a larger proportion of them should be classified in one of the three groups. We need to develop a method that will put into groups unclassified individuals of the universe that are "similar" to those already classified.

One possible solution to this problem is to use discriminant analysis. A wide number of explanatory variables in the questionnaire are available to characterize each of the three core groups. Using the core groups, we can develop a discriminant analysis rule that would classify individuals on the basis of their explanatory variables. We can then apply this rule to classify a portion of the unclassified individuals in the universe. People who will be "too far" from one of the groups will remain unclassified.

Section 2 of the paper describes the construction of the core groups as well as the explanatory variables. In Section 3, we show some results of a preliminary canonical discriminant analysis in order to reduce the number of variables in the model. In Section 4, we explore different discriminant techniques applicable to our problem and present results on the selected procedure. Finally, in Section 5 we give summary remarks and present some alternatives to resolve our problem.

2. Core groups and explanatory variables

The core groups were obtained using a protocol of classification based on suggestions given by different Canadian association of mental disabilities. We established a deterministic rule from answers to questions in the mental disability portion of the questionnaire which is a part of the screening section.

Using this protocol, there is no overlap between the MH and LD group but there is some overlap between the P and the MH group and between the P and the LD group. These intersections will be labelled

MH+P and LD+P. If we use the survey weights, the classification of the universe in the different groups is 45,000 for MH, 49,000 for MH+P, 154,000 for LD, 76,000 for LD+P, 359,000 for P and 1,897,000 for the unclassified individuals with sample sizes of 527, 827, 1409, 792, 3,088 and 13,487 respectively. This protocol classifies 26% of the universe.

We want to find some variables that would explain the differences or discriminate between the five core groups. For all individuals in our universe we have available all the variables in the screening portion of the questionnaire (Section A) as well as all variables in the follow-up portion of the questionnaire. In addition to variables collected in the survey itself, we also have available several variables from the 1991 census long form.

We decided to initially look at a large number of variables including some of the variables already used in the core group definition. As long as a variable does not define a group deterministicly, we can use it in the discrimination function.

First, a few classification variables were included: age, sex, marital status, education level, labour force status and whether the person was part of the "YES" sample or the "NO" sample.

Second, some physical disability variables were used, such as the level of difficulty for people to speak and being understood which is often a problem for the MH group and also a global physical disability score.

Third, a series of variables were taken from the mental health portion of Section A. Among those, we have questions on difficulties with certain academic learning areas, difficulties in specific day to day situations, symptoms of psychiatric conditions, diagnoses of specific psychiatric conditions by a health professional, etc.

Finally, a number of derived variables from the follow-up portion of the questionnaire were also used. This includes variables related to medication taken, intensity of pain and discomfort experienced by the person, level of dependency toward daily activities, number of visits over the last 3 months to mental health specialists/workers, whether the person discontinued his education because of a health problem, level of ability to read and write reported by the person, balance between positive and negative feelings using the Bradburn scale of emotional feelings (1969), etc.

Having created a number of explanatory variables, the next step was to compare the distribution of each variable between the 5 different groups of mental disability (MH only, MH+P, LD only, LD+P and P only). For this purpose and the rest of the analysis, unweighted data were used. The main reason for this is that, as opposed to do inference, we want to use discriminant analysis to determine the "geometric proximity" of the unclassified observations with respect to the classified observations. In fact, lower classification error rates were obtained using unweighted data.

From those distributions, we can see that individuals in the MH group are relatively young (unweighted mean of 30 years old), mostly single, fairly limited, have little education, report being fairly happy (based on perception question and based on the Bradburn scale of emotional feelings) and have difficulties in most learning areas.

Individuals in the LD group have the same age and sex characteristics as the LD group, they are mostly single also but some are married, they have more education, they feel less happy and they have difficulties in some but not most learning areas.

The P group is quite distinct from the other ones, having in general more education then the other groups, being older, usually married, using more medication, feeling worse in general and being the only group with proportionally more females. Their level of limitation is highly variable.

In general the combined groups LD+P and MH+P are more limited than the pure groups LD and MH in almost every respect. Also, the 2 MH groups are more limited than the 2 LD groups in general.

Very clearly the five groups are quite distinct from one another. The mixed groups MH+P and LD+P could not be combined with the pure groups since those individuals have in general much more problems. In fact, it seems unlikely that any individual not already classified would be similar to individuals in the mixed groups.

3. Canonical discriminant analysis

It is of course unadvisable to put all explanatory variables, many of them being highly correlated, in a "classical" discriminant analysis. In addition the choice of techniques to be used would be limited by the size of the problem. A possible approach would be to use a data reduction technique such as a principal component analysis (PCA). This method would find linear combinations of the original data that would best summarize the data in terms of total variance contained in the first principal components. This method does not take into account, however, the discriminant power of each variable in the linear combinations.

It would seem appealing to find a few linear combinations of the original variables that would best summarize between-class variation (differences between the groups of mental disability) as opposed to total variation in PCA. This is the purpose of canonical discriminant analysis. Having reduced the size of the problem, we could then use a discriminant analysis method on the linear combinations of the original variables.

The technique of canonical discriminant analysis first derives the linear combination that has the highest possible multiple correlation with the groups. This linear combination is called the first canonical variable. The second canonical variable is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the groups. The canonical variables are not, however, orthogonal. The process is repeated until the number of canonical variables equals the number of variables or the number of groups minus one, whichever is smaller (up to 4 canonical variables in our problem).

We shall now present some results. The R^2 coefficient between the first canonical variable and the class variable is 0.70 accounting for 62% of the total between class variation. The R^2 coefficient of the second canonical variable is 0.54 for a cumulative percentage of total between class variation of 93%. The R^2 for the third canonical variable is 0.18 contributing for only an additional 6% of the between class variation. The contribution of the last canonical variable is minimal and was discarded for the rest of the analysis.

In terms of interpretation, the first canonical variable is positively correlated with characteristics found in the MH and LD group. Since most characteristics are measures of limitation, the MH group will obtain a higher score on this canonical variable than the LD group. This canonical variable is also negatively correlated with variables like education, age, reading and writing ability, level of emotional feeling, etc, which have large values in the P group. This canonical variable tends to dissociate the three main groups.

The second canonical variable seems to measure the intensity of the limitation and should separate the MH+P group from the MH group and the LD+P group from the LD group. No easy interpretation of the third canonical variable was found.

In order to validate these interpretations, a plot of the first two canonical variables is presented in figure 1. Each observation is identified by the numbers 1 to 5 corresponding to MH, MH+P, LD, LD+P and P. The means of each group are also identified. We also projected the unclassified observations into the space of those canonical variables to obtain an idea of their similarity with respect to the different groups. In the top half of figure 1, only the core group observations are plotted and in the bottom half we added the unclassified observations identified by the number "0". In order to avoid excessive overprinting, only a sample of 1000 observations appears on each plot. The relative position of each group with respect to the other ones seems to confirm our interpretations. The unclassified observations are grouped in the bottom left corner, indicating that they will probably be grouped with the LD or the P group. Some other ones appear on the right side indicating that those individuals escaped from the MH definition but showed similar characteristics. The unclassified observations seem to be too far from the combined groups MH+P and LD+P to be classified into those groups.

4. Discriminant analysis on canonical variables

Several discriminant analysis procedures are available to classify individuals into groups based on the canonical variables obtained in the previous step. We restricted the possibilities to the techniques available in SAS (1990), namely the procedures available in PROC DISCRIM. When multivariate normal distribution applies within each group, parametric methods can be used with either linear or quadratic discriminant functions. Quadratic discriminant function are most suitable when the individual within-group covariance matrices are different from one another. When no assumption can be made about the distribution within each group, non-parametric methods can be used to first estimate group specific densities. These methods include the kernel and the k-nearest-neighbour methods. See for instance Rosenblat (1956) and Parzen (1962). The kernel approach, however, was not found suitable for our application due to excessive computer time required by this approach.

The k-nearest-neighbour method fixes the number, k, of training set points (observations already classified that were used to develop the model) for each multidimensional observation y to be classified. The method finds the radius $r_k(y)$ that is the distance from y to the k^{th} nearest neighbour in the training set point in the metric chosen. The k smallest distances are saved and, from those, let k_t represent the number of distances associated to group t and n_t the size of group t. The estimated group t density at y is

$$f_t(y) = \frac{k_t}{n_t v_k(y)}$$

where $v_k(t)$ is the volume of the ellipsoid bounded by

$$\left\{ z \mid (z-y)' V^{-1} (z-y) = r_k^2(y) \right\}$$

The matrix V is the pooled within-class covariance matrix which in our application is the identity matrix. This is because in canonical discriminant analysis, data are transformed so that the pooled within-class covariance matrix is the identity matrix. In this case, the Euclidean metric is used and the formula reduces to the volume of a p-dimensional sphere.

For all the methods, parametric or non-parametric, with the estimated group-specific densities and their associated prior probabilities, $q_t(y)$, we can evaluate the posterior probability estimates of group membership given the explanatory variables, $p(t|y)$, using Bayes' theorem:

$$p(t \mid y) = \frac{q_t f_t(y) \; / \; n_t}{\sum_u q_u f_u(y) \; / \; n_u}$$

Each observation is then classified in the group for which it has the largest posterior probability.

Several criteria were used to select a method. A simple criterion to evaluate the performance of a discriminant technique is to look at the classification error-rate estimate. To avoid any bias, we decided to use sixth seventh of the sample to develop the discriminant function (the training sample) and one seventh to test it (the test sample). The group of membership is known for each observation in the test sample. Therefore, from the test sample, we can calculate the number of cases incorrectly classified by the model to obtain an unbiased estimate of the classification error rate. The process was repeated using seven different training and test samples. We then calculated the means and the medians of the estimated classification error-rate over the different samples.

With respect to this criterion, the non-parametric approach performed slightly better than the parametric one. The overall classification error rate was around 20% for most values of k. The optimal value for k is difficult to determine. The larger the value of k, the smoother is the estimated density. However, excessive values of k will tend to favour the classification into the large groups to the detriment of the small groups, which is highly undesirable. Comparable results were obtained for values of k between 5 and 25.

A probably more important criterion is the performance of the different methods with respect to the newly classified observations. The new classified observations should have similar characteristics to the core group in which they were classified. Four different methods were compared using this method: the nearest neighbour methods with k=5, k=10 and k=25 and the normal method with quadratic distance function.

Each method was applied to the full set of 6,643 core group observations which will constitute the training sample. The next step is to classify a portion of the unclassified observations in our universe. In order to do this, we use the group specific density estimates obtained from the training set of observations, observe the explanatory variables for each observation to be classified and estimate the posterior probability of membership for each group. The observation is then classified according to the largest such probability.

If we put no restriction, all observations will be classified. One way to control this is to classify only observations for which the largest posterior probability of membership is larger than a specified threshold value. We can also use different threshold values for different groups based on the posterior probability distribution of the newly classified observations in each group. Fixing the threshold value for a group, we can look at the observations classified in that group and compare their characteristics with those of the core group. If characteristics are similar, we can stop, and if not, we can try to increase the threshold value and compare the characteristics again. Each time we increase the threshold value we obtain observations that are closer to the core group but we are also classifying fewer observations.

Using this criterion, we found that the normal method classified observations with characteristics of the P group in the LD group and vice-versa, confounding the two groups. With the nearest-neighbour approach, poor results were obtained for the MH group for k=5. Even though, this method classifies the maximum number of observations in the MH group, the characteristics of those observations are not those of the core MH group. The best compromise seems to be achieved with k=25 with a threshold p-value of 0.85 for the LD group, 0.5 for the P group and no threshold values for the MH group. With this empirical method, 179 new observations were classified in the MH group, 1,237 observations in the LD group and 1,592 observations in the P group.

In general newly classified individuals had slightly lower mental scores than the core groups but higher physical scores than the core groups. This is expected since most of the newly classified did not have a "yes" to the mental screening questions (most of those who had a "yes" to those questions were already part of the core groups) but had at least a "yes" to the physical screening questions since they were screened-in.

If we combine these new groups with the core groups, we would obtain weighted estimates of 57,000 for MH, 310,000 for LD, 688,000 for P and still 49,000 for MH + P and 76,000 for LD + P. The proportion of the universe that is classified is now 46%. The sizes of the LD and P group were multiplied approximately by 2, and, by 1.3 for the MH group. There is still 1,400,000 individuals who remain unclassified which does not seem unreasonable.

5. Summary remarks and conclusion

A protocol of classification was given based on suggestions from different associations of mental disabilities. This protocol partially classified the group of people with potential indication of mental disabilities. Five groups were obtained, that is the mental handicap group (MH), the learning disability group (LD), the psychiatric disability group (P) and two intersections between MH and P and between LD and P. We are fairly confident that each person classified in this preliminary step is in the appropriate group. Given these preliminary groups or core groups and a number of explanatory variables, we established a rule to classify individuals using the explanatory variables in a discriminant analysis. Two major steps were involved. The first step was to reduce the number of explanatory variables by taking linear combinations that best summarize the between-group variability, using canonical discriminant analysis. The second step was to create a rule that classifies individuals on the basis of their first three canonical variable scores, using a 25-nearest neighbour discriminant analysis method. This method classified an extra 20% of our universe. If we include the core groups, 46% of the universe is now classified. For the rest of them, we have an indication on which group they most resemble to but we do not have enough evidence to classify them in a specific group.

The method used has several limitations. The most important one is the fact that we are not totally sure that the core groups are correctly classified. In a typical discriminant analysis, these a priori groups are known exactly. Secondly, the canonical discriminant analysis is supposed to be performed on quantitative variables only. In our case, some of our variables are dummy variables, others are ordinal, others are interval variables and others are semi-continuous. It would have been possible to transform those variables using optimal and non-optimal transformation in the context of canonical discriminant analysis (van der Burg and de Leeuw, 11983). However, some of those transformations, especially the optimal ones, could be fairly difficult to interpret. Third, the choice of the "best" method to be applied is not obvious and different methods can produce very different results when applied to the unclassified observations. Fourth, if we put no restriction on the method, all the unclassified observations will be classified based on their largest posterior probability of membership. The selection of threshold posterior probability values for each group remains somewhat subjective. Finally, all methods used did not take the survey weights into account. The proper use of weights incorporating the sampling design of the survey into the methods remains an open problem. Nonetheless, we judge that the technique selected is a useful tool to classify individuals with characteristics similar to those in the core groups.

Other approaches could also be used. It would be possible to construct a logistic regression model to predict probabilities of membership in each group from a number of explanatory variables. The model could then be applied to the unclassified individuals. Again data reduction techniques should be used, and, it would also be conceivable to apply a logistic regression model on canonical variables. However, in an application where we try to classify individuals into groups, discriminant analysis appears to be a more natural and also a more flexible approach than logistic regression.

REFERENCES

Bradburn, N.M. (1969). *The Structure of Psychological well-being*. Chicago: Aldine Publishing Co.

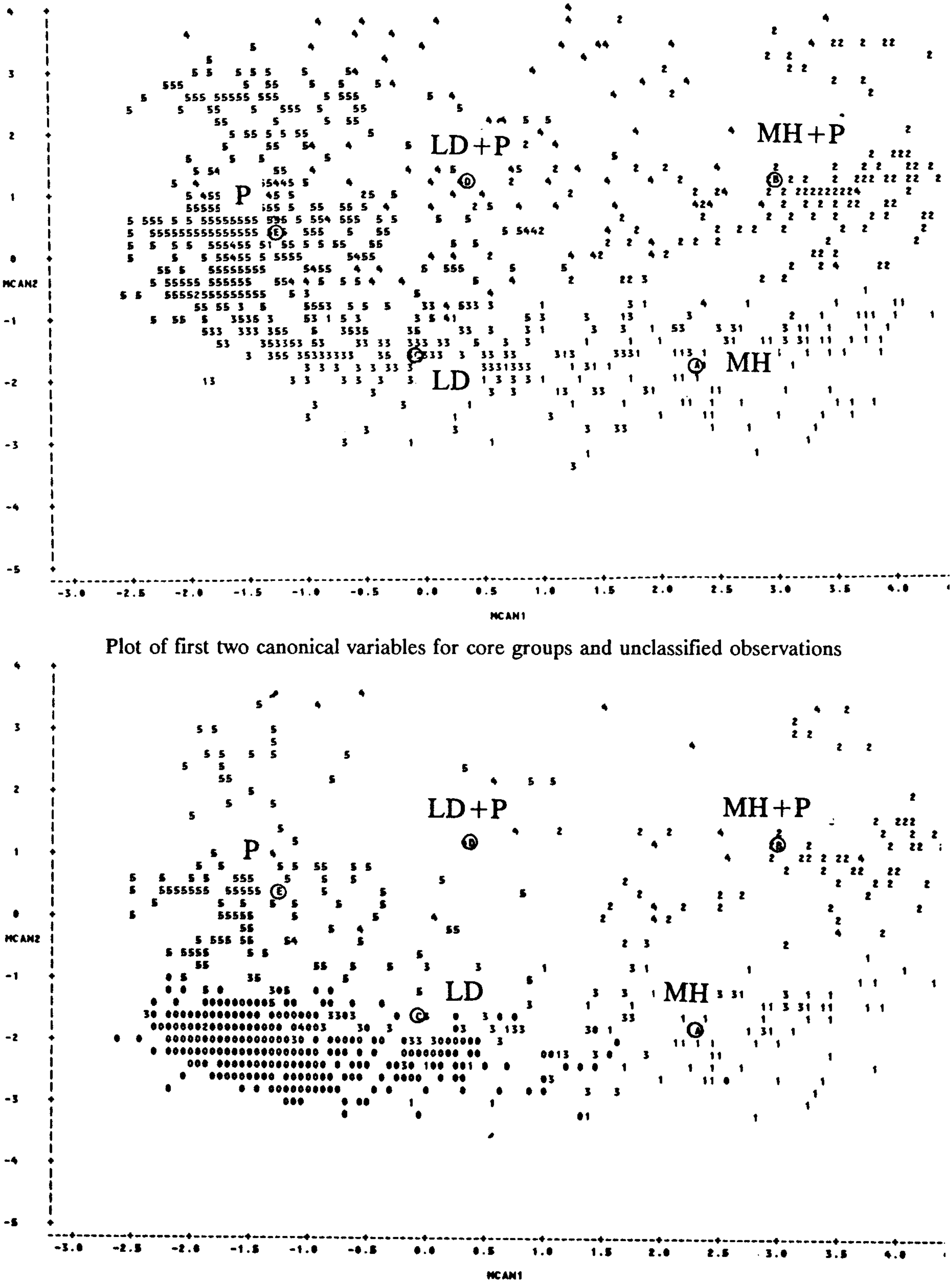
Parzen, E. (1962). On Estimation of a probability density Function and Mode. *Annals of Mathematical Statistics*, 33, 1065-1076.

Rosenblatt, M. (1956). Remarks on some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics*, 27, 832-837.

SAS Institute Inc. (1990). *SAS User's Guide: Statistics, Version 6, Fourth Edition, Volume 1 and Volume 2.*

van der Burg, E. and de Leeuw, J. (1983). Non - linear Canonical Correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54-80.

FIGURE 1



Plot of first two canonical variables for core groups

Plot of first two canonical variables for core groups and unclassified observations