KM 28(1988)
pag 137-148

FINDING GROUPS OF CONSUMERS WITH SIMILAR PREFERENCE WEIGHTS USING CLUSTERWISE LINEAR REGRESSION

Michel Wedel

and

Cor Kistemaker[1]

SUMMARY

In this paper an extension of the Clusterwise Linear Regression method is proposed, a FORTRAN computer program for the application to preference data was developed. The method solves the problem of finding a given number of clusters of subjects that attribute the same importance to perceived product dimensions in the formation of product preference. It is particularly useful when consumers' evoked sets of products are small and multicollinearity plays a role in fitting preference models at the individual level. An application to data on elderly people's preferences for meat products is given.

---

1 TNO-CIVO Institute of Toxicology and Nutrition
P.O. Box 360 3700 AJ Zeist, The Netherlands.

INTRODUCTION AND REVIEW

Consumer choice behaviour towards frequently purchased products is purposive, but the choice process is bounded by limitations of information on choice alternatives and time available for making choices. Not the objective product attributes are important in the choice process but the evaluative criteria of consumers based on the subjective perception of product attributes. In determining preferences for products consumers trade off the evaluative criteria (perceived product dimensions) against eachother.

An important issue in modelling the formation of consumers' preferences for products is the recognition of the influence of individual differences (Carroll 1972, Bettman 1979). In external analyses (Carroll 1972) perceived product dimensions are related to consumers' preference statements. The perceived product dimensions can be obtained by e.g. multidimensional scaling of product similarity data or factor analysis or partial least squares analysis of product ratings on attribute scales (Shepard 1972, Wierenga 1980, Martens and Martens 1986). Fitting of preference models at the individual level (using multiple linear regression of stated preferences on product dimensions) yields estimates of the importances of the product dimensions. In the analyses subjects are often grouped on the basis of the preference parameters with existing hierarchical or divisive methods to detect 'natural' segments within the population.

When a consumers evoked set of products is small, the model matrix in the regression may not be of full rank and the preference parameters are not estimable. This problem is frequently solved by deleting a sufficient number of terms from the model, so that the remaining parameters can be estimated. A serious disadvantage of this approach is that the estimates of the preference parameters in the reduced model are biased when the model terms excluded are good predictors of preference. As an alternate solution to the above problem Urban (1975) suggests to group the subjects (e.g. on the basis of their preferences) and to fit the model thought to be appropriate within groups across subjects and products. This procedure provides sufficient degrees of freedom in most cases, but it is based on the assumption that the segments are homogeneous with respect to the preference parameters (an assumption that cannot be checked).

Even if the individual model matrices are of full rank there are often few degrees of freedom for estimation which results in very unreliable estimates of the parameters (Mason et al. 1975). Grouping individuals on the basis of the estimated preference parameters as proposed by Urban and Hauser (1977) can then hardly be expected to yield meaningful clusters.

Apart from the solutions indicated above one might remove all subjects from the data with which singularity occurs, but this may result in a considerable reduction of the study population and affects the representativeness of the results.

The hierarchical clustering method Judgement analysis (JAN, Blottenberg and Christal 1968, Lutz 1977) also suffers from the problems outlined above. This method starts with single subject clusters and combines two previously defined clusters iteratively in such a way that the loss in overall predictive efficiency of the regression of preference judgements on attributes (measured by R^2) is minimized. It was shown by Adler and Kafry (1980) that JAN is identical to well known clustering techniques applied to the predicted individual preferences and thus equivalent to a clustering based on the estimated preference weights.

Spath (1979, 1982) described a method for clusterwise linear regression for solving the problem of finding a given number of clusters of observations, such that the overall error sum of squares of the regression within clusters is minimal. In this paper we propose an extension of this method. It yields maximum likelihood estimates of the grouping factor that partitions subjects into segments with homogeneous preference parameters (given the number of segments present). The method circumvents the problems connected with multicollinearity in the estimation of the individual preference parameters and thus permits preference models to be fitted which would be severely overparameterized at the individual level.

THE METHOD

Let:

y_i = the (J_ix1) vector of preference scale values of individual i for J_i products (i=1...n) ,

X_i = the (J_ix P) model matrix for individual i, the P columns are functions of the product dimensions, depending on the model that is appropriate for the analysis ,

β_i = the (Px1) vector of preference parameters of individual i.

Consider an external analysis of the preference data in which the measured preferences y_i are related to the product dimensions. We consider the following model:

$$(1) \qquad y_i = X_i\beta_i + \varepsilon_i \,, \qquad \varepsilon_i \sim N(0, I\sigma_i^2) \,, \qquad i = 1...n \,,$$

where ε_i is a vector of error terms. Consider the case where the evoked set of products J_i is so small for some i, that $J_i < P$. Now the coefficients β_i are not estimable.

Assume each of the y_i to arise from one of K groups, assume K known and assume the parameter vectors β_i of the individual preference models to be the same for the set of n_k individuals (denoted by C_k) in group k. ($\beta_i = \beta_k$, $i \in C_k$). Now let :

N_k = the number of observations in segment k, $N_k = \sum_{i \in C_k} J_i$,

y_k = the (N_kx1) partitioned vector of preference scale values of individuals in segment k, consisting of the n_k subvectors y_i , $i \in C_k$,

X_k = the (N_k x P) partitioned model matrix, consisting of the n_k submatrices X_i, $i \in C_k$,

β_k = the (Px1) vector of preference parameters in segment k.

If $\sigma_i = \sigma_k$ for all $i \in C_k$, the model for the N_k observations in group k is:

$$(2) \qquad y_k = X_k\beta_k + \varepsilon_k \,, \qquad \varepsilon_k \sim N(0, I\sigma_k^2) \,, \qquad k = 1...K \,,$$

If $P < N_k$ in each of the k groups, the X_k are of full rank P and the β_k are estimable given the partition of subjects into K segments.

The objective is to find the partition of the subjects into the K groups. Analogous to Scott and Symons (1971) the maximum likelihood estimate of the grouping factor γ, can be derived ($\gamma = \gamma_1 \ldots \gamma_n$; $\gamma_i = k$ for $i \in C_k$). Assuming the partition of subjects into K segments known, the log-likelihood function for the parameters β_k, σ_k is given by:

$$(3) \qquad -1/2 \sum_{k=1}^{K} (y_k - X_k\beta_k)'(y_k - X_k\beta_k)/\sigma_k^2 - 1/2 \sum_{k=1}^{K} N_k \ln(\sigma_k^2)$$

For each possible partition into K groups the likelihood is maximized by the ordinary least squares estimates of β_k and σ_k, $\hat{\beta}_k$ and $\hat{\sigma}_k$. It follows by substitution of $\hat{\beta}_k$ and $\hat{\sigma}_k$ in (3) that the M.L. estimate of γ is that grouping of the n subjects into K groups that minimizes:

$$(4) \qquad \prod_{k=1}^{K} \hat{\sigma}_k^{2(N_k/N)}$$

where $N = \sum_{k=1}^{K} N_k$. If $\sigma_k = \sigma$ for all k, the M.L. estimate of the grouping factor is that grouping that minimizes $\hat{\sigma}$, the pooled residual mean square (RMS) of the regressions within the K groups (which is equivalent to maximizing R^2). For $N_k = 1$ and $\sigma_k = \sigma$ the clusterwise regression problem of Spath (1978) minimizing the L_2 norm is obtained.

Although estimates of the parameters with clusterwise regression are Maximum Likelihood estimates the asymptotic properties do not apply, the number of parameters estimated being close to the number of observations. As the distribution of the minimum RMS is unknown the significance of the regressions within clusters cannot be tested with the usual t- and F- tests, but simplified Monte Carlo significance tests could be used. In Monte Carlo test procedures the outcome of the test is determined by the rank of a statistic derived from the observed data, relative to the values of that statistic derived from random samples (the reference set). The reference set is generated in accordance with the hypothesis being tested. For the simplified Monte Carlo test the reference set consists of M -1 samples and the nulhypothesis is to be rejected if the testcriterion from the observed data is greater (or less) than $M-M(\alpha/2)$ or more of the values from the reference set (α is the level of significance of the two-sided test, and both M and $M(\alpha/2)$ are integers). The power of the test increases with M (Hope 1968). The significance of the regressions within clusters obtained with Clusterwise regression can be tested with e.g. a Monte Carlo permutation test, in which the reference set is obtained by permuting the observed preference scores randomly among products for each subject.

The partition minimising the criterion (4) can be found by comparing all possible partitions of subjects into K segments. For large numbers of subjects the computational time required would be excessive, and therefore in practice, the transfer algorithm of Banfield and Bassil (1977) can be used to obtain the partition. This algorithm starts from a given classification of the individuals, tests each possible transfer of one subject and swop of two subjects, and executes them (if they improve the criterion value) until no more improvement on the criterion (4) can be made. When the Banfield and Bassil algorithm is applied to minimize (4), transfers from clusters with P observations or less are not permitted.

The Banfield and Bassil algorithm was implemented in a FORTRAN computer program RMSCLUST by the authors. The program starts from a random or preset classification. The criterion (4) can be minimized both with and without the assumption of equal σ_k. As with most divisive methods, optimal classification found may not be unique and it may not be a global optimum. Banfield and Bassil suggest to use a classification with greater than K clusters as a starting point and to work down to the desired K clusters to help avoiding local optima, or else to use a set of different random starting classifications. The RMSCLUST program can start from a given number of clusters and automatically works down to the given number of desired clusters. The program output includes monitoring of the numbers of transfers and swops tried and executed as well as the changes in the criterion values during the iteration process. An option for generating a reference set for a Monte Carlo permutation test is included in the program ($J_i > 5$; $i=1...n$).

AN APPLICATION

In a nationwide random sample of 199 subjects ranging from 65–80 years of age, data were collected on preferences and perceived product characteristics with respect to 11 different meat products (Wedel et al.

1986). Subjects were asked to rank photographs of the meat products in order of their preference. Preference orders were then rescaled to range from 0, corresponding to lowest preference, to 1, corresponding to highest preference. 20 Attributes were evaluated and reduced to four perceptual dimensions, using factor analysis (see e.g. Hauser and Koppelman 1979). These dimensions, labeled quality, fatness, exclusiveness and convenience respectively, were to be related to preferences to obtain estimates of the preference weights according to a linear model[2]. Multicollinearity did not play an important role in fitting the linear model at the individual level, as the number of observations (J_i=11) is substantially larger than the number of parameters (P=5) to be estimated. Clusterwise regression was applied to these data (n=199). The algorithm was started with a random classification of 8 clusters and worked down to 2 final clusters, minimising the RMS. 50346 Transfers were tried and 853 executed, while 122627 swops were tried and only 3 executed (this required 7234 CPU seconds on a VAX 11/750). Figure 1 shows a plot of the RMS of the cluster solutions against the number of clusters. As for more than 5 clusters the RMS hardly decreased with increasing number of clusters, we decided that the 5 cluster solution gave a sufficient approximation to the data for our purposes. The five cluster solution had a RMS of 4.69. Clusterwise regression was applied 25 times more with different random starting classifications, dividing the same data into 5 clusters. The resulting RMS's found were inbetween 4.69 and 4.77. Five times a solution with a RMS of 4.69 was found, three times the solution was identical to the one described above, two times a different solution was obtained in which only 9 subjects were classified differently. Consequently the classification found does not seem to be a local optimum, but it appears to be a rather flat global optimum.

---

2 The assumption of independence of the error terms is not tenable when rank ordered data are used as a dependent variable at the individual level. In the present data, however,ties were permitted, and when models are fitted within groups across several individuals, the assumption will be more nearly met. In fact the parameter estimates have the ML properties if the error terms are independent and have fixed variance for larger numbers of observations.

FIGURE 1 - POOLED RESIDUAL MEAN SQUARE (RMS) OF CLUSTERWISE REGRESSION SOLUTIONS AGAINST THE NUMBER OF CLASSES

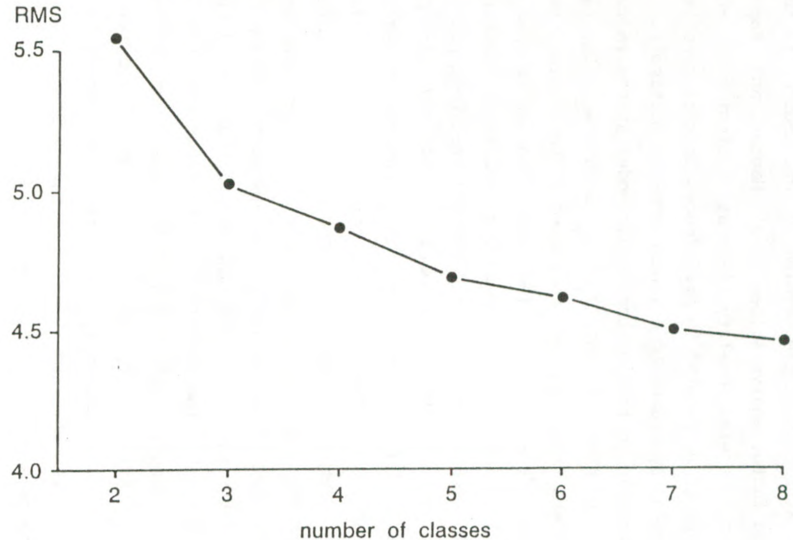

Table 1 shows the coefficients, estimated across subjects and products within the clusters obtained. To evaluate the significance of the regression coefficients 40 datasets were generated by permutation of the preference scores within individuals, and 199 values of the t-statistic were obtained for each dimension (40 times 5 segments, one deleted at random). The 2.5 th and 97.5 th percentiles of the distribution of the t-values of the reference set are given in table 1. The effects of quality in all segments, fatness in segments 3 and 4, exclusiveness in segment 5, and convenience in segment 1 are significant at the 5% level by the simplified Monte Carlo test procedure.

Table 1. Preference weights and t-values* of four perceptual dimensions of meat in 5 segments obtained with clusterwise regression.

Segment:		1	2	3	4	5
n:		37	32	56	33	41
Quality						
estimate		0.30	0.14	0.20	0.18	0.19
t-value	(-8.3; 7.5)	16.0	15.9	22.0	16.4	17.5
Fatness						
estimate		-0.05	-0.04	-0.08	-0.12	0.04
t-value	(-7.1; 7.3)	-4.4	-4.0	-9.4	-9.9	3.6
Exclusiveness						
estimate		0.04	0.03	0.07	-0.09	-0.10
t-value	(-8.6; 8.4)	3.6	2.3	7.2	-7.5	-10.8
Convenience						
estimate		0.10	0.05	0.03	0.05	0.05
t-value	(-7.8; 8.4)	8.5	4.9	3.6	4.8	5.2

* The 2.5 th and 97.5 th percentiles of the distribution in the reference set are given in brackets.

For comparison a clustering algorithm was applied to the preference weights estimated at the individual level, minimizing the determinant of the pooled within group covariance matrix (the det(W) method yields M.L.

estimates of the grouping parameters, given the number of segments and assuming the preference weights to follow a Multivariate Normal distribution with identical Covariance matrices within segments, Scott and Symons 1977). This method is implemented in the statistical package GENSTAT (Alvey et al, 1977), and also uses the Banfield and Bassil algorithm. The partition obtained with clusterwise regression was used as a starting classification. The percentages of subjects remaining in the clusterwise regression solution was 40% (for clusters 1 to 5: 40, 44, 41, 47 and 22% respectively). The det(W) method yielded (about 5-10%) smaller standard deviations of individual coefficients within segments, while the clusterwise regression solution had greater predictive efficiency: the R^2 for the regression across subjects and products was 0.367 for the det(W) solution, while R^2 was 0.524 for the clusterwise regression solution.

Subsequently, a random sample of 5 products was taken for each subject and the data on preferences and perceived product characteristics subjected to clusterwise regression, using the final five cluster solution obtained from the full dataset as an initial classification. The percentage of subjects remaining in the initial classification was 57% (for clusters 1 to 5: 71, 54, 42, 41 and 84% respectively). The R^2 was 0.616 . Methods clustering the individual estimates of the parameters could not be applied for comparison as no reliable estimates of individual preference weights could be obtained from the reduced dataset.

CONCLUDING REMARKS

The clustering method described in this paper can be used when an overparameterized preference model is to be fitted to individual data. It was shown that if in the subsequent analyses individuals are to be clustered on the basis of their estimated preference parameters, clusterwise regression can be applied to obtain a clustering of individuals, circumventing the problem of multicollinearity at the individual level. Clusterwise regression is a divisive method and as such burdened with problems of convergence to a unique global optimum.

In practice the assumption that coefficients for subjects within the same segment are identical might be somewhat unrealistic. If the preference parameters within each segment are assumed to follow a multivariate Normal distribution, clusterwise regression can be applied and least squares estimates of the mean preference parameters and the error variance can be shown to be unbiased, however they are not Maximum Likelihood estimates.

We conclude that the clusterwise regression is useful as an exploratory data analysis technique, especially if multicollinearity plays a role in fitting regression models at the individual level.

Acknowledgements. The authors gratefully acknowledge the support of the Minestry of Agriculture and Fisheries for supplying the funds for this reasearch project, Mrs. K.F.A.M. Hulshof and Drs. Ir. M.R.H. Lowik for their efforts in data collection, Ir. J.T.N.M. Thissen and Drs. E. Schoen for their valuable comments, and Mrs. E. Aarnink for assistance in dataprocessing.

REFERENCES

Adler, I. and P. Kafry (1980). "Capturing and clustering judges policies," Organizational Behavior and human performance 25 (3), 384-394.

Alvey et al. (1977). "GENSTAT a general statistical program," Rothamsted experimantal station: Harpenden, England.

Banfield, C.F. and L.C. Bassil (1977). "A transfer algorithm for non-hierarchical classification," Applied Statistics, 26 (Algorithm AS113), 206-210.

Bettman, J.R. (1979). "An information processing theory of consumer choice," Addison Wesly, Massachusets.

Blottenberg, R.A. and R.E. Christal (1968). "Grouping criteria, a method which retains maximum predictive efficiency," The journal of experimental education 36, 28-34.

Carroll, J.D. (1972). "Individual differences and Multidimensional scaling," Multidimensional scaling, R.N. Shepard, A.R. Romney, S.B. Nerlove (eds.) Seminar Press: New York.

Hauser, J.R. and F.S. Koppelman (1979). "Alternative perceptual mapping techniques: relative accuracy and usefulness," Journal of Marketing Research 16, 495-506.

Hope, A.C.A. (1968). "A simplified Monte Carlo significance test procedure," Journal of the Royal Statistical Society, Series B 30, 582-598.

Lutz, J.G. (1977). "The multivariate analogue of JAN," Educational and Psychological Measurement 37, 37-45.

Martens, M. and H. Martens (1986). "Partial Least Squares," Statistical Procedures in Food research, J.R. Piggott (ed.). Elsevier applied science: London.

Mason, R.L., R.F. Gunst and J.T. Webster (1975). "Regression analysis and problems of multicollinearity," Communications in Statistics 4 (3), 277-292.

Scott, A.J. and M.J. Symons (1971). "Clustering methods based on Likelihood ratio criteria," Biometrics 27, 387-397.

Shepard, R.N. (1972). "A taxonomy of some principal types of data and of multidimensional methods for their analysis," Multidimensional scaling, R.N. Shepard, A.R. Romney, S.B. Nerlove (eds.). Seminar Press: New York.

Spath, H. (1979). "Algorithm 39, Clusterwise Linear Regression," Computing, 22, 367-373.

Spath, H. (1982). "Algorithm 48, A fast algorithm for Clusterwise linear regression," Computing, 29, 175-181.

Urban, G.L. (1975). "Perceptor, a model for product positioning," Management Science 8, 858-871.

Urban, G.L. and J.R. Hauser (1977). "A normative methodology for modelling consumer response to innovation," Operations Research 4, 579-617.

Wedel, M., K.F.A.M. Hulshof and M.R.M. Lowik (1986). "Determinants of choice behavior of elderly people in the Netherlands with respect to meat products," Contemporary research in marketing, proceedings of the XV-th annual conference of the European Marketing Academy: Helsinki 1986.

Wierenga, B. (1980). "Multidimensional models for the analysis of consumer's perceptions and preferences with respect to agricultural and food products," Journal of agricultural Economics 1, 83-97.

Ontvangen: 30-09-1987
Geaccepteerd: 18-12-1987