Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 154 (2004) 779–786
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

Interfaces with Other Disciplines

Market segmentation by maximum likelihood clustering using choice elasticities

Harald Hruschka [a,*], Werner Fettes [b], Markus Probst [c]

[a] *Department of Marketing, University of Regensburg, Universitätsstrasse 31, D- 93053 Regensburg, Germany*
[b] *debis Systemhaus GEI, Munich, Germany*
[c] *msg Systeme, Ismaning, Germany*

Received 23 May 2001; accepted 23 September 2002

Abstract

We determine market segments by clustering households on the basis of their average choice elasticities across purchases and brands w.r.t. price, sales promotion and brand loyalty. The cluster analysis technique used is a maximum likelihood method which allows varying size and orientation and assumes constant volume. Elasticities originate from choice models with alternatively linear and nonlinear utility functions. Choice models are estimated on the basis of household scanner data. Segments are interpreted by means of multiple discriminant analysis and multinomial logit models whose predictors are elasticities of predictors and external variables (i.e. number of purchases, number of brands bought, income and household size), respectively.
© 2002 Published by Elsevier B.V.

Keywords: Marketing; Market segmentation; Cluster analysis; Choice models

1. Introduction

The main focus of this paper is the determination of market segments based on household-specific choice elasticities by means of a maximum likelihood clustering technique. Households are heterogeneous w.r.t. choice elasticities, i.e. relative changes in choice probabilities for brands of a product group divided by relative changes in marketing instruments and other predictors which influence brand choice. Homogeneous choice models (i.e. models with constant parameters across households) are able to reflect this kind of heterogeneity if some of the predictors vary across households. [1]

Section 2 introduces household-specific elasticities and describes the maximum likelihood clustering technique which uses these elasticities as segmentation criteria. Section 3 presents choice models with linear and nonlinear utility functions together with the computation of average choice

* Corresponding author. Tel.: +49-941-943-2279/2277; fax: +49-941-943-2278.
E-mail address: harald.hruschka@wiwi.uni-regensburg.de (H. Hruschka).

[1] This research is supported by a grant of the Deutsche Forschungsgemeinschaft (DFG).

0377-2217/$ - see front matter © 2002 Published by Elsevier B.V.
doi:10.1016/S0377-2217(02)00807-X

elasticities w.r.t. price, sales promotion and brand loyalty on the basis of estimated parameters. Section 4 contains results of an empirical study. It gives estimation results of choice models and concentrates on interpreting market segments discovered by the cluster analysis method. Discriminant analysis serves to assess the importance of elasticities for the formation of segments. Moreover, we estimate multinomial logit models to explain segment membership by means of external household descriptors (number of purchases, number of brands bought, income and household size).

2. Household heterogeneity and maximum likelihood clustering based on choice elasticities

We use average household-specific choice elasticities as clustering variables. Choice elasticities are defined by the ratio of the relative change in choice probability and the relative change in the value of a certain predictor. Predictors considered are reference price, loss, gain, display, feature and loyalty (for more details see Section 3). More exactly, the point elasticity of the ith predictor $c_{x_i,h,m,t}$ (with $i = 1,\ldots,6$) for household h, brand m at purchase occasion t w.r.t. choice of brand m can be written as:

$$\frac{\partial p(y_{h,m,t}|B,\{\vec{c}_{x,h,t}\})}{\partial c_{x_i,h,m,t}}\,\frac{c_{x_i,h,m,t}}{p(y_{h,m,t}|B,\{\vec{c}_{x,h,t}\})} \tag{1}$$

$y_{h,m,t}$ denotes a binary purchase indicator (equal to one if household h purchases brand m at occasion t, else zero), $\vec{c}_{x,h,m,t}$ the vector of predictors of brand m at occasion t of household h, $\vec{c}_{x,h,t}$ the set of predictors of all brands at occasion t of household h. Vector B consists of the parameters of one of the choice models described in Section 3.

For each household we compute the arithmetic mean of the elasticities of each predictor across all brands and all purchases of this household. Elasticity E_{ih} symbolizes the average elasticity of household h for predictor i. Elasticities referring to a household are collected in a household-specific elasticity vector $E_h = (E_{1h},\ldots,E_{6h})$.

Heterogeneity of households w.r.t. elasticities can be measured by the total sum of squares TSS of elasticities E_{ih} ($\overline{E}_i$ is the average elasticity of predictor i over the total number of households denoted by H):

$$\mathrm{TSS} = \sum_{i=1}^{6}\sum_{h=1}^{H}(E_{ih}-\overline{E}_i)^2 \quad \text{with} \quad \overline{E}_i = \frac{1}{H}\sum_{h=1}^{H}E_{ih} \tag{2}$$

Households belonging to the same market segment should respond to changes in predictors in a similar way, i.e. possess similar average elasticities for the predictors considered. On the other hand, it should be feasible to discriminate between different market segments (i.e. they should have different elasticity vectors). We search for a fixed partition of households.

Assuming multivariate normal parameters of subpopulations (mean vector ϵ_k, covariance matrix $\sum_k$) cluster memberships are determined by maximizing the likelihood (Banfield and Raftery, 1993):

$$\text{const.}\quad \prod_{k=1}^{G}\prod_{h\in I_k}\left|\sum_k\right|^{-1/2} \times \exp\left(-1/2(E_h-\epsilon_k)'\sum_k^{-1}(E_h-\epsilon_k)\right) \tag{3}$$

G denotes the number of clusters, I_k the index set of households belonging to cluster k and H_k the number of households in cluster k.

We replace ϵ_k by its maximum likelihood estimate, the average elasticity vector EC_k across households belonging to cluster k which is defined as follows:

$$\mathrm{EC}_k = \frac{1}{H_k}\sum_{h\in I_k}E_h \tag{4}$$

This leads to the concentrated log likelihood which is equivalent to expression 3:

$$\text{const.} - 1/2\sum_{k=1}^{G}\left(\mathrm{tr}\left(W_k\sum_k^{-1}\right)+H_k\log\left|\sum_k\right|\right) \tag{5}$$

W_k is the cross-product matrix for cluster k:

$$W_k = \sum_{h\in I_k}(E_h-EC_k)(E_h-EC_k)' \tag{6}$$

Moreover, maximizing log likelihood is equivalent to minimizing the following expression:

$$\sum_{k=1}^{G} H_k \log|W_k/H_k| \tag{7}$$

The eigenvalue decomposition of the covariance matrix is:

$$\sum_k = D_k \Lambda_k D_k' = \lambda_k A_k D_k' \tag{8}$$

D_k denotes the matrix of eigenvectors, Λ_k the diagonal matrix of eigenvalues, λ_k the maximum eigenvalue. D_k determines the orientation, λ_k the size (volume) and A_k the shape of cluster k.

The eigenvalue decomposition of the cross-product matrix is:

$$W_k = L_k \Omega_k L_k' \tag{9}$$

Applications of the most general model with different cluster-specific size, orientation and shape matrix suffer from its lack of parsimony (Symons, 1981). Therefore we use an intermediate model with different size and orientation, but constant shape matrix A. For this model criterion 7 can be simplified to objective S^*:

$$S^* = \sum_{k=1}^{G} H_k \log(S_k/H_k) \quad \text{with } S_k = \operatorname{tr}(A^{-1}\Omega_k) \tag{10}$$

We apply an agglomerative hierarchical clustering algorithm to find local minima of S^* and decide on the number of clusters r for which the approximate weight of evidence F_r is maximum. This measure is based on the Bayes factor (the ratio of posterior to prior odds) for $G = r$ against $G = s$. It is defined as follows:

$$F_r = \begin{cases} 0 & \text{for } r = 1 \\ \sum_{t=1}^{r-1} f_t & \text{for } r \geqslant 2 \end{cases} \tag{11}$$

with

$$f_r = 2(l_{k'} + l_{k''} - l_k - 3/2 + \log(pH_{r,r+1})2\delta_r$$

$l_{k'}$, $l_{k''}$, l_k denote the maximized log likelihoods for the clusters k', k'' that are merged and the cluster k resulting from merger during each step of the hierarchical algorithm, respectively. p is the number of elements in the elasticity vector, $H_{r,r+1}$ the number of observations in the merged cluster and δ_r the decrease in the number of parameters by merging clusters.

For the number of clusters chosen this way the classification obtained by the hierarchical algorithms is changed by iterative relocation, i.e. moving households from one segment to another if this improves criterion S^*.

3. Choice models and computation of elasticities

Brand choice models are based on the assumption that consumers purchase that brand out of a choice set which they perceive to have maximum utility. Utility is conceived to be additively made up of a deterministic component and a random term. Assuming each household chooses the brand perceived to have the largest utility which is formed by adding an iid type I extreme value distributed error term to a deterministic component leads to the multinomial logit model (McFadden, 1973; McFadden, 1980; Corstjens and Gautschi, 1983). According to this model the conditional choice (purchase) probability of brand m at occasion t by household h is:

$$p(y_{h,m,t}|B, \{\vec{c}_{x,h,t}\}) = \frac{\exp(V_{h,m,t})}{\sum_{m^* \in \mathscr{M}_{h,t}} \exp(V_{h,m^*,t})} \tag{12}$$

$V_{h,m,t}$ denotes the deterministic utility of brand m at occasion t for household h, $\mathscr{M}_{h,t}$ the set of brands available at the outlet visited by household h at occasion t.

For the predominant linear specification (to be brief we refer to it as MNL model in the following) deterministic utility is written as:

$$V_{h,m,t} = \vec{\beta}_1 \cdot \vec{c}_{x,h,m,t} + \vec{D}_m \cdot \vec{\beta}_3 \tag{13}$$

$\vec{D}_m$ is a $M-1$ dimensional vector of zero-one dummy variables (M is the number of brands analyzed) which only for brand $m > 1$ attains the value one, i.e. $\vec{D}_m = (0, \ldots, 0, 1, 0, \ldots, 0)$.

To obtain a flexible nonlinear alternative of this specification we approximate deterministic utility by means of a feedforward multilayer perceptron with Q hidden units which is known to approximate any continuous multivariate function and its derivatives with the desired level of precision given

a sufficient number of hidden units each following the binary logistic function g (Hornik et al., 1989; Ripley, 1993). We specify deterministic utility as linear combination of the values of Q hidden units which themselves are formed by nonlinear transformations of linearly combined predictors' values:

$$V_{h,m,t} = \sum_{j=1}^{Q} \left(\beta_{2,j} \cdot g\left(\vec{\beta}_{1,j} \cdot \vec{c}_{x,h,m,t} \right) \right) + \vec{D}_m \cdot \vec{\beta}_3 \qquad (14)$$

Putting deterministic utilities formed according to Eq. (14) into the basic multinomial logit equation (12) finally gives the combined model which we call neural net–multinomial logit (NN–MNL) model.

Predictors of both the MNL and the NN–MNL model are:

- brand dummy variables $\vec{D}_m$
- reference price $c_{r,h,m,t}$
- price loss $\max(c_{p,h,m,t} - c_{r,h,m,t}, 0)$
- price gain $\max(c_{r,h,m,t} - c_{p,h,m,t}, 0)$
- display $c_{d,h,m,t}$ (binary)
- feature $c_{f,h,m,t}$ (binary)
- loyalty $c_{l,h,m,t}$

The first suffix of the symbols indicates the predictor (r reference price, p observed price, d display, f feature, l loyalty). $c_{p,h,m,t}$ is the price observed at the point of sale.

Aside from brand dummy variables all the predictors vary over households, brands and purchase occasions. Prices as well as display and feature variables can be obtained directly from the purchase data, whereas reference prices and loyalties have to be estimated.

Following the seminal paper of Guadagni and Little (1983) we measure loyalty values by exponentially smoothing past purchases for each household. Reference prices constitute internal prices to which households compare observed prices (Winer, 1988). Reference prices reflect the expected price level of a brand which we set equal to a one-period forecast obtained by a reference price model. We study two alternative models of the reference price mechanism, a linear and a nonlinear reference price model. The latter is specified as neural net (NN) of the feedforward perceptron type with one layer of three hidden units and logistic functions for hidden units. Predictors of both models consist of brand-specific dummy variables, prices lagged maximally three periods and a time index.

High reference prices are associated with lower choice probabilities. Observed prices below the reference price (which households perceive as gains) stimulate purchases, i.e. increase choice probability. Observed prices above the reference price (which households perceive as losses) may deter from purchasing and therefore decrease choice probability. Prospect theory predicts asymmetric effects, i.e. that consumers respond more to losses than to gains (Kahneman and Tversky, 1979; Winer, 1988).

Estimation of NN–MNL choice models and NN reference price models consists of two steps, stochastic gradient descent followed by BFGS, a quasi-Newton optimization method (a detailed description of the NN–MNL model and its estimation procedure can be found in Hruschka et al., 1999). Estimation of MNL models only requires the BFGS step. Linear reference price models are estimated by OLS.

We derive the following closed form expression for elasticities on the basis of the NN–MNL model:

$$\begin{aligned}(1 - p(y_{h,m,t}|B, \{\vec{c}_{x,h,t}\}))c_{x_i,h,m,t} \\ \times \sum_{j=1}^{Q} \left(\beta_{2,j} \cdot g\left(\vec{\beta}_{1,j} \cdot \vec{c}_{x,h,m,t} \right) \right. \\ \left. \times \left(1 - g\left(\vec{\beta}_{1,j} \cdot \vec{c}_{x,h,m,t} \right)\right)\beta_{1,j,i} \right)\end{aligned} \qquad (15)$$

It is fairly obvious that the following well-known expression for elasticities of the MNL model is a special case of formula 15:

$$(1 - p(y_{h,m,t}|B, \{\vec{c}_{x,h,t}\}))c_{x_i,h,m,t}\beta_{1,j,i} \qquad (16)$$

For binary predictors (i.e. feature, display) we consider the change in choice probabilities achieved by using the respective sales promotion instrument:

$$\begin{aligned}p(y_{h,m,t}|B, \{\vec{c}_{x,h,t}\}, c_{x_i,h,m,t} = 1) \\ - p(y_{h,m,t}|B, \{\vec{c}_{x,h,t}\}, c_{x_i,h,m,t} = 0)\end{aligned} \qquad (17)$$

Choice elasticities are postulated to be positive for sales promotion variables (display, feature), brand

loyalties and gains. On the other hand, effects of reference prices, observed prices as well as losses are expected to be negative. The effect of losses on choice probability should be greater than that of gains.

Parameters of the choice models introduced here do not vary across the population and do not follow a mixing distribution. Nonetheless these models are able to reflect heterogeneity w.r.t. elasticity, because they have household-specific independent variables like brand loyalties and reference prices (Wedel et al., 1999).

4. Empirical study

We analyze purchase data of the six largest brands in terms of market share for one product group (ketchup) acquired in a scanner panel. For a time span of 123 weeks the data refer to households making at least ten purchases. This way 811 households remain for analysis.

Estimation and evaluation of choice models is based on ten random assignments of households to estimation and test data sets. NN–MNL models with 10 hidden units achieve the best average log likelihood values across these 10 test data sets. In view of their performance on the test data sets we only consider the linear reference price model for the MNL model, the nonlinear reference price model for the NN–MNL model with 10 hidden units in the following (see Table 1).

The total sums of squares TSS of elasticity vectors for the MNL model and the NN–MNL model with 10 hidden units amount to 0.360 and 0.702, respectively. These results indicate that the NN–MNL model implies more heterogeneity of households w.r.t. elasticities than the MNL model.

For household-specific elasticity vectors computed on the basis of the MNL model we obtain the maximum value of the approximate weight of evidence F_r for $r = 4$ segments. The same approach recommends $r = 7$ segments for elasticity vectors derived from the NN–MNL model with 10 hidden units, which seems natural given the higher heterogeneity implied by this choice model. Final partitions with four or seven segments are found by iterative relocation.

Multiple linear discriminant analysis with segment membership as dependent variable serves to assess the importance of the individual choice elasticities for reference price, gain, loss, feature, display and loyalty. Both for elasticities derived from the MNL and the NN–MNL model two discriminant functions suffice to recover more than 96% of the variance (i.e. of the among group sum of squares). Table 2 also contains product moment correlations of individual elasticities with the first two discriminant functions.

These correlations show that elasticities for loss, reference price and gain are the most important predictors for segments obtained on the basis of the MNL model. The same conclusion can be drawn for elasticities for reference price and loss in the case of segments derived from the NN–MNL model. Therefore we restrict interpretation of the segments to these individual elasticities.

Absolute average elasticities for the four segments obtained on the basis of the MNL model are given in Table 3. In segment 4 high reference price elasticity goes together with the lowest loss elasticity, but the highest gain elasticity. The largest segment 1 is characterized by medium reference price elasticity, low loss elasticity and medium gain elasticity. Households of segment 2 have medium reference price elasticity, very high loss elasticity

Table 1
Log likelihood of choice models on test data

Choice model	Reference price model	
	Linear	Nonlinear
MNL	−954.22	−954.54
NN–MNL		
$Q = 3$	−925.11	−933.30
$Q = 10$	−889.61	−885.43

Table 2
Correlation of elasticities with discriminant functions

	MNL		NN–MNL	
	Function 1	Function 2	Function 1	Function 2
Reference price	0.301	0.885	0.868	0.467
Loss	0.796	0.509	−0.375	0.889
Gain	−0.322	0.811	−0.283	−0.102
Display	0.187	0.208	−0.092	0.487
Feature	0.199	0.236	−0.203	0.314
Loyalty	0.025	0.253	−0.054	0.190
Variance explained	77.4%	97.2%	77.8%	96.7%

Table 3
Average absolute elasticities (MNL)

Segment number	Reference price	Loss	Gain	Relative segment size
4	2.09	0.34	0.093	6.2
1	1.68	0.41	0.034	78.9
2	1.62	1.37	0.023	8.0
3	1.45	0.53	0.017	6.9

and low gain elasticity. Segment 3 comprises households with the lowest reference price elasticity, rather low loss elasticity and low gain elasticity.

Table 4 contains absolute average elasticities for the four segments obtained on the basis of the NN–MNL model. These results show two segments with high reference price elasticities (segment 7 with small loss elasticity, segment 5 with somewhat higher loss elasticity). Two segments have medium reference price elasticities (segment 1, the largest segment, with low loss elasticity and segment 3 with high loss elasticity). Three segments consist of households with small reference price elasticities (segment 6 with the highest loss elasticity, segment 2 with small loss elasticity, segment 4 with high loss elasticity).

Table 4
Average absolute elasticities (NN–MNL)

Segment number	Reference price	Loss	Relative segment size
7	2.88	0.20	6.8
5	2.80	0.29	6.9
3	2.23	0.59	8.3
1	2.16	0.20	55.0
4	1.66	0.57	6.2
2	1.46	0.22	10.5
6	1.24	0.73	6.4

We analyze the seven segments derived for the elasticities of the NN–MNL choice model by means of multinomial logit models (Maddala, 1983). They specify the probability of household h to belong to segment k conditional on a vector x_h of external variables not used for clustering as:

$$P(k|x_h) = \begin{cases} \frac{1}{1+\sum_{j=2}^{G} \exp(\beta_j \cdot x_h)} & \text{for } k = 1 \\ \frac{\exp(\beta_k \cdot x_h)}{1+\sum_{j=2}^{G} \exp(\beta_j \cdot x_h)} & \text{for } k = 2, \ldots, G \end{cases} \tag{18}$$

These models differ from their discrete choice relative, the MNL model of Section 3, by not including a latent variable like utility. They possess the following household-specific predictors:

- number of purchases (purchases)
- number of brands bought (brands)
- low or high income (income)
- college education (education)
- household size (size)
- number of children (children)

We study three different versions of the logit model of equation (18), called A, B and C. Model A ignores interaction effects. Model B comprises pairwise interaction effects of income and education

Table 5
MNL model of segment membership (model C)

Segment	Constant	Purchases	Brands	Income	Size	Brands × purchases	Brands × income	Brands × size	Size × income
2	−8.49	14.19	12.17	−2.49	9.27	−28.01	3.57	−21.30	1.94
	−3.43	2.91	2.77	−2.20	1.69	−2.64	2.24	−1.91	0.95
3	−1.56	−12.00	−1.29	−0.69	−3.02	18.69	0.62	7.45	0.47
	−0.62	−2.19	−0.36	−0.58	−0.61	2.24	0.44	1.02	0.20
4	−5.54	2.05	9.81	2.71	2.61	−10.85	−2.25	−14.63	−5.32
	−1.98	0.30	2.19	2.05	0.45	−0.91	−1.14	−1.50	−1.89
5	−4.25	2.53	5.23	−0.31	−2.68	−8.76	1.44	2.32	−1.59
	−1.64	0.41	1.27	−0.30	−0.49	−0.77	1.02	0.26	−0.70
6	−2.71	−18.38	4.73	0.62	10.65	17.37	−0.37	−22.78	−3.54
	−0.81	−2.38	0.91	0.33	1.88	1.40	−0.13	−2.35	−1.06
7	5.18	−22.74	−5.83	0.50	−6.35	24.03	0.48	5.19	−1.80
	1.84	−2.61	−1.28	0.54	−1.22	1.71	0.31	0.54	−0.88

First line: coefficient; second line: *t*-value.

with number of purchases and number of brands bought as well as of household size and children with income and education. Model C omits the main effects of income, education and number of children. It includes pairwise interactions of the number of brands bought with number of purchases, income, household size as well as of household size with income. Maximum likelihood estimation provides log likelihood values for models A, B, C of −1121.56, −1094.73 and −1090.02, respectively. The numbers of parameters of these models amount to 42, 90 and 54. In accordance with likelihood ratio tests of models B vs. A and models C vs. A we present results of model C only.

Looking at coefficients of model C with absolute *t*-values greater than 2.0 we arrive at the following descriptions of the segments (Table 5). Segment 2 households make many purchases or buy many brands (but note that membership probability becomes lower if they both make many purchases and buy many brands), have either lower income or higher income and buy many brands. Households in segment 3 make either few purchases or buy many brands frequently. Segment 4 consists of households who buy many brands and have higher income. Households of segment 5 cannot by characterized by the predictors studied. Segment 6 households make few purchases, buy few brands and have small household size. Segment 7 households purchase very infrequently.

References

Banfield, J.D., Raftery, A.E., 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49, 803–821.

Corstjens, M.L., Gautschi, D.A., 1983. Formal choice models in marketing. Marketing Science 2, 19–56.

Guadagni, P.M., Little, J.D.C., 1983. A logit model of brand choice calibrated on scanner data. Marketing Science 2, 203–238.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 3, 359–366.

Hruschka, H., Fettes, W., Probst, M., 1999. Artificial neural net-multinomial logit model. A Semiparametric Approach to Analyze Brand Choice, Discussion Paper, Faculty of Economics, University of Regensburg, Regensburg, Germany.

Kahneman, D., Tversky, A., 1979. An analysis of decision under risk. Econometrica 47, 363–391.

Maddala, G.S., 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

McFadden, D., 1973. Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

McFadden, D., 1980. Econometric models for probabilistic choice among products. Journal of Business. 53, S13–S34.

Ripley, B.D., 1993. Statistical aspects of neural networks. In: Barndorff-Nielsen, O.E., Jensen, J.L., Kendall, W.S. (Eds.), Networks and Chaos—Statistical and Probabilistic Aspects. Chapman & Hall, London, pp. 40–123.

Symons, M., 1981. Clustering criteria and multivariate normal mixtures. Biometrics 73, 35–43.

Wedel, M., Kamakura, W., Arora, N., Bemmaor, A., Chiang, J., Elrod, T., Johnson, R., Lenk, P., Neslin, S., Poulsen, C.S., 1999. Discrete and continuous representations of unobserved heterogeneity in choice modeling. Marketing Letters 10, 219–232.

Winer, R.S., 1988. Behavioral perspective on pricing. In: Devinney, t.m. (Ed.), issues in Pricing. Lexington Books, Lexington, MA, pp. 35–57.