Modelling customer preference heterogeneity to iPad attributes using a Finite Mixture procedure

Liberato Camilleri, Keith Lia
Department of Statistics and Operations Research
University of Malta
Msida (MSD 06)
Malta
E-mail: liberato.camilleri@um.edu.mt

KEYWORDS
Market Segmentation, Latent Class Mixture Model, EM algorithm, Maximum Likelihood Estimation

ABSTRACT

The identification of segments in strategic market planning has long been recognized as a powerful tool to understand consumer behaviour. An approach that has managerial appeal in addressing market heterogeneity is by assuming that customers can be grouped in a number of unobserved homogeneous segments where customers in each cluster have similar purchasing behaviours. This paper describes the different procedures in affecting market segmentation focusing more on the Finite Mixture approach, while the application addresses heterogeneity issues in customer preferences when purchasing iPads given demographic and product-related predictors.

1. INTRODUCTION

Traditionally, market segmentation have been conducted either by using priori segmentation in which the number of segments are determined before the data collection or post hoc segmentation in which the segments are identified by forming groups of consumers that are homogeneous along a set of measured characteristics. One of the most used post hoc methods is the two-stage approach in which a conjoint regression model is fitted for each respondent and utilities (regression coefficients) are estimated for each level of each attribute for every person. Segments are then generated by conducting cluster analysis of the individual-level utilities. The main problem with the two-stage approach is that different clustering techniques produce different segments in which the initial utility estimation method and the subsequent cluster analysis optimize different and unrelated objective functions.

In response to the limitations of these traditional clustering methods, several integrated conjoint segmentation methods were proposed where the estimation and the segmentation stage are conducted concurrently. Hagerty (1985) proposed a Q-type factor analysis to partition the respondents and showed that the method reduces the variance of individual parameter estimates without unduly increasing the bias of the estimates. Kamakura (1988) uses the same general approach by pooling respondents who are similar in terms of their conjoint full-profile responses, but employs an agglomerative clustering algorithm. He showed that his approach improves predictive accuracy at the individual respondent level. (Ogawa 1987) presented a stochastic logit framework to model rank order responses. The model uses a hierarchical, non-overlapping clustering method and estimation and segmentation are conducted concurrently. (DeSarbo et al., 1989) proposed a clusterwise regression procedure that uses a simulated annealing algorithm for optimization. (Spath 1982) proposed a non-hierarchical, clusterwise regression procedure to identify homogeneous groups in terms of the relationship between dependent and independent variables. (Wedel and Kistemaker 1989) proposed a generalization of the clusterwise regression by extending Spath's method to handle more than one observation per individual. Their procedure uses an exchange algorithm, developed by Banfield and Bassil to maximize the likelihood and yields non-overlapping, non-hierarchical segments. (Wedel and Steenkamp 1991) used a fuzzy clusterwise regression algorithm to partition the data by minimizing the residual sum of squares criterion, which represents the sum of the distances of subjects from the regression equations in all clusters.

The development of new techniques for segmentation in the area of finite mixture (latent class) models stands out to be the most far-reaching developments in the early 90's. The work of (Kamakura and Russell 1989), (DeSarbo et al., 1992) and (Wedel and DeSarbo 1995) brought major changes in market segmentation applications in theory and practice. Finite mixture models address heterogeneity through a discrete distribution where estimation is carried out by maximizing the likelihood function. The main advantage of these models is that they address market heterogeneity by assuming a number of unobserved clusters.

Managers seem to be comfortable with the idea of market segments, and the models tend to do well in identifying useful groups. Another advantage of latent class models is that they enable statistical inference where estimation and segmentation are carried out simultaneously. A study conducted by Vriens, Wedel, and Wilms (1996) found that finite mixture models had the best overall performance of nine conjoint clustering methods (which included both post hoc and integrated conjoint segmentation methods) in terms of parameter recovery, segment membership recovery and predictive accuracy.

Recent changes in the market environment presented new challenges and opportunities for market segmentation. The introduction of micro marketing, direct marketing and mass customisation enabled marketers to customize their products or services to very small groups of customers. This implied that estimation and predicted responses to marketing variables had to be conducted at the individual level rather than the segment level. Bayesian estimation methods in marketing have gained popularity in the last ten years and are used extensively in various marketing problems. Besides providing a set of techniques that allow for the development and analysis of complex models they can estimate models at the individual level in which heterogeneity is addressed through a continuous rather than a discrete distribution. While the conceptual appeal of Bayesian methods have long been recognized, the recent popularity arises from computational and modeling breakthroughs. Hierarchical Bayesian estimation was rarely used in the past due to the fact that it could only be applied to simple models since the class of models for which the posterior inference could be computed was no larger than the class of models for which exact sampling results were available. The technical problems in applying the method to complex models seemed insurmountable.

During the last ten years, simulation methods, particularly Markov chain Monte Carlo (MCMC) methods have overcome these computational constraints for a wide range of marketing models. The classic work of (Roberts and Casella 2004), (Gelman et al., 2004) and (Rossi et al., 2006) contributed considerably towards this shift in interest in Bayesian estimation. A study conducted by (Andrews, Ansari and Currim 2002) compares the relative efficiency of Finite Mixture and Hierarchical Bayes conjoint analysis models in terms of fit, prediction, and parameter recovery. The authors show that both modelling techniques are equally effective in recovering individual-level parameters and predicting rating evaluations. They found that the two modelling techniques produce good parameter estimates both at the individual and segment levels. Moreover, the authors show that the two models are robust to violations of underlying assumptions and that traditional individual-level models tend to overfit the data.

2. FINITE MIXTURE MODEL FRAMEWORK

Let the random variables $\mathbf{y}_j = (y_{jk})$ for $j = 1,\ldots,n$ and $k = 1,\ldots,K,$ belong to a super-population which constitutes a mixture of a finite number (I) of sub-populations in proportions $\pi_1,\ldots,\pi_I$, where it is not known in advance from which class a particular vector of observation arises. The probabilities π_i follow the constraint:

$$\sum_{i=1}^{I}\pi_i = 1\ ,\pi_i \geq 0, i = 1,\ldots,I \qquad (1)$$

Assume that the conditional probability density function of y_{jk} given that y_{jk} comes from class i, takes the form:

$$f_{jk|i}(y_{jk}\,|\,\theta_{ijk},\lambda_i) = \exp\left\{\frac{y_{jk}\theta_{ijk} - b(\theta_{ijk})}{a(\lambda_i)} + c(y_{jk},\lambda_i)\right\} \qquad (2)$$

for specific functions $a(.)$, $b(.)$ and $c(.)$ conditional upon class i and y_{jk} are independently distributed with canonical parameters θ_{ijk} and means μ_{ijk}. The dispersion parameter λ_i is assumed to be a known constant over observations in class i, while $a(\lambda_i) > 0$. The predicted value μ_{ijk} is linked to the linear predictor η_{ijk} through the link function $g(\cdot)$ such that in class i:

$$\eta_{ijk} = g(\mu_{ijk}) \qquad (3)$$

where the linear predictor comprises P covariates $\mathbf{X}_1,...,\mathbf{X}_p$ where $\mathbf{X}_p = (\mathbf{X}_{jkp})$, $p = 1,\ldots,P$ and the parameter vectors $\beta_i = (\beta_{ip})$ in class i.

$$\eta_{ij} = \sum_{p=1}^{P} X_{jp}\beta_{ip} \qquad (4)$$

Conditional upon class i, a generalized linear model consists of a specification of the distribution of the response variable y_{jk}, a linear predictor, η_{ijk} and a function $g(.)$ which links the random component to the systematic component. The unconditional probability density function of an observation vector $\mathbf{y}_j$ can then be expressed in the finite mixture model form:

$$f_j(y_j\,|\,\Phi) = \sum_{i=1}^{I}\pi_i\prod_{k=1}^{K} f_{jk|i}(y_{jk}\,|\,\beta_i,\lambda_i) \qquad (5)$$

where $\Phi' = (\pi',\beta',\lambda')$, $\pi = (\pi_1,\ldots,\pi_I)'$, $\lambda = (\lambda_1,\ldots,\lambda_I)'$ and $\beta = (\beta'_1,\ldots,\beta'_I)'$. To estimate the parameter vector Φ we formulate the likelihood for Φ:

$$L(\Phi;y) = \prod_{j=1}^{n} f_j(y_j\,|\,\Phi) \qquad (6)$$

An estimate of the parameter vector Φ is obtained by maximising the above likelihood equation with respect Φ subject to the constraint (1), using the EM algorithm (Dempster, Laird and Rubin 1977). Once an estimate of Φ is obtained, estimates of the posterior probability α_{ij}, that observation *j* comes from the latent class *i* can be calculated for each observation vector $\mathbf{y}_j$ by using Bayes' theorem given by:

$$\alpha_{ij}(y_j|\Phi) = \frac{\pi_i \prod_{k=1}^{K} f_{jk|i}(y_{jk}|\beta_i,\lambda_i)}{\sum_{i=1}^{I}\pi_i \prod_{k=1}^{K} f_{jk|i}(y_{jk}|\beta_i,\lambda_i)} \qquad (7)$$

The EM Algorithm iterates between an expectation E-step and a maximization M-step. To derive the EM Algorithm, we introduce unobserved data z_{ij} indicating if observation *j* belongs to latent class *i*, such that $z_{ij}=1$ if *j* comes from class *i* and $z_{ij}=0$ otherwise. It is assumed that these z_{ij} are independent and identically distributed and have a multinomial distribution.

$$f_j(\mathbf{z}_j|\pi) = \prod_{i=1}^{I}\pi_i^{z_{ij}} \qquad (8)$$

where the vector $\mathbf{z}_j = (z_{ij},\ldots,z_{Ij})'$. We denote the matrix $(\mathbf{z}_1,\ldots,\mathbf{z}_n)'$ by $\mathbf{Z}$ and the matrix $(\mathbf{X}_1,\ldots,\mathbf{X}_p)$ by **X**. It is assumed that the observed data y_{jk} given unobserved data $\mathbf{z}_j$ are conditionally independent and that y_{jk} given $\mathbf{z}_j$ has the density function:

$$f(y_{jk}|\mathbf{z}_j) = \prod_{i=1}^{I} f_{jk|i}(y_{jk}|\beta_i,\lambda_i)^{z_{ij}} \qquad (9)$$

So the observations y_{jk} comprise the incomplete data set and the unknown observations z_{ij} are treated as missing data. Hence the complete data set combines **X** and **Z** and the complete-data log-likelihood can be formed by using the equations (8) and (9).

$$\ln L_c(\Phi;\mathbf{y},\mathbf{Z}) = \sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{i=1}^{I} z_{ij}\ln f_{jk|i}(y_{jk}|\beta_i,\lambda_i) + \sum_{j=1}^{n}\sum_{k=1}^{K}\sum_{i=1}^{I} z_{ij}\ln\pi_i$$

The complete log-likelihood $\ln L_c(\Phi;\mathbf{y},\mathbf{Z})$ is maximized using an iterative EM algorithm. In the E-step the complete log-likelihood is replaced by its expectation calculated on the basis of the provisional estimates of Φ from the previous iteration. In the M-step the expectation of the complete log-likelihood is maximized with respect to the parameter vector Φ to obtain new updated parameter estimates. The E-step and M-step are then alternated repeatedly until the iterative procedure converges and no further improvement in the likelihood function is possible. Dempster, Laird and Rubin (1977) proved that the EM algorithm provides monotone increasing values of the complete log likelihood.

In the E-Step the expectation of the complete log-likelihood is calculated with respect to the conditional distribution of the unobserved data **Z** given the observed data **y** and provisional estimates of Φ. $E[\ln L_c(\Phi;\mathbf{y},\mathbf{Z})]$ can be obtained by replacing z_{ij} in $\ln L_c(\Phi;\mathbf{y},\mathbf{Z})$ by their expected values, $E(z_{ij}|\mathbf{y},\Phi)$. To obtain this expectation, we first calculate the conditional distribution of $\mathbf{y}_j$, given **Z**, which is:

$$f(\mathbf{y}_j|\mathbf{Z},\Phi) = \prod_{i=1}^{I}\left(\prod_{k=1}^{K} f_{jk|i}(y_{jk}|\beta_i,\lambda_i)\right)^{z_{ij}} \qquad (10)$$

By using Bayes' theorem, we can derive the conditional distribution of z_{ij} given $\mathbf{y}_j$ by using equations (10) and (8), which in turn can be used to calculate the required conditional expectation given by:

$$E(z_{ij}|\mathbf{y}_j,\Phi) = \frac{\pi_i \prod_{k=1}^{K} f_{jk|i}(y_{jk}|\beta_i,\lambda_i)}{\sum_{i=1}^{I}\pi_i \prod_{k=1}^{K} f_{jk|i}(y_{jk}|\beta_i,\lambda_i)} \qquad (11)$$

This is identical to the posterior probability $\alpha_{ij}(y_j|\Phi)$ in equation (7). Estimates of the posterior probabilities $\hat{\alpha}_{ij}$ are obtained by evaluating equation (11) using the current estimates of β and λ.

The M-step maximizes the expectation of the complete log-likelihood with respect to the parameter vector Φ after replacing the unobserved data **Z** in $\ln L_c(\Phi;\mathbf{y},\mathbf{Z})$ by their current expected values $\hat{\alpha}_{ij}$:

$$E[\ln L_c(\Phi;\mathbf{y},\mathbf{Z})] = \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{n}\hat{\alpha}_{ij}\ln f_{jk|i}(y_{jk}|\beta_i,\lambda_i) + \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{n}\hat{\alpha}_{ij}\ln\pi_i$$

The maximization of $E[\ln L_c(\Phi;\mathbf{y},\mathbf{Z})]$ with respect to π subject to the constrain (1), is solved by maximizing the augmented function:

$$\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{n}\hat{\alpha}_{ij}\ln\pi_i - \mu\left(\sum_{i=1}^{I}\pi_i - 1\right) \qquad (12)$$

where μ is a Langrangian multiplier. Setting the derivative of (12) with respect to π_i equal to zero and solving for π_i

$$\hat{\pi}_i = \sum_{j=1}^{n}\hat{\alpha}_{ij}/n \qquad (13)$$

Maximization of $E[\ln L_c(\Phi;\mathbf{y},\mathbf{Z})]$ with respect to $\boldsymbol{\beta}$ and λ is equivalent to independently maximizing each of the following *I* expressions:

$$L_i^* = \sum_{j=1}^{n}\sum_{k=1}^{K}\hat{\alpha}_{ij}\ln f_{jk|i}(y_{jk}|\beta_i,\lambda_i) \qquad (14)$$

3. APPLICATION

The finite mixture model was used to identify factors that influence the customer choices when buying iPads and identify the product attributes that most influence the consumers in buying the product. In this application, the three selected iPad attributes included the price, capacity and connectivity. This survey was designed and devised on Kwik Survey (an online survey questionnaire) where a number of iPads profiles having distinct attributes were generated and these profiles had to be assessed on a 7-point Likert scale where 1 corresponds to 'Not worthy' and 7 corresponds to 'Extremely worthy'. A rating scale was selected since it expresses the intensity of a preference better than a ranking scale. The target population for this survey were university students. The respondents were asked to provide demographic information, including their gender, age and whether they owned an iPad.

The three selected iPad attributes included the capacity of the iPad (16GB, 32GB and 64GB), connectivity (Wi-Fi and Wi-Fi plus 3G) and price (€500, €600, €700 and €800). These three attributes were chosen on the merit that they are found in literature to be the most pertinent when compared to the other attributes, such as colour and size. A full-profile method and full factorial design were chosen for the data collection method yielding a total of 24 distinct profiles. The sample of 364 participants who completed the online questionnaire included a larger proportion of females (55.5%) than males. Around 70% of the university students had less than 24 years and only a third owned an iPad.

To identify the optimal number of segments, the finite mixture model was fitted several times each time changing the number of segments from 1 to 4. For each solution the BIC criterion was computed. Table 1 shows that the three-segment solution is the one which minimizes the criterion.

Number of segments K	Deviance (-2 log L)	Number of parameters d	BIC
1	21192	7	21233
2	19203	14	19286
3	19133	21	19257
4	19097	28	19262

Table1: BIC value for each segment solution

4. RESULTS OF FINITE MIXTURE ANALYSIS

Posterior probabilities were computed for each respondent and each person was allocated to the segment with highest posterior probability. 212 respondents were allocated to segment 1, 111 students to segment 2 and the remaining 41 participants to segment 3. Segment 1 included a larger proportion of females, aged between 17 and 19 years and owned an iPad. Segment 2 comprised a larger proportion of males, aged between 20 and 23 years and owned an iPad. Segment 3 included a larger proportion of males, aged at least 24 years and did not own an iPad. Table 2 displays the parameter estimates and standard errors for each segment solution.

Parameter estimates	Standard Error	Term
3.739	0.033	Segment(1)
5.137	0.045	Segment(2)
1.424	0.075	Segment(3)
2.478	0.036	Price(1).Segment(1)
0.740	0.048	Price(1).Segment(2)
0.824	0.080	Price(1).Segment(3)
1.538	0.036	Price(2).Segment(1)
0.504	0.048	Price(2).Segment(2)
0.387	0.080	Price(2).Segment(3)
0.625	0.036	Price(3).Segment(1)
0.209	0.048	Price(3).Segment(2)
0.080	0.080	Price(3).Segment(3)
-1.950	0.031	Capacity(1).Segment(1)
-0.865	0.041	Capacity(1).Segment(2)
-0.069	0.069	Capacity(1).Segment(3)
-0.874	0.031	Capacity(2).Segment(1)
-0.304	0.041	Capacity(2).Segment(2)
-0.254	0.069	Capacity(2).Segment(3)
-0.709	0.025	Connectivity(1).Segment(1)
-0.395	0.034	Connectivity(1).Segment(2)
-0.130	0.057	Connectivity(1).Segment(3)

Table 2: Parameter estimates and standard errors

Segmentation is effective if it is identifiable and accessible. These segments are meaningless if they are not described and defined. Figures 1, 2 and 3, show the mean rating scores provided by respondents in different segments for different profile manifestations categorized by the levels of capacity, connectivity and price.

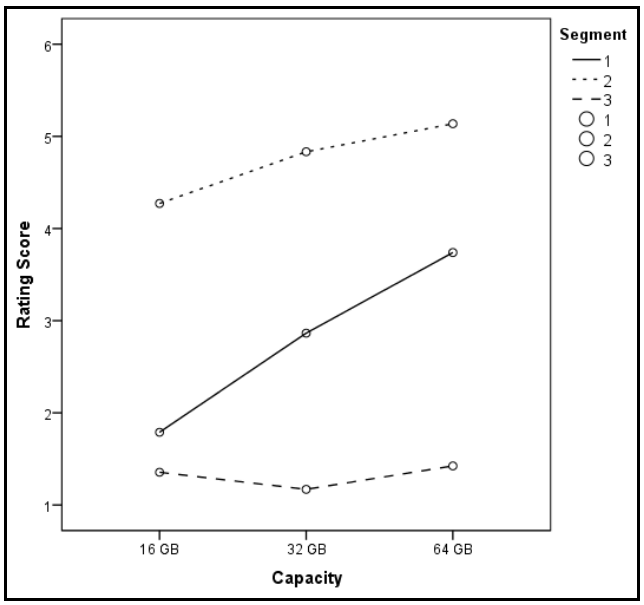

Figure 1: Mean rating score by cluster membership and iPad capacity

Respondents in Segment 1 worth iPads more if they have higher capacity, faster connectivity and are less expensive. Respondents in Segment 2 behave similarly to those in Segment 1 because they value iPads more if they have higher capacity, faster connectivity and are cheaper in price. Though, on average, they are providing higher rating scores

and are discriminating less between the iPad attributes categories since changes in their mean rating scores are less conspicuous compared to those in Segment 1. Respondents in Segment 3 are providing very low rating scores. They are not price sensitive and hardly discriminate between the iPad features since their mean rating scores vary marginally for different profile manifestations.

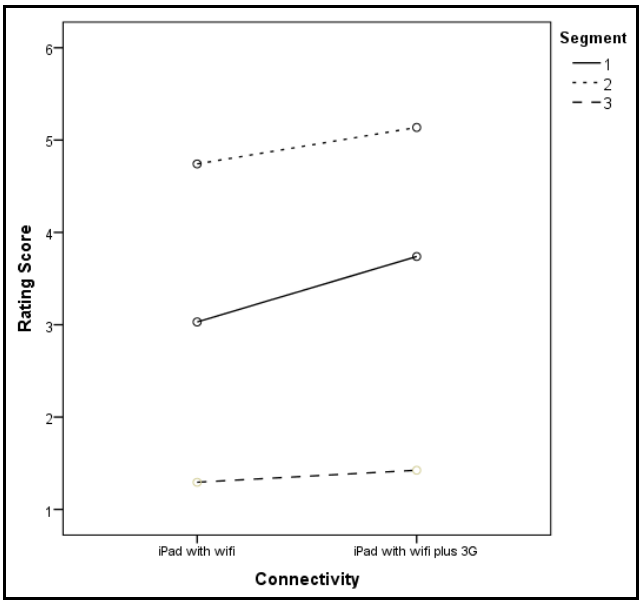

Figure 2: Mean rating score by cluster membership and iPad connectivity

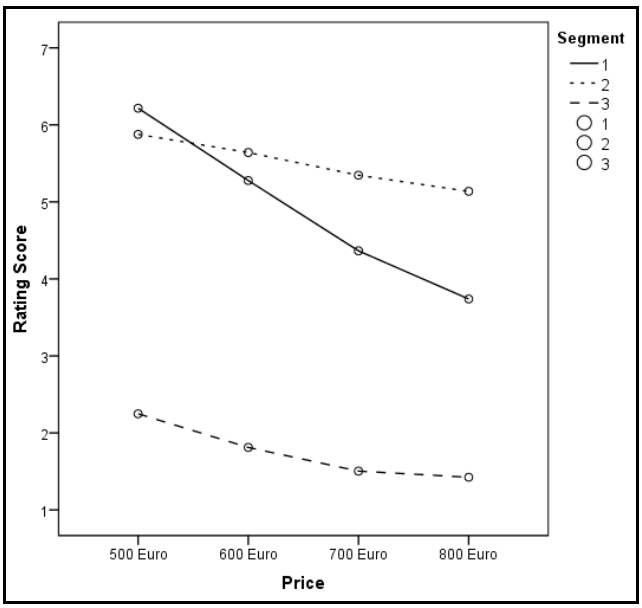

Figure 2: Mean rating score by cluster membership and iPad price

REFERENCES

Andrews, R.L, Ansari, A. and Currim, I.S. (2002), Finite Mixture Conjoint versus Hierarchical Bayes, *Journal of Marketing Research*, 19, 1, 87-98.

Dempster, A., Laird, N. and Rubin, D. (1977), Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society*, B, 39, 1-38.

DeSarbo, W.S., Oliver, R.L. and Rangaswamy, A. (1989), A Simulated Annealing Methodology for Clusterwise Linear Regression, *Psychometrika*, 54, 707-736.

DeSarbo, W., Wedel, M., Vriens, M. and Ramaswamy, V. (1992), Latent class metric conjoint analysis, *Marketing Letters*, 3, 3, 273-288.

Gelman, A., Carlin, J. and Rubin, D. (2004), *Bayesian Data Analysis* (2nd edn). Chapman & Hall CRC

Hagerty, M.R. (1985), Improving the predictive Power of Conjoint Analysis: Use of Factor and Cluster Analysis *Journal of Marketing Research*, 22, 168-184.

Kamakura, W. (1988), A Least Squares Procedure for Benefit Segmentation with Conjoint Experiments, *Journal of Marketing Research*, 25, 157-167.

Kamakura, W. and Russell, G. (1989), Probabilistic choice model for market segmentation and elasticity structure, *Journal of Marketing Research*, 26, 379-390.

Ogawa, K. (1987), Approach to Simultaneous Estimation and Segmentation in Conjoint Analysis, *Marketing Science*, 6, 66-81.

Roberts, C, and Casella G. (2004), *Monte Carlo Statistical Methods* (2nd edn). New Your Springer-Verlag.

Rossi, P., Allenby, G. and McCulloch, R. (2006), *Bayesian Statistics and Marketing* (2nd edn). John Wiley & Sons

Spath, H. (1982), A Fast Algorithm for Clusterwise Linear Regression, *Computing*, 29, 175-181.

Vriens, M., Wedel, M. and Wilms, T. (1996), Conjoint Segmentation Methods A Monte Carlo Comparison, *Journal of Marketing Research*, 23, 73-85.

Wedel, M. and DeSarbo, W.S. (1995), Mixture likelihood Approach for Generalized Linear Models, *Journal of Classification*, 12, 1-35.

Wedel, M. and Kistemaker, C. (1989), Consumer Benefit Segmentation using Clusterwise Linear Regression, *Journal of Research Marketing*, 6, 45-49.

Wedel, M. and Steenkamp, J.B. (1991), Clusterwise Regression Method for Simultaneous Fuzzy Market Structuring and Segmentation, *Journal of Market Research*, 28, 385-396.

AUTHOR BIOGRAPHY

LIBERATO CAMILLERI studied Mathematics and Statistics at the University of Malta. He received his PhD degree in Applied Statistics in 2005 from Lancaster University. His research specialization areas are related to statistical models, which include Generalized Linear models, Latent Class models, Multi-Level models and Random Coefficient models. He is presently a lecturer in the Statistics department at the University of Malta.