INVESTIGATING THE FACTORS WHICH AFFECT THE PERFORMANCE OF THE EM ALGORITHM IN LATENT CLASS MODELS

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

KEYWORDS
Latent class model, Market segmentation, EM algorithm, Monte Carlo simulation

ABSTRACT

Latent class models have been used extensively in market segmentation to divide a total market into market groups of consumers who have relatively similar product needs and preferences. The advantage of these models over traditional clustering techniques lies in simultaneous estimation and segmentation, which is carried out using the EM algorithm. The identification of consumer segments allows target-marketing strategies to be developed.

The data comprises the rating responses of 262 respondents to 24 laptop profiles described by four item attributes including the brand, price, random access memory (RAM) and the screen size. Using the facilities of R Studio, two latent class models were fitted by varying the number of clusters from 2 to 3.

The parameter estimates obtained from these two latent class models were used to simulate a number of data sets for each cluster solution to be able to conduct a Monte-Carlo study, which investigates factors that have an effect on segment membership and parameter recovery and affect computational effort.

1. INTRODUCTION

Latent class models (LCM) differ from standard regression models because they accommodate discrete latent variables. In layman terms, LCM assume that the heterogeneous observations in a sample arise from a number of homogenous subgroups (segments) mixed in unknown proportions. The main inferential goals of LCM are to identify the number of segments and simultaneously estimate the regression model parameters for each segment; and classify the individuals in their most likely segment. The characteristics of each segment can be deduced based on the demographic information of the members within each segment. In the past decade, LCM has increased in popularity, particularly in market segmentation, which is mainly due to technological advancements, rendering complex LCM computationally feasible, even on large data sets.

2. THEORETICAL FRAMEWORK

Latent class models assume that the population consists of *S* segments having unknown proportions $\pi_1, \pi_2, ..., \pi_S$. These proportions must satisfy the following two constraints:

$$\pi_s \geq 0 \ \forall s, \text{ and } \sum_{s=1}^{S} \pi_s = 1 \quad (1)$$

The conditional probability density function of the responses $\mathbf{Y}_i$, given that $\mathbf{Y}_i$ comes from segment s is given by:

$$\mathbf{Y}_i \sim f_{i|s}(y_i \,|\boldsymbol{\theta}_{si}, \varphi_s) \ , \quad (2)$$

where, the conditional density function is assumed to be a mixture of segment-specific densities, $f_{ik|s}(y_i \,|\theta_{sik}, \varphi_s)$. These component mixtures are assumed to be independent within the latent classes, such that:

$$f_{i|s}(y_i \,|\boldsymbol{\theta}_{si}, \varphi_s) = \prod_{k=1}^{K} f_{ik|s}(y_i \,|\theta_{sik}, \varphi_s) \quad (3)$$

If $\mathbf{Y}_i$ has conditional multivariate normal distribution then $f_{i|s}(y_i \,|\boldsymbol{\theta}_{si}, \varphi_s) = f_{i|s}(y_i \,|\mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s)$ can be expressed as:

$$(2\pi)^{-K/2} \left|\boldsymbol{\Sigma}_s\right|^{-1/2} \exp\left[-\frac{1}{2}\left(y_i - \mathbf{X}_i\boldsymbol{\beta}_s\right)'\boldsymbol{\Sigma}_s^{-1}\left(y_i - \mathbf{X}_i\boldsymbol{\beta}_s\right)\right] \quad (4)$$

The unconditional probability density function of $\mathbf{Y}_i$, given the vector of unknown parameters $\boldsymbol{\Omega}' = (\boldsymbol{\pi}', \boldsymbol{\beta}', \boldsymbol{\Sigma})$, is:

$$f_i(y_i \,|\boldsymbol{\Omega}) = \sum_{s=1}^{S} \pi_s f_{i|s}(y_i \,|\mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s) \quad (5)$$

The likelihood function $L(\boldsymbol{\Omega}, \mathbf{y}_i) = \prod_{i=1}^{N} f_i(y_i \,|\boldsymbol{\Omega})$ is given by:

$$\prod_{i=1}^{N}\left[\sum_{s=1}^{S} \pi_s f_{i|s}(y_i \,|\mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s)\right] \quad (6)$$

The likelihood function, formulated by equation (6), is used to estimate the parameter vector $\boldsymbol{\Omega}$. The estimate $\hat{\boldsymbol{\Omega}}$, is obtained by using the maximum likelihood (ML) technique, in particular, through the use of the EM algorithm. Using Bayes' theorem, the posterior probability $\alpha_{is}(\mathbf{y}_i, \boldsymbol{\Omega})$ can be computed using the parameter estimates $\hat{\boldsymbol{\Omega}}$.

$$\alpha_{is}(\mathbf{y}_i, \boldsymbol{\Omega}) = \frac{\pi_s f_{i|s}(y_i | \mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s)}{\sum_{s=1}^{S} \pi_s f_{i|s}(y_i | \mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s)} \quad (7)$$

The procedure updates the parameter estimates iteratively, and when it ultimately converges, the posterior probabilities given by (7) will be used to assign each respondent to that segment with the largest posterior probability.

3. THE EM ALGORITHM

Dempster *et al.*, (1977) are credited with presenting the EM algorithm in its current form, where parameters of a mixture distribution are estimated by using the concept of incomplete data. The central idea behind the EM algorithm is to augment the data by including unobserved, referred to as missing, data, which comprises of unknown 0-1 indicators indicating whether a respondent belongs or not to a particular segment. Hence, instead of maximizing the likelihood via standard optimization methods, the expected complete-data log-likelihood function is maximized using the EM algorithm.

Let z_{is} be the unknown 0-1 indicator variables representing the unobserved data, which are assumed to be independent and identically multinomially distributed.

$$f(z_i, \pi) = \prod_{s=1}^{S} \pi_s^{z_{is}} \quad (8)$$

where, $\mathbf{z}_i = (z_{i1}, ..., z_{is})$ and $\mathbf{z} = (\mathbf{z}_1', ..., \mathbf{z}_s')$. Since z_{is} is considered as missing data, the complete-data likelihood function, $L_c(\boldsymbol{\Omega}, \mathbf{y}_i, \mathbf{z})$ is given by:

$$L_c(\boldsymbol{\Omega}, \mathbf{y}_i, \mathbf{z}) = \prod_{s=1}^{S} \prod_{i=1}^{N} \left(\pi_s f_{i|s}(y_i | \mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s) \right)^{z_{is}} \quad (9)$$

The complete log-likelihood function $\log[L_c(\boldsymbol{\Omega}; \mathbf{y}_i, \mathbf{z})]$ is:

$$\sum_{s=1}^{S} \sum_{i=1}^{N} \left[z_{is} \log(\pi_{is}) + z_{is} \log\left(\pi_s f_{i|s}(y_i | \mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s) \right) \right] \quad (10)$$

In the E-step, the expectation of the complete log-likelihood function, given by (10) is calculated with respect to the conditional distribution of the missing data, given both the observed data and the initial estimates of $\boldsymbol{\Omega}$. Since the complete-data log-likelihood function is linear in z_{is}, the expectation $E\left[\log\left(L_c(\boldsymbol{\Omega} | \mathbf{y}_i, \mathbf{z}) \right) \right]$ is obtained by replacing the z_{is} by their conditional expectation, given the observed data.

$$\sum_{s=1}^{S} \sum_{i=1}^{N} \left[E(z_{is} | \mathbf{y}_i, \boldsymbol{\Omega}) f_{i|s}(y_i | \mathbf{X}_i, \boldsymbol{\beta}_s \boldsymbol{\Sigma}_s) + E(z_{is} | \mathbf{y}_i, \boldsymbol{\Omega}) \log(\pi_s) \right]$$

where $E(z_{is} | \mathbf{y}_i, \boldsymbol{\Omega})$ is given by:

$$\frac{\hat{\pi}_s f_{i|s}(y_i | \mathbf{X}_i, \hat{\boldsymbol{\beta}}_s \hat{\boldsymbol{\Sigma}}_s)}{\sum_{s=1}^{S} \hat{\pi}_s f_{i|s}(y_i | \mathbf{X}_i, \hat{\boldsymbol{\beta}}_s \hat{\boldsymbol{\Sigma}}_s)} = \hat{\alpha}_{is} \quad \text{where} \quad \hat{\pi}_s = \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_{is} \quad (11)$$

These posterior probabilities are updated iteratively by replacing the estimates of $\hat{\boldsymbol{\beta}}_s$ and $\hat{\boldsymbol{\Sigma}}_s$ obtained from the previous iteration.

4. APPLICATION

Two latent class models were fitted to identify factors that influence the customer choices when buying laptops and identify the product attributes that most influence the consumers in buying the product. In this application, the four selected laptop attributes included the brand (HP, Asus); the price (€500, €600, €700); RAM (4GB, 8GB) and the screen size (12 inch, 15 inch). This survey was designed and devised on Survey Monkey (an online survey questionnaire) where a number of laptop 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 'Very worthy'. A rating scale was selected since it expresses the intensity of a preference better than a ranking scale. A full-profile method and full factorial design were chosen for the data collection method yielding a total of 24 distinct profiles. 69.8% of 262 participants who completed the online questionnaire were females, 74.4% were university students and 73.7% were less than 30 years. All participants owned a laptop. The first latent class model assume a 2-segment solution and the second assume a 3-segment solution. The parameter estimates of the two latent class models will be used in a simulation study, described in section 5, to investigate factors that affect the performance of the EM algorithm.

For the 2-segment solution, 175 (66.8%) respondents were allocated to segment 1 and the remaining 87 (33.2%) were allocated to segment 2. Respondents in both segments rated HP laptops more than Asus; rated cheaper laptops more than expensive ones; rated 4GB RAM laptops less than 8GB RAM; and rated 12 inch screen laptop less than 15 inch screen. However, participants in segment 2 are discriminating more between the brands, prices, screen sizes and random access memories compared to participants in segment 1. Table 1 displays the parameter estimates and standard errors for the 2-segment solution.

Parameter	Segment1		Segment2	
	Est.	S.E.	Est.	S.E.
Intercept	5.18	0.05	3.66	0.08
Brand (HP)	0.07	0.04	0.39	0.06
Brand (Asus)	0		0	
Price (€500)	0.37	0.05	0.56	0.08
Price (€600)	0.31	0.05	0.20	0.07
Price (€700)	0		0	
RAM (4GB)	-0.72	0.04	-0.82	0.06
RAM (8GB)	0		0	
Size (12inch)	-0.58	0.04	-1.29	0.06
Size (15inch)	0		0	

Table 1: Parameter estimates for the 2-segment solution

For the 3-segment solution, 117 (44.7%) respondents were allocated to segment 1, 24 (9.2%) respondents were allocated

to segment 2 and the remaining 121 (46.1%) were allocated to segment 3. Respondents in both segments rated HP laptops more than Asus; rated cheaper laptops more than expensive ones; rated 4GB RAM laptops less than 8GB RAM; and rated 12 inch screen laptop less than 15 inch screen. However, participants in segment 1 are discriminating more between the prices and screen sizes; participants in segment 2 are discerning more between the brands; and participants in segment 3 are discriminating more between the random access memories. Table 2 displays the parameter estimates and standard errors for the 3-segment solution.

Parameter	Segment1		Segment2		Segment3	
	Est.	S.E.	Est.	S.E.	Est.	S.E.
Intercept	4.47	0.03	2.37	0.07	5.32	0.03
Brand (HP)	0.19	0.03	0.46	0.06	0.11	0.03
Brand (Asus)	0		0		0	
Price (€500)	0.51	0.03	0.08	0.07	0.43	0.03
Price (€600)	0.28	0.03	-0.04	0.07	0.32	0.03
Price (€700)	0		0		0	
RAM (4GB)	-0.53	0.03	-0.63	0.06	-0.91	0.03
RAM (8GB)	0		0		0	
Size (12inch)	-1.27	0.03	-0.60	0.06	-0.43	0.03
Size (15inch)	0		0		0	

Table 2: Parameter estimates for the 3-segment solution

Number of segments *S*	Deviance (-2 log *L*)	Number of parameters *P*	BIC
2	10868	12	21858
3	10545	18	21273

Table 3: BIC value for the 2- segment and 3-segment solutions

Table 3 displays the deviances, number of parameters and BIC values of the two-segment and three-segment solutions. Figures 1 to 4 provide graphical displays of the mean rating scores grouped by segment and laptop attributes. Respondents in segments 1 and 2 are price sensitive but not brand sensitive, while respondents in segment 3 are brand sensitive but not price sensitive. Respondents in all three segments prefer 8GB RAM and 15 inch screen laptops more than 4GB RAM and 12 inch screen laptops.

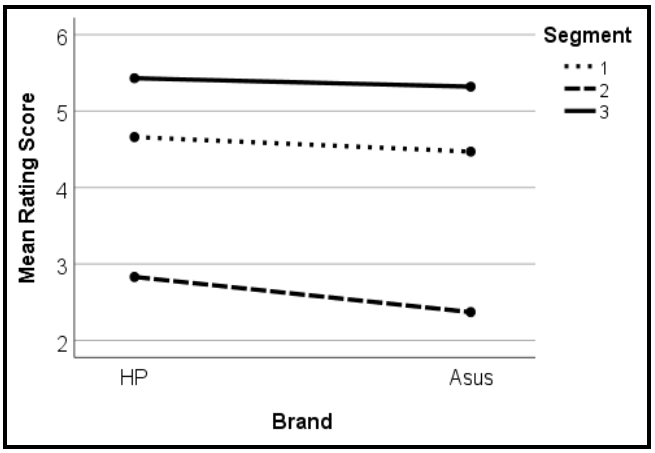

Figure 1: Mean rating scores grouped by segment and brand

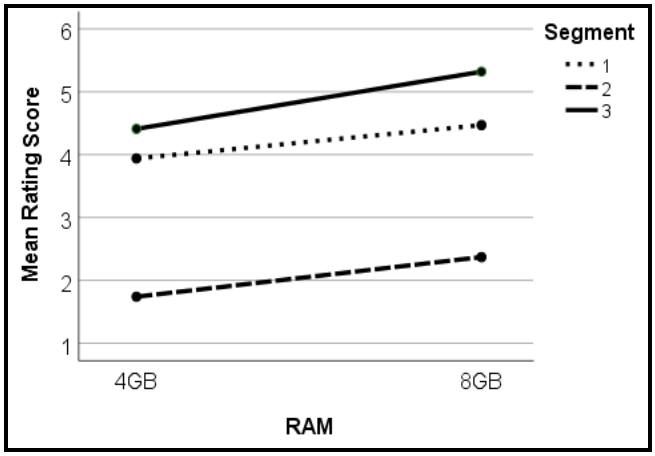

Figure 2: Mean rating scores grouped by segment and RAM

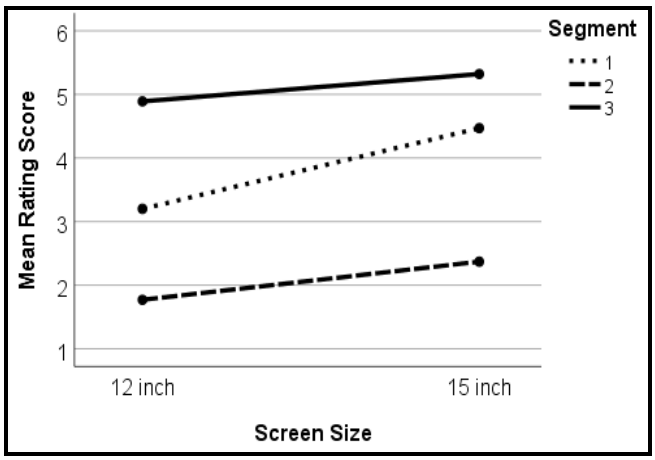

Figure 3: Mean rating scores grouped by segment and size

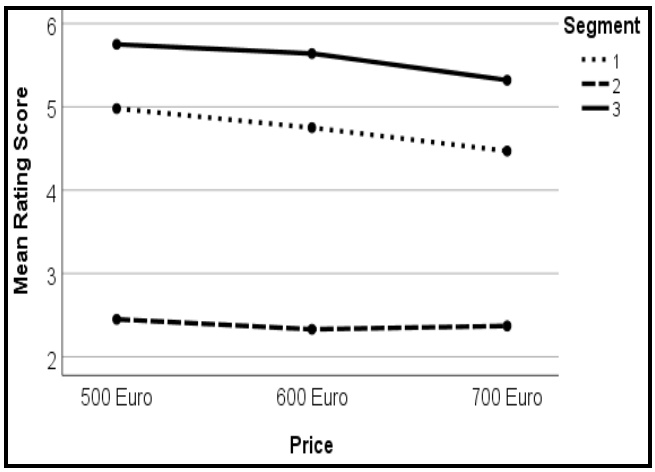

Figure 4: Mean rating scores grouped by segment and price

5. MONTE CARLO SIMULATION

A further task was to examine the performance of latent class models by modifying a number of factors. Three of the factors that are highlighted in literature as having potential effect on model performance include:

- Number of simulated respondents
- Number of segments
- Size of perturbation parameter σ_i^2 of the error terms.

The above three factors reflect a variation in conditions in many applications which are expected to affect the performance of the model fit. The design used in the study was 3 x 2^3 full factorial design, which yielded 24 observations. The following four measures are normally used to assess computational effort, parameter recovery, predictive power, goodness of fit and segment membership recovery. The root-mean-squared error between the true and estimated parameters is a measure of parameter recovery. β_p and $\hat{\beta}_p$ are the true and estimated parameters, where *P* is the number of parameters.

$$RMS\left(\hat{\boldsymbol{\beta}}\right) = \left[\sum_{p=1}^{P} \frac{\left(\beta_p - \hat{\beta}_p\right)^2}{P}\right]^{\frac{1}{2}} \tag{13}$$

The root-mean-squared error between the true and estimated segment membership probabilities is a measure of segment proportion recovery. π_s and $\hat{\pi}_s$ are the true and estimated segment membership probabilities, where *S* is the number of segments.

$$RMS\left(\hat{\boldsymbol{\pi}}\right) = \left[\sum_{s=1}^{S} \frac{\left(\pi_s - \hat{\pi}_s\right)^2}{S}\right]^{\frac{1}{2}} \tag{14}$$

The root-mean-squared-error between the true and predicted responses is a measure of the predictive power. y_{ik} and y_{ik} are the true and estimated responses, where *N* and *K* are the number of hypothetical subjects and the number of profiles assessed by each subject.

$$RMS\left(y\right) = \left[\sum_{i=1}^{N}\sum_{k=1}^{K} \frac{\left(y_{ik} - y_{ik}\right)^2}{N.K}\right]^{\frac{1}{2}} \tag{15}$$

In order to assess the factors that affect the performance of latent class models, synthetic data sets were generated, where the simulation was devised to mimic the laptop application. To allocate hypothetical subjects to segments the proportion π_s of members in each segment was specified, satisfying the constraint these proportions sum to 1. This was carried out by first generating *N* uniformly distributed pseudo-random real values in the range [0,1] and then by computing the cumulative probabilities $q_s = \sum_{j=1}^{s} \pi_j$. Every subject whose corresponding value was in the range (q_{s-1}, q_s) was allocated to segment *s*. This gives a random segment allocation to each hypothetical subject. To simulate the subjects' rating responses, the linear predictors and the corresponding parameters $\boldsymbol{\beta}_k$ were specified for the *S* segments. Moreover, the design and the linear predictor were set the same as in the application. Given the segment allocation of each member, synthetic data values were generated for each subject. These values were then perturbed by adding an error term having a normal distribution. Six specified cut-points α_r were used to convert these values to rates ranging from 1 to 7. Values in the range (α_{r-1}, α_r) were converted to rate *r*. This gives a random rating category allocation to each profile by each hypothetical subject.

The number of simulated respondents was varied at three levels (100, 300 and 500). It is expected that a greater number of simulated subjects improve the precision of the estimated segment-level parameters. The number of segments was also varied at two levels (2 and 3 segments) because these represent the range of segments commonly found in segmentation applications. It is expected that a greater number of segments deteriorate the precision of the estimated segment-level coefficients as a greater number of model parameters have to be estimated. The error terms were assumed to be normally distributed and the parameter σ_i^2 was set to 0.1, 0.5 and 1. It is expected that a larger perturbation value reduces the precision of the estimated segment-level parameters since there will be less cohesion in each segment and lower segment separation.

Number of subjects	Perturbation value	Number of segments	$RMS\left(\hat{\boldsymbol{\beta}}\right)$
100	0.1	2	0.2402
300			0.2396
500			0.2376
100	0.5		0.2716
300			0.2545
500			0.2606
100	1.0		0.3270
300			0.3138
500			0.3327
100	0.1	3	0.2427
300			0.2401
500			0.2341
100	0.5		0.3940
300			0.3674
500			0.3276
100	1.0		0.4049
300			0.3679
500			0.3490

Table 4: Parameter recovery using simulated data

Number of subjects	Perturbation value	Number of segments	$RMS\left(\hat{\boldsymbol{\pi}}\right)$
100	0.1	2	0.0475
300			0.0260
500			0.0249
100	0.5		0.0485
300			0.0260
500			0.0246
100	1.0		0.0721
300			0.0607
500			0.0547
100	0.1	3	0.0638
300			0.0288
500			0.0257
100	0.5		0.0638
300			0.0295
500			0.0276
100	1.0		0.0780
300			0.0689
500			0.0557

Table 5: Segment proportion recovery using simulated data

Ten data sets were generated for each factor level combination according to the number of subjects, number of segments and the perturbation value. Each simulated data set was re-fitted using a latent class model.

Number of subjects	Perturbation value	Number of segments	$RMS(y)$
100	0.1	2	1.3920
300			1.3945
500			1.3952
100	0.5		1.4364
300			1.4420
500			1.5088
100	1.0		1.4833
300			1.4195
500			1.5260
100	0.1	3	1.5620
300			1.7948
500			1.4629
100	0.5		1.5634
300			1.5828
500			1.4799
100	1.0		1.5633
300			1.5931
500			1.5061

Table 6: Assessing predictive power using simulated data

The $RMS(\hat{\boldsymbol{\beta}})$ $RMS(\hat{\boldsymbol{\pi}})$ and $RMS(\hat{y})$ values shown in tables 4, 5 and 6 were computed after permuting the parameters and predicted responses to match estimated and true segments optimally. All the three measures were averaged over these ten data sets.

Number of subjects	Perturbation value	Number of segments	Segment membership recovery
100	0.1	2	100%
300			100%
500			100%
100	0.5		100%
300			99.86%
500			99.46%
100	1.0		98.80%
300			97.92%
500			96.42%
100	0.1	3	100%
300			100%
500			99.94%
100	0.5		99.92%
300			97.67%
500			97.96%
100	1.0		93.80%
300			97.20%
500			91.36%

Table7: Segment membership recovery using simulated data

The percentage number of subjects that are correctly classified into their true segments is a measure of segment membership recovery. Table 7 displays the percentage number of subjects, averaged over the ten data sets, which are correctly classified into their true segments. It should be noted that after assigning each hypothetical subject to a segment with highest posterior probability these segments were permuted to maximize match with the true segments.

6 CONCLUSIONS

In general, the percentage of correctly classified hypothetical subjects in their true segment improves with a decrease in the number of segments and a reduction in the perturbation value; however, it is unaffected by changes in sample size. Parameter recovery improves with a decrease in the perturbation value, a decrease in the number of segments, and an increase in the sample size. Predictive power improves with a decrease in the perturbation value, however it is unaffected by changes in the number of segments or sample size. Segment proportion recovery improves with an increase in sample size, a decrease in the number of segments and a decrease in the perturbation value. The results corroborate with the findings of Camilleri and Portelli (2007); Wedel and DeSarbo (1995); and Vriens, Wedel and Wilms (1996).

REFERENCES

Camilleri, L., Portelli, M. (2007). Segmenting the heterogeneity of tourist preferences using a Latent Class model combined with the EM algorithm, *Proceedings of the 6th APLIMAT International Conference, Bratislava*. 343-356.

Dempster, A.P., Laird, N.M. Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm, *Journal of the Royal Statistical Society*, B, 39, 1-38.

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

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

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, Item Response models, Generalized Estimation Equations models, Multilevel models, Structural Equations models and Survival models. He is presently an associate professor in the department of Statistics and Operations Research at the University of Malta.