Heterogeneous linear mixed models applied to the segmentation of long-term liking data

Marijn Hazelbag

Unilever Confidential

Supervisors:
Claire Boucon, Unilever R&D Vlaardingen
Prof. Paul Eilers, Erasmus MC

Specialization: Statistical Science

Unilever

Universiteit Leiden

Mathematical Institute Leiden

Contents

1 Material and methods **4**

1.1 Material 4

1.2 Heterogeneous linear mixed model 4

1.3 Model selection criteria 6

1.4 Tests for model parameters 7

1.5 Software 7

A Latent process model and transformation **7**

A.1 Introduction 7

A.2 Method 7

B Model specification in R **8**

Abstract

We will discuss the use of heterogeneous linear mixed models to analyze long term liking data. Random effects are modelled as a mixture of Gaussian distributions, highlighting segments of consumers with similar response patterns through time. A posteriori investigation of these segments will be performed using cross-tabulation. Suggestions for future study designs will be made, as these models allow for missing data and measurements on unbalanced time points.

Keywords : longitudinal mixed model hlmm package lcmm psychological scale data ceiling effects curvelinearity latent class regression random coefficient regression finite mixture distributions maximum likelihood estimation

1 Material and methods

1.1 Material

Five personal care products with highly different sensory properties, such as color, fragrance and texture were used. Long term liking data was collected for 330 subjects, which were selected on basis of age, income, education, concern about the use of a personal care product and current use of this type of product in order to fit the target group. The groups were balanced with respect to the main recruitment criteria. The design was parallel, indicating that each respondent got to rate only one of the five products. Subjects were asked to rate their liking on the use of a product for one month at day 1, 3, 7, 14, 21 and day 28. This data was collected on a continuous line scale (0 - 88) [1] with anchor points at the beginning, at the end and at regular intervals. Additional information about the consumers was collected. Neophobia is either high or low, based on a segmentation of consumers regarding general neophobia and product neophobia. Continuation of use was questioned at day 28 where respondents could choose from: 1 continue using this product, 2 go back to my usual brand and 3 try another brand. Since there were only two respondents who choosed 3, we decided to take 2 and 3 together. Expected benefit of the product was registered at day 1.

1.2 Heterogeneous linear mixed model

We will start by explaining the homogeneous linear random coefficient model: Let $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{ij})$ be the response vector for subject i at occasion j. With $i = (1, 2, ..., I)$ and $j = (1, 2, ..., J)$. The linear mixed model in equation 1 for the $j \times 1$ response vector Y_i, is defined as:

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i \tag{1}$$

Where X_i is a $j \times p$ design matrix for the p-vector of fixed effects. Z_i is a $j \times q$ design matrix for the q-vector of random effects u_i, which represents the subject specific regression coefficients. β. For example we may want to specify a random intercept and a random slope. Giving respondent i an individual intercept which deviates from the general intercept β_0 by u_{0i} and an individual slope which deviates from β_1 by u_{1i}. The errors ϵ_i are assumed to be normally distributed with diagonal covariance matrix $\sigma^2 I_J$ and are assumed to be independent from the vector of random effects u_i. This example is illustrated below for three respondents who were followed during three days.

[1] Originally planned to measure on a 100mm line scale, which was reduced to 88mm upon printing.

$$\left[\begin{array}{c} Y_{11}\\ Y_{12}\\ Y_{13}\\ \hline Y_{21}\\ Y_{22}\\ Y_{23}\\ \hline Y_{31}\\ Y_{32}\\ Y_{33} \end{array}\right] = \left[\begin{array}{cc} 1 & 1\\ 1 & 2\\ 1 & 3\\ \hline 1 & 1\\ 1 & 2\\ 1 & 3\\ \hline 1 & 1\\ 1 & 2\\ 1 & 3 \end{array}\right] \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} + \left[\begin{array}{cccccc} 1 & 0 & 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 2 & 0 & 0\\ 1 & 0 & 0 & 3 & 0 & 0\\ \hline 0 & 1 & 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0 & 2 & 0\\ 0 & 1 & 0 & 0 & 3 & 0\\ \hline 0 & 0 & 1 & 0 & 0 & 1\\ 0 & 0 & 1 & 0 & 0 & 2\\ 0 & 0 & 1 & 0 & 0 & 3 \end{array}\right] \begin{bmatrix}u_{01}\\ u_{02}\\ u_{03}\\ u_{11}\\ u_{12}\\ u_{13}\end{bmatrix} + \left[\begin{array}{c} \epsilon_{11}\\ \epsilon_{12}\\ \epsilon_{13}\\ \hline \epsilon_{21}\\ \epsilon_{22}\\ \epsilon_{23}\\ \hline \epsilon_{31}\\ \epsilon_{32}\\ \epsilon_{33} \end{array}\right]$$

In an homogeneous mixed model, equation 1, u_i is normally distributed with mean μ and covariance matrix D i.e.

$$u_i \sim N(\mu, D) \tag{2}$$

In order to extend this to the heterogeneous linear mixed model, in equations 3 and 4, u_i is assumed to follow a mixture of G multivariate Gaussians with different means (μ_g) and covariance matrix D which can optionally be class-specific D_g.

$$u_{ig} \sim N(\mu_g, D_g) \tag{3}$$

$$\text{where: } D_g = w_g^2 D$$

$$\text{with class proportional parameter: } w_g^2$$

Each class of the mixture has a probability π_g which suffices the following conditions:

$$0 \leq \pi_g \leq 1 \forall g = 1, G \text{ and} \sum_{g=1}^{G} \pi_g = 1 \tag{4}$$

In this model we can have covariates in three different parts of the model. That is why we split up matrix X_i in three parts: matrix X_{1i} contains the variables for the class-membership part, matrix X_{2i} contains the variables for the common effects over classes and X_{3i} contains the covariates for the class-specific effects. Considering G latent homogeneous classes, we define discrete latent variable $c_i = g$ if subject i belongs to class g, where every subject belongs to only one class, namely the class to which the subject has the highest posterior probability. The probability of latent class membership is explained according to covariates X_{1i}, according to multinomial logistic regression:

$$0 \leq \pi_{ig} = P(c_i = g|X_{1i}) = \frac{e^{\xi_{0g} + X_{1i}\xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l} + X_{1i}\xi_{1l}}} \tag{5}$$

The general formulation of the heterogeneous linear mixed model is:

$$(Y_i|c_i = g) \ = \ Z_i u_{ig} + X_{2i}\beta + X_{3i}\gamma_g + \epsilon_i \tag{6}$$

pdfpageduration

where Z_i =the $j \times q$ design matrix for q-vector of random effects u_{ig}
where X_{2i} =the $j \times r$ design matrix for the r-vector of common effects over classes β
where X_{3i} =the $j \times s$ design matrix for the s-vector of class-specific effects γ_g

Given that respondent i belongs to group g, the model is merely the homogeneous linear random coefficient model we saw in equation 1. In other words, the model can be seen as a finite mixture of LMM's. Estimation of $\theta_G = (u_{ig}, \beta, \gamma_g)$ for a fixed number of latent classes G is obtained by maximizing the likelihood:

$$L(\theta_G) = \sum_{i=1}^{N} \ln \left(\sum_{g=1}^{G} \ P(c_i = g|X_{1i}, \theta_G) \times \ \phi_{ig}(Y_i|c_i = g; X_{2i}, X_{3i}, Z_i, \theta_G) \right) \tag{7}$$

$$\phi_{ig} \text{ pdf of} MVN(X_{2i}\beta + X_{3i}\gamma_g + Z_i\mu_g, Z_i B_g Z_i' + \sigma_\epsilon^2 I_{ni})$$

The likelihood is maximized using a modified Marquardt optimization algorithm (1) (**?**). For a given number of classes G this method simultaneously finds the parameters for the trajectories and the multinomial logit part. Doing this for different numbers of G allows us to indentify the optimal number of classes.

1.3 Model selection criteria

Optimal number of classes is found by minimizing BIC (2),see equation 8. As BIC performed best of all IC's considering class enumeration (3). Where k is the number of free parameters to be estimated.

$$\text{BIC} = -2 \ln L + k \ln(n) \tag{8}$$

Wedel and Kamagurka (4) suggest the measure of entropy to quantify the degree of separation in the estimated posterior probabilities. Entropy is calculated in the following way:

$$E_G = 1 - \frac{\sum\limits_i \sum\limits_g (-\hat{p}_{ig} ln(\hat{p}_{ig}))}{n \ ln(G)} \tag{9}$$

Where p_{ik} is the posterior probability of respondent i to belong to class k, and K is the total number of classes. If a respondent has intermediate posterior probabilities different classes, this indicates that the model is uncertain about the classification of this respondent. Values of entropy range between 0 and 1, where 0 indicates classification uncertainty and 1 perfect separation.

1.4 Tests for model parameters

Wald-tests are used for testing the model parameters. The ML estimate $\hat{\theta}$ is compared to a reference value θ_0, usually 0. Where the assumption is that the difference is approximately normally distributed and the square of the difference, see equation 10 below, is compared to a χ^2 distribution. The second assumption is that the standard error of the parameter value is known (while it is actually estimated). The wald-test is liberal.

$$\frac{(\hat{\theta} - \theta_0)^2}{Var(\hat{\theta})} \tag{10}$$

1.5 Software

We use Proust-Lima and Jacqmin-Gadda recently developed R package lcmm for estimation of heterogeneous linear mixed models and latent process heterogeneous mixed models. Within this package function hlme estimates latent class mixed models assuming a gaussian outcome. Their function lcmm extends this approach to handle non Gaussian quantitative and ordinal outcomes.

A Latent process model and transformation

A.1 Introduction

In (5) Proust-Lima et al. conclude: "To distinguish the impact of a covariate on the initial level of a scale from its impact on the change in quantitative scale scores over time (not only psychometric tests but also scales evaluating quality of life or activities of daily living), mixed models that account for their metrologic properties should be preferred over the LMM." In the long term liking data there may be varying sensitivity of the liking scale to a change in the underlying appreciation of the product, this is called curvelinearity. We will apply a latent process homogeneous mixed model in order to account for this.

A.2 Method

In order to allow for transformations of the response, we can extend this to the latent process heterogeneous linear mixed model as follows:

$$h(Y_{ij};\eta) = a + b\Lambda_{ij} + \epsilon_{ij} \tag{11}$$

With a and b parameters needing to be estimated that replace β_0 and the variance of u_{0i}.

$$\Lambda_{ij} = Z_{ij}u_{ig} + X_{2ij}\beta + X_{3ij}\gamma_g \tag{12}$$

The log-likelihood of interest is the log-likelihood of the outcomes in their natural scale, and thus includes the Jacobian of the transformation h. It is given by:

$$L(y;\theta) = L(\hat{y};\theta) + ln(J(y;\theta)) = \sum_{i=1}^{N} L(y;\theta) + \sum_{i=1}^{N} \ln(J(y_i;\theta)) \tag{13}$$

Where θ is the complete vector of parameters containing the transformation parameters $\eta = (\eta_1 ... \eta_n)$, the fixed parameters β, class specific parameters γ and the variance covariance paramters D. The parameters for the transformation are estimated simultaneously with all the other parameters. The Beta Cumulative Distribution Function can take very different shapes, including concave convex and sigmoid. Lcmm uses a four parameter beta transformation.

B Model specification in R

After downloading the lcmm package for R. We can use: library(lcmm) in order to use the functions within the package. Data should be in long format and should be sorted with respect to the subject variable (This with respect to keeping track of the class-membership of the respondents.). The model from section **??** is specified in the following way:
hlme(fixed= liking ∼ Day*Product, random=∼ 1+Day, data= mydataset, subject='RespNr')

The model from section **??** is specified in the following way:
hlme(fixed= liking ∼ Day, mixture=∼ 1 + Day, random=∼ 1 + Day, ng=3, data= nschin, subject= 'RespNr', nwg= TRUE)

The model from section **??** is specified in the following way:
hlme(fixed= liking ∼ Day + Product, mixture=∼ 1 + Day, random=∼ 1 + Day, ng=3, data= nschin, subject= 'RespNr', nwg= TRUE, classmb=∼neophobia)

References

[1] Cecile Proust Lima and Helene Jacqmin-Gadda, Estimation of linear mixed models with a mixture of distributions for the random effects. Computational Methods Programs Biomed, 1992, 78(2), 165-173.

[2] Leroux, B.G., Consistent estimation of a mixing distribution. The Annals of Statistics, 2005, 20(3), 1350-1360.

[3] Nylund, K.L. and Asparouhov, T. and Muthen, B.O., Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural Equation Modeling, 2007, 14(4), 535-569.

[4] Nylund, K.L. and Asparouhov, T. and Muthen, B.O., Market segmentation: Conceptual and methodological foundations. Springer, 2000, 8.

[5] Proust-Lima, C. and Dartigues, J.F. and Jacqmin-Gadda, H., Misuse of the Linear Mixed Model When Evaluating Risk Factors of Cognitive Decline. American journal of epidemiology, 2011, 174, 1077-1088.