The Need for Market Segmentation in Buy-Till-You-Defect Models

E. Korkmaz, D. Fok, R. Kuik

Buy-till-you-defect [BTYD] models are built for companies operating in a non-contractual setting to predict customers' transaction frequency, amount and timing as well as customer lifetime. These models tend to perform well, although they often predict unrealistically long lifetimes for a substantial fraction of the customer base. This obvious lack of face validity limits the adoption of these models by practitioners. Moreover, it highlights a flaw in these models. Based on a simulation study and an empirical analysis of different datasets, we argue that such long lifetime predictions can result from the existence of multiple segments in the customer base. In most cases there are at least two segments: one consisting of customers who purchase the service or product only a few times and the other of those who are frequent purchasers. Customer heterogeneity modeling in the current BTYD models is insufficient to account for such segments, thereby producing unrealistic lifetime predictions.

We present an extension over the current BTYD models to address the extreme lifetime prediction issue where we allow for segments within the customer base. More specifically, we consider a mixture of log-normals distribution to capture the heterogeneity across customers. Our model can be seen as a variant of the hierarchical Bayes [HB] Pareto/NBD model. In addition, the proposed model allows us to relate segment membership as well as within segment customer heterogeneity to selected customer characteristics. Our model, therefore, also increases the explanatory power of BTYD models to a great extent. We are now able to evaluate the impact of customers' characteristics on the membership probabilities of different segments. This allows, for example, one to a-priori predict which customers are likely to become frequent purchasers.

The proposed model is compared against the benchmark Pareto/NBD model (Schmittlein, Morrison, and Colombo 1987) and its HB extension (Abe 2009) on simulated datasets as well as on a real dataset from a large grocery e-retailer in a Western European country. Our BTYD model indeed provides a useful customer segmentation that allows managers to draw conclusions on how customers' purchase and defection behavior are associated with their shopping characteristics such as basket size and the delivery fee paid.

Keywords: Buy-till-you-defect models, segmentation, mixture of normals, Bayesian estimation, customer base analysis.

1 Introduction

The majority of online retailers operates under a non-contractual setting where customers can stop buying from the company without letting the company know. Although defection by customers is unobserved, it needs to be taken into account if the company wants to generate accurate predictions of individual behavior. Such predictions in turn can help to improve returns on marketing actions by better distributing the limited marketing budgets. Therefore, in the literature a lot of attention is given to modeling customer behavior in non-contractual settings.

Buy-till-you-defect [BTYD] models capture the customer's transaction frequency and timing for companies operating under a non-contractual setting. The common modeling approach is to assume stochastic arrival processes (with steady and heterogeneous rates) for each customer's purchase and defection behavior. While a customer is active (the defection has not arrived yet), her transactions arrive according to the assumed arrival process. Usually a Poisson arrival process is assumed as this requires only limited data on a customer's purchase history. On the population level, the heterogeneity over the customer base is modeled by assuming some standard continuous probability distribution.

Among various BTYD models, the Hierarchical Bayes extension (Abe 2009) of the well-known Pareto/NBD model (Schmittlein, Morrison, and Colombo 1987) stands out. Hereafter we will call this model the "HB model". This model explicitly takes into account the dependency between the behavioral parameters driving the two arrival processes on purchase and defection. More precisely, across the population the rates of the two processes are assumed to follow a joint log-normal distribution. This allows obtaining the correlation between purchase and defection rates. In a situation where this correlation is non-zero, the HB model outperforms other BTYD models in terms of forecasting performance (Korkmaz, Kuik, and Fok 2012).

However, existing BTYD models have a common drawback. In many situations these models generate unexpectedly long lifetime predictions for customers. The predictions are sometimes so extreme that the customer is predicted to remain active in the customer base of the company for at least another thousand years. As we will show in this paper, the HB model on a dataset from an online grocery retailer yields extreme lifetime predictions for a substantial group of customers.[1] This extreme lifetime prediction problem has also been observed by Wübben and

[1] Note that the Pareto/NBD, BG/NBD and PDO models generate extreme lifetime predictions on this dataset as well. We focus on the HB model as it performs the best on this data compared to the other BTYD models. This is due to a significant and strong correlation between the purchase and defection parameters.

Wangenheim 2008 in their empirical validation study. First of all, the extreme predictions indicate that the models could be improved as such predictions are obviously off. Second, such predictions substantially lower the face validity of the BTYD models, making it more difficult to get these models to be used in practice. On the technical level, the extreme predictions may be difficult to explain as it seems a counter-intuitive phenomenon for hierarchical models. One may expect that the multivariate normal heterogeneity distribution would shrink outlying customers toward the center of the population. This would normally result in fewer extremes.

To date, there is not a clear explanation in the literature on the reasons behind the extreme lifetime predictions. Even though there are some models that focus solely on the defection process (Fader, Hardie, and Lee (2005), Jerath, Fader, and Hardie (2011)), the lifetime predictions (one of the two major outputs of the BTYD models) are still not reasonable enough that they can be directly used for managerial decision making.

Our explanation for this phenomenon consists of two parts. First, the data is not very informative on the lifetime of a specific individual. We only observe consumer behavior on a limited time interval and we cannot observe defection directly. Second, the customer base likely contains a number of segments. At least two segments are expected: the customers who only purchase the service/product a few times, and the customers who become frequent buyers. This leads to a multi-modal heterogeneity distribution, which cannot be fitted well using any of the current models. In fact, in case of the log-normal distribution (corresponding to the HB model), the variance is forced to be large in order to capture the one-time users as well as the more regular users. The fact that this inflates the customer lifetimes is not sufficiently penalized through the fit of the model as we only observe the customers for a limited time period. This phenomenon will also lead to biased estimates for individual level parameters. In sum, more attention should be paid to heterogeneity modeling for the BTYD models, especially in the case where multiple customer segments exist.

In what follows, we further investigate the reasons behind the extremely long lifetime estimates. Based on our findings, we propose a new BTYD model that overcomes the lifetime estimation problem. In this new model we propose a mixture of log-normals distribution as the heterogeneity distribution. We show that this not only improves the direct usability of lifetime predictions, but also substantially increases the explanatory power of these models.

Based on our model building, simulation and empirical studies, our contribution is twofold.

First, as the mixture of normals heterogeneity distribution can accommodate multi-modal, heavy-tailed and skewed distributions, we obtain better lifetime predictions than the ones from the Pareto/NBD and the HB models. This is especially true for datasets where there exists inherent multimodality. Second, in line with Van Oest and Knox (2011), Reinartz and Kumar (2000), and Schmittlein and Peterson (1994), we show that different customer segments may exhibit different patterns concerning purchase and defection behavior. We also show that customer characteristics can be linked to this segmentation to gain more insight on the customer base.[2] Using data coming from an online retailer in a Western European country, we illustrate the added explanatory power of the proposed model. The customer characteristics explain how the segments differ from each other.

In this paper, we raise two research questions: (1) Does a heterogeneity distribution that accommodates multimodality lead to better predictive performance of the BTYD model?; and (2) Can we relate segments in the customer base to certain customer characteristics? Especially the second question may be very relevant in practice. Segmentation through latent classes is an important method not only for predictive but also for descriptive studies (Cooil, Aksoy, and Keiningham 2008). If firms are able to predict the segment to which a customer belongs, they can allocate their limited marketing resources in a more efficient way. Based on the predicted segment membership, the customer can be assigned a particular treatment. In other words, effective segmentation allows a company to determine which customers they should try to serve and how to best position their products and services for each segment. Moreover, by a better understanding of the customer base through the relationship between segments and the observable customer characteristics, the company may be able to predict the behavior for a new customer based only on the shopping characteristics from her first purchase.

In the next two sections we briefly review the HB model and give an initial analysis of the extreme lifetime prediction problem. In Section 4, we present the mixture of normals model, hereafter called MHB model, including estimation details. Prediction results from a simulation study showing the contribution of the MHB model compared to its benchmark HB model are presented in Section 5. Section 6 presents the results of our empirical study. General conclusions are discussed in Section 7.

[2]This extends the results of Van Oest and Knox (2011) who show using a modified BG/NBD model that customer complaints can be indicators of customer defection.

2 The Hierarchical Bayes BTYD Model

All BTYD models describe the transaction behavior of individuals $i = 1, \ldots, N$ over a time period starting at the first transaction for each individual. As the time of the first purchase of the individuals usually do not coincide, each individual is observed for a different length of time. We measure time relative to the first purchase. Hence, for each customer $t = 0$ corresponds to the time of the first purchase. We denote the total observation time for customer i as T_i.

In BTYD models, customer i remains active for a stochastic and unobserved lifetime which is denoted by $t_{\Delta,i}$. The Pareto/NBD (Schmittlein, Morrison, and Colombo 1987) and the HB model (Abe 2009) have the same individual level assumptions: The customer makes purchases according to a Poisson process with rate λ_i until the lifetime ends (defection occurs), and her lifetime $t_{\Delta,i}$ has an exponential distribution with rate μ_i. The observed customer data is denoted by the vector $[x_i, t_{x,i}, T_i]$, where x_i represents the number of repeat purchases, and $t_{x,i}$ represents the time of the last observed purchase.[3] Using these distributional assumptions, we obtain[4]

$$\begin{aligned} \text{Prob}(X_i = x|\lambda_i, t_{\Delta,i}, T_i) &= e^{-\lambda_i(t_{\Delta,i} \wedge T_i)} \frac{(\lambda_i(t_{\Delta,i} \wedge T_i))^x}{x!}, \\ \pi(t_{\Delta,i}|\mu_i, T_i) &= \mu_i e^{-\mu_i t_{\Delta,i}} \ . \end{aligned} \tag{1}$$

The purchase and the defection rates are assumed to be distributed according to some standard distributions across the population. While Schmittlein, Morrison, and Colombo (1987) assume two independent gamma distributions for the Pareto/NBD model, Abe (2009) relaxes the independence assumption by employing a bivariate log-normal distribution in his hierarchical Bayesian extension of the Pareto/NBD model. In the HB model, it is also possible to incorporate observed customer characteristics. These characteristics for individual i are collected in a $(1 \times R)$ row vector D_i. This vector does not contain a constant. Using the row vector $\theta_i = [\log(\lambda_i), \log(\mu_i)]$ the HB model suggested by Abe 2009 specifies

$$\theta_i|\beta, \Gamma, \Delta \sim N(\beta + D_i\Delta, \Gamma), \tag{2}$$

where β is a (1×2) vector of intercepts, Δ is an $(R \times 2)$ matrix of coefficient parameters and Γ denotes a (2×2) variance-covariance matrix.

[3] Thanks to the memorylessness property on the inter-arrival time distribution, $[x_i, t_{x,i}, T_i]$ summarizes customer i's full history without loss of information.

[4] The value $(t_{\Delta,i} \wedge T_i)$ is the minimum of $t_{\Delta,i}$ and T_i.

3 An initial investigation of the lifetime prediction problem

To understand whether the extreme lifetime prediction problem stems from an inherent characteristic of the HB model, or from a lack of fit of the model, we conduct an initial simulation study.[5] For this purpose, we generate data exactly matching the assumptions of the model, that is, Poisson arrivals combined with an exponential lifetime for the individuals, and a bi-variate log-normal for the heterogeneity distribution. For now we assume that customer characteristics are not available. The four steps of the data generation process are as follows:

1. Fix the hyper-parameters (β and Γ) to some *known* values:

   We choose the following values, $\beta^*_\lambda = \log(0.08)$ and $\beta^*_\mu = \log(0.04)$.[6] The variance-covariance matrix is chosen to be equal to the identity matrix.

2. Draw behavioral parameters θ^*_i for $i = 1, \ldots, N$ according to the heterogeneity distribution:

   Draw $\theta^*_i \sim \pi(\theta_i|\beta^*, \Gamma^*)$ from the multivariate normal distribution. Here we take $N = 1,000$.

3. Draw lifetimes, $t^*_{\Delta,i}$ for $i = 1, \ldots, N$ according to the specified lifetime distribution:

   Draw $t^*_{\Delta,i} \sim \pi(t_{\Delta,i}|\theta^*_i)$ from an exponential distribution with rate parameter e^{θ_i} for customer i.

4. Draw the number of repeat transactions x_i and the last purchase time $t_{x,i}$, given an observation period T_i, lifetime $t^*_{\Delta,i}$ and behavioral parameters θ^*_i:

   For $i = 1, \ldots, N$, draw $x_i, t_{x,i} \sim \pi(x_i, t_{x,i}|t^*_\Delta, \theta^*_i, T_i)$.[7] We fix the observation period length T_i to 154 days.

We next apply Markov Chain Monte Carlo [MCMC] simulation to obtain estimates of parameters from the generated data. In this ideal setting we do not find any evidence of extreme lifetimes using the HB model. Contrary to common findings on real data, all lifetime predictions are reasonable and they tend to shrink towards the center of the data. Figure 1 contrasts the predictions against the true, simulated lifetimes. In the plot on the right hand-side we zoom in on shorter lifetimes where we observe that the HB model can retrieve the true values of the lifetime to a large extent.

[5] All calculations throughout the paper are performed using MATLAB R2011b.

[6] Note that, if no covariate data is used, or in case covariates are mean-centered, β values give the mean of the log behavioral parameters.

[7] See the details of this sampling process in the 5^{th} step of generating data for MHB model testing (for segmented data) given in Appendix C.

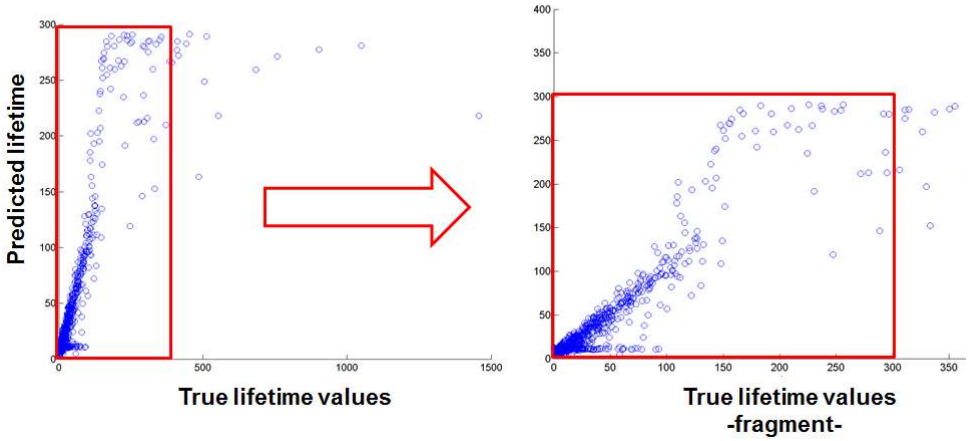


Figure 1: Lifetime predictions from the HB model versus true lifetimes on a generated dataset

Based on this simulation study, we conclude that the HB model gives reasonable lifetime predictions, if it is applied to a dataset that satisfies all model assumptions. The extreme lifetime predictions that are obtained for real data are, therefore, most likely due to a violation of one of the model assumptions. This conclusion is the very motivation of this paper. We believe that the HB model's fit problem stems from the fact that the log-normal distribution (or the gamma distribution for the Pareto/NBD model) does not accurately capture the true population distribution. The true distribution is likely to be multi-modal, as the population contains various types of customers. The existence of individuals with very short lifetimes leads to a thick right-hand tail of the log defection rate distribution; and due to the symmetry of the normal distribution we also obtain a thick left-hand tail. For the individuals in this part of the distribution, we might erroneously conclude that their defection rate is virtually zero (leading to infinitely long lifetime predictions). All in all, we need to capture the multimodality in the data to avoid drawing wrong conclusions on the customer level.

4 Mixture HB BTYD Model

Based on our earlier motivation, we propose to model customer heterogeneity in a way that allows for latent classes. We label this model the mixture HB [MHB] model. We propose two different variants of the MHB model. In the first variant, a-priori segment probabilities are independent of customer covariates. In the second, we allow covariates to influence the segment probabilities. In the mixture model literature such covariates are called concomitant variables.

In principle one would be able to obtain better predictive performance with the second model that accommodates concomitant variables.

4.1 MHB Model without Concomitant Variables

To allow for a multi-modal heterogeneity distribution, we replace the multivariate normal distribution over the log purchase and log defection rates by a mixture of K multivariate normal distributions.[8] One can also view this as a distribution that allows for K segments in the population. However, within a segment customers may still differ from each other. The mixtures of normals approach provides a great deal of flexibility. First, it may capture a distribution with multiple modes. Next, it could capture a distribution with fat tails if one of the components is a normal component with a large variance. The mixture of normals approach has become quite popular in marketing due to its flexibility and the potential interpretation of each mixture component as representing a 'segment'. Finally, the parameters in these models are relative easy to estimate (Rossi, Allenby, and McCulloch 2005).

More formally, we write the heterogeneity distribution as

$$\begin{aligned}
\theta_i &= D_i\Delta + \eta_i, \\
\eta_i &\sim \mathrm{N}(\beta_{s_i}, \Gamma_{s_i}), \\
s_i &\sim \mathrm{Multinomial}_K(p),
\end{aligned}$$

where s_i indicates the segment to which customer i belongs to. With each segment (or component) we associate a mean vector and a variance-covariance matrix, namely β_k and Γ_k, $k = 1, \ldots, K$. The vector p contains the K segment probabilities where their values sum up to 1.

The proposed model is visualized in Figure 2.[9] The joint distribution of the observable data

[8] Data examination shows us that there are generally two major segments in the customer base of grocery e-tailers, namely frequent and incidental buyers. However in the model we present here, we do not fix the number of latent components.

[9] Figure 2 helps us to easily identify the direct dependency relationships between neighboring parameters. Note that the joint distribution of the observable data and all latent variables and parameters in Equation (3) holds since $(x_i, t_{x,i}, T_i), t_{\Delta,i}, z_i$ are independent of $p, \Delta, \beta_{s_i}, \Gamma_{s_i}$ given θ_i.

and all latent variables and parameters can be decomposed as

$$\pi(\{(x_i, t_{x,i}), t_{\Delta,i}, z_i, \theta_i, s_i\}_{i=1}^N, \Delta, \{\beta_k, \Gamma_k\}_{k=1}^K, p)$$
$$= \prod_{i=1}^{N} \left[\pi((x_i, t_{x,i})|t_{\Delta,i}, z_i, \theta_i)\, \pi(t_{\Delta,i}|z_i, \theta_i)\, \pi(z_i|\theta_i)\, \pi(\theta_i|\Delta, \beta_{s_i}, \Gamma_{s_i})\, \pi(s_i|p)\right] \times$$
$$\pi(\Delta)\, \pi(p) \prod_{k=1}^{K} \left[\pi(\beta_k|\Gamma_k)\, \pi(\Gamma_k)\right]. \quad (3)$$

The observables are x_i, $t_{x,i}$ and T_i. The variables z_i and $t_{\Delta,i}$ relate to the unobserved defection process. z_i is a latent binary indicator denoting whether customer i is active ($z_i = 1$) or inactive ($z_i = 0$) at the end of the calibration period (T_i). The latent lifetime is given by $t_{\Delta,i}$. The set of values $(x_i, t_{x,i}), (t_{\Delta,i}, z_i), \theta_i, s_i$ are distributed independently across individuals when conditioned on $(\Delta, p, \{\beta_k, \Gamma_k\}_{k=1}^K)$.

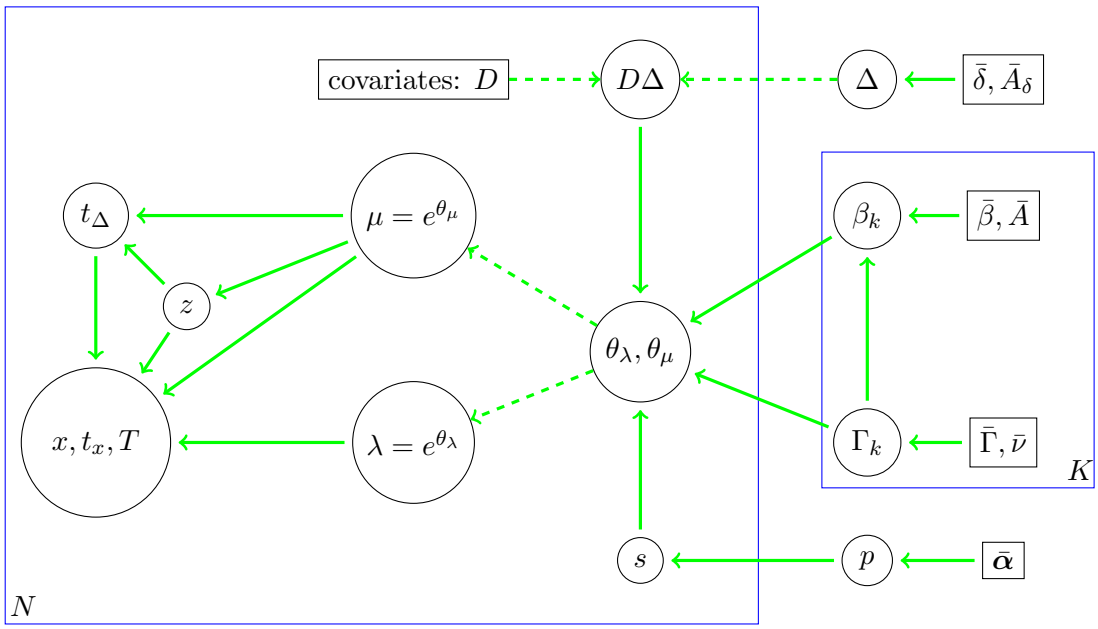


Figure 2: Customer purchase and defection behavior model. Constant values are enclosed by rectangles. Each variable in the big box is of dimension N, representing each customer. Each value in the smaller box is of dimension K, representing each latent component. The value of the indicator variable $s \in \{1, \cdots, K\}$ picks one out of K components with β_k and Γ_k; $k = 1, \ldots, K$. The covariates, D, are assumed not to include an intercept. The intercept is modeled through β_k. The dashed lines represent deterministic relations.

As said, D_i is the observable characteristics (covariate) row vector of an individual and does not include an intercept. We follow the advice by Rossi, Allenby, and McCulloch (2005, Page 144) to mean-center all covariates, so that the mean of θ for the average customer is entirely determined

by the mixture component means (β_k). Therefore $\mathbb{E}[\theta_i|D_i = \overline{D}, p, \{\beta_k\}_{k=1}^K] = \sum_{k=1}^K p_k\beta_k$.

We choose the standard conditionally conjugate priors to complete the model specification, that is,

$$\begin{aligned}
\text{vec}(\Delta) = \delta &\sim N(\bar{\delta}, A_\delta^{-1}),\\
p &\sim \text{Dirichlet}(\alpha),\\
\beta_k|\Gamma_k &\sim \text{N}(\bar{\beta}, \Gamma_k \otimes \bar{A}),\\
\Gamma_k &\sim \text{IW}(\bar{\Gamma}, \bar{\nu}).
\end{aligned}$$

IW denotes the Inverse Wishart distribution. A discussion on setting the values of the prior parameters is presented in Section 6.

Bayesian inference

The posterior distribution for all parameters and latent variables is not available in closed form. We use MCMC sampling for inference on the parameters and the latent variables for the MHB model. More specifically, we use a Metropolis within Gibbs sampler (see Hastings (1970) and Geman and Geman (1984)). The sampler uses the latent variables z_i and $t_{\Delta,i}$. We present the main steps of the sampler below, details of the sampling procedure are given in Appendix A.
MCMC sampler for MHB model:

[0] Set initial values for $\theta_i, i = 1, \ldots, N$, and repeat the following.

[1a] Generate $z_i|x_i, t_{x,i}, T_i, \theta_i$ according to the being active probability $\frac{\lambda_i}{\lambda_i+\mu_i e^{(\lambda_i+\mu_i)(T_i-t_{x,i})}}$ (as given in Equation (3) in Schmittlein, Morrison, and Colombo (1987)), for $i = 1, \ldots, N$.

[1b] Generate $t_{\Delta,i}|x_i, t_{x,i}, T_i, z_i, \theta_i$ using an exponential distribution with rate $(\mu_i + \lambda_i)$ truncated to $(t_{x,i}, T_i)$ if $z_i = 0$; and an exponential distribution with rate μ_i truncated to (T_i, ∞) if $z_i = 1$ (see Equation (8)).

[2a] Calculate $\tilde{p}_{ik}|\theta_i, D_i, \Delta, \beta_k, \Gamma_k, p_k$, the conditional posterior membership probabilities of customer i for component k using Equation (10) in Appendix A.

[2b] Generate $s_i|\tilde{p}_i$, the indicator variable for the segment to which the customer i belongs by drawing from a multinomial distribution with parameters $\tilde{p}_i = [\tilde{p}_{i1}, \cdots, \tilde{p}_{iK}]$.

[3] Generate $\beta_k|\theta, \Delta, s, \Gamma_k$ and $\Gamma_k|\theta, \Delta, s$ for each latent class k using a multivariate normal regression update (see Rossi, Allenby, and McCulloch (2005, Page 34)). Note that $\pi(\beta_k, \Gamma_k|\theta, \Delta, \{s_i\}_{i=1}^N)$ does not depend on rates θ_i for those customers that do not belong to the component k. Let $\theta^{(k)}$ be the matrix of behavioral parameters for those customers who belong to segment k, that is, $\theta^{(k)} = \{\theta_i\}_{i:s_i=k}^N$. Then

$$\begin{aligned}\pi(\beta_k, \Gamma_k|\theta, \Delta, \{s\}_{i=1}^N) &= \pi(\beta_k, \Gamma_k|\theta^{(k)}, \Delta) \\ &\propto \pi(\theta^{(k)}, \Delta, \beta_k, \Gamma_k) \\ &= \pi(\theta^{(k)} - D^{(k)}\Delta|\beta_k, \Gamma_k)\, \pi(\beta_k|\Gamma_k)\, \pi(\Gamma_k) \qquad (4)\end{aligned}$$

[4] Generate $\Delta|\theta, \beta, \Gamma, s$, the regression coefficients over the whole population, using a standard multivariate regression update; $\Delta \sim \pi(\Delta|\theta, \beta, \Gamma, s)$. For this step, the data should be pooled from K components (see Rossi, Allenby, and McCulloch (2005, Page 148)). Details on Δ sampling are provided in Appendix A.

[5] Draw p conditional on $\{s_i\}_{i=1}^N$. This conditional distribution is a Dirichlet, that is, update on the membership probabilities of the components: $p|\{s_i\}_{i=1}^N \sim \text{Dir}(\alpha_1 + \sum_{i=1}^N I[s_i = 1], \ldots, \alpha_K + \sum_{i=1}^N I[s_i = K])$, where $I[A]$ denotes an indicator function which equals one if condition A is true, and zero otherwise.

[6] Generate $\theta_i|t_{x,i}, x_i, T_i, z_i, t_{\Delta,i}, \beta_{s_i}, \Gamma_{s_i}$ with a Gaussian random-walk Metropolis Hastings [MH] algorithm, for $i = 1, \ldots, N$. The step size in the random-walk MH algorithm is set by applying an adaptive MH method in the burn-in phase (Gilks, Richardson, and Spiegelhalter 1996).

4.2 Mixture Model with Concomitant Variables

In the previous section, the prior segment probability was equal for all customers. This implies that without a purchase history we cannot distinguish the different types of customers. In this section we extend the model using concomitant variables such that the prior segment probabilities depend on customer characteristics.

We replace the common vector p by an individual specific vector p_i. To relate these probabilities to customer characteristics we build on the multinomial probit [MNP] model. As is common

in the MNP model we introduce latent customer specific "utilities" for each segment. These utilities are denoted by u_{ik}, for $i = 1, \ldots, N$ and $k = 1, \ldots, K$, and they may depend on the concomitant variables C_i as

$$u_{ik} = C_i \omega_k + \varepsilon_{ik}, \tag{5}$$

where $\varepsilon_{ik} \sim N(0, 1)$ and C_i contains a constant next to L concomitant variables. Finally we set ω_K to a vector of zeros (with length $(L + 1)$) for identification (Paap and Franses 2000). Given the utilities, the segment to which a customer belongs to is completely determined. The customer is assigned to the segment that has the highest utility, that is,

$$s_i = \text{argmax}_k u_{ik}. \tag{6}$$

The model is visualized in Figure 3. Every relationship in Figure 3 is defined in terms of probability distributions (solid arrows) or in a deterministic way (dashed arrows). Note that the probabilities of belonging to a segment depend on the distribution of the utilities. This latter distribution is a function of the MNP model's coefficients $\omega_1 \ldots, \omega_K$.

The joint distribution of the data and parameters now becomes,

$$\begin{aligned}
&\pi(\{(x_i, t_{x,i}), t_{\Delta,i}, z_i, \theta_i, s_i, u_i\}_{i=1}^N, \Delta, \{\beta_k, \Gamma_k\}_{k=1}^K, \omega) \\
&= \prod_{i=1}^{N} \left[\pi((x_i, t_{x,i})|t_{\Delta,i}, z_i, \theta_i)\, \pi(t_{\Delta,i}|z_i, \theta_i)\, \pi(z_i|\theta_i)\, \pi(\theta_i|\Delta, \beta_{s_i}, \Gamma_{s_i})\, I[s_i = \text{argmax}_k u_{ik}]\, \pi(u_i|\omega)\right] \\
&\qquad \times \pi(\Delta)\, \pi(\omega) \prod_{k=1}^{K} \left[\pi(\beta_k|\Gamma_k)\, \pi(\Gamma_k)\right], \quad (7)
\end{aligned}$$

where $u_i = (u_{i1}, \ldots, u_{iK})$ and $\omega = (\omega_1, \ldots, \omega_K)$. Both in Equation (3) and Equation (7), the dependence of densities on prior parameters has been suppressed.

Bayesian inference

We again use a Metropolis within Gibbs sampler to obtain the posterior conditional densities for each of the parameters. Note that to satisfy the *irreducibility* requirement of the Markov chain the sampler needs to skip the deterministic relationships between parameters. Therefore, we do not sample the segment indicators s_i; these are determined through the utilities u_{ik} as in Equation (6).

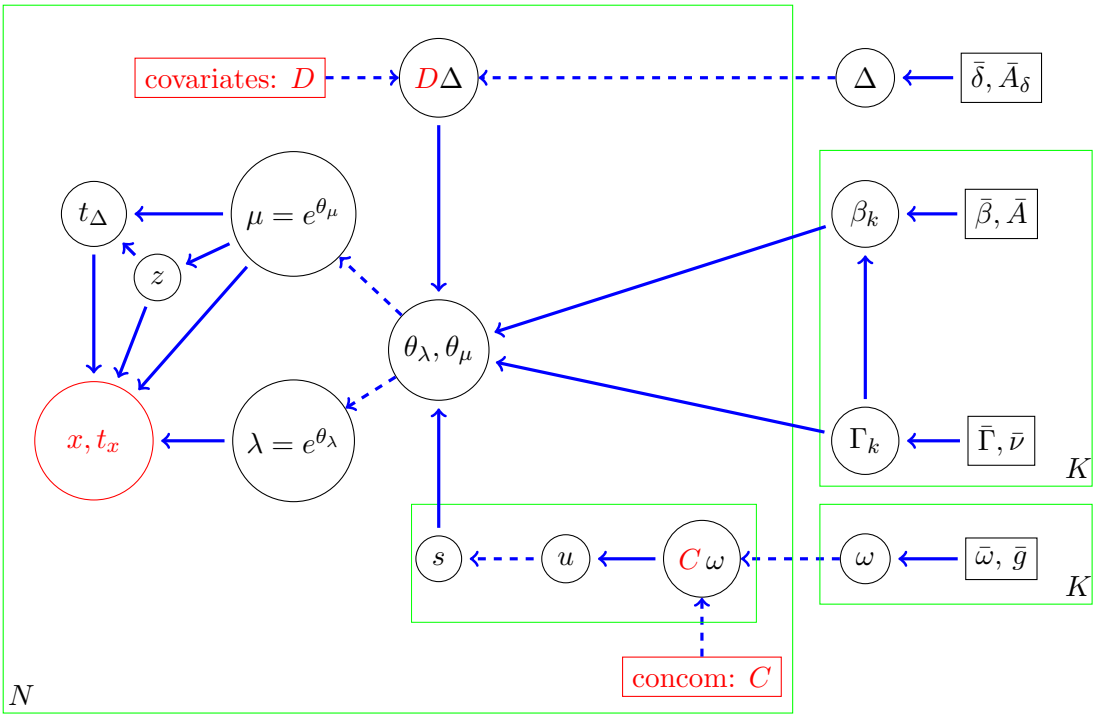


Figure 3: Customer purchase and defection behavior model with latent classes. Constant values are enclosed by rectangles. Each variable in the big box is of dimension N, representing each customer. Each data structure in the smaller boxes on the right hand side of the figure is of dimension K, representing different latent components. The matrices of the inner box are of dimension $(N \times K)$. Red color represents observables. The dashed line represents a deterministic relation rather than a probabilistic one.

The resulting sampler is very similar to the one for the previous model. The only difference is in the assignment of customers to different latent components. Therefore, only the second and the third steps of the Gibbs Sampler are different in this sampler. In these steps we update the utility values for each customer and the component-specific probit coefficients ω. The other steps of the sampler are identical to those given under MHB model without concomitant variables. The MCMC sampler becomes:

[0] Set initial values for $\theta_i, i = 1, \ldots, N$, and repeat the following.

[1a] Generate $z_i | t_{x,i}, x_i, T_i, \theta_i$.

[1b] Generate $t_{\Delta,i} | t_{x,i}, x_i, T_i, z_i, \theta_i$.

[2a] Generate $u_i | C_i, \omega, D_i, \Delta, \theta_i, \beta, \Gamma$, the utility row vector of customer i for the latent segments.

[2b] Update the segment indicators $s_i | u_i$ that assign customers to one of the K components according to the component that has the highest utility value.

[3] Generate $\omega|u$, the latent component specific coefficients using a standard multivariate normal regression update.

[4] Generate $\beta_k|\theta, \Delta, s, \Gamma_k$ and $\Gamma_k|\theta, \Delta, s$ for each latent class k.

[5] Generate $\Delta|\theta, \beta, \Gamma, s$ using a standard multivariate update after pooling data from K components.

[6] Generate $\theta_i|t_{x,i}, x_i, T_i, z_i, t_{\Delta,i}, \beta_k, \Gamma_k$ with a Gaussian random-walk MH algorithm.

The details of the sampling procedures for the nodes ω and $\boldsymbol{u}$ are presented in Appendix B.

5 Model Testing on Generated Data

In order to evaluate the performance of the proposed BTYD models with latent classes, we start by testing them on generated datasets. We generate data based on some *known* parameter values and next see whether we can retrieve those values using the models. This also provides a test to see if our implementation of the MCMC sampler is done properly and converges fast. This approach is especially crucial as some events are unobservable. In our case the segment allocation and the actual lifetime are not observable in a real-life setting. Furthermore, we assess the effects of misspecification, that is, using HB instead of MHB model.

We present the data generation process and some statistics on the generated dataset in Section 5.1. Following that, we present the prediction performance of each model under comparison (MHB models with and without concomitant variables, and the HB model). In Section 5.3, we give a robustness analysis of the proposed models by testing all models' predictive performance on a generated data with a unimodal heterogeneity distribution.

5.1 Data Generation

Considering $N = 1,000$ customers and $K = 2$ latent components, we generate a transaction dataset for $T = 200$ days following the three major steps. Details of the data generation, including the exact parameter values, are given in Appendix C.

1. Allocate customers to components $(s_i^*|\omega^*)$:

Fix the component specific regression coefficient matrix to its *true* value ω^*; generate *true* utilities such as $u^* = C\,\omega^* + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, 1)$; and assign each customer to the component with the the highest utility.[10]

2. Generate customer specific behavioral parameters $\theta_i^* | \beta_{s_i}^*, \Gamma_{s_i}^*$:

   Fix the true hyper-parameter values β^* and Γ^* for each of the components; generate *true* behavioral parameters for each customer by sampling from a MVN distribution such as $\theta_i^* \sim \pi(\theta_i | \beta_k^*, \Gamma_k^*)$.

3. Generate customer lifetime $(t_\Delta^* | \theta^*)$ and transaction data $((x, t_x) | \theta^*, t_\Delta^*, T)$:

   Draw $t_{\Delta,i}^* \sim \pi(t_{\Delta,i} | \theta_i^*)$ from an exponential distribution with the rate parameter of $\theta_{\mu,i}^*$. Given an observation period T and lifetime $t_{\Delta,i}^*$, generate number of transactions and the time of the last purchase based on Poisson purchase arrivals.[11]

The data generation is in line with Section 3, apart from the segmentation of customers. We generate one, uninformative covariate D from a standard uniform distribution. As we mean-center all covariate data, it does not affect the mean values of the (component-specific) hyper-parameters. The concomitant variable C is on the other hand chosen to be informative. In order to keep things simple, for the first half of the data, the concomitant variable is set to 1 and for the other half to -1. Note that randomness is introduced on customers' assignment to components by the utility generation in the first step of generating data.

Table 1 shows some descriptive statistics on the generated data. In this dataset we can easily distinguish the two different components, namely Segment 1 with loyal customers and Segment 2 with customers who quickly stop buying. The final two rows show that the concomitant variable cannot perfectly determine the segment allocation.

5.2 Estimation Evaluation Scheme

In this section we compare the predictive performance of the three models: the HB model proposed by Abe (2009), the MHB model without concomitant variables, and the MHB model with such variables. We run all the models on the generated data and compare the results on both population and individual levels. For all the hierarchical Bayes models under comparison,

[10] We fix ω^*, the $((L+1) \times K)$ MNP probit coefficient matrix to $\left[\begin{smallmatrix} 0.1 & 0 \\ 0.8 & 0 \end{smallmatrix}\right]$ where $L = 1$ is the number of concomitant variables.

[11] See the details of sampling process $x, t_x | \theta^*, t_{\Delta,i}^*, T$ in the 5^{th} step given in Appendix C.

Table 1: Descriptive statistics on the generated data with two components

	All customers	Segment 1	Segment 2
# of customers	1000	528	472
Avg. # of transaction (x)	126.79	238.94	1.34
Std. # of transaction (x)	215.36	247.38	0.80
Avg. last purchase time (t_x)	94.03	171.82	7.01
Std. last purchase time (t_x)	92.62	55.12	20.87
% concomitant (1)	50	68	29
% concomitant (-1)	50	32	71

the MCMC simulation has run $200,000$ iterations of which the last $40,000$ (with a thinning factor of 10) have been used for posterior inference. Markov chain convergence was monitored using trace plots of posterior draws.

5.2.1 Population level comparison

The MHB models can directly be compared to each other as they are both applied to the 2-segment case. However, the HB model cannot be directly compared with the mixture models on the population level due to the smaller number of parameters. We report the true values of segment specific intercept vectors (β_k^*) as well as the posterior mean predictions from the HB and MHB models in Table 2. Note that the values in parenthesis give the standard deviation of the posterior draws for each parameter. The second and the third rows of Table 2 presents the posterior means and standard deviations of the segment specific intercepts (β_k) from the MHB models. These mean β_k values give the population level means of the behavioral parameters (θ vector) for each segment. Based on these two rows, we conclude that both of the MHB models perform well in recovering the *true* parameter values presented in the first row. As expected, the mean estimates for the HB model (presented in the last row of the same table) are in between the MHB model's segment specific estimates.

The true values of segment specific variance-covariance matrix Γ_k^*, and the posterior mean of its predictions from the MHB and HB models are presented in Table 3. Again as the HB model accommodates only one component, there is only one variance-covariance matrix prediction from this model. The most striking result from these tables is the huge difference in the variance of the log defection rate across the models (see $\Gamma_{2,2}$ values). This already hints at a potential cause of extreme lifetime predictions. We will further discuss this in the next section.

Table 2: True (segment specific) intercept vectors (β_k) and their posterior means from MHB and HB models on generated data. As the HB model accommodates one mode, there is only one β prediction from this model. Note that the first element of β is the mean of log purchase rates (θ_λ), and the second is the mean of log defection rates (θ_μ).

	β_1		β_2	
TRUE	0	−6.908	−4.605	−2.996
MHB (without con.)	−0.033	−6.906	−4.585	−2.972
	(0.039)	(0.169)	(0.232)	(0.230)
MHB (with con.)	−0.016	−6.878	−4.687	−2.976
	(0.036)	(0.147)	(0.192)	(0.172)
HB	−1.357	−4.248	-	-
	(0.095)	(0.203)		

Table 3: True (segment specific) variance-covariance matrices (Γ_k) and their posterior mean from MHB and HB models on generated data. As HB model accommodates one mode, there is only one Γ prediction from this model.

	Γ_1	Γ_2
TRUE	$\begin{pmatrix} 0.640 & 0 \\ 0 & 0.640 \end{pmatrix}$	$\begin{pmatrix} 0.640 & 0 \\ 0 & 0.640 \end{pmatrix}$
MHB (without con.)	$\begin{pmatrix} 0.670 & 0.044 \\ 0.044 & 1.250 \end{pmatrix}$	$\begin{pmatrix} 0.837 & 0.113 \\ 0.113 & 0.748 \end{pmatrix}$
MHB (with con.)	$\begin{pmatrix} 0.677 & 0.028 \\ 0.028 & 1.190 \end{pmatrix}$	$\begin{pmatrix} 0.864 & 0.057 \\ 0.057 & 0.787 \end{pmatrix}$
HB	$\begin{pmatrix} 2.76 & -5.28 \\ -5.28 & 19.04 \end{pmatrix}$	-

5.2.2 Individual level comparison

We next compare the model configurations based on the individual level predictions. We focus on the predictions of the purchase and defection rates as well as the predicted lifetime. We measure the predictive performance using the mean absolute error (MAE) and the correlation between the predicted and the true values. Table 4 summarizes the results.

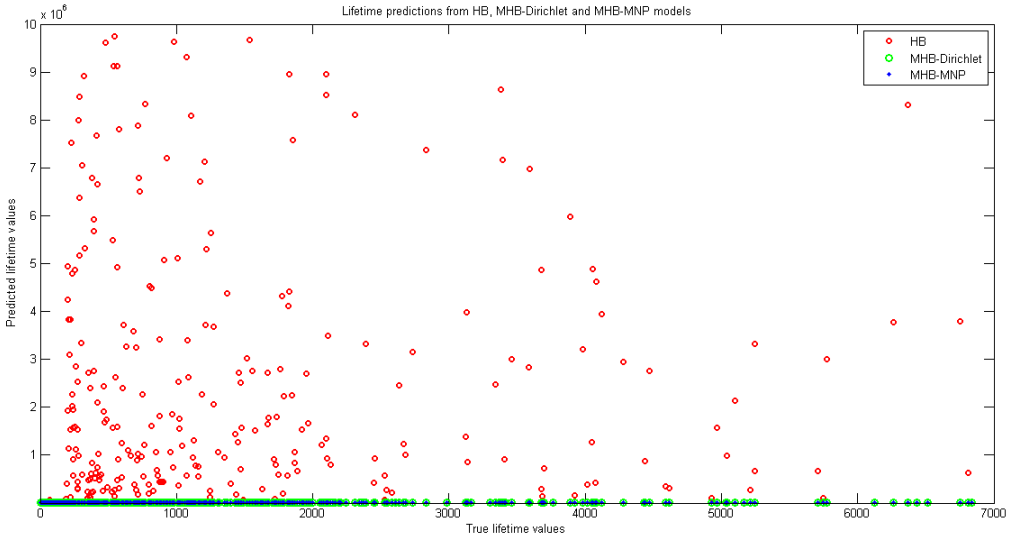
Table 4: Comparison of the models on the individual metrics (MAE and correlation between true values and predicted means) on generated data

		HB	MHB (without con.)	MHB (with con.)
Purchase rate (λ)	MAE	0.086	0.045	0.044
	CORR	0.996	0.997	0.997
Defection rate (μ)	MAE	108,658	0.024	0.023
	CORR	0.035	0.547	0.549
Lifetime	MAE	77,381,052	902	852
	CORR	0.026	0.523	0.526

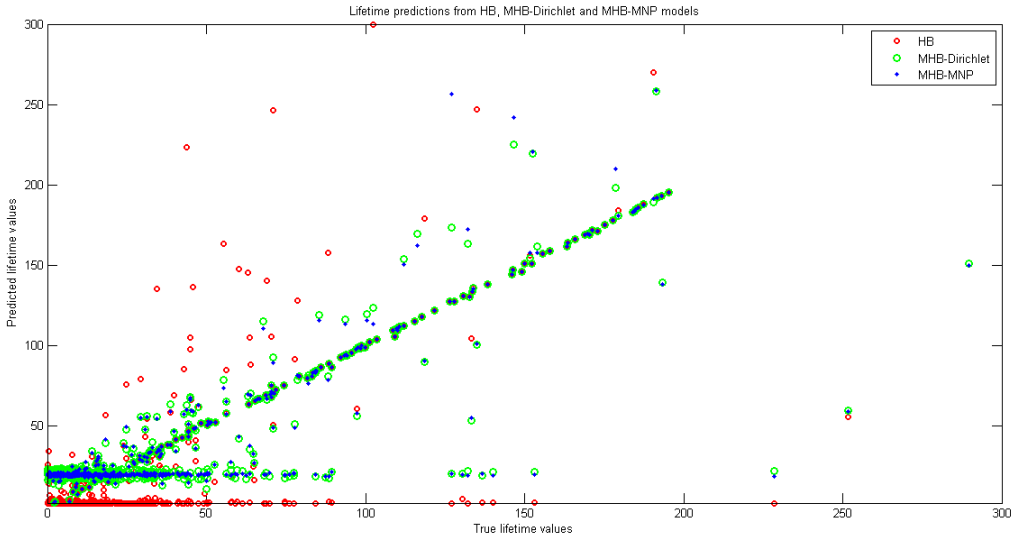
Note that 99.9% of the customers are assigned to their true components for both MHB models.

Table 4 shows that all models perform relatively well on predicting the purchase rate λ. However, the MAE for the HB model is about twice as large as that for the MHB models. When it comes to predicting the defection rate μ and the lifetime, there are enormous differences between the HB and the MHB models. The MHB models predict these measures relatively well, especially considering the fact that we cannot observe the defection. The performance of the HB model clearly demonstrates the earlier mentioned phenomenon of extreme predictions. The predictive performance on the lifetime is illustrated in Figure 4a where it is very easy to observe the extremely long lifetime predictions for the HB model. Figure 4b gives a small fragment of Figure 4a where the axes are limited to the 0 to 300 range. The lifetime predictions based on the HB model hardly show a relation with the true values.

The conclusion from these experiments is quite clear. The MHB models perform well on generated data where there are multiple customer segments. Assuming a unimodal heterogeneity distribution as is done in the HB model can lead to very poor predictive performance on defection and lifetime. In fact the performance is so poor that we observe very extreme lifetime predictions in this case. This confirms our suggestion that such extreme predictions in earlier applications of BTYD models are due to multimodality. We will further investigate this on real data in Section 6.



(a) Scatter plot showing the extreme lifetime predictions from the HB model. Note the vertical scale.



(b) A small fragment of the upper scatter plot - axes limited to 300.

Figure 4: Scatter plots showing the difference in customer lifetime predictions between HB and MHB models on generated data.

5.3 Robustness Analysis on MHB Model - Testing on a unimodal data

We also study the performance of the MHB model relative to the HB model in case the customer base has a unimodal heterogeneity distribution. For this purpose, we have generated new data.[12] Table 5 shows some statistics on this data.

Table 5: Descriptive statistics on the (uni-modal) generated data

# of customers	1000
Avg. # of purchases	5.613
Std. # of purchases	8.965
# of customers with no repeat purchase	367
Avg. last purchase time (t_x)	26.085
Max. last purchase time (t_x)	153.92
Observation time (T)	154

Tables 6 and 7 present the posterior means of the population level parameters from the three models together with the true parameter values. Based on these tables, we conclude that if the MHB model is applied to a dataset where the heterogeneity distribution is unimodal, it does not deteriorate the estimates. All customers are simply assigned to one of the components, leaving the other empty. As a result the predictive performance of the MHB models is only slightly worse than that of the HB model, see Table 8. This loss in predictive performance can entirely be attributed to the fact that MHB model contains more parameters.

6 Empirical Study

In this section, we test our MHB model on real-life data.[13] We first present the explanatory contribution of the MHB model by revealing the segments in the customer base as well as by showing how these segments differ from each other. Next, we compare the predictive performance

[12]The data has been generated in four steps:

1. Fix β^* (hyper-parameters): $\beta_\lambda = \log(0.08)$ and $\beta_\mu = \log(0.04)$. The variance covariance matrix Γ is chosen to be equal to the identity matrix.
2. For $i = 1, \ldots, N$: Draw $\theta_i^* \sim \pi(\theta_i|\beta^*)$ from multivariate normal distribution.
3. For $i = 1, \ldots, N$: Draw $t^*_{\Delta,i} \sim \pi(t_{\Delta,i}|\theta_i^*)$ from an exponential distribution with the rate parameter of e^{θ_μ}.
4. For $i = 1, \ldots, N$: Draw $x_i, t_{x,i} \sim \pi(x_i, t_{x,i}|y_i^*, \theta_i^*)$.

[13]Note that we do not include the MHB model without concomitant variables in this section due to two reasons. First of all, this model is dominated by its counterpart model with concomitant variables due to the ability of explaining how the segments differ from each other. Second, in order to provide a concise overview of the predictive results from the models in comparison, we include only the MHB model with concomitant variables together with the benchmark Pareto/NBD and the HB models.

Table 6: True values and posterior means of β using the MHB (with and without concomitant variables) and the HB models. As the second component from the MHB models becomes empty, β_2 values are not reported.

	β	
TRUE	$-2,526$	-3.219
MHB (without con.) β_1	-2.420	-3.357
	(0.064)	(0.076)
MHB (with con.) β_1	-2.483	-3.293
	(0.064)	(0.067)
HB	-1.357	-4.248
	(0.106)	(0.115)

Table 7: True values and posterior means of Γ using the MHB (with and without concomitant variables) and the HB models. As the second component from the MHB models becomes empty, Γ_2 values are not reported.

	Γ
TRUE	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
MHB (without con.) Γ_1	$\begin{pmatrix} 1.040 & 0.052 \\ 0.052 & 0.991 \end{pmatrix}$
MHB (with con.) Γ_1	$\begin{pmatrix} 0.996 & -0.017 \\ -0.017 & 0.990 \end{pmatrix}$
HB	$\begin{pmatrix} 1.095 & -0.043 \\ -0.043 & 0.947 \end{pmatrix}$

Table 8: Comparison of the models on the individual metrics (MAE and correlation between true values and predicted means) on generated data

		HB	MHB* (without con.)	MHB** (with con.)
Purchase rate (λ)	MAE	0.061	0.073	0.062
	CORR	0.849	0.798	0.848
Defection rate (μ)	MAE	0.040	0.040	0.042
	CORR	0.376	0.372	0.343
Lifetime	MAE	17.165	17.293	17.448
	CORR	0.783	0.782	0.769

* 99.7% of the customers is assigned to Component 1.
** 100% of the customers is assigned to Component 1.

of the MHB model against the benchmark Pareto/NBD and HB models. To provide a fair judgment on the performance of the models in consideration, we focus on out-of-sample predictive power.

The dataset we consider contains daily transaction data of an online grocery retailer in a Western European country (called OG hereafter). We base our analysis on a random set of 1460 customers who started buying from the company in January 2009. We ignore all Sundays as OG does not provide delivery on that day. The data contains the initial and the repeat purchase information of each customer over a period of 309 days. To estimate the model parameters, we use the transaction data of all customers over the first 154 days, leaving a 155 day holdout period for model validation. The transaction data contains information on the number of shopping items, the Euro values of the shopping basket and the delivery fee, the number of discounted items in the basket and also the percentage discount rate of each basket. Table 9 presents some descriptive statistics. According to this table an average customer purchases 11 times in the calibration period. However, this number drops to 9 in the validation period mostly because of customers who have left the company by then. On average, the first transaction of customers contains a basket made up of 64 items of which 6 come with a discount. The average initial basket is worth 126 Euros after discount and the delivery fee is 7 Euros.

Table 9: Descriptive statistics for the OG dataset

# of customers	1460
Available time frame	309 days
Time split (in-sample/out-of-sample)	154/155
Zero repeaters in estimation period (%)	174 (12%)
Zero repeaters in holdout period (%)	295 (20%)
Zero repeaters in estimation and holdout periods (%)	135 (9%)
# of purchases in estimation period (all)	16,252
# of purchases in holdout period	12,827
Avg. # purchases per customer in estimation period (std.)	11.13 (10.76)
Avg. # purchases per customer in holdout period (std.)	8.79 (10.78)
Avg. observation time T (std.)	143.76 (7.39)
Avg. recency rate $((T - t_x)/T)$	0.27
Avg. # of items in the first purchase (std.)	64.34 (40.67)
Avg. # of discounted items in the first purchase (std.)	5.93 (8.14)
Avg. basket value after discount -in €- (std.)	125.73 (71.51)
Avg. discount rate of the basket (%)	4.08%
Avg. delivery fee of the first purchase -in €- (std.)	6.97 (1.37)

We use the number of items in the basket together with the basket value and the delivery fee

from the initial purchase as explanatory factors in our MHB model. These variables are used as covariate and as concomitant variables. We standardize the covariate vector so that the β_k vector represents the average values of the log of the purchase and defection rate for the k^{th} component. Moreover, we applied a log transformation on the number of items in the initial shopping basket as this variable is highly skewed.[14]

There are two points that one needs to pay attention to when applying the MHB model. The first concerns the number of latent components that refers the number of segments in the customer base. To set the number of mixture components, we run the MHB model with different numbers of latent components and choose the optimum one based on the number of customers assigned to each component (Frühwirth-Schnatter 2006). If additional segments become too small, we stop adding segments. We do not use likelihood-based measures as obtaining the marginal likelihood is computationally very challenging, even in the basic BTYD model. As an alternative one may choose the number of segments based on out-of-sample predictive performance. However, in our case we would then have to split our data in three parts, to leave one part for a fair comparison against the alternative HB model. Although there is a growing literature on Bayesian analysis of mixtures when the number of components are unknown (Richardson and Green (1997), Stephens (2000), Hurn, Justel, and Robert (2003), Dellaportas and Papageorgiou (2006), Nobile and Fearnside (2007)), we leave this issue for further research.

Secondly, in order to apply the MHB model, we need to set the prior distributions. In many Bayesian applications, the prior is chosen to be uninformative by setting a very large variance so that the prior will not affect the posterior. However, for the MHB model, setting a very diffuse prior on the Γ_k has a major impact on the posterior distribution of behavioral parameters as well as on the group membership parameters. We, therefore, set $\nu_0 = J + 30$ and $\Gamma_0 = \nu_0 I$, where $J = 2$ represents the number of behavioral parameters for an individual customer (see Rossi, Allenby, and McCulloch (2005, Page 150)). We have carried out a simulation study where we set different prior degrees of freedom. The results confirm that setting a too diffuse prior leads to unstable estimates. Setting the prior degrees of freedom to $J + 30$ seems to be informative enough to obtain stable results without the prior influencing the posterior results too much.

To obtain posterior results, we apply our Metropolis within Gibbs sampler as presented in Section 4.2. The MCMC steps are repeated for $400,000$ iterations of which the last $40,000$ were

[14] Our computational experiments revealed that a highly skewed covariate might cause very unstable estimations.

used to infer the posterior distribution of parameters. Convergence was monitored visually and checked with the Geweke test (Geweke et al. 1991). For each of the hyper-parameters, the Geweke convergence diagnostic concludes that the two non-overlapping parts of the Markov chain[15] are from the same posterior distribution.

For our dataset from OG, we end up with two segments, with a general customer share of 41% and 59%. When we increase the number of components to three, one of the component covers only 4 customers, while the others contain the rest in a balanced share. Similarly for the four-component case, the two additional components together cover only 1% of the whole customer base. A detailed discussion of the results from MHB models with three or four segments is presented in Appendix D. One noteworthy conclusion is that the MHB model with two latent components gives better out-of-sample predictions than the ones with three or four latent components on this particular dataset. In general one may also expect to find two major segments: the frequent buyers and those who try the service only a couple of times and quit very early.

We first investigate the differences between the two identified segments. To this end we first allocate each individual to one of the segments based on the posterior segment membership probabilities. Next we take a look at descriptive statistics of the resulting two groups. Table 10 shows these statistics. The first component (41%) clearly contains customers who buy more frequently (on average 19.3 times) and more recently from the company. The difference between the end of the observation period and the last purchase time is evidently much higher for the second component (59.93 vs. 7.30 days as 'average recency' in Table 10 shows). Conversely, the customers in the second component ordered only a couple of times (on average 3.75 times) and these orders took place a long time ago. Next to the differences between segments on shopping frequency (x) and recency ($T - t_x$), we gain further insight on the additional variables. We see clear differences between segments on the average number of shopping items, average basket value, average delivery fee and the price sensitivity of customers. It seems that the frequent buyers on average have smaller shopping baskets both in value and in number of items, and pay higher amounts for the delivery service. We can, therefore, conclude that these customers are less price sensitive as they do not mind to often pay high delivery fee. The lower average discount rate on their baskets reveals the same fact as well. On the other hand, there is a major group of customers who uses the service provided by OG to buy once in a while in bigger quantities.

[15] We chose the two non-overlapping parts of the Markov chain as the first 0.1 proportion of the chain just after the burn-in iterations and the last 0.5 proportion of the chain.

These customers tend to pay less in delivery fees and they seek more discount. On this particular dataset, we clearly see two distinct segments in the customer base with different willingness to pay on home delivered groceries. All in all, besides providing predictions on purchase frequency and customer lifetime like the other BTYD models do, our proposed MHB model further provides an inherent segmentation where we can distinguish segments also on additional variables. Below, we elaborate on the difference between segments by checking the posterior results for the regression coefficients (ω) appearing in the segment membership MNP model.

Table 10: Descriptive statistics on the two segments obtained from MHB model

	Segment 1	Segment 2
# of customers	599	861
% of customers	41.03	58.97
Avg. observation time T	147.33	141.27
Avg. last purchase time t_x	140.04	81.34
Avg. recency $(T - t_x)$	7.30	59.93
Avg. # of purchases x	19.31	3.75
Avg. # of items in the basket	59.75	67.54
Avg. basket value (in €)	106.06	139.41
Avg. delivery fee (in €)	7.19	6.81
Avg. # of discounted items	5.03	6.56
Avg. discount rate of basket (%)	0.03	0.05

The MHB model allows us to make inference on the differences across the segments based on the concomitant variables. We have included three concomitant (and covariate) variables, namely the log number of items, basket value and delivery fee from the initial purchases of customers. Table 11 shows the posterior mean and 95% highest posterior density region (HPDR) for the coefficients ω in the MNP choice model. Based on the highest posterior density region from the posterior draws on ω, we conclude that components substantially differ from each other on all of the concomitant variables included. Table 11 confirms the conclusions from Table 10 such as that Segment 1 is less likely than Segment 2 (at the average value of the concomitant variables) through the intercept (-0.435), and the customers from the first component buy in smaller amounts and pay higher fees. This inference can be extended by adding any available information into the model.

Table 12 and Table 13 present the posterior means of the segment specific means and variances of the log purchase and log defection rates. These tables again support our previous findings. The posterior mean on log purchase rate is higher for the first component than that of the

Table 11: Posterior mean and 95% highest posterior density region on ω

	Mean ω		HPDR	
Intercept	−0.435*	0	−0.812	−0.129
Log # of items	1.002*	0	0.285	1.621
Basket value	−0.014*	0	−0.021	−0.007
Delivery fee	0.190*	0	0.067	0.346

* Indicates that 0 is not contained in the 95% HPDR.

Recall that we restrict ω_2 (referring the second segment) to zeros vector. Therefore, the coefficients in our MNP sub-model are evaluated relative to each other.

second component (−2.221 vs. 3.811) which says that customers in the first component buy more frequently. The result on the log defection rate is also intuitive as the customers in Segment 1 are more loyal and have longer lifetimes.

Table 12: Segment-specific posterior mean (and standard deviation) of the log purchase and log defection rates for the two-component MHB model with concomitant variables

	β	
MHB Component 1	−2.221	−10.419
	(0.055)	(0.917)
MHB Component 2	−3.811	−7.272
	(0.093)	(0.308)

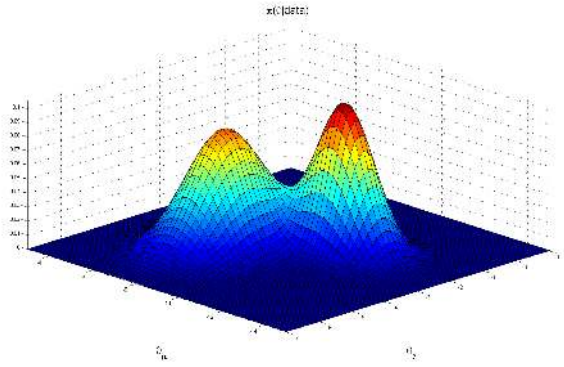
Table 13: Posterior mean variance-covariance within segments (Γ_k) for the two-component MHB model with concomitant variables

	Γ_1	Γ_2
MHB	$\begin{pmatrix} 0.299 & 0.017 \\ 0.017 & 1.275 \end{pmatrix}$	$\begin{pmatrix} 1.004 & 0.029 \\ 0.029 & 1.260 \end{pmatrix}$

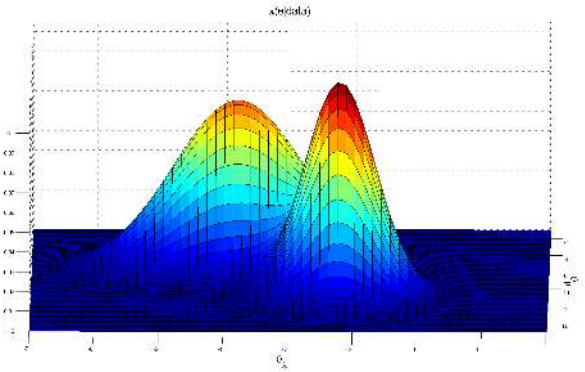
We now have a look at the shape of the heterogeneity distribution. We visualize the posterior distribution with the plots in Figure 5. These plots are created by using the segment sizes, the mean values of β_k and Γ_k and the "`gmdistribution`" function in MATLAB. The multimodality on the heterogeneity distribution is very clear from these figures.

It is also interesting to compare the heterogeneity distribution from the MHB model against the one from the HB model. We, therefore, present the posterior means of the hyper-parameters β and Γ in Table 14 for the HB model[16] and show the shape of the heterogeneity distribution

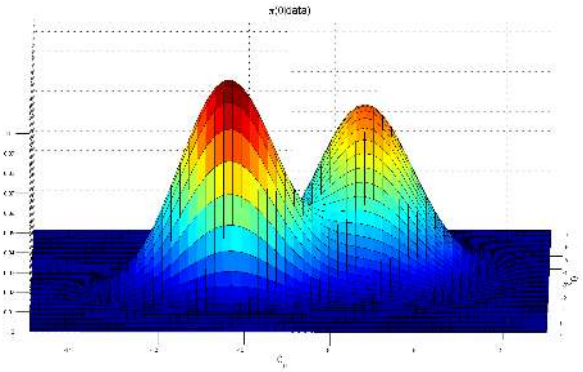
[16] All the MCMC settings are the same for the HB and MHB models.



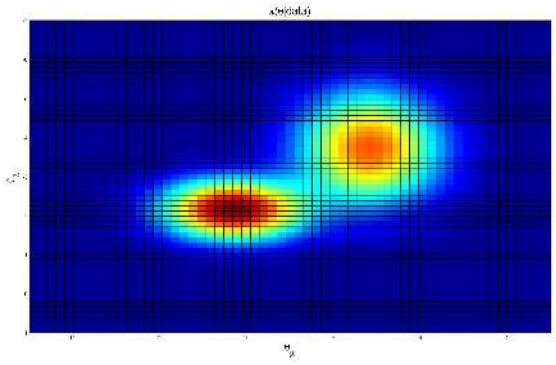
(a) Bivariate Gaussian mixture heterogeneity distribution



(b) From θ_λ perspective



(c) From θ_μ perspective



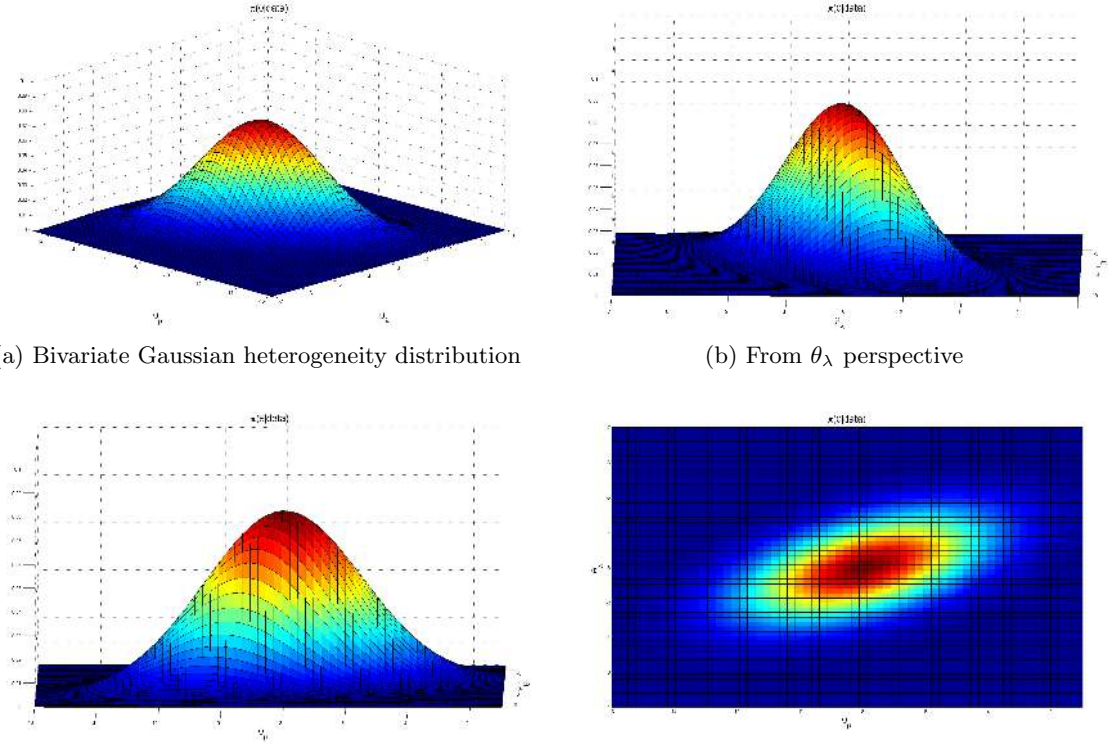
(d) Contour plot of the heterogeneity distribution

Figure 5: The shape of the posterior heterogeneity distribution (bivariate Gaussian mixture distribution)

over the whole population in Figure 6. As the HB model tries to fit a unimodal distribution, we see higher variance on the heterogeneity distribution, especially on the log defection parameter which ultimately will cause extreme lifetime predictions. The heterogeneity distribution of the HB model masks the bi-modal structure over the behavioral parameter's distribution.

Table 14: Posterior mean of the intercept vector β and the variance-covariance matrix Γ for the HB model

	HB	
β	-3.062	-8.083
	(0.036)	(0.929)
Γ	$\begin{pmatrix} 1.016 & -1.339 \\ -1.339 & 6.369 \end{pmatrix}$	



(a) Bivariate Gaussian heterogeneity distribution

(b) From θ_λ perspective

(c) From θ_μ perspective

(d) Contour plot of the heterogeneity distribution

Figure 6: The shape of the posterior heterogeneity distribution (bivariate Gaussian distribution) for OG

Next we closely look at the correlation between the log rates within each segment.[17] The HB model outperforms earlier BTYD models in the case where there is a correlation between the

[17] As emphasized by Abe (2009), it makes most sense to look at the estimated correlations without any covariates for the HB and MHB models. Therefore, Table 15 reports the posterior mean correlations between the behavioral parameters for a model without covariates.

log purchase and log defection rates (Korkmaz, Kuik, and Fok 2012). Table 15 shows that for the HB model we obtain a significant correlation (−0.596). This fact can easily be observed on Figure 6d. For the MHB model, we do not find evidence for correlation between behavioral parameters within each segment even though one can observe such correlation on the overall customer base (see Figure 5d). Apparently the correlation has now been taken up in the segment structure.

Table 15: Posterior mean and 95% highest posterior density region of correlations between log purchase and log defection rates

	$\rho_{\theta_\lambda \theta_\mu}$		
	Mean	95% HPDR	
HB	−0.596*	−0.789	−0.364
MHB Segment 1	−0.013	−0.303	0.285
MHB Segment 2	0.001	−0.172	0.176

* Indicates that 0 is not contained in the 95% HPDR.

Finally, we move on to the predictive performance. Pareto/NBD model parameters are estimated differently than the MHB and HB models. The hyper-parameters of this model are estimated by Maximum Likelihood Estimation [MLE]. In order to estimate the behavioral rates for every individual, we use a Metropolis-Hastings within Gibbs sampler as discussed in Korkmaz, Kuik, and Fok (2012). To provide a fair comparison, we did not incorporate any covariates for the HB and MHB models as the Pareto/NBD model cannot accommodate such additional information. Table 16 presents statistics on the out-of-sample predictions of the number of transaction as well as lifetime predictions for the MHB, HB and Pareto/NBD models. For the predicted number of transactions we can measure the predictive performance. We use MSE, MAE and the correlation between predicted means and observed values. For the predicted lifetime value, we cannot evaluate the performance as the actual lifetime cannot be observed. Instead, we present the mean and median prediction in days.

Table 16 shows that the hierarchical Bayes models (HB and MHB) outperform the standard Pareto/NBD model. This finding matches the results in earlier papers and the fact that we found a significant correlation between behavioral parameters (see Table 15). The HB and MHB models perform very similarly on the out-of sample number of transaction predictions. However, the HB model tends to perform slightly better in predicting the number of purchases on all three measures.

When it comes to lifetime metric, there is a clear difference among the models' predictions. The

Table 16: Out-of-sample predictions from the Pareto/NBD, HB and MHB models

MODELS	# of purchases			lifetime	
	CORR	MSE	MAE	Mean	Median
MHB	0.922	19.172	2.866	7.23E+3	2.15E+3
HB	0.924	18.581	2.774	8.17E+45	4.80E+3
Pareto/NBD	0.921	21.556	3.005	5.30E+130	4.11E+9

Pareto/NBD[18] model and the HB model both produce extremely long mean lifetime predictions. Whereas the mean lifetime prediction from the MHB model is around 20 years. We also check the median posterior results on lifetime predictions as they results in less extreme values. The median lifetime for the Pareto/NBD model is still extremely long. For the HB model it is 16 years, meanwhile the results from the MHB model is 7 years. Based on these results we can say that the lifetime predictions obtained from the MHB model can directly be used as a customer loyalty index for managerial decision making. This is in sharp contrast to the results from the other models.

7 Conclusions

The contribution of our paper is twofold. First, we propose a new BTYD model that addresses the extreme lifetime prediction problem of the current BTYD models. If the current BTYD models are applied on datasets where the heterogeneity distribution is multi-modal, one is very likely to obtain extreme lifetime predictions. The main reason for this is that the assumed heterogeneity distribution very poorly fits reality. As a result the variance in the distribution is inflated and extreme lifetime predictions are generated. In other words, if there are different segments in the population, the standard BTYD models should not be used. We have substantiated this claim through a simulation experiment as well as through a real-life application. Using a mixture of normals as the heterogeneity distribution yields better predictive results on both lifetime and number of transaction compared to two major benchmark models, namely the Pareto/NBD and the HB models.

Second, our model increases the explanatory power of BTYD models. We provide a way to distinguish between different groups of customers using explanatory variables. This not only

[18]The hyper-parameter estimations of the Pareto/NBD model on defection rate are $s = 0.04$ and $\beta = 38.24$ (shape and scale parameters of the gamma heterogeneity distribution). The estimated average defection rate for the Pareto/NBD model is given by $s/\beta = 0.001$. As the shape parameter s is less than 1, analytically the expected lifetime value of a random customer from the cohort diverges to infinity.

gives a better perspective on the customer base, but also provides information to managers on customers without prior purchase history. For instance, if a transaction from a new customer to OG contains small basket size and if this new customer pays relatively high delivery fee, it is more likely that she will continue buying from OG than another new customer who orders in a bigger quantity and pays less in delivery fee. We believe that our MHB model provides a solid methodology to empirically investigate what kind of customer characteristics relate to the lifetime or shopping frequency of customers.

As a future extension, the MHB model can be further developed to endogenize the number of segments. The current version of the model does not treat the number of latent components as a model parameter. However, there is a growing literature on finding the number of latent components within the parameter estimation process. The reversible jump Markov Chain Monte Carlo (RJMCMC) method may be useful here, see Richardson and Green 1997; Stephens 2000; Nobile and Fearnside 2007 and Dellaportas and Papageorgiou 2006. The model-specific set-up of this method, however, requires further investigation as incorporating RJMCMC in the proposed complex model is not straightforward. Alternatively one may build on the Dirichlet Process Prior as in Rasmussen (1999), Ishwaran and James (2002) and McAuliffe, Blei, and Jordan (2006).

We also advocate further testing of this model on other datasets. The lifetime estimates resulting from BTYD models have not been used a lot in the past. The main reason for this is the poor performance of those estimates. We believe that this situation has changed with our proposed model. We, therefore, hope to see more applications of these models to predict customer lifetime.

Appendix A Sampling steps of the MCMC for the MHB model without concomitant variables

1. **Nodes z and t_Δ.**

   In this subsection the focus is on data and parameters of a single customer. We suppress in our notation the conditioning on T_i which is assumed throughout the subsection. We wish to compute

$$\begin{aligned}\pi(t_{\Delta,i}, z_i|x_i, t_{x,i}, \lambda_i, \mu_i, \varpi) &= \pi(t_{\Delta,i}|z_i, x_i, t_{x,i}, \lambda_i, \mu_i)\pi(z_i|x_i, t_{x,i}, \lambda_i, \mu_i) \\ &= \pi(t_{\Delta,i}|z_i, t_{x,i}, \lambda_i, \mu_i)\pi(z_i|t_{x,i}, \lambda_i, \mu_i)\end{aligned}$$

   where ϖ signals parameters other than written explicitly. The right hand side shows that the conditional probability does not depend on the ϖ parameters. $t_{\Delta,i}$ is the defection time. Considering the functional dependence of the distribution of the time of defection, $t_{\Delta,i}$, of a customer conditioned on the data $(x_i, t_{x,i})$ and parameters (λ_i, μ_i) of that customer, we have

$$\pi(t_{\Delta,i}|x_i, t_{x,i}, \lambda_i, \mu_i) \propto \pi(t_{\Delta,i}, x_i, t_{x,i}|\lambda_i, \mu_i) = \pi(x_i, t_{x,i}|t_{\Delta,i}, \lambda_i, \mu_i)\, \pi(t_{\Delta,i}|\lambda_i, \mu_i)$$

   and

$$\pi(x_i, t_{x,i}|t_{\Delta,i}, \lambda_i, \mu_i) = \pi(x_i|t_{x,i}, t_{\Delta,i}, \lambda_i, \mu_i)\, \pi(t_{x,i}|t_{\Delta,i}, \lambda_i, \mu_i) \propto \pi(t_{x,i}|t_{\Delta,i}, \lambda_i, \mu_i)\ .$$

   So $\pi(t_{\Delta,i}|x_i, t_{x,i}, \lambda_i, \mu_i) \propto \pi(t_{x,i}|t_{\Delta,i}, \lambda_i, \mu_i)\, \pi(t_{\Delta,i}|\lambda_i, \mu_i) \propto I_{[t_{x,i},\infty)}(t_{\Delta,i})\, e^{-\lambda_i(t_{\Delta,i}\wedge T_i)}\, e^{-\mu_i t_{\Delta,i}}$ and

$$\pi(t_{\Delta,i}|x_i, t_{x,i}, \lambda_i, \mu_i) = \frac{I_{[t_{x,i},\infty)}(t_{\Delta,i})\, e^{-\lambda_i(t_{\Delta,i}\wedge T_i)}\, e^{-\mu_i t_{\Delta,i}}}{C(x_i, t_{x,i}, \lambda_i, \mu_i)} \tag{8}$$

   with the constant $C(x_i, t_{x,i}, \lambda_i, \mu_i)$ determined as

$$C(x_i, t_{x,i}, \lambda_i, \mu_i) = \int_{t_{x,i}}^{\infty} e^{-\lambda_i(t_{\Delta,i}\wedge T_i)}\, e^{-\mu_i t_{\Delta,i}}\, \mathrm{d}t_{\Delta,i} = \frac{e^{-(\lambda_i+\mu_i)t_{x,i}} - e^{-(\lambda_i+\mu_i)T_i}}{\lambda_i+\mu_i} + \frac{e^{-(\lambda_i+\mu_i)T_i}}{\mu_i}\ .$$

   Once we have the conditional distribution of $t_{\Delta,i}$ we can easily find the (discrete) distribution of the binary variable z_i indicating whether the customer is active at T_i (corresponding to $z_i = 1$) or not (corresponding to $z_i = 0$). The value of z_i is determined as $z_i = I_{[T_i,\infty)}(t_{\Delta,i})$ and then

$$\begin{aligned}\mathrm{Prob}(z_i = 1|x_i, t_{x,i}, \lambda_i, \mu_i) &= \frac{\int_{T_i}^{\infty} e^{-\lambda_i T_i}\, e^{-\mu_i t_{\Delta,i}}\, \mathrm{d}t_{\Delta,i}}{C(x_i, t_{x_i,i}, \lambda_i, \mu_i)} = \frac{\frac{e^{-(\lambda_i+\mu_i)T}}{\mu_i}}{\frac{e^{-(\lambda_i+\mu_i)t_{x,i}} - e^{-(\lambda_i+\mu_i)T_i}}{\lambda_i+\mu_i} + \frac{e^{-(\lambda_i+\mu_i)T_i}}{\mu_i}} \\ &= \frac{1}{\frac{\mu_i}{\lambda_i+\mu_i}\left(e^{(\lambda_i+\mu_i)(T_i - t_{x_i,i})} - 1\right) + 1}\ .\end{aligned} \tag{9}$$

See Abe (2009) and Schmittlein, Morrison, and Colombo (1987) for Equation (9). The distribution $\pi(t_{\Delta,i}|z_i, t_{x,i}, \lambda_i, \mu_i)$ is now the distribution given in Equation (8) truncated to the interval $(t_{x,i}, T_i)$ if $z_i = 0$, and to the interval (T_i, ∞) if $z_i = 1$.

2. **Node** s.

   Draw indicator variables for latent class membership, for each customer i;

   $s_i \sim \pi(s_i|\theta_i, \Delta, \beta_{s_i}, \Gamma_{s_i}, p_k) \propto \pi(\theta_i - D_i\Delta|\beta_k, \Gamma_k)\, p_k$. This can be done in two steps:

   a) Calculate the conditional membership probabilities for each customer and each component as

   $$\widetilde{p}_{ik} = \frac{p_k\, \varphi(\theta_i - D_i\Delta|\beta_k, \Gamma_k)}{\sum_{\ell=1}^{K} p_\ell\, \varphi(\theta_i - D_i\Delta|\beta_\ell, \Gamma_\ell)}, \tag{10}$$

   where $\varphi(\cdot)$ is the multivariate normal density.

   b) Draw the indicator variables of customer i from the multinomial distribution with the parameters of membership probabilities to each groups: $s_i \sim \text{Multinomial}_K(\widetilde{p}_i)$ where $\widetilde{p}_i = [\widetilde{p}_{i1}, \cdots, \widetilde{p}_{iK}]$.

3. **Nodes** $\boldsymbol{\beta}$ **and** $\boldsymbol{\Gamma}$.

   Draw hyper-parameters for each latent class k; $(\beta_k, \Gamma_k) \sim \pi(\beta_k, \Gamma_k|\theta, \Delta, s)$. Note that the value of the quantity $\pi(\beta_k, \Gamma_k|\theta, \Delta, s)$ does not depend on rates θ for those customers that do not belong to the class indicated by s. Let $\theta^{(k)}$ be the rates for those customers for which the class indicator variable has value k: $\theta^{(k)} = \{\theta_i\}_{i:s_i=k}^{N}$. Then, according to Equation (4) on Page 12,

   $$\pi(\beta_k, \Gamma_k|\theta, \Delta, s) = \pi(\theta^{(k)} - D^{(k)}\Delta|\beta_k, \Gamma_k)\, \pi(\beta_k|\Gamma_k)\, \pi(\Gamma_k)$$

   This comes down to the linear regression update:

   a) **Node** $\boldsymbol{\beta}$.

      The conditionally conjugate prior for the intercept (or mean) of each class is given as

      $$\beta_k|\Gamma_k \sim \text{N}(\bar{\beta}, \bar{\Gamma} \otimes \bar{A})$$

      where $\bar{\beta}$ stands for the location parameter, and $\bar{A}$ stands for the shape parameter determining the tightness of the prior.

      The posterior density for $\text{vec}(\beta_k)$ is sampled from a normal distribution with a mean $\text{vec}(\widetilde{\beta_k})$ where $\widetilde{\beta_k} = (\iota'\iota + \bar{A})^{-1}(\iota'(\theta^{(k)} - D^{(k)}\Delta) + \bar{\beta}\bar{A}$ and a variance of $\Gamma_k \otimes (\iota'\iota + \bar{A})^{-1}$. ι is a vector of ones with the size of the number of customers in the k^{th} component.

   b) **Node** $\boldsymbol{\Gamma}$.

The conjugate prior on the covariance structure of each latent class is

$$\Gamma_k \sim \mathrm{IW}(\bar{\Gamma}, \bar{\nu}),$$

where $\bar{\Gamma}$ gives the location parameter, $\bar{\nu}$ gives the degrees of freedom.

The posterior density for Γ_k is sampled from the inverse Wishart distribution with the scale matrix of $\bar{\Gamma}_k + ((\theta^{(k)} - D^{(k)}\Delta) - \iota\widetilde{\beta})'((\theta^{(k)} - D^{(k)}\Delta) - \iota\widetilde{\beta}) + (\widetilde{\beta} - \bar{\beta})'\bar{A}(\widetilde{\beta} - \bar{\beta})$ and the degrees of freedom $\bar{\nu} + N^k$.

4. **Node $\boldsymbol{\Delta}$**.

The regression coefficient matrix (without an intercept) over the customer base has the following conjugate prior

$$\mathrm{vec}(\Delta) = \delta \sim \mathrm{N}(\bar{\delta}, \bar{A}_\delta)\ .$$

The posterior density for $\mathrm{vec}(\Delta)$ is again a normal distribution with mean $(X'X + A_\delta)^{-1}(X'y + A_\delta\bar{\delta})$ and variance $((X'X) + A_\delta)^{-1}$ where

$$X'X = \sum_k \Gamma_k^{-1} \otimes D'^{(k)} D^{(k)}$$
$$X'y = \mathrm{vec}\left(\sum_k D'^{(k)}\theta^{(k)}\Gamma'^{-1}_k\right)$$

Details of Δ sampling:

As this model does not distinguish the slope among different components, the regression coefficients are drawn over the whole population; $\Delta \sim \pi(\Delta|\theta, \beta, \Gamma, s)$. In these expressions we consider data for all customers.

At this stage we use the mean β and variance-covariance matrix Γ of each component, parameter values θ for each customer. Besides, we have the information on covariates D and the prior distribution on regression coefficients $\delta = \mathrm{vec}(\Delta)$ which is given as $N(\bar{\delta}, A_\delta^{-1})$.

We create a linear regression model that covers customer data in all segments. In order to be able to pool data from K components, we collect the multivariate regression models across the components. To do so, we standardize all equations.

- Customer data should be updated (θ^*) by shifting the mean of the normal mixture on the basis of observations coming from the covariate information D.

$$\theta^* = \theta - D\Delta \tag{11}$$

- For component k, we have the multivariate regression model given as:

$$\theta^{*(k)} = \iota\beta_k + \varepsilon^{(k)}, \text{ where } \text{vec}(\varepsilon^{(k)}) \sim N(0, \Gamma_k \otimes I) \tag{12}$$

Now we write all MVR models coming from each component in a way that error is standardized: First we write the regression models in a way that Δ are the regression coefficients,

$$\begin{aligned}
\theta^{*(k)} &= \iota\beta_k + \varepsilon^{(k)} \\
\theta^{(k)} - D^{(k)}\Delta &= \iota\beta_k + \varepsilon^{(k)} \\
\theta^{(k)} - \iota\beta_k &= D^{(k)}\Delta + \varepsilon^{(k)} \\
\text{vec}(\theta^{(k)} - \iota\beta_k) &= \text{vec}(D^{(k)}\Delta) + \text{vec}(\varepsilon^{(k)}),
\end{aligned}$$

given that $\text{vec}(\varepsilon^{(k)}) \sim N(0, \Gamma_k \otimes I)$ and using the property of $\text{vec}(ABC) = (C' \otimes A))\text{vec}(B)$, we obtain

$$\text{vec}(\theta^{(k)} - \iota\beta_k) = (I \otimes D^{(k)})\text{vec}(\Delta) + \text{vec}(\varepsilon^{(k)}) \tag{13}$$

It is time to standardize the error for the MVR model of each component. In order to standardize it, we use one of the characteristics of the covariance matrix. The variance-covariance matrix Γ_k is a positive definite matrix which implies that there exist a (non-unique) matrix M^k such that $\Gamma_k = M^{k'}M^k$.

We use one of the properties of the variance-covariance matrix to standardize the error term. Let A be a random vector with $\text{Cov}(A) = \Gamma$. If an N matrix is multiplied with A, the covariance of the new structure is expressed as $\text{Cov}(NA) = N\Gamma N'$. Using this property, in order to scale the variance-covariance matrix of each component to unit covariance, we need to multiply Γ_k with the inverse of its decomposition, so that $M'^{-1}M'MM^{-1} = I$. In the following equations the k index, showing the mixture component, is ignored for simplicity:

$$\begin{aligned}
(M'^{-1} \otimes I)\text{vec}(\theta - \iota\beta) &= (M'^{-1} \otimes I)(I \otimes D)\text{vec}(\Delta) + U \\
(M'^{-1} \otimes I)\text{vec}(\theta - \iota\beta) &= (M'^{-1} \otimes D)\text{vec}(\Delta) + U
\end{aligned} \tag{14}$$

U represents the unit covariance structure.

- In Equation (14), if we write the expressions as $y = (M'^{-1} \otimes I)\text{vec}(\theta - \iota\beta)$, $X = (M'^{-1} \otimes D)$, $\delta = \text{vec}(\Delta)$, and then we have the regression model $y = X\delta + U$. After stacking all the

regression models from the mixture components, we deal with a standard normal regression update, where errors are independent and of unit size. Δ can be sampled from a normal distribution with mean $(X'X + A_\delta)^{-1}(X'y + A_\delta\bar{\delta})$ and variance $((X'X) + A_\delta)^{-1}$. Note that the matrix M is not used in this sampling process.

The moments mentioned can be calculated efficiently as follows:

$$X'X = \sum_k {\Gamma_k}^{-1} \otimes {D'}^{(k)} D^{(k)}$$
$$X'y = \text{vec}\left(\sum_k {D'}^{(k)} \theta^{(k)} {\Gamma_k}'^{-1}\right)$$

5. **Node $\boldsymbol{p}$.**

   Draw $p \sim \pi(p|s)$. Dirichlet update: $\upsilon \sim \text{Dir}(\bar{\boldsymbol{\alpha}} + \#)$. Here $\#_k = |\{i|s_i = k\}|$. We set $\bar{\boldsymbol{\alpha}}$ as 1.

6. **Node θ.**

   Draw, for each customer i, a new value for $\theta_i \sim \pi(\theta_i|x_i, t_{x,i}, y_i, z_i, \Delta, \beta_k, \Gamma_k, s_i)$. Note that

$$\pi(\theta_i|x_i, t_{x,i}, y_i, z_i, \Delta, \beta_k, \Gamma_k, s_i) \propto \pi(x_i, t_{x,i}, y_i, z_i, \theta_i\Delta, \beta_k, \Gamma_k, s_i)$$

   and that

$$\pi(x_i, t_{x,i}, y_i, z_i, \theta_i\Delta, \beta_k, \Gamma_k, s_i) \propto \pi(x_i, t_{x,i}, y_i, z_i|\theta_i)\pi(\theta_i|\beta_k + D_i\Delta, \Gamma_k)$$

   Sampling of θ_i requires the Metropolis Hastings algorithm.

Appendix B Sampling steps for the MNP model with concomitant variables

1. **Node ω.**

   The conjugate prior on the latent component-specific regression coefficients is $\omega_k \sim \text{N}(\bar{\omega}, \bar{g})$. ω_k is dimension of $((L+1) \times 1)$ where L is the number of concomitant variables. It describes the effect of concomitant variables on each of the latent classes. The draws from the posterior distribution can be obtained by a standard regression update process.

$$u_{ik} = C_i\,\omega_k + \varepsilon_{ik}$$

where $\varepsilon_{ik} \sim \mathrm{N}(0, \mathbb{I}_{\mathbb{K}})$, $\mathbb{I}_{\mathbb{K}}$ is the identity matrix of dimension K. The normal regression update on the component specific regression coefficients:

$$(\omega|u) \sim \mathrm{N}((C'C + A)^{-1}(C'u + A\bar{\omega}), (C'C + A)^{-1}) \; .$$

Note that for identification, we restrict $\omega_K = 0$.

2. **Node $\boldsymbol{u}$**.

In order to assign each customer to a latent component, we use latent utility variable u. The selector function $\varsigma(u)$ determines which component each customer is assigned to, that is,

$$\varsigma(u_i) = k, \text{ if } u_{ik} > u_{ij} \text{ for all } j \neq k,$$

where $u_{ik} = C_i\,\omega_k + \varepsilon_{ik}$ is the utility of customer i being assigned to the latent component k. C_i is the row vector of component-invariant behavioral characteristics (concomitant variables) of customer i (together with an intercept), ω_k is the component specific regression coefficients, and ε_{ik} is the stochastic error term.

The probability of customer i being a member of component k is equal to

$$\begin{aligned}
\mathrm{Prob}(s_{ik} = 1) &= \mathrm{Prob}(u_{ik} \geq u_{ij}, \text{ for all } j \text{ in } (K-1) \text{ components}) \\
&= \mathrm{Prob}(u_{ij} - u_{ik} \leq 0, \quad \text{all } j \neq k) \\
&= \mathrm{Prob}(\varepsilon_{ij} - \varepsilon_{ik} \leq C_i\,(\omega_k - \omega_j), \quad \text{all } j \neq k) \\
&= \mathrm{Prob}(\widetilde{\varepsilon}_{ikj} \leq C_i\,\widetilde{\omega}_{kj}, \quad \text{all } j \neq k)
\end{aligned}$$

where $\widetilde{\varepsilon}_{ikj} = \varepsilon_{ij} - \varepsilon_{ik}$ and $\widetilde{\omega}_{kj} = (\omega_k - \omega_j)$.

To allocate customers to latent components, we need to sample from the $u_{ik} = C_i\,\omega_k + \varepsilon_{ik}$ where $\varepsilon_i = [\varepsilon_{i1}, ..., \varepsilon_{iK}] \sim \mathrm{N}(0, \mathbb{I}_{\mathbb{K}})$. As mentioned in McCulloch and Rossi 1994, direct draws from truncated multivariate normal random vectors are difficult to accomplish efficiently due to the very high rejection frequencies. The insight from McCulloch and Rossi 1994 is to recognize that one can define a Gibbs sampler by breaking each draw of u into a sequence of K univariate truncated normal draws by cycling through the u vector (one-at-a-time sampling or one dimensional sampling).

Considering customer i, we need to draw from

$$u_i \sim \pi(u_i|\cdots, \theta, \beta, \Gamma, \omega, \cdots) \propto \pi(\theta_i|\Delta, \beta_{\varsigma(u_i)}, \Gamma_{\varsigma(u_i)})\,\pi(u_i|C_i\,\omega) \; . \tag{15}$$

We need to take into account the discrete jumps that may happen through $\varsigma(u)$ as this results in new parameter values on β, Γ and ω. We separately investigate each component of Equation (15).

- $\pi(\theta|\Delta, \beta_{\varsigma(u)}, \Gamma_{\varsigma(u)})$: The dependency here is interceded through $\varsigma(u)$. Recall that

$$\varsigma(u_i) = k, \text{ if } u_{ik} > u_{ij} \text{ for all } j \neq k(\text{ or if } s_{ik} = 1) .$$

  We focus on $\pi(s|u)$,

$$\pi(s_k|u_k, u_{-k}) = I(s_k = 1)I(u_k > \max(u_{-k})) + I(s_k \neq 1)I(u_k < \max(u_{-k}))$$

  writing $\max(u_{-k}) = u_o$,

$$\pi(\theta|\Delta, \beta_{\varsigma(u)}, \Gamma_{\varsigma(u)}) = I(u_k > u_o)\pi(\theta|\Delta, \beta_k, \Gamma_k) + I(u_k < u_o)\pi(\theta|\Delta, \beta_o, \Gamma_o)$$

  For the sampling of u, as $\pi(\theta|\Delta, \beta_{\varsigma(u)}, \Gamma_{\varsigma(u)})$ assumes different values on cones of $\mathbb{R}^{\mathbb{K}}$, we need to deal with a normal density that is scaled differently in these cones.

- $\pi(u|C\,\omega)$: Utilities have a multivariate Normal distribution, that is,

$$\pi(u_i|C_i\,\omega) \propto e^{-1/2(u_i-\bar{u})'\Sigma^{-1}(u_i-\bar{u})},$$

  where $\bar{u} = c_i\omega$.

So the conditional density of utilities can be written as

$$\begin{aligned}\pi(u_k|\theta, \boldsymbol{\beta}, \boldsymbol{\Gamma}, u_{-k}, \omega) \propto (&I(u_k > u_o)\,|\Gamma_k|^{-1/2} e^{-1/2(\theta-(\beta_k+D\Delta))(\Gamma_k)^{-1}(\theta-(\beta_k+D\Delta))'} \\ &+ I(u_k < u_o)\,|\Gamma_o|^{-1/2} e^{-1/2(\theta-(\beta_o+D\Delta))(\Gamma_o)^{-1}(\theta-(\beta_o+D\Delta))'})\; e^{-1/2(u_k-\bar{u})^2}, \quad (16)\end{aligned}$$

where we omit the index i for clarity.

This is a combination of two truncated normal distributions, see Figure 7. We write Ω_r as the scaling factor of the truncated normal distribution on the right,

$$\Omega_r = |\Gamma_k|^{-1/2} e^{-1/2(\theta-(\beta_k+D\Delta))(\Gamma_k)^{-1}(\theta-(\beta_k+D\Delta))'}$$

where $u_k < u_o(\max(u_{-k}) = u_o)$; and Ω_l as the scaling factor of the other truncated normal distribution

$$\Omega_l = |\Gamma_o|^{-1/2} e^{-1/2(\theta-(\beta_o+D\Delta))(\Gamma_o)^{-1}(\theta-(\beta_o+D\Delta))'}$$

where $u_k > u_o$.

Then,

$$\pi(u_k|\theta, \boldsymbol{\beta}, \boldsymbol{\Gamma}, u_{-k}, \omega) \propto (I(u_k > u_o)\Omega_r + I(u_k < u_o)\Omega_l)\, e^{-1/2(u_k-\bar{u})^2} \,. \tag{17}$$

The normalization constant is easily computed. Let ϕ be the density function of the normal distribution with mean $\bar{u}$ and variance 1. Then, the final version for the sampling distribution is

$$\pi(u_k|\theta, \boldsymbol{\beta}, \boldsymbol{\Gamma}, u_{-k}, \omega) = \frac{\Omega_r I(u_k > u_o) + \Omega_l I(u_k < u_o)}{(1-\Phi(u_o-\bar{u}_k))\Omega_r + \Phi(u_o-\bar{u}_k)\Omega_l}\phi(u_o-\bar{u}_k) \,. \tag{18}$$

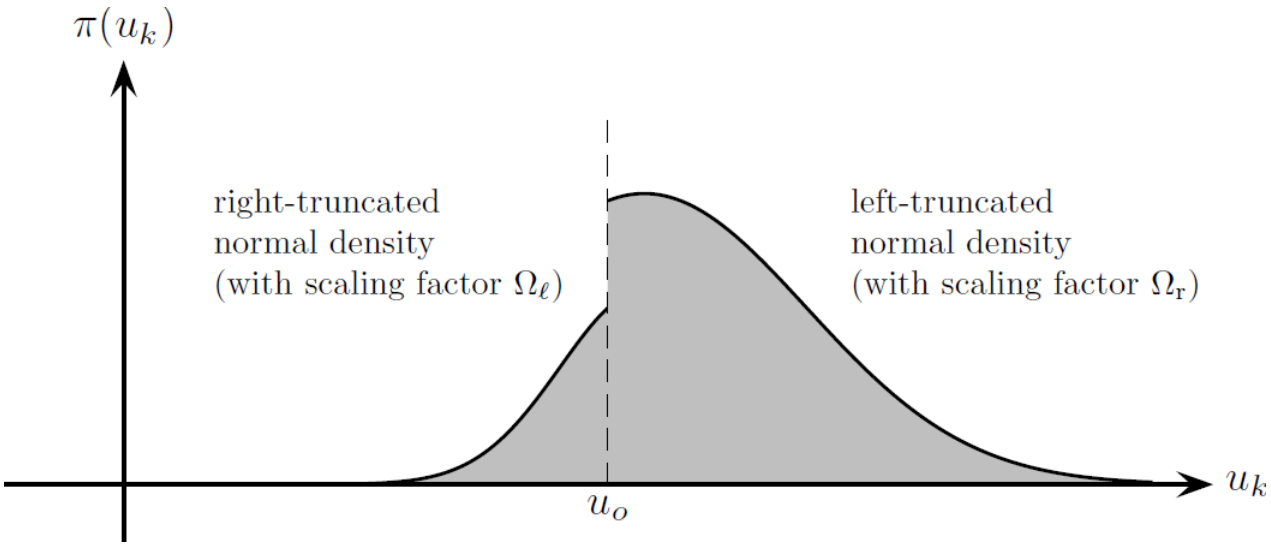


Figure 7: The sampling density for the utilities.

The sampling is now done by applying the following to all utility components:

- Sample $U \sim$ Uniform$[0,1]$ to determine which truncated normal distribution to sample from.
- If $U < \frac{\Omega_l \Phi(u_o-\bar{u}_k)}{\Omega_l \Phi(u_o-\bar{u}_k)+\Omega_r(1-\Phi(u_o-\bar{u}_k))}$, then truncate to the right and sample from the left side of the truncated normal distribution. In particular, set

  $$u_k^{\text{new}} = \Phi^{-1}(\Phi(u_o-\bar{u}_k)U') + \bar{u}_k$$

  where $U' \sim$ Uniform$[0,1]$.
- If $U > \frac{\Omega_r(1-\Phi(u_o-\bar{u}_k))}{\Omega_l \Phi(u_o-\bar{u}_k)+\Omega_r(1-\Phi(u_o-\bar{u}_k))}$, then truncate to the left and sample from the right side of the truncated normal distribution. In particular, set

$$u_k^{\text{new}} = \Phi^{-1}\left((1-\Phi(u_o-\bar{u}_k))U' + \Phi(u_o-\bar{u}_k)\right) + \bar{u}_k \ .$$

Appendix C Data generation for MHB model testing

Consider $N = 1,000$ customers and $K = 2$ latent components. We generate a single covariate data, D $(N \times 1)$, for all customers from a standard uniform distribution. We create another customer characteristics matrix including an intercept and a concomitant variable, C $(N \times L)$ where $L = 2$. In order to keep it simple, for the first half of the population the concomitant variable is set to 1 and for the other half it is set to -1. The transaction data of customers are generated in five steps:

1. Fix the component specific regression coefficient matrix, ω^* $(L \times K)$ to $\left[\begin{smallmatrix} 0.1 & 0 \\ 0.8 & 0 \end{smallmatrix}\right]$. Using the concomitant matrix together with the ω^* matrix, we generate utilities, u^*, using the normally distributed error term.[19] More specifically, we use the following utility generation form: $u^* = C\,\omega^* + \varepsilon$, where $\varepsilon \sim \mathrm{N}(0, \mathbb{I})$. Note that the used parameter values are chosen to balance the random and deterministic components of utilities. Given the *true* utility values u^*, customers are assigned to each component,

$$s_i^* = k, \text{ if } u_{ik}^* > u_{ij}^* \text{ for all } j \neq k.$$

   Based on this procedure, we add randomness on assigning customers to their *true* components. In our sample 52.8% of the customers is assigned to segment 1.

2. Fix the hyper-parameter values β^* and Γ^* for each of the components: We aim to generate a customer dataset that has $K = 2$ distinct groups or in other words that has a bi-modal heterogeneity distribution over the customer base. As the covariate data, D, is standardized, the β vector represents the average values of parameters of interest (log of purchase and defection parameters) for each component. Our main concern is on distinguishing between the components. We, therefore, use a rather different set of parameters for each component. We set $\beta_1^* = [\log(1), \log(1/1000)]$ and $\beta_2^* = [\log(1/100), \log(1/20)]$. The (2×2) covariance matrices Γ_k^* are chosen to be equal to $0.64 \times \mathbb{I}$ for each of the components.

3. Generate behavioral parameters $\theta_i^* \sim \pi(\theta_i | \beta_{s_i}^*, \Gamma_{s_i}^*)$ for each of the customers: Conditional on the membership to one of the two components, customer's behavioral parameters are generated from normal distributions independently given the associated hyper-parameters.

4. Generate lifetime $t_{\Delta,i}^*$ for each of the customers: For $i = 1, \cdots, N$, draw $t_{\Delta,i}^* \sim \pi(t_{\Delta,i} | \theta_i^*)$. As the lifetime is distributed according to an exponential distribution with the rate parameter of e^{θ_μ}, this step is evident.

[19] The proposed model employs a MNP sub-model to assign customers to latent components.

5. Generate repeat transaction frequency x_i and the last transaction time in calibration period $t_{x,i}$ for each of the customers: For $i = 1, \cdots, N$, draw $x_i \sim \pi(x_i|t^*_{\Delta,i}, \theta^*_i)$. Transaction data basically contains two elements: transaction number x_i and the time of the last transaction $t_{x,i}$. Note that the time of the first order t_0 and the total observation time T are fixed ($t_0 = 0, T = 200$) and they are common across the customers. The sampling scheme of transaction data $(x_i, t_{x,i})$, given the defection time $t_{\Delta,i}$ and the parameters θ_i is the following:[20]

   Let $(V_l)_{l=1,2,\ldots}$ be iid exponentially distributed with mean $1/\lambda$. Put $E_x = \sum_{l=1}^{x} V_l$. Then E_x has an Erlang-x distribution: the sum of x independent exponential distributions with average $1/\lambda$. Write $\hat{t}_\Delta = \min(t_\Delta, T)$ where $\hat{t}_\Delta$ is the effective time of defection. Now, for $x \geq 1$, we can compute

$$\begin{aligned}\pi(x, t_x|t_\Delta, \theta) &= \pi(E_x = t_x, V_{x+1} + E_x > \hat{t}_\Delta) = \pi(E_x = t_x)\,\pi(V_{x+1} > \hat{t}_\Delta - t_x|E_x = t_x) \\ &= \pi(E_x = t_x)\,\pi(V_{x+1} > \hat{t}_\Delta - t_x) = \frac{\lambda^x t_x^{x-1}}{(x-1)!} e^{-\lambda t_x}\, e^{-\lambda(\hat{t}_\Delta - t_x)} = \frac{\lambda^x t_x^{x-1}}{(x-1)!}\, e^{-\lambda \hat{t}_\Delta}\end{aligned}$$

   Performing the integral t_x over the interval $(0, \hat{t}_\Delta)$ leads to[21]

$$\text{Prob}(x, t_x \leq t|t_\Delta, \theta) = \frac{\lambda^x t^x}{x!} e^{-\lambda \hat{t}_\Delta}$$

   for $t < \hat{t}_\Delta$ and $x \neq 0$. Clearly, $\text{Prob}(x = 0, t_x \leq t|t_\Delta, \theta) = e^{\lambda \hat{t}_\Delta}$, and for $t < T$

$$\begin{aligned}F(t) &\equiv \text{Prob}(t_x \leq t|t_\Delta, \theta) \\ &= \begin{cases} 0 & \text{if } t < 0 \\ \sum_{x=0}^{\infty} \frac{\lambda^x t^x}{x!} e^{-\lambda \hat{t}_\Delta} & \text{if } 0 \leq t < \hat{t}_\Delta \\ 1 & \text{if } t \geq \hat{t}_\Delta \end{cases} \\ &= \begin{cases} 0 & \text{if } t < 0 \\ e^{-\lambda(\hat{t}_\Delta - t)} & \text{if } 0 \leq t < \hat{t}_\Delta \\ 1 & \text{if } t \geq \hat{t}_\Delta \end{cases}\end{aligned}$$

   and for $s \in [0, 1]$,

$$F^{-1}(s) = \begin{cases} 0 & \text{if } s \leq e^{-\lambda \hat{t}_\Delta} \\ \hat{t}_\Delta + \ln(s)/\lambda & \text{if } s > e^{-\lambda \hat{t}_\Delta} \end{cases}$$

   All this leads to the following sampling scheme for recency-frequency (RF) data.

[20] We drop the i index in the following derivations for the sake of simplicity on notation.

[21] And in turn to

$$\text{Prob}(x = 0|\theta) = \int_0^\infty e^{-\lambda \hat{t}_\Delta} \mu e^{-\mu t_\Delta} \mathrm{d}t_\Delta = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda+\mu)T}\,.$$

a) Draw $t_\Delta \sim \text{EXP}(\mu)$.

b) Draw $U \sim \text{U}[0,1]$. Put

$$t_x = \begin{cases} 0 & \text{if } U \leq e^{-\lambda \hat{t}_\Delta} \\ \hat{t}_\Delta + \ln(U)/\lambda & \text{if } U \geq e^{-\lambda \hat{t}_\Delta} \end{cases}$$

c) Put

$$x = \begin{cases} 0 & \text{if } t_x = 0 \\ 1 + \text{POISSON}(\lambda t_x) & \text{if } t_x > 0 \end{cases}$$

Appendix D Setting the number of components for the MHB model with concomitant variables

Table 17 shows the out-of-sample prediction accuracy of the MHB model for different numbers of components. The MHB model with 2 components performs best in predicting the number of purchases in the validation period. As discussed earlier, our main criterion of determining the optimum number of components is the number of members within each group (Frühwirth-Schnatter 2006). Based on this criterion, we decide that the optimum number of components is 2 with a general customer share of 59% and 41% for the two segments. When the number of components increases to 3, one of the component covers only 4 customers (0.2%) of the customer base while the others contain the rest of it in a balanced share. For the 4 component case, two additional components cover only around 1% of the customers.

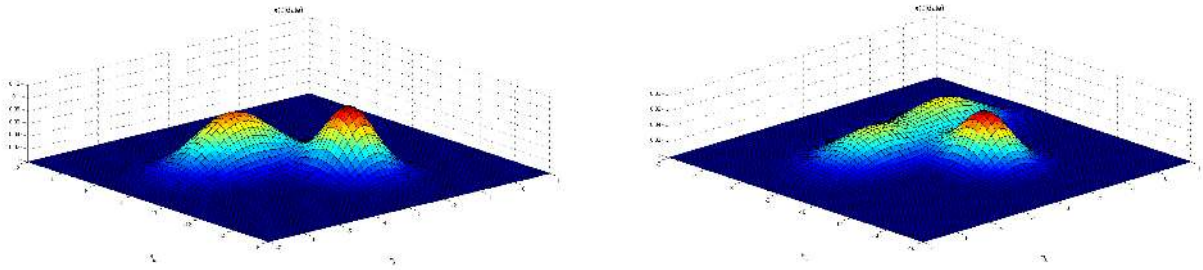
We did not use the Bayesian counterparts of likelihood based model comparison methods, i.e. the marginal likelihood comparison, because of the lack of the closed-form solution to the marginal likelihood. Schwarz criterion is not used either, because it is not evident that the regularity conditions for deriving Schwarz's criterion through asymptotic expansions actually hold (Frühwirth-Schnatter 2006).

Table 17: The out-of-sample prediction performance of the MHB model with concomitant variables on different number of components on OG data

MHB Model	# of purchase			# of customers (%) in each component			
	Correlation	MSE	MAE	Comp1	Comp2	Comp3	Comp4
2-Component	0.9208	19.556	2.851	599 (41%)	861 (59%)	-	-
3-Component	0.9207	19.654	2.860	601 (41%)	855 (59%)	4 (0.2%)	-
4-Component	0.9200	19.738	2.857	602 (41%)	839 (58%)	15 (1%)	4 (0.2%)

Figure 8 shows the heterogeneity distribution for the OG data using the MHB model with 3 or 4 components. The plot given in Figure 8a is not different that the MHB model with 2 components where there are only two peaks, i.e. the additional component does not capture a different (heterogeneous) characteristic. However, when 4 components are forced on the MHB model, we observe three peaks on OG data (see Figure 8b). Despite this additional peak in the 4 component model, which may capture

different characteristics of the heterogeneity distribution, this model clearly deteriorates out-of-sample estimation results. Note that, this model performs the worst in out-of-sample predictions. We therefore opt for the 2 component model in this paper.



(a) Bivariate Gaussian mixture heterogeneity distribution from MHB model with 3 components

(b) Bivariate Gaussian mixture heterogeneity distribution from MHB model with 4 components. Note the vertical scale.

Figure 8: The shape of the posterior heterogeneity distribution (bivariate Gaussian mixture distribution) over the online retailer's customer base when the MHB model is run with 2 and 3 components.

References

Abe, M. (2009). "Counting Your Customers One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model." In: *Marketing Science* 28.3, pp. 541–553.

Cooil, B., L. Aksoy, and T.L. Keiningham (2008). "Approaches to customer segmentation." In: *Journal of Relationship Marketing* 6.3-4, pp. 9–39.

Dellaportas, Petros and Ioulia Papageorgiou (2006). "Multivariate mixtures of normals with unknown number of components." In: *Statistics and Computing* 16.1, pp. 57–68.

Fader, P.S., B.G.S. Hardie, and K.L. Lee (2005). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model." In: *Marketing Science*, pp. 275–284.

Frühwirth-Schnatter, Sylvia (2006). *Finite mixture and Markov switching models*.

Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 6, pp. 721–741.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

Hastings, W.K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." In: *Biometrika*, pp. 97–109.

Hurn, Merrilee, Ana Justel, and Christian P Robert (2003). "Estimating mixtures of regressions." In: *Journal of Computational and Graphical Statistics* 12.1, pp. 55–79.

Ishwaran, Hemant and Lancelot F James (2002). "Approximate Dirichlet Process computing in finite normal mixtures." In: *Journal of Computational and Graphical Statistics* 11.3.

Jerath, K., P.S. Fader, and B.G.S. Hardie (2011). "New Perspectives on Customer 'Death' Using a Generalization of the Pareto/NBD Model." In: *Marketing Science* 30.5, pp. 866–880.

Korkmaz, E., R. Kuik, and D. Fok (2012). "'Counting your customer': When will they buy next? An empirical validation of probabilistic customerbase analysis models." In:

McAuliffe, Jon D, David M Blei, and Michael I Jordan (2006). "Nonparametric empirical Bayes for the Dirichlet process mixture model." In: *Statistics and Computing* 16.1, pp. 5–14.

McCulloch, R. and P. E. Rossi (1994). "An exact likelihood analysis of the multinomial probit model." In: *Journal of Econometrics* 64.? Pp. 207–240.

Nobile, Agostino and Alastair T Fearnside (2007). "Bayesian finite mixtures with an unknown number of components: The allocation sampler." In: *Statistics and Computing* 17.2, pp. 147–162.

Paap, Richard and Philip Hans Franses (2000). "A dynamic multinomial probit model for brand choice with different long-run and short-run effects of marketing-mix variables." In: *Journal of Applied Econometrics* 15.6, pp. 717–744.

Rasmussen, Carl Edward (1999). "The infinite Gaussian mixture model." In: *NIPS*. Vol. 12, pp. 554–560.

Reinartz, W.J. and V. Kumar (2000). "On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing." In: *The Journal of Marketing*, pp. 17–35.

Richardson, Sylvia and Peter J Green (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." In: *Journal of the Royal Statistical Society: series B (statistical methodology)* 59.4, pp. 731–792.

Rossi, P., G.M. Allenby, and R. McCulloch (2005). *Bayesian statistics and marketing*. John Wiley and Sons, Ltd.

Schmittlein, D.C., D.G. Morrison, and R. Colombo (1987). "Counting your customers: Who are they and what will they do next?" In: *Management Science*, pp. 1–24.

Schmittlein, D.C. and R.A. Peterson (1994). "Customer base analysis: An industrial purchase process application." In: *Marketing Science*, pp. 41–67.

Stephens, Matthew (2000). "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods." In: *Annals of Statistics*, pp. 40–74.

Van Oest, Rutger and George Knox (2011). "Extending the BG/NBD: A simple model of purchases and complaints." In: *International Journal of Research in Marketing* 28.1, pp. 30–37.

Wübben, M. and F. Wangenheim (2008). "Instant customer base analysis: Managerial heuristics often 'get it right'." In: *Journal of Marketing* 72.3, pp. 82–93.

ERIM Report Series *Research in Management*	
ERIM Report Series reference number	ERS–2014–006–LIS
Date of publication	2014–04–24
Version	24–04–2014
Number of pages	47
Persistent URL for paper	http://hdl.handle.net/1765/51244
Email address corresponding author	rkuik@rsm.nl
Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus University Rotterdam PO Box 1738 3000 DR Rotterdam, The Netherlands Phone: +31104081182 Fax: +31104089640 Email: info@erim.eur.nl Internet: http://www.erim.eur.nl
Availability	The ERIM Report Series is distributed through the following platforms: RePub, the EUR institutional repository Social Science Research Network (SSRN) Research Papers in Economics (RePEc)
Classifications	The electronic versions of the papers in the ERIM Report Series contain bibliographic metadata from the following classification systems: Library of Congress Classification (LCC) Journal of Economic Literature (JEL) ACM Computing Classification System Inspec Classification Scheme (ICS)