A Validation of and Extension to a Non-Parametric Approach to Buy-'Til-You-Die Models

Akhil Ganti
Advisor: Peter Fader

May 1, 2019

Abstract

In this paper, a non-parametric approach to Buy-'Til-You-Die (BTYD) models, which are a class of probability models used to capture the purchasing habits of customers, is investigated. The theory behind BTYD models and Dirichlet processes as the method of non-parameterization is discussed, and model specifications and results are given thereafter. A reflection on the managerial applications of this work follows.

1 Introduction

Since the time the Pareto/NBD framework was proposed as a method of describing repeat-buying behavior, in which customer purchasing habits are modeled where the number of transactions at the customer-level is Poisson distributed (with heterogeneity in transaction rate) and each customer has an exponentially-distributed unobserved lifetime (with heterogeneity also in the dropout rates), various extensions and variations have been proposed over the years, all of which have come to comprise the class of models known as Buy-'Til-You-Die. One such model is the Beta-Geometric/Beta-Binomial (BG/BB) model, which can be thought of as a discrete version of the Pareto/NBD model, where customers have the opportunity to purchase in discrete intervals (e.g. at the end of every month), and the opportunity for "death" happens right after. Another such example is the periodic-death opportunity (PDO) model, where the continuous nature of transactions from the Pareto/NBD remains the same but the opportunity for death occurs at discrete intervals across the calendar year (i.e. not tied to transaction times). One of the enduring features of all the models in this class, however, is the presence of parameterization - that is, the modeling of heterogeneity for both transaction rates and dropout rates is done according to a fixed distribution (e.g. gamma distributions in the Pareto/NBD case) that produces point estimates and with no prior. While this has allowed for an intuitive and easily communicable motivation, the downside is that there is a loss of flexibility in the shapes that the distributions can take on.

This paper examines the consequences of removing this limitation, which is accomplished by assuming Dirichlet process priors and applying this to the three aforementioned models. In particular, the application of this non-parametric approach to the Pareto/NBD model will be based on the work done in an unpublished paper by Quintana and Marshall, such that the section will serve as a validation of their results.

Section 2 discusses the theory behind the BTYD framework. Section 3 gives a brief introduction to Dirichlet processes and the stick-breaking representation explicitly used in the model specifications. Section 4 details the non-parameterized models themselves, with comparisons to results from the original papers. Section 5 discusses the managerial implications of this work with potential use cases. Finally, section 6 ends with an overview of the paper and potential avenues for further inquiry.

2 BTYD Models

The Buy-'Til-You-Die class is used to model the purchasing characteristics and habits of customers, which is then used to predict customer lifetime value. The baseline story across the different variations has two primary components:

- Customers, while "alive," have some propensity for repeating the action in question (e.g. purchasing, donating, etc.), and there is heterogeneity in the extent to which customers repeat this action

- Customers churn, or "die," at some point in time, and there is heterogeneity in how long it takes for customers to churn

Evaluating customer lifetime value involves the computation of metrics such as the probability of a customer being alive and the expected number of future transactions for a given customer.

The most widely used BTYD model is the Pareto/NBD model, introduced in Schmittlein et al. (1987), which is discussed further in detail in below. However, others have been introduced that are based on different underlying stories/assumptions about customer behavior. The PDO and BG/BB models discussed below are two such variants, and others include the BG/NBD and SBB-G/B models.

2.1 Pareto/NBD

The Pareto/NBD model applies to non-contractual settings where purchases are continuous. As such, purchases can happen at any time and it is unknown when exactly customers churn. The model is based on several assumptions:

- Customers are "alive" for some lifetime after which they "die" and become permanently inactive
- While alive, a customer's transactions are modeled by a Poisson process with transaction rate λ:

$$P(X(t) = x|\lambda) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \quad x = 0, 1, 2...$$

- There is heterogeneity in transaction rates across customers, which is modeled with a gamma distribution with shape r and scale α:

$$g(\lambda|r, \alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)}$$

- Each customer has an unobserved lifetime of length τ that is exponentially distributed with dropout rate μ:

$$f(\tau|\mu) = \mu e^{-\mu\tau}$$

- There is heterogeneity in dropout rates across customers, which is modeled with a gamma distribution with shape s and scale β:

$$g(\lambda|s, \beta) = \frac{\beta^s \mu^{s-1} e^{-\mu\beta}}{\Gamma(s)}$$

- The transaction and dropout rates vary independently across customers

An NBD model for the distribution of the number of transactions results from the second and third assumptions, whereas a Pareto-II model for the distribution of lifetimes results from the fourth and fifth assumptions. Taken together, the model's likelihood function for the latent parameters λ and μ can be written as:

$$\begin{aligned} L(\lambda,\mu|x,t_x,T) &= L(\lambda|x,t_x,\tau T)P(\tau > T|\mu) \\ &\quad + \int_{t_x}^{T} L(\lambda|x,T,\text{inactive at}\tau \in (t_x,T])f(\tau|\mu)d\tau \\ &= \frac{\lambda}{\lambda+\mu}(\mu e^{-(\lambda+\mu)t_x} + \lambda e^{-(\lambda+\mu)T}) \end{aligned}$$

where x is the number of repeat transactions, t_x is the time of the most recent transaction, and T is the last time in the observation period.

However, because λ and μ are not explicitly known for each customer, the distributions for λ and μ are used to take the expectation over $L(\lambda,\mu|x,t_x,T)$:

$$\begin{aligned} L(r,\alpha,s,\beta|x,t_x,T) &= \int_0^\infty \int_0^\infty L(\lambda,\mu|x,t_x,T)g(\lambda|r,\alpha)g(\mu|s,\beta)d\lambda d\mu \\ &= \frac{\Gamma(r+x)\alpha^r\beta^s}{\Gamma(r)}\{\frac{1}{(\alpha+T)^{r+x}(\beta+T)^s} + (\frac{s}{r+s+x})A_0\} \end{aligned}$$

where, if $\alpha \geq \beta$:

$$A_0 = \frac{{}_2F_1\left(r+s+x, s+1; r+s+x+1; \frac{\alpha-\beta}{\alpha+t_x}\right)}{(\alpha+t_x)^{r+s+x}} - \frac{{}_2F_1\left(r+s+x, s+1; r+s+x+1; \frac{\alpha-\beta}{\alpha+T}\right)}{(\alpha+T)^{r+s+x}}$$

and if $\alpha \leq \beta$:

$$A_0 = \frac{{}_2F_1\left(r+s+x, r+x; r+s+x+1; \frac{\beta-\alpha}{\beta+t_x}\right)}{(\beta+t_x)^{r+s+x}} - \frac{{}_2F_1\left(r+s+x, r+x; r+s+x+1; \frac{\beta-\alpha}{\beta+T}\right)}{(\beta+T)^{r+s+x}}$$

2.2 PDO

The periodic death opportunity, or PDO, model follows a similar story to that of the Pareto/NBD but with a key difference in the customer churn/death process. Namely, rather than modeling the death opportunities as occurring in continuous time, the PDO model models them as occurring at fixed, discrete intervals (for example, customers have the opportunity to churn every 3 days). Thus, while the first three assumptions from the Pareto/NBD remain the same, the fourth and fifth assumptions change as follows:

- The random variable Ω represents the (unobserved) time at which a customer dies, such that every τ units of time, the customer dies with probability θ. The probability a customer has died by time t is:

$$P(\Omega \leq t|\theta, \tau) = 1 - (1-\theta)^{\lfloor t/\tau \rfloor}$$

- There is heterogeneity in churn probability θ across customers, which is modeled with a beta distribution:

$$f(\theta|a,b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

As before, the transaction rates and dropout probabilities vary independently across customers.

One interesting aspect of the PDO model is that, as τ approaches zero, it collapses into the Pareto/NBD model, as this essentially means that the customer can die at any moment. As such, the Pareto/NBD model serves as a subset of the PDO model.

From the above assumptions, the model's likelihood function for the latent parameters λ and θ is as follows:

$$\begin{aligned} L(\lambda, \theta, \tau | x, t_x, T) = \lambda^x e^{-\lambda T}(1-\theta)^{\lfloor T/\tau \rfloor} &+ \delta_{\lfloor T/\tau \rfloor > \lfloor t_x/\tau \rfloor} \\ &\cdot \sum_{j=1}^{\lfloor T/\tau \rfloor - \lfloor t_x/\tau \rfloor} \lambda^x e^{-\lambda(\lfloor t_x/\tau \rfloor + j)\tau} \theta (1-\theta)^{\lfloor t_x/\tau \rfloor + j - 1} \end{aligned}$$

where x is the number of repeat transactions, t_x is the time of the most recent transaction, and T is the last time in the observation period.

Since each customer's λ and θ are not explicitly known, the distributions for λ and θ are used to take the expectation over $L(\lambda, \theta, \tau|x, t_x, T)$, which results in the following likelihood function:

$$\begin{aligned} L(r,\alpha,a,b,\tau|x,t_x,T) &= \int_0^1\int_0^\infty L(\lambda,\theta,\tau|x,t_x,T)f(\lambda|r,\alpha)f(\theta|a,b)d\lambda d\theta \\ &= \frac{\Gamma(r+x)\alpha^r}{\Gamma(r)}[(\frac{1}{\alpha+T})^{r+x}\frac{B(a,b+\lfloor T/\tau\rfloor}{B(a,b)} + \delta_{\lfloor T/\tau\rfloor > \lfloor t_x/\tau\rfloor} \\ &\quad \cdot \sum_{j=1}^{\lfloor T/\tau\rfloor - \lfloor t_x/\tau\rfloor}\{(\frac{1}{\alpha+(\lfloor t_x/\tau\rfloor+j)\tau})^{r+x} \\ &\quad \cdot \frac{B(a+1,b+\lfloor t_x/\tau\rfloor+j-1}{B(a,b)}\}] \end{aligned}$$

2.3 BG/BB

Whereas the Pareto/NBD and PDO models focused on transactions that can happen at any point in time, regardless of the nature of the death process, the philosophy of the beta-geometric/beta-Bernoulli, or BG/BB, model focuses on discrete transactions in a non-contractual setting. Thus, the majority of the assumptions differ from those presented thus far, though one can see how they are discrete analogs:

- While alive, a customer's transactions are modeled by a Bernoulli distribution with probability p:

$$P(Y_t = 1|p, \text{alive at } t) = p, \quad 0 \leq p \leq 1$$

- There is heterogeneity in purchase probabilities across customers, which is modeled with a beta distribution:

$$f(p|\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

- At the beginning of every transaction opportunity, a customer dies with probability θ

- There is heterogeneity in death probabilities across customers, which is modeled with a beta distribution:

$$f(\theta|\gamma,\delta) = \frac{\theta^{\gamma-1}(1-\theta)^{\delta-1}}{B(\gamma,\delta)}$$

Similar to before, the transaction and death probabilities are independent across customers.

From the above assumptions, the model's likelihood function for the latent parameters p and θ is:

$$\begin{aligned} L(p,\theta|x,t_x,n) = p^x(1-p)^{n-x}(1-\theta)^n \\ + \sum_{i=0}^{n-t_x-1} p^x(1-p)^{t_x-x+i}\theta(1-\theta)^{t_x+i} \end{aligned}$$

where x is the number of repeat transactions, t_x is the time of the most recent transaction, and n is the number of transaction opportunities.

Because each customer's latent parameters p and θ are unknown, the distributions for p and θ are used to take the expectation over $L(p,\theta|x,t_x,n)$, which results in the following likelihood function:

$$\begin{aligned} L(\alpha,\beta,\gamma,\delta|x,t_x,n) &= \int_0^1\int_0^1 L(p,\theta|x,t_x,n)f(p|\alpha,\beta)f(\theta|\gamma,\delta)dpd\theta \\ &= \frac{B(\alpha+x,\beta+n-x)}{B(\alpha,\beta)}\frac{B(\gamma,\delta+n)}{B(\gamma,\delta)} \\ &\quad + \sum_{i=0}^{n-t_x-1}\frac{B(\alpha+x,\beta+t_x-x+i)}{B(\alpha,\beta)} \\ &\quad \cdot \frac{B(\gamma+1,\delta+t_x+i)}{B(\gamma,\delta)} \end{aligned}$$

3 Dirichlet Processes

3.1 Introduction

The Dirichlet process (DP) is a stochastic process that produces a distribution over distributions, whereby each draw from the process creates a distribution. In the following explanation of DPs, the treatment given by Teh (2017) is extensively referred to.

The DP is essentially an infinite-dimensional generalization of Dirichlet distributions. To demonstrate, assume a mixture model with K components:

$$\begin{array}{ll} \pi|\alpha \sim \mathrm{Dir}(\frac{\alpha}{K},...,\frac{\alpha}{K}) & \theta_k^*|H \sim H \\ z_i|\pi \sim \mathrm{Mult}(\pi) & x_i|z_i,\{\theta_k^*\} \sim F(\theta_k^*) \end{array}$$

Here, π represents the mixing proportion, α is the Dirichlet prior hyperparameter, and H is the base prior distribution over the component parameters θ^*_K, which parameterizes $F(\theta)$. This construction leads to an infinite mixture model as $K \to \infty$, from which the DP ultimately is derived. Notably, with an infinite mixture model, the number of components does not need to be predefined.

3.2 Definition

A random distribution G has a DP prior if all its marginal distributions are Dirichlet distributed. With a base distribution H over the parameter space Θ and the hyperparameter α (known as the concentration parameter), $G \sim \mathrm{DP}(\alpha, H)$ if:

$$(G(A_1), ...G(A_r)) \sim \mathrm{Dir}(\alpha H(A_1), ..., \alpha H(A_r))$$

where $A_1, ..., A_r$ is any finite measurable partition over Θ.

3.3 Stick-Breaking Construction

The stick-breaking construction is one common way to understand and represent DPs, and it is the method that is explicitly used as the prior for the models discussed later in this paper. With this, $G \sim \mathrm{DP}(\alpha, H)$ means that:

$$\begin{aligned} \beta_k &\sim \mathrm{Beta}(1, \alpha) & \theta^*_k &\sim H \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) & G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta^*_k} \end{aligned}$$

Here, G can be understood as a weighted sum of point masses, where the weights π_k are constructed as follows. Consider a stick of length 1, and assign to π_1 the stick length that is broken off at β_1. From the remaining part, break the stick at β_2 and assign to π_2 the length of this broken portion. Continue recursively until all π_k values are assigned in this manner.

4 Empirical Analysis

In the Pareto/NBD and PDO models below and their subsequent analyses, the CDNOW dataset was used since it was also used by Quintana and Marshall (for the Pareto/NBD model) and Fader et. al. (for the PDO model), thus maintaining consistency. For the implementation and analysis of the BG/BB model, the donations dataset used in the original paper is used here. Finally, STAN software was used to perform MCMC sampling.

4.1 Pareto/NBD

As mentioned previously, the development for the non-parametric Pareto/NBD model is done as a validation to the unpublished work presented in Quintana and Marshall (2014), which is henceforth referred to as QM. However, the methodology used in this paper (both for this model and the subsequent two models) contains changes that were made to their approach, so an overview of their methodology will be presented first, followed by a description of the changes.

4.1.1 QM Model Specification

Let (x_i, t_{x_i}, T_i), for $i = 1, ..., n$, represent the fully relevant transaction history for each customer, where x_i is the number of transactions done by customer i, t_{x_i} is the time of the last transaction, and T_i is the time at which the observation period ends (such that $t_{x_i} \leq T_i$). Note that $x_i = 0$ implies that $t_{x_i} = 0$.

The central assumptions of the Pareto/NBD model (i.e. λ_i and μ_i are gamma-distributed, etc.) remain, so the likelihood for the ith customer given his/her λ_i and μ_i is:

$$p(x_i, t_{x_i}, T_i|\lambda_i, \mu_i) = \frac{\lambda_i^{x_i}\mu_i}{\lambda_i+\mu_i}e^{-(\lambda_i+\mu_i)t_{x_i}} + \frac{\lambda_i^{x_i+1}}{\lambda_i+\mu_i}e^{-(\lambda_i+\mu_i)T_i}$$

Note that in the case that $x_i = t_{x_i} = 0$, the above likelihood collapses to:

$$p(x_i = 0, t_{x_i} = 0, T_i|\lambda_i, \mu_i) = \frac{\mu_i+\lambda_i e^{-(\lambda_i+\mu_i)T_i}}{\lambda_i+\mu_i}$$

In this case, whenever $\lambda_i > 0$ and $T_i > 0$, the likelihood is maximized for $\mu_i = \infty$. As a result, QM introduce an explicit π parameter such that, with probability π, a given customer may a priori churn immediately after time $t = 0$. Their complete model specification is thus as follows:

Let $\theta_i = (\lambda_i, \mu_i)$. A hierarchical model is defined such that the probability defined above is at the top level, and the parameters $\theta_1, ...\theta_n$ are defined by the following mixture:

$$\theta_1, ...\theta_n | \pi, F \sim \pi\delta_{(0,c)}(\cdot) + (1 - \pi)F(\cdot)$$

where, a priori, $\pi \sim \text{Beta}(a, b)$ is independent of $F \sim \text{DP}(M, F_0(\omega))$.

Here, in the zero-inflated case, $\lambda_i = 0$ and μ_i is an arbitrarily large constant c to still allow for the possibility that a customer in this group will make a transaction at some later time. $F_0(\omega)$ is defined as the four-dimensional distribution:

$$F_0(\omega) = \text{Unif}(r|r_0, r_1) \times \text{Unif}(\alpha|\alpha_0, \alpha_1) \times \text{Unif}(s|s_0, s_1) \times \text{Unif}(\beta|\beta_0, \beta_1)$$

QM also implemented least-square clustering, which estimates the clustering of observations based on realizations from the posterior clustering distribution. This allows for the number of clusters to be produced as a side effect of the process rather than requiring the number to be pre-defined. In their analysis, QM fit the model using Markov chain Monte Carlo (MCMC) sampling with the following values: $a = b = 1$ to produce a uniform prior on π between 0 and 1; $M = 1$; $r_0 = \alpha_0 = s_0 = \beta_0 = 0.5$; $r_1 = \alpha_1 = s_1 = \beta_1 = 10000$, which allows for generally unconstrained support; and $c = 10000$ for the zero-cluster (which implies $\mu = 10000$). In addition, in constructing the DP, truncation is applied such that only approximately $k = 25$ values are imputed (see section 3.3).

4.1.2 Changes Made

The specification implemented by QM remains generally the same in the applied ideas for the modeling in this paper but with a few key changes. They are as follows:

- Rather than using the least-squares clustering method separate from the model, the clustering methodology here is built in to the model itself. This is achieved by representing r, α, s, and β as vectors instead of scalar values, whose size is equal to the level of truncation in the DP (k in section 3.3). Since the truncation level is a user-defined parameter, this approach is similar to pre-defining the number of clusters as one might do with the k-means algorithm. However, the key difference is that, with the DP, an arbitrarily large k can be chosen (to mimic the weighted sum whose limit goes to infinity), and the resulting π_k values can be used to determine the relative size and importance of the k clusters that are produced.

With this approach, the MCMC sampling produces a n-by-k matrix of mean λ and μ values for each customer and each cluster (n customers and k clusters), from which customers are ad-hoc assigned to clusters based on which corresponding λ and μ values maximize the log-likelihood function.

In the actual application of this method, the value k was forced to be kept to a small value due to computational reasons. To exemplify, when applied to the PDO model that is discussed later, a value of $k = 5$ led to a total run-time of almost 70 hours. However, although it seems that such a small value would not well approximate the DP in its limit of infinity, it turns out such truncations are arbitrarily accurate and still work in practice (see Campbell et al. (2019)). Note that the limits of the uniform priors in the definition of F_0 remain the same.

- The second main change made in the current model is that a π parameter is not used to directly model a zero-cluster. Instead, it is assumed that the previously explained re-parameterization of r, α, s, and β will subsume such a cluster and preclude the necessity for an explicit definition. As a result of this, the modeling of λ_i and μ_i changes as well. Namely, rather than being a weighted mixture based on π, they are instead gamma-distributed at the cluster-level.

Although not a change between the current and QM models, one thing to note is that the log-likelihood for the Pareto/NBD model was constructed on given λ_i and μ_i values as opposed to the four gamma-distribution parameters. The main reason for this was that implementing the hypergeometric function as one of the subroutines led to issues with gradient calculation.

4.1.3 Results

Figure 1 presents the distributions of the cluster-level parameters, figure 2 shows the predicted posterior distributions of the latent variables λ and μ, figure 3 shows the distribution of customers in the various clusters, and table 1 presents the summary statistics for the key parameters of the model.

As can be seen, the posterior distribution for λ has a form similar to that of a gamma distribution, whereas the distribution of μ is much more spread out and somewhat bi-modal. Overall, however, the results are surprising since they not only defy the form of the gamma distribution (specifically for μ) but they are also somewhat inconsistent from the results produced by QM, which can be seen in figure 4. One explanation for this, however, is the aforementioned choice to not model an explicit zero-cluster and to instead let the Dirichlet process prior account for that. Additionally, the parameters themselves across the clusters are fairly divergent from the point estimates attained in Pareto/NBD model.

Table 2 presents several goodness of fit statistics to compare the current model with that of QM and the original Pareto/NBD models. Across the given metrics, with correlation, mean absolute error (MAE), and root mean squared error (RMSE) representing out-of-sample fit, the current model outperforms both QM and the original Pareto/NBD models. It is also worth noting that, despite the relative increase in 16 parameters (from modeling 5 clusters) from QM to the current model, a likelihood-ratio test produces a p-value of virtually zero, which indicates significance.

One major pitfall of this non-parametric approach, is that it is unable to predict well for a randomly chosen customer with no prior information. This occurs because the cluster-level parameters are unable to be weighted well enough (specifically, according to the weights given by the DP prior) to produce accurate single point estimates. In this case, it is better to use the point estimates computed from the maximum-likelihood estimation given by the base Pareto/NBD model. However, one potential solution around this would be to a posteriori fit a gamma distribution to λ and μ (e.g. using MLE) to back out the parameter values, though such a method leads to a lower log-likelihood than the aforementioned "base" method.

4.2 PDO

The PDO model was constructed based on the work presented in Jerath et al. (2011) and was generally in line with the theory laid out in section 2.2. The main change is that, similar to the Pareto/NBD model, r, α, a, and b are implemented as vectors with length equal to the size of the truncated DP. This size is also the desired upper bound on the number of clusters.

As in the original paper, only one τ value is imputed, rather than modeling different values for each cluster. The primary reason for this was computational - as mentioned above, it had already taken approximately 70 hours to complete MCMC sampling without doing so. In addition, in contrast to the Pareto/NBD above, the log-likelihood function here was constructed on the underlying distributions' parameters, as opposed to the latent variables, since the aforementioned gradient computation error was not encountered.

4.2.1 Results

Table 3 presents the summary statistics for the key parameters of the model, figure 5 presents the distributions of the cluster-level parameters, figure 6 shows the predicted posterior distributions of the latent variables λ and μ, and figure 7 shows the distribution of customers in the various clusters.

Here again, the posterior distribution for λ has a form similar to that of a gamma distribution, while the beta-distributed μ has a long left tail with a mean of approximately 0.90 and a median of approximately 0.90. Compared to the original PDO model, which produced a log-likelihood of -9,585.6, this non-parametric approach produced a log-likelihood of -9,534.351, indicating a significant improvement based on the likelihood ratio test. As with the Pareto/NBD model, however, the model does not perform well for unconditional expectations.

4.3 BG/BB

The BG/BB was implemented based on the work presented in Fader et al. (2009) and in line with the theory laid out in section 2.3. The main change is that, similar to the Pareto/NBD model, α, β, γ, and δ are implemented as vectors with length equal to the size of the truncated DP. This size is also the desired upper bound on the number of clusters. Additionally, similar to the PDO model, the log-likelihood is calculated from the distribution parameters rather than the latent variables.

4.3.1 Results

Table 4 presents the summary statistics for the key parameters of the model, figure 8 presents the distributions of the cluster-level parameters, figure 9 shows the predicted posterior distributions of the latent variables p and θ, and figure 10 shows the distribution of customers in the various clusters. Compared to the original BG/BB model, which produced a log-likelihood of -33,225.6, this method resulted in a log-likelihood of -23,021.68, indicating a significant improvement based on the likelihood ratio test. Again, however, as with the Pareto/NBD and PDO models, this model does not perform well for unconditional expectations.

5 Economic and Business Applications

The most direct consequence from the improvement shown by these non-parametric models is that they can be used by managers and decision-makers to better understand the behavior and underlying characteristics of their business' customers. For example, with the non-parameterized Pareto/NBD model, the distribution of μ is fairly unlike a gamma distribution, and such a quirk would not have been captured by the original model. Since all the analyses conducted here are in the context of already having some information at the customer-level, these results naturally lend themselves to a more customer-centric approach, in which managers can better utilize different marketing strategies in a more targeted nature.

As a result of these model developments, more accurate and precise estimates for customer lifetime value (CLV) can be made at the individual level. Rather than relying on point estimates from a distribution across the entire customer base to calculate CLV from expectations, this non-parametric approach, particularly through the clustering information it provides, can provide segmentation information about the customers. From this, managers can then more accurately calculate threshold values for how much to spend retaining these customers.

As a corollary to this result, it can be seen how this new methodology would be particularly useful in settings where customer behavior is fundamentally not well-modeled by the assumptions underlying these models. An example of this arises in the secondary retail dataset used by QM, in which the the posterior predictions for λ and μ diverge from the form of a gamma distribution.

6 Conclusions & Future Work

In this paper, a more effective, non-parametric approach to modeling customer transaction behavior was validated and extended to variants in the Buy-'Til-You-Die class of models. A methodology of using the stick-breaking representation of Dirichlet processes as the means of non-parameterization in place of an underlying, fixed distribution was demonstrated to produce significantly better predictive results both in and out of sample. Lastly, a framework for understanding the clustering/segmentation of the customers based on these models was briefly presented.

The work presented in this paper can be thought of as an additional proof-of-concept, in that it opens up many different possibilities of in-depth investigation and study. One such example is to explore how the aforementioned issue with the weighting scheme can be resolved to allow for modeling of and predictions on randomly chosen with no prior transaction history. In addition, since the precise values of the scale parameters in the gamma distribution are actually not of great importance, another possible area of study would be to fix those values (for example, such that $\alpha = \beta \approx 10$ in the Pareto/NBD model, which is what the original point estimates are) to allow for a greater degree of interpretation in the resulting cluster-level parameters.

Ultimately, as there does not currently exist a large quantity of literature in this area of non-parametric probability models in marketing (particularly with Dirichlet processes), there is much study to be done as well as a variety of potential applications to be further studied.

References

Campbell, T., Huggins, J. H., How, J. P., and Broderick, T. (2019). Truncated random measures. *Bernoulli*, 25(2):1256–1288.

Fader, P., Hardie, B., and Shang, J. (2009). Customer-base analysis in a discrete-time noncontractual setting. *SSRN Electronic Journal.*

Fader, P. S., Hardie, B. G. S., and Lee, K. L. (2005). “counting your customers” the easy way: An alternative to the pareto/nbd model. *Marketing Science*, 24(2):275–284.

Jerath, K., Fader, P., and Hardie, B. (2011). New perspectives on customer death using a generalization of the pareto/nbd model. *SSRN Electronic Journal.*

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2004). Assessing heterogeneity in discrete choice models using a dirichlet process prior. *Review of Marketing Science*, 2(1).

Ma, S.-H. and Liu, J.-L. (2007). The mcmc approach for solving the pareto/nbd model and possible extensions. *Third International Conference on Natural Computation (ICNC 2007).*

Quintana, F. A. and Marshall, P. (2014). A bayesian non-parametric pareto/nbd model: Individual and cluster analysis1.

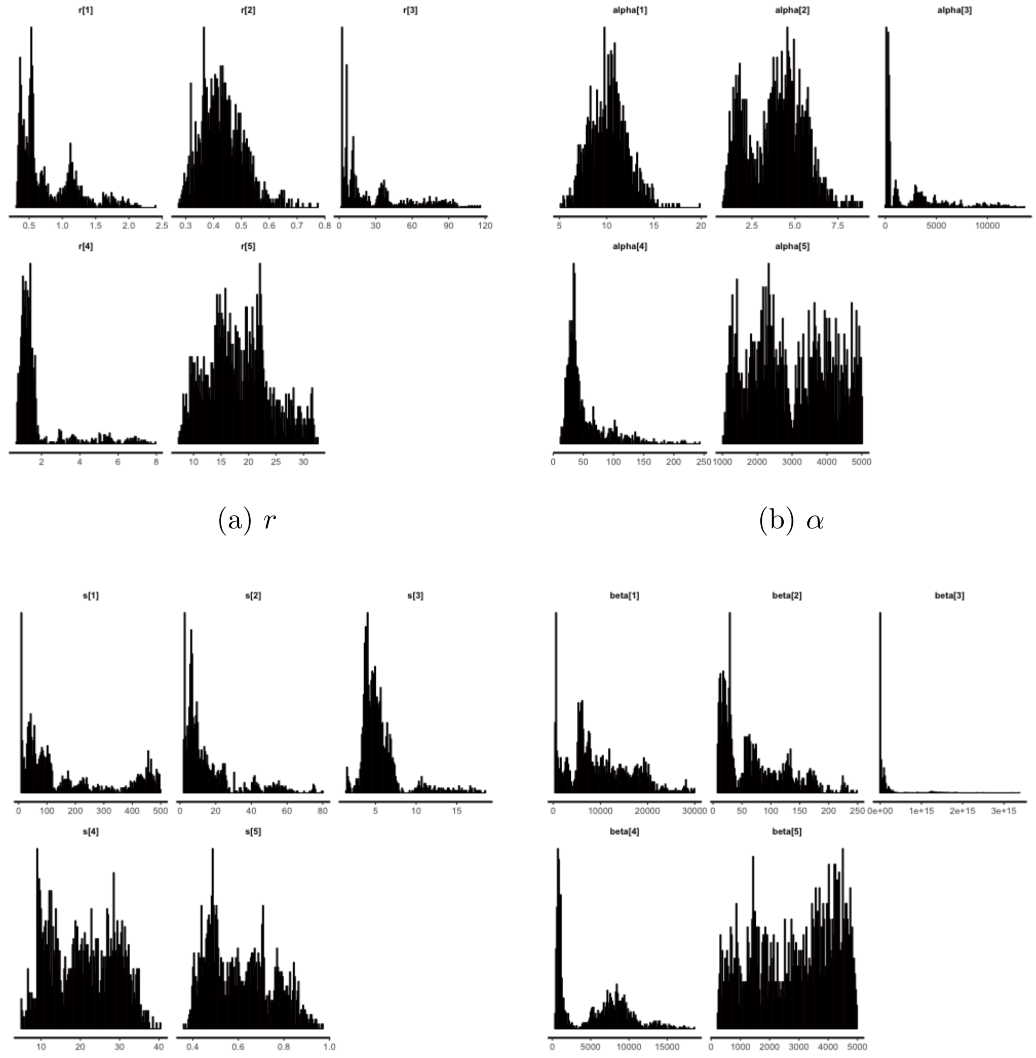
Schmittlein, D. C., Morrison, D. G., and Colombo, R. (1987). Counting your customers: Who-are they and what will they do next? *Management Science*, 33(1):1–24.

Teh, Y. W. (2017). Dirichlet process. *Encyclopedia of Machine Learning and Data Mining*, page 361–370.

Appendix

Cluster	π	r	α	s	β
1	0.453	0.761	10.212	171.461	9.778×10^3
2	0.224	0.432	3.909	15.441	$6.906\,805 \times 10^1$
3	0.0673	27.374	2300.219	5.420	1.112×10^{14}
4	0.227	1.721	49.596	20.673	5.172×10^3
5	0.0271	18.279	3001.578	0.611	2.858×10^3

Table 1: Summary statistics for key parameters of the current Pareto/NBD model



(a) r

(b) α

(c) s

(d) β

Figure 1: Cluster-level distributions of key parameters of the current Pareto/NBD model

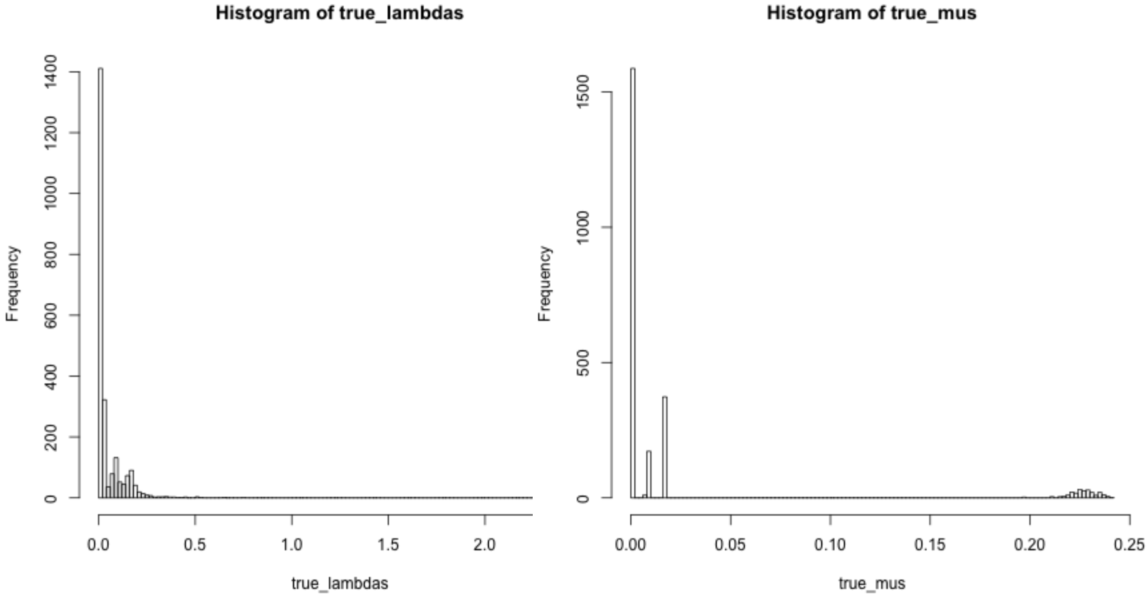


Figure 2: Posterior distributions for λ and μ of the current Pareto/NBD model

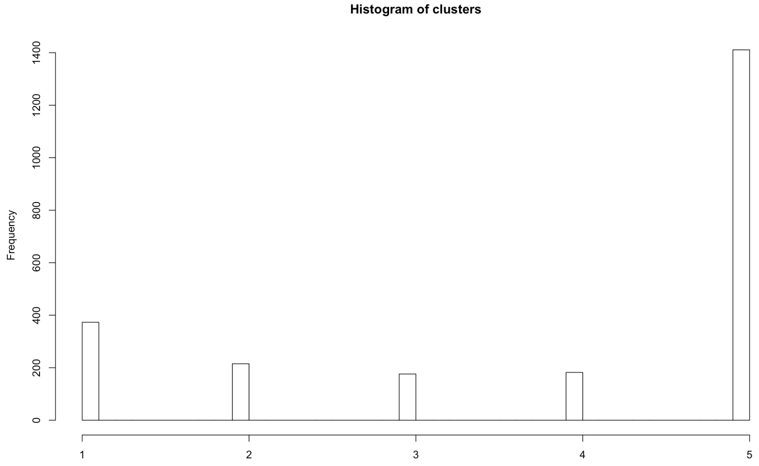


Figure 3: Distribution of customers among clusters of the current Pareto/NBD model

Statistics	Current	QM	Pareto/NBD
Log-Likelihood	−7852.3	−8588.5	−9595.0
Correlation	0.752	0.631	0.630
MAE	0.536	0.668	0.755
RMSE	1.555	1.599	1.604

Table 2: Model comparisons for goodness-of-fit of the current Pareto/NBD model

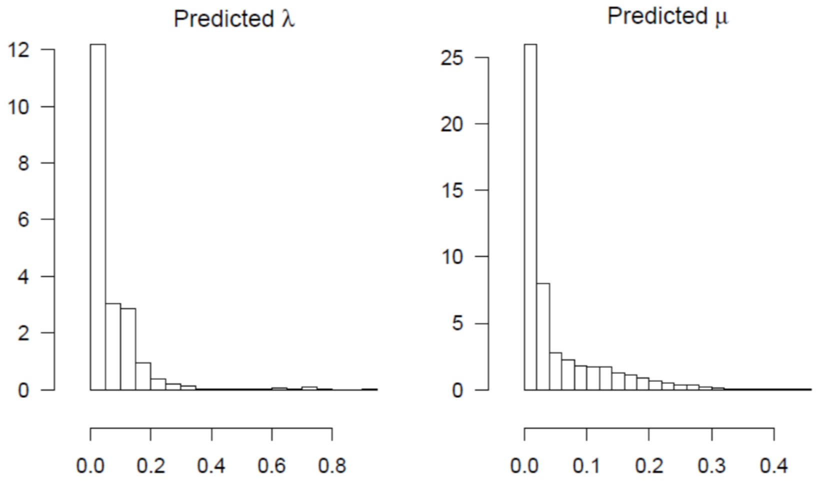
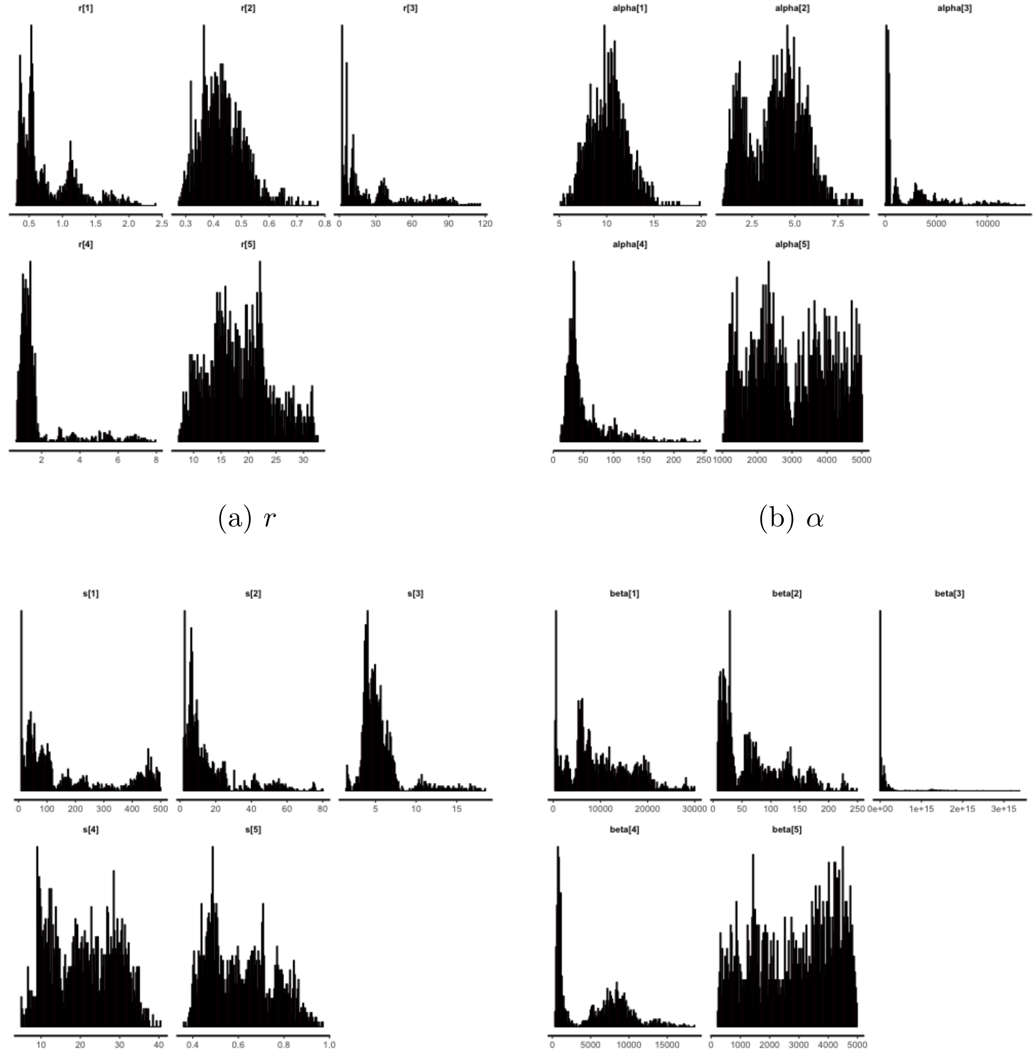


Figure 4: Posterior distributions for λ and μ from QM

Cluster	π	r	α	a	b	τ
1	0.671	0.273	1.804	1.813	0.148	6.504
2	0.117	0.340	5.355	3.523	0.167	6.504
3	0.186	1.005	0.580	6.716	1.551	6.504
4	0.005	4.209	0.364	0.376	0.289	6.504
5	0.021	0.467	2.531	0.731	0.534	6.504

Table 3: Summary statistics for key parameters of the current PDO model



(a) r

(b) α

(c) s

(d) β

Figure 5: Cluster-level distributions of key parameters of the current PDO model

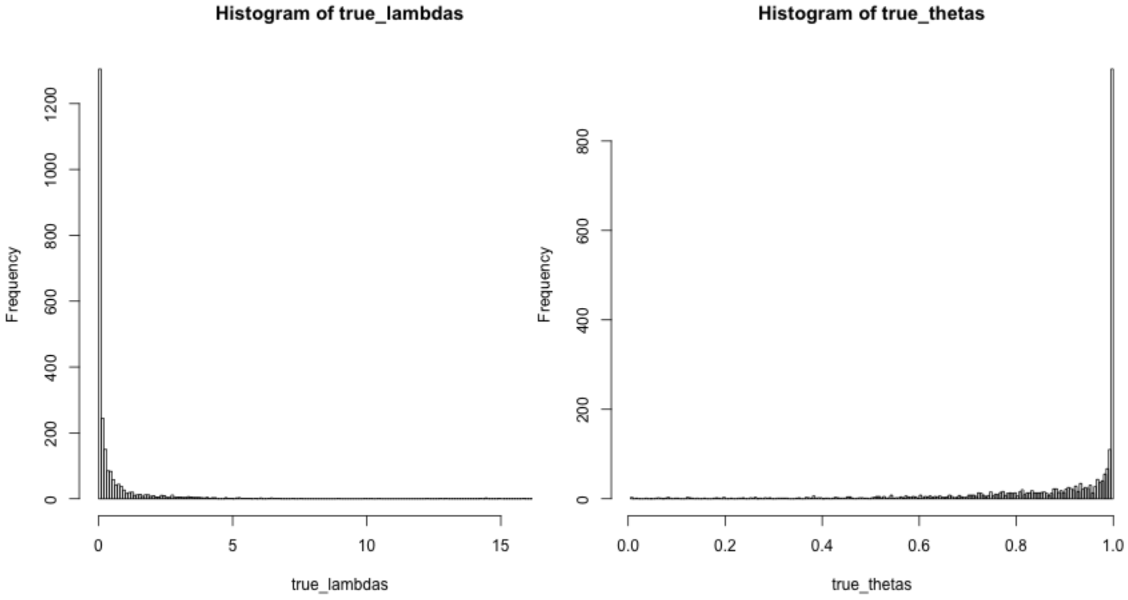


Figure 6: Posterior distributions for λ and θ of the current PDO model

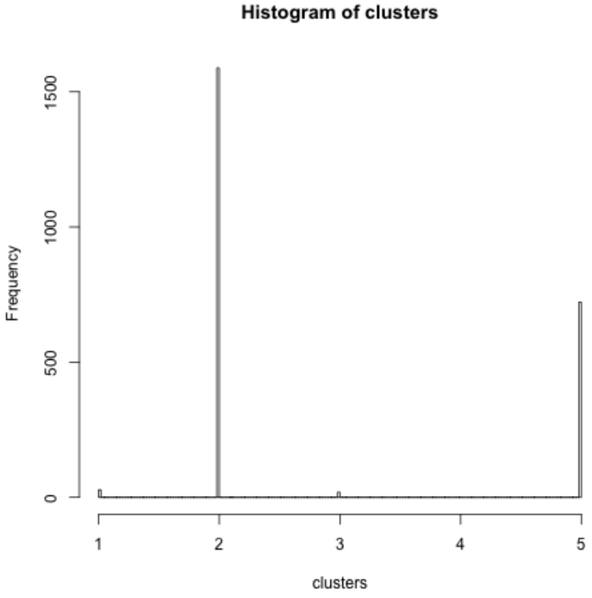
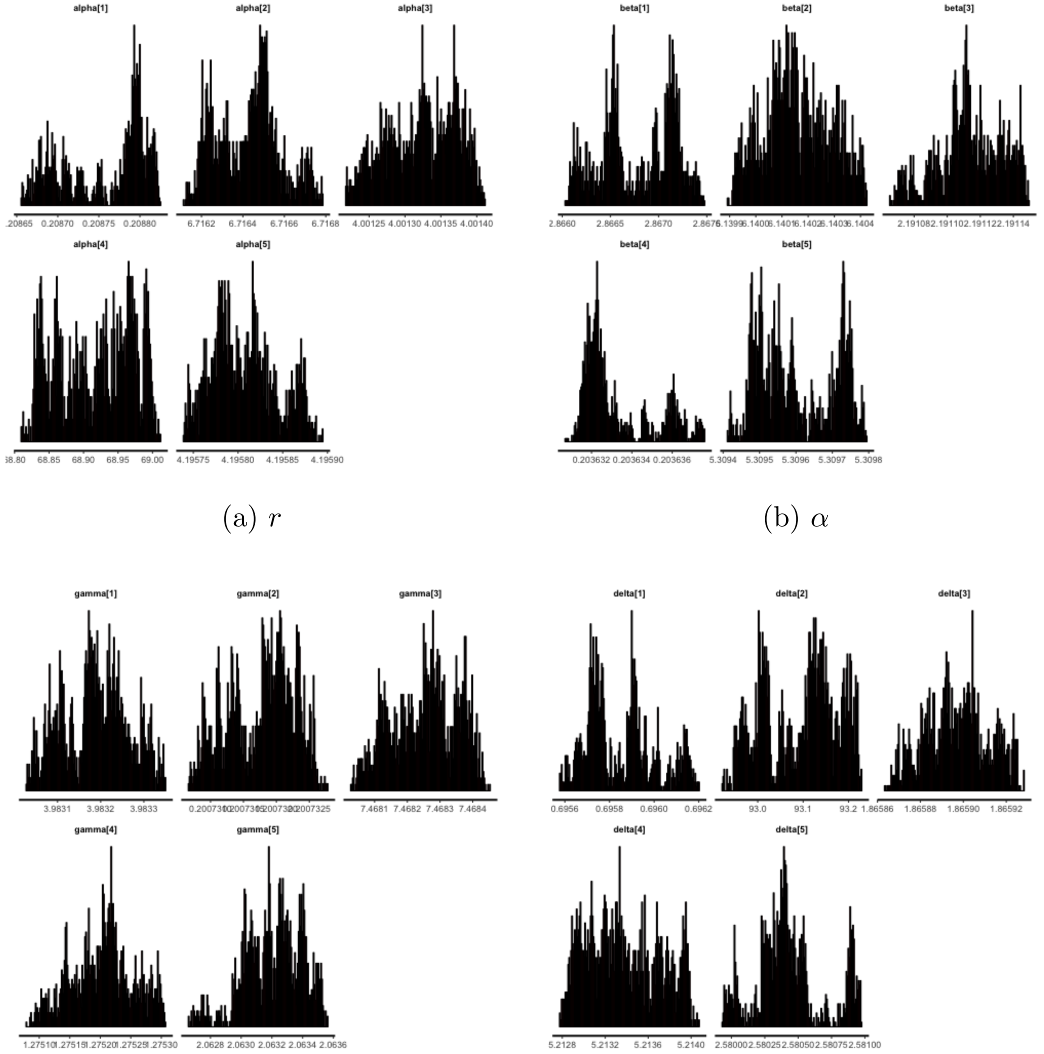


Figure 7: Distribution of customers among clusters of the current PDO model



(a) r

(b) α

(c) s

(d) β

Figure 8: Cluster-level distributions of key parameters of the current BG/BB model

Cluster	π	α	β	γ	δ
1	0.177	0.209	2.867	3.983	0.696
2	0.635	6.716	6.140	0.201	93.094
3	4.001	2.191	7.468	6.716	1.866
4	0.019	68.918	0.204	1.275	5.213
5	0.005	4.196	5.310	2.063	2.580

Table 4: Summary statistics for key parameters of the current BG/BB model

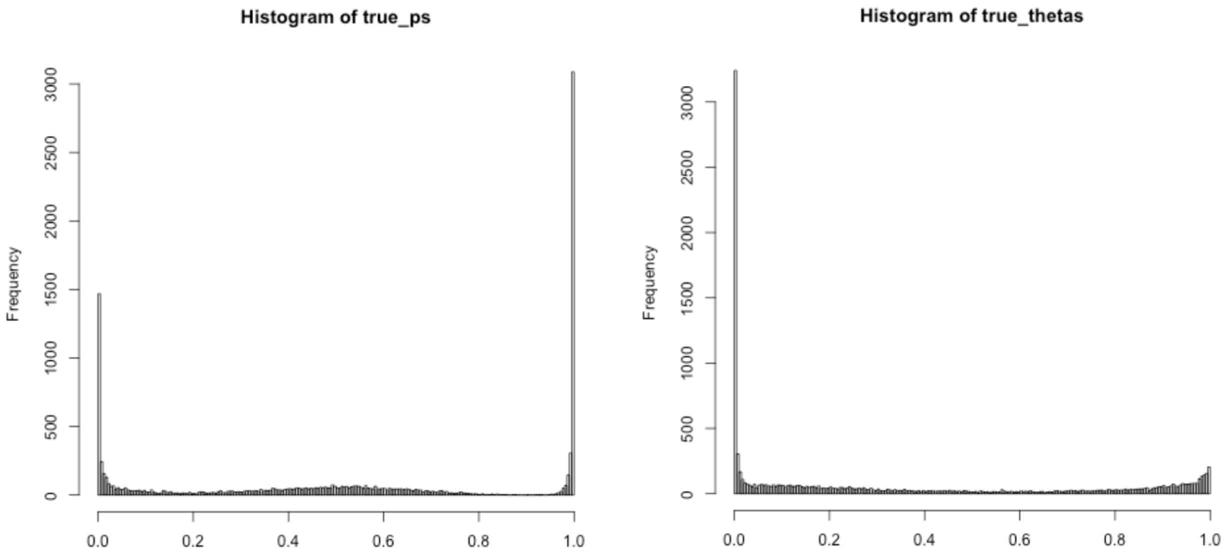


Figure 9: Posterior distributions for λ and θ of the current BG/BB model

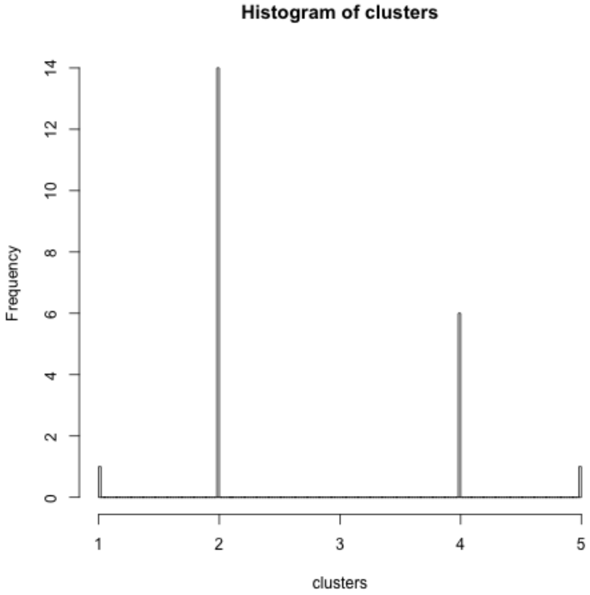


Figure 10: Distribution of customers among clusters of the current BG/BB model