Nonparametric Hierarchical Bayesian Modeling in Non-contractual Heterogeneous Survival Data

Shouichi Nagano  
NTT Service Evolutions Labs  
nagano.shouichi@lab.ntt.co.jp

Yusuke Ichikawa  
NTT Service Evolutions Labs  
ishikawa.yusuke@lab.ntt.co.jp

Noriko Takaya  
NTT Service Evolutions Labs  
takaya.noriko@lab.ntt.co.jp

Tadasu Uchiyama  
NTT Service Evolutions Labs  
uchiyama.tadasu@lab.ntt.co.jp

Makoto Abe  
The University of Tokyo  
abe@e.u-tokyo.ac.jp

ABSTRACT
An important problem in the non-contractual marketing domain is discovering the customer lifetime and assessing the impact of customer’s characteristic variables on the lifetime. Unfortunately, the conventional hierarchical Bayes model cannot discern the impact of customer’s characteristic variables for each customer. To overcome this problem, we present a new survival model using a non-parametric Bayes paradigm with MCMC. The assumption of a conventional model, logarithm of purchase rate and dropout rate with linear regression, is extended to include our assumption of the Dirichlet Process Mixture of regression. The extension assumes that each customer belongs probabilistically to different mixtures of regression, thereby permitting us to estimate a different impact of customer characteristic variables for each customer. Our model creates several customer groups to mirror the structure of the target data set. The effectiveness of our proposal is confirmed by a comparison involving a real e-commerce transaction dataset and an artificial dataset; it generally achieves higher predictive performance. In addition, we show that preselecting the actual number of customer groups does not always lead to higher predictive performance.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—Data Mining; G.3 [Mathematics of Computing]: Probability and Statistics—Survival Analysis, Probabilistic Algorithms

Keywords
CRM, Model choice, Non-parametric Bayes, MCMC

1. INTRODUCTION
The concepts of customer relationship management (CRM) have been recently gaining wide attention in business and academia [1][2]. This approach focuses on allocating resources to support business activities in order to gain a competitive advantage. CRM focuses on managing the relationship between a company and its current and prospective customers. A good relationship with the customer leads to higher customer value.

Estimating customer lifetime and the impact of the customer’s characteristic variables on profitable lifetime is an important goal in CRM marketing. Historically, survival analysis has usually been carried out by applying statistical models. In the non-contractual marketing domain, we cannot observe the customer’s dropout, e.g., e-commerce site, brick-and-mortar shop and free web service, customers can halt their flow of transactions with no explicit notification of their dropout.

This problem was first recognized by Schmittlein, Morrison, and Colombo [3]. They proposed ParetoNBD model; it estimates profitable lifetime using Recency-Frequency(RF) data. RF data includes purchase frequency, day of the first purchase, and day of the last purchase. This model is attracting the attention of researchers and practitioners because of its increasing importance in new types of marketing, such as CRM, and One-to-One Marketing. Their work is highly regarded and follow up research has been conducted [4][5][6][7][8]. Abe [8] proposed a hierarchical Bayes extension to the Pareto/NBD model to estimate the impact of the customer’s characteristic variables on profitable lifetime duration. The hierarchical Bayes model(HB model), whose lifetime parameter is a function of customer characteristics, can achieve this in one step.

HB model assumes a single functional relationship between lifetime parameters and the customer characteristics; so a single set of coefficients (impact of the customer’s characteristic variables) is estimated using data from all customers in the sample. These coefficients are effective in supporting marketing decisions for average, but they do not...
support customized marketing decisions for individual customer. This is because the HB model pools the impact over all customers. The most recent trend in CRM is for personalized actions, e.g., promotions or recommendations, for each customer, so estimating the impact of the customer’s characteristic variables one by one is an important goal. It permits the identification of the most effective customers in terms of increasing lifetime. For example, Abe [8] discovered the marketing knowledge that keeping a food corner fully stocked is effective in decreasing the dropout rate for the store. Unfortunately, this knowledge cannot identify the customers who could be prevented from dropping out through promotion of the food corner.

A simple solution is a model that includes as many coefficients as their customers; each customer has one coefficient. However, this raises the identification problem (degree of freedom problem) in estimation. This is because a set of customer parameters and a set of customers’ characteristic variables are needed to estimate a set of customer coefficients.

To overcome this problem, we propose a new model that can estimate the impact of the customer’s characteristic variables one by one using a non-parametric Bayes paradigm. Our model is based on an HB model that includes a multiple coefficient to which each customer belongs probabilistically.

The key feature of this model is the mathematical presentation of a dynamic coefficient distribution; it is based on the Dirichlet Process Mixture (DPM). DPM is a non-parametric Bayes model that can estimate both coefficients (parameters) and the number of coefficients, in a natural Bayesian paradigm. Accordingly, this model can provide dynamic coefficients without prior setting of the number of coefficients or coefficient parameters. In other words, our model makes the following assumptions: Finite coefficients are sampled from a potentially infinite set of unobserved coefficients, and each customer can be assigned to each coefficient probabilistically as in soft clustering.

Our model has 2 merits at the practical level, and these merits are the innovations of our research. At first, our model offers greater accuracy with trustful prior distributions using multiple coefficients. Secondly, our model can discern the impact of the customer’s characteristic variables for each customer.

The effectiveness of our proposal is confirmed by a comparison involving a real E-commerce transaction dataset and an artificial dataset.

The next section introduces related works on survival analysis and mixture distribution analysis. Section 3 describes the proposed model and compares it against the conventional model. Section 4 explains how our model uses the MCMC method for making the estimations. Section 5 presents experimental result conducted on three datasets of various types; the model’s performance is compared to that of the conventional model. Section 6 presents empirical analyses conducted on real e-commerce dataset. Section 7 presents the discussions followed by the conclusions in Section 8.

2. RELATED WORKS

Schmittlein et al [7] calibrate a Pareto/NBD model separately for each segment specified by the SIC code. The proposed model, by including segmentation variables in a hierarchical manner, allows estimation of all segments simultaneously, thereby increasing the degrees of freedom. The model can also incorporate non-nominal explanatory variables.

Reinartz et al [9] and Abe [8] proposed new models for estimating lifetime and the impact of the customer’s characteristic variables. Reinartz and Kumar proposed a 2-step model, paretoNBD for lifetime and regression to discover the impact of the customer’s characteristic variables. Abe proposed a hierarchical Bayes model and MCMC estimation which sets up the customer’s characteristic variables as a prior distribution. Both approaches can provide lifetime and impact, but the latter approach, whose dropout parameter is a function of customer characteristics, can achieve these in one step, thus providing correct error assessments for statistical inference. Accordingly, our model is based on the hierarchical Bayes approach.

Conventional models require that the number of components be given in preliminary step. (e.g., Schmittlein [7] sets the segmentation number in a preliminary step, while Reinartz et al [9] and Abe [8] set the number of components to 1.)

Our model sets the number of components as unknown, and each customer is assigned to each component probabilistically. This means model complexity (the number of component) must be estimated from the given data. Estimation of K, the number of components, is a special kind of model choice problem, for which there is a number of possible solutions [13]:

Approach 1 The number of components is decided after parameter estimation, using entropy distance or Kullback-Leibler (KL) divergence [14].

Approach 2 The number of components is estimated as a parameter by the DPM average over all possible set of mixture cardinalities according to a particular prior.

We focus on the latter, because it exemplifies more naturally the Bayesian paradigm and offers a much wider scope for inference, including model averaging in the non-parametric approach to mix estimation.

Approach 1 above pertains strongly to the testing perspective, the entropy distance approach being based on the KL divergence between a K component mixture and its projection on the set of K−1 mixtures, in the same spirit as Dupuis and Robert [16]. Additionally, this solution can not provide correct error assessments for statistical inference in practical tasks.

Our approach, the non-parametric extension of the HB model, can identify the impact of the customer’s characteristic variables. If the number of components is always estimated to be 1, HB and our model are equivalent.
3. CONVENTIONAL MODELS AND OUR MODEL

3.1 Model Assumptions

This section describes the assumptions of the conventional HB model (Conventional hereafter) and the proposed model (Proposed). Conventional has 3 assumptions as follows.

Assumption 1 Poisson purchases. While active, each customer makes purchases according to a Poisson process with rate $\lambda$.

Assumption 2 Exponential life time. Each customer remains active for an exponential distributed distribution (lifetime) with dropout rate $\mu$.

Assumption 3 Individuals’ purchase rate $\lambda$ and dropout rate $\mu$ follow a multivariate lognormal distribution.

Assumptions 1 and 2 are identical to the behavioral assumptions of Pareto/NBD model, and their validity has been confirmed by other researchers. Assumption 3 is the additional assumption of HB model. Proposed replaces assumption 3 as follows.

Assumption 3a Individuals’ purchase rate $\lambda$ and dropout rate $\mu$ follow a mixture of multivariate lognormal distributions.

To determine the impact of the customer’s characteristic variables one by one, the model needs to set groups that use different parameters for their lognormal distributions. Unlike Conventional, Proposed sets multiple customer groups, and each group has a different lognormal distribution. In Proposed, a customer belongs to all groups probabilistically (like soft clustering).

3.2 Mathematical Notations

Following past research [7], Figure 1 depicts our notations of recency and frequency data $(x, t, T)$. Lifetime starts at time 0 when the first transaction occurs, and/or the membership starts, and customer transactions are monitored until time $T$. $x$ is the number of repeat transactions observed in time period $(0, T)$, with the last purchase (rth repeat) occurring at $t_r$. Hence, recency is defined as $T - t_r$. $\tau$ is an unserved customer lifetime. Using these mathematical notations, the preceding model assumptions can be expressed as follows:

$$P(x|\lambda) = \begin{cases} \sum_{r=0}^{\infty} \frac{(\lambda T)^r e^{-\lambda T}}{r!} & (\tau > T) \\ \sum_{r=0}^{\infty} \frac{(\lambda x)^r e^{-\lambda x}}{r!} & (\tau \leq T) \end{cases}$$

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

$$\left[ \frac{\log(\lambda)}{\log(\mu)} \right] \sim MNV \left( \theta_0 = \begin{bmatrix} \theta_\lambda \\ \theta_\mu \end{bmatrix}, \Gamma_0 = \begin{bmatrix} \sigma^2_\lambda & \sigma^2_\mu \\ \sigma^2_\mu & \sigma^2_\mu \end{bmatrix} \right)$$

$\text{MVN}$ denotes a multivariate normal distribution. $\lambda$, the Poisson distribution parameter, is the purchase frequency per period while the customer is active. $\mu$, the parameter of exponential distribution, is the dropout rate.

A model that link purchase rates and dropout rates, $\lambda$ and $\mu$, to customer characteristic can offer insights into the frequency of transactions and increasing the lifetime. The approach of Conventional is to use the logarithm of $\lambda$ and $\mu$ as a linear regression as follows. Where index $i$ is added to emphasize that the rate parameters are for customer $i$.

$$\left[ \frac{\log(\lambda_i)}{\log(\mu_i)} \right] = \theta_i = \beta_i d_i + \epsilon$$

where $\epsilon \sim MNV(0, \Gamma_0)$

$d_i$ is a $G \times 1$ column vector that contains $G$ characteristics of customer $i$. $\beta$ is a $G \times 2$ parameter vector and $\epsilon$ is a $2 \times 1$ error vector that is normally distributed with mean 0 and variance $\Gamma_0$. This formulation replaces $\theta_0$ in the previous section with $\beta d_i$. When $d_i$ contains only a single element of 1 (i.e., an intercept only), this model reduces to the previous no-covariate case.

Proposed, a customer belongs to all groups probabilistically (like soft clustering).

$$P(h_i = k|\lambda, \mu) \propto \sum_{n(k)} \frac{n(k)}{N + n} P(\lambda_i, \mu_i|h_i = k) \quad (\text{if } k \neq \text{new})$$

$$P(\lambda_i, \mu_i|h_i = k) \sim MNV(\lambda_i, \mu_i|\beta d_i, \Gamma_0)$$

We set a unique distribution for $P(\lambda, \mu|h_i = \text{new})$. 

\[ \sigma^2_\lambda = \sigma^2_\mu = \sigma^2_\mu \]
The likelihood function for RF data \((x, t_x, T)\) is given by
the following simple expression. \(z\) is an indicator function
defined as 1 if a customer is active at time \(T\) and 0 otherwise.
Another latent variable is a dropout time, \(y_i\), when \(z = 0\).
See Abe 2009 for details.

\[
L(x, t_x, T|\lambda, \mu, z, y) = \frac{\lambda^y e^{-\lambda t_x}}{\Gamma(y)} \mu^{1-y} e^{-(\lambda + \mu)(zT + (1-z)y)} 
\tag{8}
\]

Because we observe neither \(z\) nor \(y\), we treat them as missing data and apply a data augmentation technique [17].
To simulate \(z\) in our MCMC estimation procedure, we can use the following expression as the probability of a customer
being active at \(T\).

\[
P(z = 1|\lambda, \mu, t_x, T) = \frac{1}{1 + (\mu/\lambda + \mu)(e^{(\lambda + \mu)(T-t_x)} - 1)} 
\tag{9}
\]

4. PROPOSED MODEL

4.1 Parameter Estimation with MCMC

We are now in a position to estimate parameters,
\(\theta_i, y_i, z_i, h_i, \forall i; \beta_h, \forall k; \Gamma_0, \nu_0, K\), by the MCMC method. To estimate the joint destiny, we sequentially generate each parameter,
given the remaining parameters, from its conditional distribution until convergence is achieved. The procedure is described below.

Step 1 Set initial value for \(\theta_i; \forall i\).

Step 2 Sample \(z_i\) according to Equation 9, for each \(i\).

Step 3 Sample \(y_i\), using truncated exponential distribution
\((t < y < T)\) for each \(i\), if \(z_i = 0\).

Step 4 Sample \(\theta_i\) with independent MH algorithm using
likelihood Equation 8, for each \(i\).

Step 5 Sample \(h_i\) and \(K\) using Equation 6 and 7 according
to DPM, for each \(i\).

Step 6 Sample \(\beta_k, \Gamma_0\) with multivariate normal regression
update.

Step 7 Iterate Step 2-Step 6 until convergence is achieved.

Each step is explained below.
Steps 2 and 3 generate \(z\) and \(y\) which are needed by Equation 8 in step 4.
In step 4, given \(z_i\) and \(y_i\), Equation 8 is used to generate
\(\lambda_i\) and \(\mu_i\), which are transformed into \(\theta_i\) by taking their
logarithms. An independent Metropolis-Hasting algorithm
is used to generate \(\lambda_i\) first then \(\mu_i\); the proposed distribution
is lognormal. Unlike Conventional, Proposed uses \(\beta_i\) of prior
distribution for each \(i\). \(\beta_{hi}\) is used for customer \(i\).

Step 5 is the additional model choice step of Proposed. Equations 6 and 7 are used to generate \(h_i\) and \(K\) with CRP.
In the first MCMC cycle, none of the \(h_i\) of customers are decided \((h_i = 0; \forall i)\), so equation 7 is not used. The number of
clusters is optimized for each CRP step using the given data.
As for step 6, see Bayesian textbooks elsewhere for details on
multivariate normal regression update [19] [20] [21]. The hyper
parameters of this lognormal distribution, \(\beta\) and \(\Gamma_0\),

5. EVALUATIONS FOR ARTIFICIAL DATA

5.1 Experimental Setup

Proposed was evaluated in terms of the qualitative effectiveness
of non-parametric coefficient division and compared to
conventional methods.
Evaluations that use artificial purchase data have 2 merits.
First, the artificial data can be generated to cover various situations and parameters, such as number of components(\(\beta\))
and degree of noise. Datasets used in the experiments are shown in "About Data". Second, artificial data can set unobservable parameters, i.e. \(\lambda_i, \mu_i, \beta_{hi}\), and activeness at the end of calibration. A real dataset was also tested, results are
in the next section, but it can not provide wide parameter coverage.

The MCMC steps were repeated for 15,000 iterations, of
which the last 5,000 were used to infer the posterior distribution
of the parameters. Convergence was monitored visually and checked with the Geweke test [18].

5.2 Evaluation metrics

The results of Proposed are compared with Conventional
and K-given model. K-given model has the number of components, \(K\),
fixed in advance. It assumes Dirichlet distributions for the prior distribution with which customer is
associated with a component. The appendix provides details.

Accordingly, Conventional and K-given model are equivalent if \( K = 1 \), and Proposed and K-given model are equivalent if the number of distributions is actually \( K \) (if \( K \) is estimated to be 1, Proposed and Conventional are also equivalent).

Proposed was compared against the benchmark models in terms of fit in the calibration period and prediction in the validation period. As disaggregate performance measures, correlation and mean squared errors (MSE) between predicted and actual, \( \lambda \), \( \mu \), and each coefficient for individual customers were used. Additionally, loglikelihood of being active at the end of calibration was compared as permitted by the artificial data.

### 5.3 About Data

The experiment used 4 types of artificial purchase data. Each dataset had 100 customers and was designed to exhibit 2 types of impact from characteristic variables. Dataset 1 has less white noise for \( \lambda \) and \( \mu \), and a mixture of 2 types of customers. 50 customers have \( h_i = 1, \beta_1 = (1, -1) \), and the other 50 customers have \( h_i = 2, \beta_2 = (-1, 1) \). White noise \( \sigma_\lambda^2 \) and \( \sigma_\mu^2 \) are 0.1. Dataset 2 has stronger white noise for \( \lambda \) and \( \mu \), and a mixture of 2 types of customers. \( \beta \) takes the same value as dataset 1. White noise \( \sigma_\lambda \) and \( \sigma_\mu \) are 1. Dataset 3 has stronger white noise for \( \lambda \) and \( \mu \), and a mixture of 4 types of customers. 50 customers have \( h_i = 1, \beta_1 = (1, -1) \), 50 customers have \( h_i = 2, \beta_2 = (1, -4) \). 50 customers have \( h_i = 3, \beta_3 = (-1, -1, 1, 0) \), and 50 customers have \( h_i = 4, \beta_4 = (-1, -4, 1, 0) \). 50 customers have \( h_i = 5, \beta_5 = (1, 0; -1, -1) \), and 50 customers have \( h_i = 6, \beta_6 = (1, 0; 1, -1, -4) \). 50 customers have \( h_i = 7, \beta_7 = (-1, 0, 1, -1) \), and 50 customers have \( h_i = 8, \beta_8 = (-1, 0; 1, -4) \). White noise \( \sigma_\lambda \) and \( \sigma_\mu \) are 1. Every dataset includes just a single type of \( d_i \), to simplify the division problem; \( \lambda \) and \( \mu \) are independent.

Datasets 1 and 2 have the same number of components, but different white noise variance, \( \Gamma_0 \). We can evaluate the models in terms of their effectiveness and robustness for \( \Gamma_0 \). Dataset 3 and 4 have a more complex tasks, since dataset 3 includes 4 components, and dataset 4 include 8 components in 2 dimensions for \( d_i \). We can evaluate model effectiveness on these complex cases.

Artificial transaction datasets 1-4 were produced by the following steps. \( (t_x, T) \) are natural numbers.

**Step 1** Set single \( d_i \) with normal distributions \(^2\) for each customer.

**Step 2** Set \( \lambda \) and \( \mu \) with \( \beta_\lambda, \beta_\mu, d_i + e \). \( e \) represents white noise.

**Step 3** Set \( T \) with unique distribution \(^3\) for each customer.

**Step 4** Set \( \tau \) with exponential distributions using parameter \( \mu \). If \( \tau > T \), the customer is alive.

\(^2\) means are 3, and covariance of 1

\(^3\) we set \( 1 < T < 365 \) for datasets 1, 2 and 4, \( 182 < T < 365 \) for dataset 3, because dataset 3 was designed to exhibit frequency dropout.

**Figure 4:** Distributions of Dataset 1

**Figure 5:** Distributions of Dataset 2

**Figure 6:** Distributions of Dataset 3 (lambda)

**Figure 7:** Distributions of Dataset 3 (mu)

**Step 5** Set \( x \) and \( t_x \) with active period \( \min(\tau, T) \) for Poisson random transactions. \( x \) is the summation of transaction number, and \( t_x \) is the period from first transaction to the last transaction.

Table 1 shows data details.

Figure 4, 5, 6 and 7 plot data distributions between \( d \) and \( \theta \).

Artificial data are unrealistic values, for example, trans-
action $x$ is repeated over one thousand times. These experiments are intended to evaluate unobserved parameters (coefficient, lifetime, $\lambda$, $\mu$).

5.4 Result

Table 2 lists the results for datasets 1-4. Mark (●) means actual number of components in K-given model. Mark (*) means the highest score. In the results for datasets 1 and 2, HB, K-given (who has actual number of components), and Proposed achieved higher evaluation scores in every criteria, MSE, correlation and loglikelihood. Looking at the results in more detail, K-given and Proposed are effective in terms of evaluating $\lambda$ (2 different coefficients). On the other hand, Conventional is effective with regard to $\mu$ evaluation. Comparing datasets 1 and 2, K-given model shows lower criteria scores, so we can confirm the robustness of Proposed in terms of $\Gamma_0$.

Figure 8, 9, 10 and 11 are histograms of the number of components in MCMC as estimated by Proposed. Mark (*) means actual number of components. Proposed estimated that 2 components were predominate in dataset 1, and that 40% of the MCMC cycle were covered by 3 components (1 additional component) in dataset 2. As white noise $\Gamma_0$ strengthened, Proposed created an additional temporary coefficient for robustness.

In the results for dataset 3, K-given model (K=2) and Proposed showed higher evaluation scores in every criteria, MSE, correlation and loglikelihood. The more components the dataset included, the lower was the criteria score achieved by Conventional. It is natural to consider that K-given model is a generalization of Conventional (HB model equals K-given model when K=1). Focusing on the actual number of components (K=4) K-given model, achieved lower criteria scores than when the false number of component (K=2) was used. This result indicates that an actual data generative model does not always fit the data if the model uses multi-stage estimation like a survival model. For example, it is difficult to estimate $\lambda$ and $\mu$ of one shot customers ($x=0$). Survival models may perceive them to have large $\mu$ or small $\lambda$. The actual number of components models often fail in coefficient estimation because of their $\lambda$ and $\mu$ value estimations. On the other hand, K=2, not the actual number of components, allowed the model to achieve higher prediction performance. This results from making the same temporary group of customers with fuzzy $\lambda$ and $\mu$ values. The same problem of components can be seen in the results for dataset 4.

The discussion section addresses this problem.

6. EVALUATIONS FOR REAL DATA

6.1 Experimental Setup

The real database contained e-commerce transactions captured over a 162 week period (26/09/09—02/11/12) at commercial website. It includes data gathered by random sampling 3,000 customers who purchased at least one item from the website during the first 81 weeks. The first 81 weeks of data were used for model calibration and the second 81 weeks of data were used for model validation.

Proposed was compared against Conventional and K-given (K=2,4,8) in both the calibration period and the validation period (prediction performance). We used as aggregate performance measures, correlation and mean squared errors

<table>
<thead>
<tr>
<th>Table 1: Details of Artificial Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>ave.</td>
</tr>
<tr>
<td>Dataset 1</td>
</tr>
<tr>
<td>Dataset 2</td>
</tr>
<tr>
<td>Dataset 3</td>
</tr>
<tr>
<td>Dataset 4</td>
</tr>
</tbody>
</table>

Figure 8: Distributions of the Number of Component in Dataset 1

Figure 9: Distributions of the Number of Component in Dataset 2

Figure 10: Distributions of the Number of Component in Dataset 3
Table 2: Artificial Data Results

<table>
<thead>
<tr>
<th></th>
<th>Dataset 1</th>
<th></th>
<th>Dataset 2</th>
<th></th>
<th>Dataset 3</th>
<th></th>
<th>Dataset 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>log(λ)</td>
<td>log(μ)</td>
<td>coef. for λ</td>
<td>coef. for μ</td>
<td>log(λ)</td>
<td>log(μ)</td>
<td>log(λ)</td>
</tr>
<tr>
<td>Conventional</td>
<td>1.93</td>
<td>0.32*</td>
<td>1.01</td>
<td>0.06</td>
<td>0.95</td>
<td>0.98*</td>
<td>-19.23</td>
<td></td>
</tr>
<tr>
<td>K-given model(K=2)</td>
<td>1.49</td>
<td>0.76</td>
<td>0.24</td>
<td>0.03*</td>
<td>0.97*</td>
<td>0.93</td>
<td>-17.26*</td>
<td></td>
</tr>
<tr>
<td>K-given model(K=4)</td>
<td>76.63</td>
<td>20.41</td>
<td>2170.13</td>
<td>3800.56</td>
<td>0.38</td>
<td>0.17</td>
<td>-22.96</td>
<td></td>
</tr>
<tr>
<td>K-given model(K=8)</td>
<td>450.39</td>
<td>383.03</td>
<td>1607.22</td>
<td>1275.50</td>
<td>0.29</td>
<td>0.41</td>
<td>-302.76</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>1.44*</td>
<td>0.74</td>
<td>0.224*</td>
<td>0.03*</td>
<td>0.97*</td>
<td>0.93</td>
<td>-17.72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conventional</td>
<td>3.98</td>
<td>1.15*</td>
<td>1.43</td>
<td>0.21</td>
<td>0.85</td>
<td>0.91*</td>
<td>-19.42</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=2)</td>
<td>3.45</td>
<td>1.24</td>
<td>0.86</td>
<td>0.10</td>
<td>0.87</td>
<td>0.86</td>
<td>-20.26</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=4)</td>
<td>107.19</td>
<td>41.35</td>
<td>25.79</td>
<td>12.77</td>
<td>0.40</td>
<td>0.27</td>
<td>-158.31</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=8)</td>
<td>157.58</td>
<td>32.83</td>
<td>46.22</td>
<td>38.34</td>
<td>0.47</td>
<td>0.15</td>
<td>-266.95</td>
</tr>
<tr>
<td>Proposed</td>
<td>3.01*</td>
<td>1.37</td>
<td>0.65*</td>
<td>0.01*</td>
<td>0.89*</td>
<td>0.85</td>
<td>-19.31*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conventional</td>
<td>2.73</td>
<td>61.17</td>
<td>1.00</td>
<td>3.33</td>
<td>0.94*</td>
<td>0.71</td>
<td>-336.70</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=2)</td>
<td>2.77</td>
<td>35.07*</td>
<td>0.75</td>
<td>2.11*</td>
<td>0.91</td>
<td>0.76*</td>
<td>-326.41*</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=4)</td>
<td>238.48</td>
<td>57.62</td>
<td>52.51</td>
<td>5.09</td>
<td>0.44</td>
<td>0.53</td>
<td>-605.41</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=8)</td>
<td>190.34</td>
<td>137.56</td>
<td>31.49</td>
<td>48.44</td>
<td>0.44</td>
<td>0.35</td>
<td>-784.34</td>
</tr>
<tr>
<td>Proposed</td>
<td>2.04*</td>
<td>42.34</td>
<td>0.47*</td>
<td>2.34</td>
<td>0.94*</td>
<td>0.67</td>
<td>-329.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conventional</td>
<td>3.35</td>
<td>52.23</td>
<td>1.01</td>
<td>3.02</td>
<td>0.77</td>
<td>0.56</td>
<td>-2120.96</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=2)</td>
<td>3.15*</td>
<td>52.52</td>
<td>0.54*</td>
<td>3.02</td>
<td>0.78*</td>
<td>0.53</td>
<td>-2446.6</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=4)</td>
<td>30.88</td>
<td>115.73</td>
<td>4.23</td>
<td>4.17</td>
<td>0.38</td>
<td>0.24</td>
<td>-2463.8</td>
</tr>
<tr>
<td></td>
<td>K-given model(K=8)</td>
<td>115.02</td>
<td>140.04</td>
<td>10.54</td>
<td>15.60</td>
<td>0.30</td>
<td>0.17</td>
<td>-2907.7</td>
</tr>
<tr>
<td>Proposed</td>
<td>22.05</td>
<td>45.47*</td>
<td>1.91</td>
<td>2.89*</td>
<td>0.46</td>
<td>0.79*</td>
<td>-2009.01*</td>
<td></td>
</tr>
</tbody>
</table>

![Figure 11: Distributions of the Number of Component in Dataset 4](image)

(MSE) between predicted and observed numbers of transactions for individual customers. As an aggregate measure, we used root mean square (RMS) between predicted and observed weekly cumulative transactions. These measures are generally used [8] in the evaluation of non-contractual survival model. Characteristics of customer \(d\) include 4 types of value, amount of discount price, e-mail membership, average price of transaction and interest. For evaluating the effect of \(d\), every model was compared to themselves with no characteristic except intercept.

6.2 About Data
Table 3 shows data details.

6.3 Result
Figure 12 shows the distribution of correlation between \(\log(\lambda)\) and \(\log(\mu)\). The average of correlations is 0.131. Table 4 shows the results for the real dataset. Conventional and Proposed yielded higher evaluation scores for the disaggregate performance measures. For aggregate measures, Conventional, K-given (\(K=4\)), and Proposed showed high and roughly equivalent evaluation scores. Conventional and Proposed have almost the same prediction performance, however, Proposed performs better at aggregate tracking. This can be seen from the time-series tracking of the cumulative number of transactions in Figure 13. The line at week 81 separates the validation from the calibration period. On the other hand, Conventional performs better at disaggregate tracking.

Compared to their dummy characteristic model, they offered superior prediction performance as confirmed by the RMS values. The result suggest the effectiveness of the chosen characteristics. In particular, K-given (\(K=4\)) and Proposed achieved greater performance than Conventional when they use the characteristics of customer \(d\) as a prior distribution.

Furthermore, Proposed can extract the impact of the customer's characteristic variables. They can be used for extracting a list of customer for which discounts are effective for extending customer lifetime or E-mail is effective for increasing transaction number.

7. DISCUSSION
We discuss the difficulty of deciding \(K\) and emphasize the importance of our proposal; the proposed model determines the number of components from the target dataset. The results for artificial dataset 3 show that the K-given model achieved a lower criteria score for the actual number of components (\(K=4\)) than for an erroneous number of compo-
Table 3: Details of Real Dataset

<table>
<thead>
<tr>
<th>frequency (x)</th>
<th>period (t_1)</th>
<th>period (T)</th>
<th>customers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ave.</td>
<td>max.</td>
<td>min.</td>
<td>ave.</td>
</tr>
<tr>
<td>2.9</td>
<td>55</td>
<td>0</td>
<td>237.4</td>
</tr>
</tbody>
</table>

Table 4: EC Data Result

<table>
<thead>
<tr>
<th></th>
<th>Disaggregate</th>
<th>Aggregate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSEL</td>
<td>Correlation</td>
</tr>
<tr>
<td></td>
<td>Calibration</td>
<td>Validation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conventional</td>
<td>2.92</td>
<td>14.63</td>
</tr>
<tr>
<td>K-given model(K=2)</td>
<td>2.51</td>
<td>16.05</td>
</tr>
<tr>
<td>K-given model(K=4)</td>
<td>4.45</td>
<td>16.31</td>
</tr>
<tr>
<td>K-given model(K=8)</td>
<td>3.06</td>
<td>15.78</td>
</tr>
<tr>
<td>Proposed</td>
<td>2.41*</td>
<td>15.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conventional</td>
<td>2.68</td>
<td>14.99</td>
</tr>
<tr>
<td>K-given model(K=2)</td>
<td>2.52</td>
<td>15.69</td>
</tr>
<tr>
<td>K-given model(K=4)</td>
<td>2.41</td>
<td>15.97</td>
</tr>
<tr>
<td>K-given model(K=8)</td>
<td>2.40</td>
<td>16.09</td>
</tr>
<tr>
<td>Proposed</td>
<td>2.55</td>
<td>15.49</td>
</tr>
</tbody>
</table>

Figure 12: Distribution of Correlation Between \( \log(\lambda) \) and \( \log(\mu) \) for EC Data

Figure 13: Weekly Time-series Tracking Plot for EC Data

In this paper, we introduced the goals of discovering customer lifetime and the impact of the customer’s characteristic variables on lifetime duration for each customer. We developed a nonparametric mixture model to achieve both goals. We extended the assumption of the conventional model, logarithm of \( \lambda \) and \( \mu \) with linear regression, by the additional assumption of DPM of regression. It assesses the structure of the target data set and determines the number of groups that yield high prediction accuracy automatically.

Experiments on artificial datasets showed the effectiveness and robustness of our model, and the results for a real data set showed the superior prediction performance of our model.

Table 5: Results of Additional Experiment on Data 3

<table>
<thead>
<tr>
<th>Model</th>
<th>coef. for ( \lambda )</th>
<th>coef. for ( \mu )</th>
</tr>
</thead>
<tbody>
<tr>
<td>HB model</td>
<td>1.01</td>
<td>2.30</td>
</tr>
<tr>
<td>K-given model(K=2)</td>
<td>1.01</td>
<td>0.13</td>
</tr>
<tr>
<td>K-given model(K=4)</td>
<td>0.17</td>
<td>0.12</td>
</tr>
<tr>
<td>K-given model(K=8)</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.17</td>
<td>0.12</td>
</tr>
</tbody>
</table>

8. CONCLUSIONS

In this paper, we introduced the goals of discovering customer lifetime and the impact of the customer's characteristic variables on lifetime duration for each customer. We developed a nonparametric mixture model to achieve both goals. We extended the assumption of the conventional model, logarithm of \( \lambda \) and \( \mu \) with linear regression, by the additional assumption of DPM of regression. It assesses the structure of the target data set and determines the number of groups that yield high prediction accuracy automatically.

Experiments on artificial datasets showed the effectiveness and robustness of our model, and the results for a real data set showed the superior prediction performance of our model.
with chosen characteristics, over the conventional HB model and the parametric Bayes (K-given) model. Additionally, we showed that actual number of components given models do not always suit multiple stage estimation.

9. REFERENCES


APPENDIX

A. K-GIVEN MODEL

DPM is not the only way to determine the impact of the customer’s characteristic variables one by one. We introduce the K-given model as one such different other model. In the K-given model we can set the number of components, K from out of model. Our model decides K from the given data structure.

Each K-given model set has a different β corresponding to the group that the user belongs to. The model takes the Dirichlet distribution as a prior distribution for h_i. The equation is as follows. α is a vector including K of 1.

\[
\begin{bmatrix}
\log(\lambda_i) \\
\log(\mu_i)
\end{bmatrix} = \theta_i = \beta h_i d_i + e
\]

where \( e \sim MNV(0, \Gamma_0) \)

\[ P(h_i | \alpha) \sim DIR(\alpha) \]

\[ P(h_i = k | \lambda, \mu) \text{ is expressed as follows.} \]

\[ P(h_i = k | \lambda, \mu) \propto \frac{n(k) + 1}{N + K - 1} P(\lambda, \mu | h_i = k) \]

where \( k \in \{1, ..., K\} \)

The model estimates \( \theta_i, y_i, z_i, h_i, \forall i; \beta_k, \forall k; \Gamma \), using the MCMC method. MCMC step 5 in subsection 4.1 is changed as follows.

Step 5-a Sampling \( h_i \) and \( K \) with using Equation 12 and 13 according to Dirichlet distribution for each \( i \).