Nonparametric Hierarchal 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
ichikawa.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 Statistic—*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

[*] 1-1 Hikarinooka Yokosuka-Shi, Kanagawa 239-0847 Japan

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

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 inferencing. 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 inferencing, including model averaging in the non-parametric approach to mix estimation [1].

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.

[1] In addition, the unusual topology of the parameter space invalidates standard asymptotic approximations of testing procedures [15]

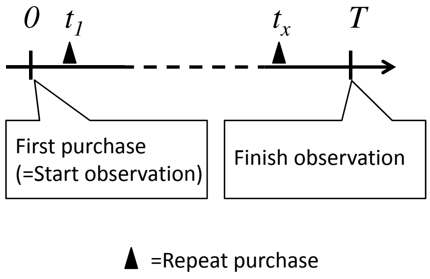

Figure 1: Notations for RF data

3. CONVENTIONAL MODELS AND OUR MODEL

3.1 Model Assumptions

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

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

Assumption 2 Exponential life time. Each customer remains active for an exponential distributed duration (lifetime) with dropout rate μ .

Assumption 3 Individuals' purchase rate λ and dropout rate μ 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 λ and dropout rate μ 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_x, 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 (xth repeat) occurring at t_x. Hence, recency is defined as $T - t_x$. τ is unobserved customer lifetime. Using these mathematical notations, the preceding model assumptions can be expressed as follows:

$$P(x|\lambda) = \begin{cases} \frac{(\lambda T)^x}{x!} e^{-\lambda T} & (\tau > T) \\ \frac{(\lambda \tau)^x}{x!} e^{-\lambda \tau} & (\tau \leq T) \end{cases} \tag{1}$$

$$f(\tau) = \mu e^{-\mu \tau} \tag{2}$$

$$\left[\begin{array}{c} log(\lambda) \\ log(\mu) \end{array} \right] \sim MNV\Big(\theta_0 = \left[\begin{array}{c} \theta_\lambda \\ \theta_\mu \end{array} \right], \Gamma_0 = \left[\begin{array}{cc} \sigma_\lambda^2 & \sigma_{\lambda\mu} \\ \sigma_{\mu\lambda} & \sigma_\mu^2 \end{array} \right]\Big) \tag{3}$$

MVN denotes a multivariate normal distribution. λ, the Poisson distribution parameter, is the purchase frequency per period while the customer is active. μ, the parameter of exponential distribution, is the dropout rate.

A model that link purchase rates and dropout rates, λ and μ, 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 λ and μ as a linear regression as follows. where index i is added to emphasize that the rate parameters are for customer i.

$$\left[\begin{array}{c} log(\lambda_i) \\ log(\mu_i) \end{array} \right] = \theta_i = \beta' d_i + e \tag{4}$$
$$\text{where } e \sim MNV(0, \Gamma_0)$$

d_i is a $G * 1$ column vector that contains G characteristics of customer i. β is a $G * 2$ parameter vector and e is a $2 * 1$ error vector that is normally distributed with mean 0 and variance Γ_0. This formulation replaces θ_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.

Note that Conventional, which uses a single β, can not determine the impact of the customer's characteristic variables one by one, therefore, Proposed sets multiple β for multiple customer groups. The model sets an infinite number of groups, and sets a different β to each group that the user belongs to. We define the Dirichlet process (DP) as a prior distribution for unknown β. Famous implementations of DP are the stick-breaking process and the Chinese restaurant process (CRP). Our model adopts the latter, because it readily suits the MCMC procedure. CRP handles a potentially infinite group mixture in theory, but it makes a number of groups according to the given data structure [12].

$h_i = c, i \in 1, ..., N, c \in 1, ..., C$ means customer i belongs to the cth group, and is assigned β_c. The equation is as follows.

$$\left[\begin{array}{c} log(\lambda_i) \\ log(\mu_i) \end{array} \right] = \theta_i = \beta'_{h_i} d_i + e \tag{5}$$
$$\text{where } e \sim MNV(0, \Gamma_0)$$
$$\text{where } h_i|\alpha \sim CRP(\alpha)$$

$P(h_i = k|\lambda, \mu)$ is expressed as follows; n is the number of customers, and $n(k)$ is number of customers for which $h_i = k$.

$$P(h_i = k|\lambda_i, \mu_i) \propto \begin{cases} \frac{n(k)}{N+\alpha-1} P(\lambda_i, \mu_i|h_i = k) & (\text{if} k \neq new) \\ \frac{\alpha}{N+\alpha-1} P(\lambda_i, \mu_i|h_i = k) & (\text{if} k = new) \end{cases} \tag{6}$$

$$P(\lambda_i, \mu_i|h_i = k) \sim MNV(\lambda_i, \mu_i|\beta_k d_i, \Gamma_0) \tag{7}$$

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

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, when $z = 0$. See Abe 2009 for details.

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

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.

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

4. PROPOSED MODEL

4.1 Parameter Estimation with MCMC

We are now in a position to estimate parameters, $\theta_i, y_i, z_i, h_i, \forall_i; \beta_k, \forall_k; \Gamma_0, K$, 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 θ_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 β_k, Γ_0 with multivariate normal regression update.

Step 7 Iterate Step2-Step6 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 λ_i and μ_i, which are transformed into θ_i by taking their logarithms. An independent Metropolis-Hasting algorithm is used to generate λ_i first then μ_i; the proposed distribution is lognormal. Unlike Conventional, Proposed uses β of prior distribution for each i. β_{h_i} 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, β and Γ_0, are estimated in a Bayesian manner with a multivariate normal prior and an inverse Wishart prior, respectively:

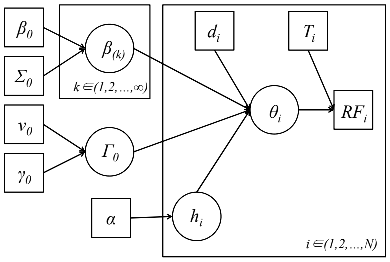

Figure 2: Graphical Model of Proposal

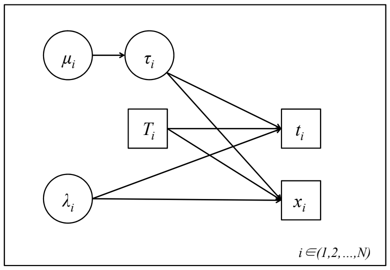

Figure 3: Graphical Model of RF Data Generation

$$\beta \sim MNV(\beta_0, \Sigma_0) \quad (10)$$

$$\Gamma_0 \sim IW(v_0, \gamma_0) \quad (11)$$

These distributions are standard in Baysian regression [19] [20] [21]. We set a non-informative prior distribution for the hyper-parameters.

Figure 2 show our graphical model. RF_i include x and t. Figure 3 show the detail of RF_i generation.

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(β) and degree of noise. Datasets used in the experiments are shown in "About Data". Second, artificial data can set unobservable parameters, i.e. λ_i, μ_i, β_{h_i}, 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, λ, μ, 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 λ and μ, and a mixture of 2 types of customers. 50 customers have $h_i = 1, \beta_1 = (1, -1)$, and the other 50 customers $h_i = 2, \beta_2 = (-1, 1)$. White noise σ^2_λ and σ^2_μ are 0.1. Dataset 2 has stronger white noise for λ and μ, and a mixture of 2 types of customers. β takes the same value as dataset 1. White noise σ_λ and σ_μ are 1. Dataset 3 has stronger white noise for λ and μ, 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 $h_i = 3, \beta_3 = (-1, -1)$, and 50 customers have $h_i = 4, \beta_4 = (-1, -4)$. white noise σ_λ and σ_μ are 1. Dataset 4 has stronger white noise for λ and μ, and a mixture of 8 types of customers; they has 2 independent dimensions in d_i. 50 customers have $h_i = 1, \beta_1 = (1, -1; -1, 0)$, 50 customers have $h_i = 2, \beta_2 = (1, -4; -1, 0)$. 50 customers have $h_i = 3, \beta_3 = (-1, -1; 1, 0)$, 50 customers have $h_i = 4, \beta_4 = (-1, -4; 1, 0)$, 50 customers have $h_i = 5, \beta_5 = (1, 0; -1, -1)$, 50 customers have $h_i = 6, \beta_6 = (1, 0; -1, -4)$, 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 σ_λ and σ_μ are 1. Every dataset includes just a single type of d_i, to simplify the division problem; λ and μ are independent.

Datasets 1 and 2 have the same number of components, but different white noise variance, Γ_0. We can evaluate the models in terms of their effectiveness and robustness for Γ_0. Dataset 3 and 4 has 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 λ and μ with $\beta'_{h_i} d_i + e$. e represents white noise.

Step 3 Set T with unique distribution [3] for each customer.

Step 4 Set τ with exponential distributions using parameter μ. If $\tau > T$, the customer is alive.

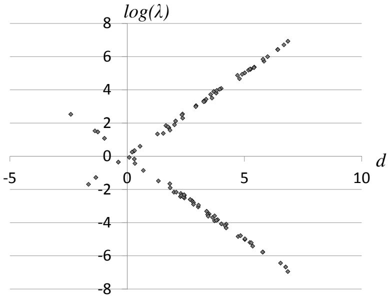

Figure 4: Distributions of Dataset 1

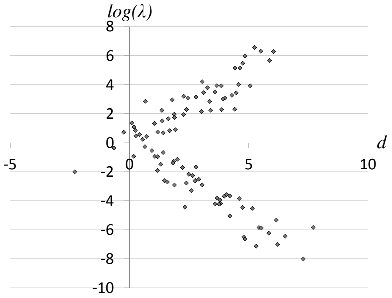

Figure 5: Distributions of Dataset 2

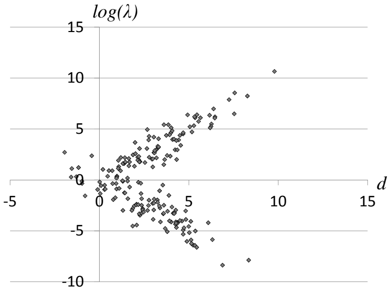

Figure 6: Distributions of Dataset 3 (lambda)

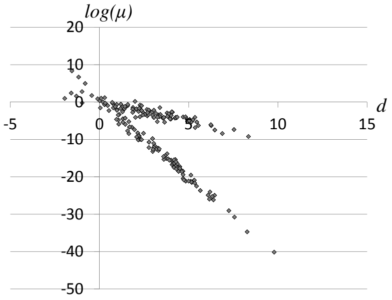

Figure 7: Distributions of Dataset 3 (mu)

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 θ.

Artificial data are unrealistic values, for example, trans-

[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.

Table 1: Details of Artificial Data Set

	frequency (x)			period (t_x)			period (T)			customers
	ave.	max.	min.	ave.	max.	min.	ave.	max.	min.	#
Dataset 1	9,280	240,884	0	26.9	354	0	189.7	359	2	100
Dataset 2	3,531	124,850	0	12.7	227	0	184.3	363	4	100
Dataset 3	70,486	8,731,537	0	79.3	360	0	271.6	364	183	200
Dataset 4	1,624	84,876	0	51.5	354	0	185.0	364	2	400

action x is repeated over one thousand times. These experiment are intended to evaluate unobserved parameters (coefficient, lifetime, λ, μ).

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 λ (2 different coefficients). On the other hand, Conventional is effective with regard to μ evaluation. Comparing datasets 1 and 2, K-given model shows lower criteria scores, so we can confirm the robustness of Proposed in terms of Γ_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 Γ_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 the λ and μ of one shot customers (x=0). Survival models may perceive them to have large μ or small λ, The actual number of components models often fail in coefficient estimation because of their λ and μ 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 λ and μ 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 disaggregate performance measures, correlation and mean squared errors

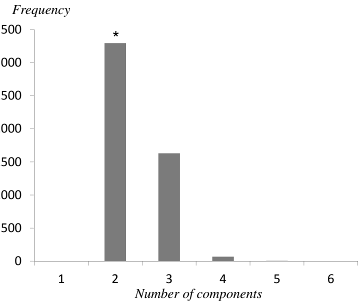

Figure 8: Distributions of the Number of Component in Dataset 1

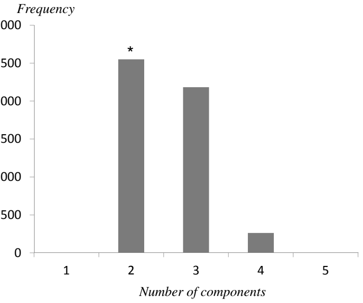

Figure 9: Distributions of the Number of Component in Dataset 2

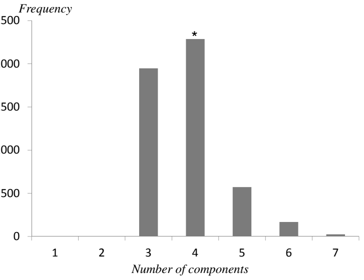

Figure 10: Distributions of the Number of Component in Dataset 3

Table 2: Artificial Data Results

	MSE				Correlation		loglikehood
	log(λ)	log(μ)	coef. for λ	coef. for μ	log(λ)	log(μ)	
Dataset 1							
Conventional	1.93	0.32*	1.01	0.06	0.95	0.98*	-19.23
K-given model(K=2 †)	1.49	0.76	0.24	0.03*	0.97*	0.93	-17.26*
K-given model(K=4)	76.63	20.41	2170.13	3800.56	0.38	0.17	-22.96
K-given model(K=8)	450.39	383.03	1607.22	1275.50	0.29	0.41	-302.76
Proposed	1.44*	0.74	0.224*	0.03*	0.97*	0.93	-17.72
Dataset 2							
Conventional	3.98	1.15*	1.43	0.21	0.85	0.91*	-19.42
K-given model(K=2 †)	3.45	1.24	0.86	0.10	0.87	0.86	-20.26
K-given model(K=4)	107.19	41.35	25.79	12.77	0.40	0.27	-158.31
K-given model(K=8)	157.58	32.83	46.22	38.34	0.47	0.15	-266.95
Proposed	3.01*	1.37	0.65*	0.01*	0.89*	0.85	-19.31*
Dataset 3							
Conventional	2.73	61.17	1.00	3.33	0.94*	0.71	-336.70
K-given model(K=2)	2.77	35.07*	0.75	2.11*	0.91	0.76*	-326.41*
K-given model(K=4 †)	238.48	57.62	52.51	5.66	0.44	0.53	-605.41
K-given model(K=8)	190.34	137.56	31.40	48.44	0.44	0.35	-784.34
Proposed	2.04*	42.34	0.43*	2.34	0.94*	0.67	-329.25
Dataset 4							
Conventional	3.35	52.23	1.01	3.02	0.77	0.56	-2120.96
K-given model(K=2)	3.18*	52.52	0.54*	3.02	0.78*	0.53	-2446.6
K-given model(K=4)	30.88	115.73	4.23	4.17	0.38	0.24	-2463.8
K-given model(K=8†)	115.02	140.04	10.54	15.60	0.30	0.17	-2907.7
Proposed	22.05	45.47*	1.91	2.89*	0.46	0.79*	-2009.01*

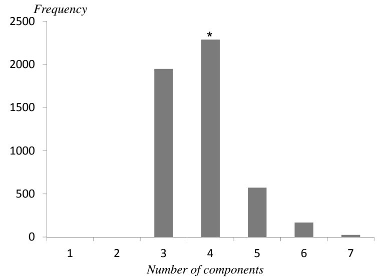

Figure 11: Distributions of the Number of Component in Dataset 4

(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 intercept. 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 show the distribution of correlation between $log(\lambda)$ and $log(\mu)$. The average of correlations is 0.131. Table 4 show 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

frequency (x)			period (t_x)			period (T)			customers
ave.	max.	min.	ave.	max.	min.	ave.	max.	min.	#
2.9	55	0	237.4	565	0	396.4	565	5	3000

Table 4: EC Data Result

	Disaggregate				Aggregate		
	MSE		Correlation		RMS		
	Calibration	Validation	Calibration	Validation	Calibration	Validation	Pooled
Conventional	2.92	14.63*	0.98*	0.65*	32.72	165.80	119.46
K-given model(K=2)	2.51	16.05	0.96	0.63	32.90	169.00	121.70
K-given model(K=4)	4.45	16.31	0.89	0.60	32.49*	165.64	119.35
K-given model(K=8)	3.06	15.67	0.94	0.63	32.71	168.96	120.57
Proposed	2.41*	15.78	0.96	0.64	32.68	164.80*	119.21*
Not using characteristic							
Conventional	2.68	14.99	0.98	0.65	32.80	169.10	121.80
K-given model(K=2)	2.52	15.69	0.97	0.63	32.85	170.10	122.50
K-given model(K=4)	2.41	15.97	0.97	0.63	32.86	170.04	122.46
K-given model(K=8)	2.40	16.09	0.96	0.63	32.90	168.96	122.65
Proposed	2.55	15.49	0.97	0.64	32.82	170.20	122.60

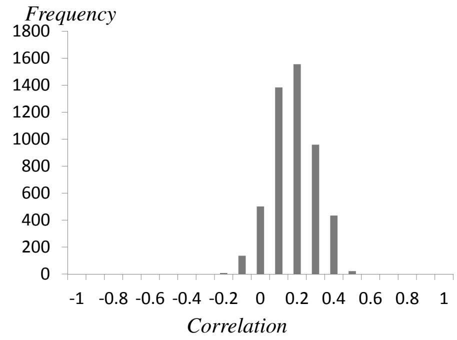

Figure 12: Distribution of Correlation Between $log(\lambda)$ and $log(\mu)$ for EC Data

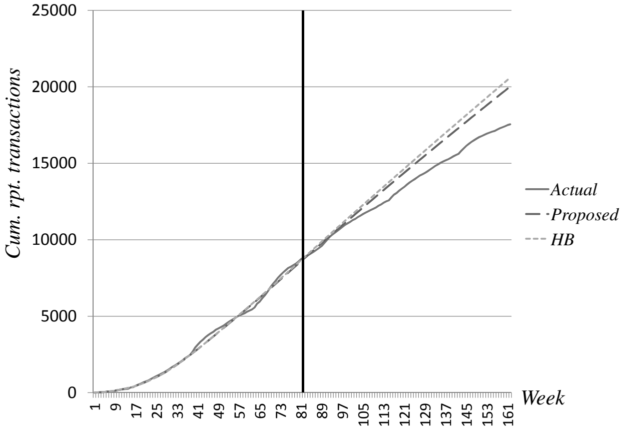

Figure 13: Weekly Time-series Tracking Plot for EC Data

nents(K=2). This raises the question of the effectiveness of multiple stage estimation.

Table 5 lists the results for dataset 3 for just β_k estimation. For this, actual λ and μ were input to MCMC, steps 5 and 6. The K-given model and Proposed achieved higher criteria scores for the actual number of components (K=4) in this additional experiment. This is different from the previous result. Thus only the proposed model could achieve high criteria scores in both cases, single and multiple estimation.

These results indicate that the survival model approach has difficulty in achieving high prediction performance without using DPM in multiple stage estimation, even if the user (researcher) has actual knowledge of customer group number.

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 λ and μ 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

	coef. for λ	coef. for μ
HB model	1.01	2.30
K-given model(K=2)	1.01	0.13
K-given model(K=4†)	0.17	0.12
K-given model(K=8)	0.18	0.12
Proposed	0.17	0.12

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

[1] Rust R. T. and T. S. Chung, Marketing models of service and relationships, Marketing Science, vol.25, no.6, pp.560–580, 2006.

[2] Sun B., Technology innovation and implications for customer relationship management, Marketing Science, vol.25, no.6, pp.594–597, 2006.

[3] Schmittlein D. C., D. G. Morrison, and R. Colombo , ”Counting your customers: Who are they and what will they do next?”, Management Science, vol.33 no.1, pp.1–24, 1987.

[4] Fader P. S., B. G. S. Hardie and K. L. Lee, Counting your customers the easy way: An alternative to the Pareto/NBD model, Marketing Science, vol.24, no.2, pp.275–284, 2005.

[5] Fader P. S., B. G. S. Hardie and K. L. Lee, RFM and CLV: Using iso-value curves for customer base analysis, Journal of Marketing Research, vol.42, no.4, pp.415–430, 2005.

[6] Reinartz W. J. and V. Kumar, On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing, Journal of Marketing, vol.64, no.4 pp.17–35, 2000.

[7] Schmittlein D. C. and R. A. Peterson, Customer base analysis: An industrial purchase process application, Marketing Science, vol.13, no.1, pp.41–67, 1994.

[8] Abe, M., ”Counting Your Customers” One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model, Marketing Science, vol.28, no.3, pp.541-553, 2009.

[9] Reinartz W. J., and V. Kumar, ”The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration”, Journal of Marketing, vol.67, no.1, pp.77–99, 2003.

[10] D. Blackwell and J. B. MacQueen, ”Ferguson distributions via Polya urn schemes”, The Annals of Statistics, vol.1, no.2, pp.353–355, 1973.

[11] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada and N. Ueda, ”Learning systems of concepts with an infinite relational model”, Proceedings of the 21st National Conference on Artificial Intelligence, 2006.

[12] J. Pitman, Combinatorial Stochastic Processes, Lecture Notes for St. Flour Summer School, Springer-Verlag, New York, 2002.

[13] Marin J. M., Mengersen K. and Robert C. P., Bayesian modelling and inference on mixtures of distributions, In Handbook of Statistics, vol.25, pp.459–507, Amsterdam, North-Holland, 2005.

[14] Mengersen K. and Robert C., Testing for mixtures: A Bayesian entropic approach (with discussion), Bayesian Statics 5, pp.255–276, Oxford. Oxford University Press, 1996.

[15] Lindsay B., Mixture Models: Theory, Geometry and applications, IMS Monographs, Hayward, CA, 1995.

[16] Dupuis J. and Robert C., Model choice in qualitative regression models, Journal of Statistical Planning and Inference, vol.111, pp.77–94, 2003.

[17] Tanner, M. A., W. H. Wong, The calculation of posterior distributions by data augmentation, theory and methods, Journal of the American Statistical Association, vol.82, no.398, pp.528–540, 1987.

[18] Geweke J., Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, Bayesian Statistics, vol.4, Oxford University Press, Oxford, pp.169–193, 1992.

[19] Congdon P., Bayesian Statistical Modelling, London, 2001.

[20] Gelman A., J. B. Carlin, H. S. Stern and D. B. Rubin, Bayesian Data Analysis. Chapman and Hall, Boca Raton, FL, 1995.

[21] Rossi P. E., G. Allenby and R. McCulloch, Bayesian Statistics and Marketing, John Wiley and Sons, London, 2005.

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.

$$\begin{bmatrix} log(\lambda_i) \\ log(\mu_i) \end{bmatrix} = \theta_i = \beta'_{h_i} d_i + e \qquad (12)$$

$$\text{where } e \sim MNV(0, \Gamma_0)$$

$$P(h_i|\alpha) \sim DIR(\alpha)$$

$P(h_i = k|\lambda, \mu)$ is expressed as follows.

$$P(h_i = k|\lambda, \mu) \propto \frac{n(k) + 1}{N + K - 1} P(\lambda, \mu|h_i = k) \qquad (13)$$

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

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.