A GENERALIZED COUNT MODEL ON CUSTOMERS' REPEAT PURCHASING

Hongyu Chen, College of Business Administration, California State University, Long Beach, Long Beach, CA 90840, 562-985-2509, hongyu.chen@csulb.edu
Ruixia Shi, Mihaylo College of Business and Economics, California State University, Fullerton, Fullerton, CA 92831, 657-278-2253, rshi@fullerton.edu
Suresh P. Sethi, Jindal School of Management, SM 30, University of Texas at Dallas, P.O. Box 830688, Richardson, TX 75083, 972-883-6245, sethi@utdallas.edu

ABSTRACT

A critique for the classic Beta-Binomial/Negative Binomial Distribution (BB/NBD) model on customers' repeat purchasing is that the Beta distribution does not accurately capture the heterogeneity among customers, and in consequence, affects the validity of the whole model. This is especially true when modelling customers' behavior for online business, where customers are invisible and highly diverse. We propose a new approach to model the diversity of customers' categorical spending. Specifically, we relax the Beta assumption in the traditional BB/NBD model to include a generalized distribution. The generalization is made possible through using the Gaussian quadrature. The extension fully captures customers' heterogeneity.

INTRODUCTION

Online retailers face a dilemma when studying their customers. From the positive side, customers leave digital footprints on every step of their purchasing processes (e.g.， their purchasing history, browsing history, wish lists, etc.), which is a source of fortune for retailers to study and analyze their customers. However, online retailers never meet their customers in person. They may be from the other side of the globe. It is a much more diverse population compared with the brick and mortar one.

The diversity among customers proposes a new challenge to researchers. Traditional models used to study customers' purchasing behavior may not be as effective in the online context. Take the problem of customers' category spending as an example, where a website provides a special category of products (e.g., shoes, clothing) and customers repeatedly visit the website and make purchases. At each visit, customers could select any brand they like. The Beta-Binomial/Negative Binomial Distribution (BB/NBD) model is probably the most well-known framework to study customers' repeat purchasing in this situation [5] [6] [10]. BB/NBD model assumes that a customer's categorical spending follows a Negative Binomial Distribution (NBD), while purchases of a focal brand are Binomially distributed. To capture customers' heterogeneity across population, the selection rate of the focal brand at each purchase follows a Beta distribution. Many previous studies have confirmed the validity of the BB/NBD model [4] [6]. Among them, the power of using NBD to model customers' categorical spending has been well tested and established [6] [9]. Dunn et al. [3, p. 256] states that "for most purposes in brand purchasing studies, the NBD tends to be accepted as robust to most observed departures from its (stationary) Poisson assumption".

However, the main reason for using Beta distribution to model customers' heterogeneity on selection rates is the mathematical convenience. Beta is the conjugate prior of Binomial distribution. As a result, the BB/NBD model is mathematically tractable and a lot of predictive results can be expressed in closed forms [4] [6] [11]. Meanwhile, Beta distribution is also quite flexible to capture customers' variations, and thus is well accepted in modeling customers' heterogeneity [2] [4] [6].

Using Beta distribution is amenable for analysis, but it has its limitation in modeling reality, especially with the advent of online retailers. In the online context, customers are invisible and highly diverse; the distributions for selection rates may have more than one mode. We may also observe spikes in the distributions. If these are the cases, Beta distributions cannot fully capture the heterogeneity among customers. Then, how to model the real purchasing behaviors of customers? And how could we quantify the difference? In an extreme situation, if the customers' heterogeneity is piecewise linear distributed, can we rectify the model?

In this paper, we propose a new approach to capture customers' heterogeneity using a general distribution. Specifically, we assume customers' selection rates may follow any arbitrary distribution, which is a big step forward on the original BB/NBD model, and can solve the heterogeneity problem effectively. However, in achieving this, we bring a general function form into our model, and the mathematical representation is no longer tractable. Therefore, we introduce the Gaussian quadrature to our model for the purpose of approximation.

Our key contribution is based on the following observation. In the BB/NBD framework, the distribution for customers' selection rates only shows up within the integral. In other words, the exact expression of the distribution does not matter. Only the integral of selection rates play an important role. We thus resort to the Gaussian quadrature to approximate the integral. Specifically, we express the whole integral of the selection rates in the BB/NBD model using the Gaussian quadrature. After introducing the Gaussian quadrature, the untraceable general distribution for selection rates is reduced to k discretized values. These discretized values could be treated as k unknown "parameters" of the distribution. We could estimate these parameters using maximum likelihood as we estimate the two parameters for Beta distribution in the BB/NBD framework. Since the Gaussian quadrature is accurate for all polynomials up to the degree of $2k-1$, the whole formula converges exponentially fast as k increases [7, p180]. As a result, we have a customer purchasing model which is general in nature as well as mathematically elegant.

The details of our approach are illustrated in the next session. We then conduct a simulation to validate the model and evaluate its performance. Concluding remarks are presented at the end of the paper.

THE MODEL

The canonic story behind our model is that a customer repeatedly visits a focus website to purchase some goods from a specific category (e.g., purchasing DVDs). She may not purchase the same focal brand (e.g., animation by Pixar) at every visit. We discretize her decision process into two steps: to visit the website for a product (e.g., DVDs), and to make the purchasing decision, i.e., whether to purchase the focal brand or not.

To quantify the first step, we follow Schmittlein et al. [9] and assume:

i. the number of purchases (n) made by a customer during a time period follows a Poisson distribution with rate λ;
ii. the purchase rate λ across customers follows a gamma distribution, a standard conjugate prior to the Poisson distribution. That is, $f(\lambda) = \alpha^r \lambda^{r-1} e^{-\lambda\alpha}/\Gamma(r)$ with the shape parameter r, the scale parameter α, and $\Gamma(\cdot)$ denoting the gamma function.

Since customers may have different purchase frequencies, to capture the heterogeneity of purchase rates among customers, we introduce the Gamma distribution in assumption *ii*. Assumptions *i* and *ii* jointly prescribe the number of total purchases (N) of a random customer to follow a negative binomial distribution (NBD). That is,

$$P(N = n|r, \alpha) = \frac{\Gamma(r+n)}{\Gamma(r)n!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^n. \qquad (1)$$

Using the NBD distribution to model customers' repeated purchasing has been verified and proved to be very robust for store-level data [3] [6] [9].

Within the category spending, we assume:

iii. at each visit, the probability that a customer selects the focal brand is *p*. Thus, the number of total purchases of the focal brand by a customer follows a binomial distribution;
iv. selection rate *p* across customers follows a general distribution with pdf $g(p)$;
v. a customer's categorical spending decision and her brand choice decision are independent with each other.

The hidden notion behind assumption *iii* is that each customer is stable on her brand preference during the observation period. In reality, customers' favorites at individual level may not be constant over time. Morrison and Schmittlein [6] refer to this as non-stationarity and discuss it extensively in their paper.

Since customers' preference towards the focal brand may be different from each other, they possibly will have different selection rates *p*. To fully capture the heterogeneity among customers, we assume that selection rates follow a general distribution in assumption *iv*. Introducing a general distribution for *p* represents an important departure from the existing studies on customers' repeat purchasing. In the traditional BB/NBD model, the selection rates are assumed to follow a Beta distribution, the conjugate prior of the Binomial distribution. As a result, the whole BB/NBD model is mathematically tractable. The resulting BB/NBD model has proved to be quite robust and has been adopted in solving many business problems [4] [6] [11]. However, this assumption is made for mathematical convenience. If there are more than two modes in the selection rates' distribution or there exist spikes, the predictive accuracy of using the BB/NBD model is doubtful. Moreover, it is very difficult to acquire customers' data to test the distributions of their selection rates in practice. For the purpose of generality and to fully capture the heterogeneity among customers, we proceed to assume that *p* follows a general distribution across the population with pdf $g(p)$ instead of assuming a specific form for *p* to be the Beta distribution as in the BB/NBD model. Since in our model, the distribution of brand purchasing probability is a general distribution, we term our model as the GB/NBD model throughout the paper.

With assumptions *i* to *v*, we can derive the distribution of a random customer's total number of purchases of the focal brand (*X*) as:

$$P(X=x)=\iint P(X=x|\lambda,p)f(\lambda)g(p)d\lambda dp=\iint \frac{(\lambda p)^x e^{-\lambda p}}{x!}f(\lambda)g(p)d\lambda dp$$

$$=\int_0^1 \frac{\alpha^r p^x \Gamma(x+r)}{x!\,\Gamma(r)(\alpha+p)^{x+r}}g(p)dp=\frac{\alpha^r \Gamma(x+r)}{x!\,\Gamma(r)}\int_0^1 \frac{p^x}{(\alpha+p)^{x+r}}g(p)dp.$$

Note that the distribution involves only the integration of $g(p)$. Although $g(p)$ might be of any form, its integral could be approximated by the Gaussian quadrature formula [7, p179]. The above equation can be expressed as

$$P(X=x)\approx \frac{\alpha^r \Gamma(x+r)}{x!\,\Gamma(r)}\sum_{i=1}^{k}\frac{p_i^x}{(\alpha+p_i)^{x+r}}g(p_i)w(i), \qquad (2)$$

where *k* is the order of the Gaussian quadrature, $w(i)$ is the weight for abscissa i, and $g(p_i)$ is the value of $g(p)$ at the Gaussian abscissa i. Since the Gaussian quadrature is accurate for all polynomials up to the degree of *2k – 1*, the above formula converges exponentially fast as *k* increases [7, p180].

After introducing the Gaussian quadrature, the general curve $g(p),\ p\in[0,1]$, is reduced to k values of $g(p_i)$. Recall that in the traditional BB/NBD model, two parameters of the Beta distribution needs to be estimated to make predictive results. Under the empirical Bayes framework, the parameters of the BB/NBD model can be estimated using maximum likelihood. After introducing the Gaussian quadrature

in our model, the general distribution $g(p), p \in [0,1]$, is reduced to k values of $g(p_i)$. If we treat $g(p_i)$ as k unknowns in equation (2), we will have a total of $k+2$ unknown parameters ($\alpha, \gamma, g(p_i)$) to be estimated given the customers' purchasing data (X). We also resort to maximum likelihood for estimating parameters

SIMULATION RESULTS

One advantage of the GB/NBD model is that it can accurately approximate any distribution of user's heterogeneity, which is very difficult to measure in reality. In this section, we conduct a series of simulation study to evaluate our model's performance when the underlying distribution deviates from the Beta distribution. Two representative distributions are tested: truncated normal and piecewise linear. Using these two distributions, we compare the simulation results of our GB/NBD model with the traditional BB/NBD model.

The setup of the simulation for the GB/NBD model is as follows:

1) There are 1,000,000 customers.
2) A purchase rate $\lambda_i \sim \Gamma(\alpha, r), i = 1 \ldots 1{,}000{,}000$ is randomly assigned for each customer.
3) For customer i, $i = 1 \ldots 1{,}000{,}000$, we conduct a random draw from the Poisson distribution with rate λ_i. The resulting number $n_i = \lambda^n e^{-\lambda}/n!$ is assigned as customer i's total category purchases.
4) The selection probability for the focal brand, i.e., $p_i \sim$ trancated normal(μ, σ^2) or $p_i \sim$ piecewise linear, $i = 1 \ldots 1{,}000{,}000$ is randomly assigned for each customer.

To get the simulation results for the BB/NBD model, we follow the first three steps for the GB/NBD model and change the distribution of the focal brand selection rate of customer i to a Binomial distribution with parameters n_i and p_i.

We estimate the parameters of the GB/NBD model using Equation (2). For the BB/NBD model, we use the following equation from Fader and Hardie [4]:

$$P(X = x) = \frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^x \frac{\Gamma(a+x)}{\Gamma(a)} \frac{\Gamma(a+b)}{\Gamma(a+b+x)} \times {}_2F_1\left(r+x, b; a+b+x; \frac{1}{\alpha+1}\right)$$

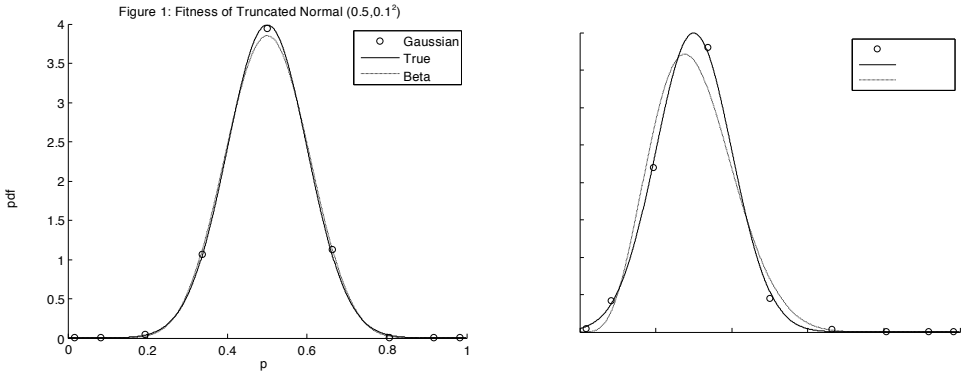
where ${}_2F_1(\cdot)$ is the Gaussian hypergeometric function [1]. A technical issue worth noting is that in the online context, the total number of category purchases n_i for customer i is known by online retailers (e.g., Amazon.com has each customer's purchasing records of any DVDs). We could estimate the Gamma parameters (r, α) first based on n_i only. Then proceed to estimate the whole model. Since we need to use a nonlinear optimization algorithm to search for the maximal value of the likelihood function, the two-step estimation strategy is less challenging.

In the simulation, we set $r = 1.8$ and $\alpha = 0.09$ for the Gamma distribution. To evaluate the accuracy of predicting the selection rates' distribution, we keep the Gamma distribution unchanged, and vary the parameters of the truncated normal distribution in step 4. Our results are shown in Figures 1-4.

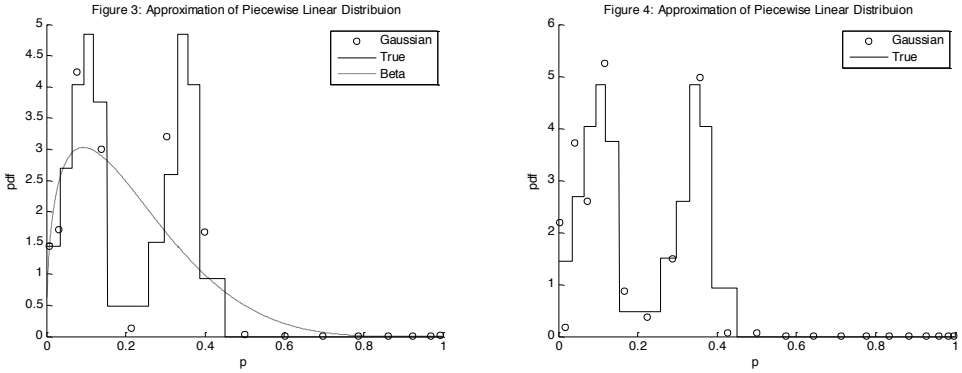
In Figure 1, the solid line is the true distribution (truncated normal distribution with $\mu = 0.5$ and $\sigma = 0.1$); the dashed line is the fitting result of the Beta distribution. As we can see, the Beta line agrees with the true distribution quite well except the part around the mean (from 0.45 to 0.55). The fitted values by GB/NBD are presented by the circle dots in the figure. It is easily seen that the fitted results by GB/NBD agree very well with true distributions. Note that the approximated values of the Gaussian quadrature are shown in scatter dots; this is because only the values at the Gaussian abscissas on the curve matter.

Figure 2 is analogous to Figure 1. It is initialized with a truncated normal distribution with $\mu = 0.3$ and $\sigma = 0.1$. The accuracy of the fitting for the Beta distribution gets worse. The disagreement of true curve

(solid line) and the Beta curve (dashed line) is visually observable. The GB/NBD (circle dots) keeps its performance and agrees with the true curve very well. Note that in Figures 1 and 2, a 9th order Gaussian quadrature is used.



In Figure 3, we test an extreme case when the underline distribution is piecewise linear (as shown by the solid curve). The estimated Beta curve fits the true distribution poorly. Compared with the Beta curve, the values predicted by the GB/NBD model are much better. The results in Figure 3 are obtained using a 15th order Gaussian quadrature. We further increase the order of the Gaussian quadrature from 15 to 21, and present the fitting results in Figure 4. The circle dots obtained from the GB/NBD model clearly track the true curve. These results evidently demonstrate the advantage of the GB/NBD model over the traditional BB/NBD model.



Next, we evaluate the predictive performance of our GB/NBD model. We simulate 1million users' purchases within a time range $(0, t)$. We split users' purchases (X) into two parts (X_1, X_2), where X_1 and X_2 denote the purchases within the time ranges $(0, t/2)$ and $(t/2, t)$, respectively. We use X_1 to estimate the *k+2* parameters for the GB/NBD model and then use the parameters estimated to predict the same user's future purchase (X'_2) within $(t/2, t)$. The prediction for the time range $(t/2, t)$ is given as

$$E(X'_2|X_1 = x) = (x + 1) * P(X_1 = x + 1)/P(X_1 = x).$$

This equation is the well-known Robbins' results [8]. The predicted results X'_2 are compared with the users' real purchases in the time range of $(t/2, t)$. To measure the difference of the predicted value and the true value, we calculate the root-mean-square error (RMSE) of the two series, where a lower RMSE value indicates a better prediction.

We follow the same procedure and obtain the predictive results for the BB/NBD model. To compare the predictive performance of GB/NBD and BB/NBD, we conduct the simulation for three different selection rates' distributions, piecewise linear and truncated normal distributions with mean 0.3 and 0.5. The standard deviations for the truncated normal distributions are kept the same at 0.1. The pdf curves

of these distributions are shown in Figures 1 to 3. The resulting RMSE values are presented in Table 1, where the numbers in the brackets are the orders of the Gaussian quadratures used in the simulation. Under the three selection rates' distributions, the RMSE value of GB/NBD is smaller than that of BB/NBD, which indicates a better prediction performance by GB/NBD. For piecewise linear cases, the RMSE by GB/NBD of using a 21 order Gaussian quadrature (6.5640) is smaller than that of using a 15 order Gaussian quadrature(6.6030), and that is to be expected. This is because a higher order Gaussian quadrature leads to a better approximation, and consequently better prediction results.

Table 1 RMSE of the Prediction Performance

	Piecewise linear	Normal(0.3, 0.1^2)	Normal(0.5, 0.1^2)
BB/NBD	7.1567	8.7461	16.8484
GB/NBD	6.5640 (21) 6.6030 (15)	7.9641(9)	14.5333(9)

CONCLUSIONS

In this paper, we extend the brand choice distribution in the traditional BB/NBD model to a generalized distribution. We assume that a customer's selection rate for the focal brand follows a general distribution as opposed to the Beta distribution in the BB/NBD model. The generalization is made possible through using the Gaussian quadrature. The extension retains the elegance of the BB/NBD framework with simple mathematic expressions. We conduct a simulation study to demonstrate the superiority of the proposed method. Our predictive results show that the GB/NBD model performs better than the BB/NBD model does in prediction accuracy.

The proposed GB/NBD model can be implemented in any situation where BB/NBD can be used. The model proposed is especially suitable in online contexts, where customers are much more diverted and the transaction volumes are much higher. As our simulation results show, when customers' selection rates of the focal brand are distributed by very complicate distributions (such as piecewise linear), the traditional BB/NBD fails in terms of both fitness and prediction accuracy, where the GB/NBD model performs well. Note that by increasing the order of the Gaussian quadrature, we always can get satisfactory results from the GB/NBD model. It is worth mentioning that even though we introduce more parameters into the model than BB/NBD does, the optimization process for the simulation of a group of 1 million customers can still finish in minutes on a modern computer (Intel i7).

REFERENCES

[1] Abramowitz, M., I. A. Stegun (eds). *Handbook of Mathematical Functions,* Dover Publications, 1972.

[2] Chatfield, C., G. J. Goodhardt. The Beta-Binomial model for customer purchasing behavior. *J. Roy. Statist. Soc. Ser. C.* 1970, 19(3), 240-250.

[3] Dunn, R., S. Reader, N. Wrigley. An investigation of the assumptions of the NBD model as applied to purchasing at individual stores. *J. Roy. Statist. Soc.* 1983, *Ser. C.* 32(3), 249-259.

[4] Fader, P. S., B. G. S. Hardie. A note on modeling underreported Poisson counts. *J. Appl. Statist.* 2000, 27(8), 953-964.

[5] Jeuland, A. P., F. M. Bass, G. P. Wright. A multibrand stochastic model compounding heterogeneous Erlang timing and multinomial choice processes. *Oper. Res.* 1980, 28(2), 255-277.

[6] Morrison, D. G., D. C. Schmittlein. Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? *J. Bus. Econ. Statist.* 1988, 6(2), 145-159.

[7] Press, W. H., S., Teukolsky, W., Vetterling, B., Flannery. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[8] Robbins, H. Prediction and Estimation for the Compound Poisson Distribution, *Proc. National Acad. Sci.* USA, 1977, Vol. 74, pp. 2670-2671.

[9] Schmittlein, D. C., A. C. Bemmaor, D. G. Morrison. Why does the NBD model work? Robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Marketing Sci. 1985,* 4(3), 255-266.

[10] Winkelmann, R. *Econometric Analysis of Count Data,* 5th ed. Springer, Berlin, 2008.

[11] Zheng, Z., P. Fader, B.PadmanabhanFrom business intelligence to competitive intelligence: Inferring competitive measures using augmented site-centric data. *Inform. Systems Res*. 2012, 23(3), 698-720.