Simon, Lydia; Adler, Jost

Worth the effort? Comparison of different MCMC algorithms for estimating the Pareto/NBD model

Journal of Business Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Simon, Lydia; Adler, Jost (2021) : Worth the effort? Comparison of different MCMC algorithms for estimating the Pareto/NBD model, Journal of Business Economics, ISSN 1861-8928, Springer, Berlin, Heidelberg, Vol. 92, Iss. 4, pp. 707-733, https://doi.org/10.1007/s11573-021-01057-6

Journal of Business Economics (2022) 92:707–733
https://doi.org/10.1007/s11573-021-01057-6

ORIGINAL PAPER

Worth the effort? Comparison of different MCMC algorithms for estimating the Pareto/NBD model

Lydia Simon[1] · **Jost Adler**[1]

Accepted: 22 July 2021 / Published online: 29 August 2021

Abstract
The Pareto/NBD model is one of the best-known models in customer base analysis. Extant literature has brought up three different Markov Chain Monte Carlo (MCMC) procedures for parameter estimation of this model. Nevertheless, three main research gaps remain. Firstly, the issue of hyper parameter sensitivity for these procedures has been disregarded even though this is crucial when dealing with small sample sizes. Secondly, present research lacks a performance comparison between the different MCMC procedures as well as with Maximum Likelihood Estimates (MLE). Thirdly, existing minimal data set requirements for this model neglect MCMC estimation procedures as they only refer to MLE. To tackle these gaps, we perform two extensive simulation studies. We demonstrate that the algorithms differ in their sensitivity towards the hyper distributions and identify one algorithm that outperforms the other procedures in all respects. In addition, we provide deeper insights into individual level forecasts when using MCMC and enhance extant data set limitation guidelines by considering not only the cohort size but also the length of the calibration period.

Keywords Customer base analysis · Pareto/NBD model · Markov Chain Monte Carlo

JEL Classification M31 · C11

1 Introduction

Customer base analysis (CBA) is an essential component of customer relationship management when looking at companies in a non-contractual setting being interested in a long-term relationship with their customers. CBA uses information on past purchases to analyse and predict transactional patterns. Classical models in this

✉ Lydia Simon
lydia.simon@uni-due.de

[1] Chair of Marketing, University of Duisburg-Essen, Lotharstr. 65, D-47057 Duisburg, Germany

field assume that the inter-purchase times of an individual customer follow a given type of distribution, where the distribution parameter(s) themselves follow another probability distribution to account for heterogeneity within the customer cohort. In addition, most of these models imply that every customer will at some point stop doing business with the company. This may e.g. happen when a customer starts purchasing from a competitor, does not need the product anymore, or literally dies. The estimation of a customer's lifetime is particularly challenging in a non-contractual setting because the timing of the dropout cannot be directly observed. Modelling lifetimes can be done in a similar way to the purchase process, involving individual dropout rates which themselves follow a heterogeneity distribution.

CBA models generally provide two measures that offer direct management benefits. The expected number of future purchases and the probability of a customer still being active at the end of the observation period are valuable metrics that enable the optimisation of marketing strategies regarding incentive or reactivation campaigns. In addition, the estimated parameters can be used to calculate the customer lifetime value (CLV) (see e.g. Fader et al. 2005b and Glady et al. 2009). On an operational level, the CLV can additionally substantiate marketing decisions like individual service offers. On the general management level, the cumulated CLV helps to determine the financial value of a company (McCarthy and Fader 2018).

The Pareto/Negative Binomial Distribution (Pareto/NBD) model was introduced by Schmittlein et al. (1987) and is one of the most acknowledged and cited CBA models in the literature. Its performance is highly regarded among researchers (see e.g. Fader et al. 2005a; Gupta et al. 2006; Batislam et al. 2007). In the early 2000s, the major point of criticism of the Pareto/NBD model was its computational complexity despite its rather simple mathematical assumptions (Fader and Hardie 2005; Jain and Singh 2002). This complexity mainly arises from the large number of hypergeometric functions that need to be calculated. In addition, determining the maximum likelihood estimates (MLE) often led to numerical problems (Ma and Liu 2007; Hoppe and Wagner 2010). This was reflected by the fact that only very few empirical validations had been performed at that time (Batislam et al. 2007). The technological progress has relaxed this issue in different respects. Not only the processing speed and methods have improved considerably, but also predefined functions and packages in publicly available software like R (R Core Team 2020) reduce the individual effort for the application of such a model and enables its utilisation as a benchmark model (Fader et al. 2005a; Jerath et al. 2011; Bemmaor and Glady 2012; Platzer and Reutterer 2016).

Over the past 10–15 years, the concept of Markov Chain Monte Carlo (MCMC) algorithms has taken root in CBA (Ma and Liu 2007; Abe 2009; Singh et al. 2009; Ma and Büschken 2011; Schweidel et al. 2014; Platzer and Reutterer 2016). Although their mathematical basis is more complex than the calculation of the MLE, these algorithms enable a new type of data analysis (Paap 2002) offering three main advantages. Firstly, MCMC procedures provide not only point estimates (i.e. the mode) of the model parameters but the entire posterior distribution of the parameters, which allows to also determine the mean, median, or standard deviation. Secondly, these procedures can estimate individual level parameters in addition to describing the heterogeneity distribution across customers. Thirdly, it is possible

to augment the timing of the individual dropouts, which gives new insights into the unobserved customer lifetime.

MCMC methods have been applied to the Pareto/NBD model using three different algorithms or algorithm frameworks, namely those by Abe (2009), Ma and Liu (2007), and Singh et al. (2009). However, the extant literature on the parameter estimation of this model is incomplete in three significant aspects. Firstly, prior research has not performed a comparison between the different MCMC algorithms or with MLE. Though MCMC algorithms provide much richer information about the model parameters than MLE, it is unclear if these algorithms perform better regarding parameter recovery or forecast accuracy and are thus worth the additional implementation effort. Secondly, the influence of the (hyper) prior distributions, i.e. the distributions of the heterogeneity parameters r, αs, and β, is left unconsidered in all three implementation methods. Robert (2007) states that the (hyper) prior distributions are the key to Bayesian inference and their determination is thus the most important step in the MCMC procedure. However, none of the authors who introduced a MCMC algorithm for the Pareto/NBD model has addressed this issue. Singh et al. (2009) use a $\Gamma(5, 5)$ distribution for each of the heterogeneity parameters without further elaboration, whereas Abe (2009) and Ma and Liu (2007) do not state which hyper parameters they employed at all. As the Pareto/NBD model is mostly used for cohorts of new customers or products and thus for small data sets, the choice of hyper distributions may be crucial for the analysis and inferences (Edwards et al. 1963). Thirdly, to the best knowledge of the authors, there exists no research on minimal data set requirements for the application of the Pareto/NBD model with MCMC. Two studies (Schmittlein and Peterson 1994; Hoppe and Wagner 2010) performed such examinations but referred to MLE only.

In this article, we contribute to extant literature by addressing these neglected issues. We use simulated data sets with known underlying parameter values to assess the hyper parameter sensitivity as well as the recovery and forecast quality of the different estimation procedures. In study 1, we compare the different MCMC algorithms (partially based on the R package BTYDplus (Platzer 2016)) with each other as well as with MLE. In order to systematically analyse the parameter estimates, especially with respect to the prior distribution sensitivity, we choose the simply structured parameter space of Fader et al. (2005), which provides three equally distributed values for each heterogeneity parameter. The results show that MCMC, and in particular Abe's algorithm, outperforms MLE. The diagnosed superiority of Abe's MCMC algorithm leads to the question whether existing data set restrictions for the practical application of the Pareto/NBD model in the literature (Schmittlein and Peterson 1994; Hoppe and Wagner 2010) can be relaxed as these refer to MLE. In study 2, we address this question by replicating the simulation study of Hoppe and Wagner (2010) and additionally incorporating Abe's algorithm. As this research question focusses on the data set properties, the parameter space for this simulation study is based on behavioural characteristics, leading to a more complex set of heterogeneity parameters. Both studies require preliminary work on the choice of hyper parameters, which gives us valuable insights into this mostly disregarded but crucial aspect of MCMC in CBA.

The remainder of this paper is organised as follows. In Sect. 2, we recall the Pareto/NBD model assumptions, describe the properties of the different MCMC algorithms, and introduce the measures we employ for their comparison. The explicit procedures, probabilities, and distributions are outlined in detail in the Technical Appendix (A). In Sect. 3, we compare the different MCMC algorithms with each other and with MLE as outlined above. In Sect. 4, we analyse the MCMC performance on different data set sizes to derive minimal requirements for cohort sizes and the length of the calibration period. In Sect. 5, we summarise and discuss our results and derive implications for researchers and practitioners alike.

2 The Pareto/NBD model

2.1 Model assumptions

The Pareto/NBD model is based on the following assumptions (Schmittlein et al. 1987). For the simplification of the notation, we remove the customer index i when regarding individual level formulas. In addition, we explain the effects of different numerical values for each of the parameters.

1. Each of the N customers goes through two stages. The "lifetime" with a firm starts with the initial purchase and ends at a non-observable time τ, where the state of inactivity is non-reversible.
2. While the customer is active, the time elapsed between two consecutive purchases follows an exponential distribution with parameter λ, i.e.

$$f(t_j - t_{j-1}) = \lambda e^{-\lambda(t_j - t_{j-1})}, \tag{1}$$

   where t_j denotes the time elapsed from the initial ($t_0 = 0$) to the jth purchase. As $E(t_j - t_{j-1}) = \frac{1}{\lambda}$, the inter-purchase times tend to be shorter for larger values of λ, leading to a higher number of purchases. Therefore, an arbitrary customer with $\lambda = 0.05$ has an average inter-purchase time of 20 time units (e.g. weeks) and, leaving the dropout process aside, we would expect him to make five purchases within 100 times units.
3. The customer's lifetime τ is exponentially distributed with parameter μ, i.e.

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

   Comparable to the purchase process, $E(\tau) = \frac{1}{\mu}$, i.e. a small μ implies a longer lifetime with the company. A customer with $\mu = 0.1$ would thus be expected to drop out after 10 time periods. Another perspective on the dropout process is the defection rate (DR). For an exponentially distributed lifetime τ, DR is a constant given by

$$DR = 1 - e^{-\mu}, \tag{3}$$

which quantifies the probability that a customer will defect within one time unit. In the above example, the probability that a customer with $\mu = 0.07$ defects within the next time unit is 6.8%.

4. The purchase rate λ varies across customers and follows a gamma distribution with parameters r and α, i.e.

$$g(\lambda|r,\alpha) = \frac{\alpha^r \lambda^{r-1} e^{-\lambda\alpha}}{\Gamma(r)}. \tag{4}$$

A direct behavioural understanding of the parameters (r,α) is very difficult. Instead, we can interpret the expected value and dispersion by considering that

$$E(\lambda) = \frac{r}{\alpha}, \sigma_\lambda = \frac{\sqrt{r}}{\alpha}, CV_\lambda = \frac{1}{\sqrt{r}}, \tag{5}$$

with σ_λ being the standard deviation and CV_λ the coefficient of variation (CV) for λ. This implies that, e.g. for $r = 0.5$ and $\alpha = 10$, the average purchase rate $E(\lambda)$ is 0.05 and λ averagely varies by $1.41 = 141\%$ within the cohort.

5. Heterogeneity in μ follows a gamma distribution with parameters s and β, i.e.

$$g(\mu|s,\beta) = \frac{\beta^s \mu^{s-1} e^{-\mu\beta}}{\Gamma(s)}. \tag{6}$$

Similar to the purchase process, a direct interpretation of (s,β) is unfeasible. Using

$$E(\mu) = \frac{s}{\beta}, \sigma_\mu = \frac{\sqrt{s}}{\beta}, CV_\mu = \frac{1}{\sqrt{s}}, \tag{7}$$

allows us to interpret that e.g. for $s = 2$ and $\beta = 20$, the average dropout parameter $E(\mu)$ is 0.1 and that it varies by $1 = 100\%$ within the cohort.

6. λ and μ vary independently across customers.

The posterior distributions of the individual and heterogeneity parameters are derived in A2.

2.2 MCMC algorithms for the Pareto/NBD model

The extant literature provides three different MCMC algorithms for the Pareto/NBD model. We present each of them in 2.2.1 to 2.2.3 and discuss their technical differences in 2.2.4.

2.2.1 Ma and Liu (2007)

Ma and Liu (2007) generate the individual parameter estimates $\{\hat{\lambda}_i, \hat{\mu}_i\}$ as well as the heterogeneity parameter estimates $(\hat{r}, \hat{\alpha}, \hat{s}, \hat{\beta})$ by applying a Gibbs sampler. Since

only the posteriors of α and β reduce to a known distribution, we draw from the combined posterior distributions (22)–(25) by using a slice sampling routine (Neal 2003) to simulate the heterogeneity parameters.For our study, we use a slice sampling routine based on the R package BTYDplus (Platzer 2016). To reduce the computational cost, the slice sampling is outsourced to C++ using the Rcpp package (Eddelbuettel and François 2011; Eddelbuettel 2013).

2.2.2 Abe (2009)

Abe (2009) also generates individual and heterogeneity parameter estimates, but additionally simulates the unobserved individual dropout times $\{\tau_i\}$ by applying a data augmentation technique (Tanner and Wong 1987). He first creates a latent indicator variable z_i, which specifies if customer i is still active in T_i by using formula (17) for $P(alive)$. Depending on the aliveness status, he draws $\hat{\tau}_i$ from the appropriate distribution shown in Figs. 1 and A3.1. The specific value of τ_i simplifies the posterior distributions for $\{\lambda_i\}$ and $\{\mu_i\}$ according to A2. Therefore, we can draw the individual parameters straight from a gamma distribution employing a common number generator. Abe's version of the posterior distribution for $\{\mu_i\}$ noted in (19) is conditional on being alive rather than conditional on the specific value of τ_i. We therefore implement a slightly different version by using $\Gamma(s+1, \beta+\tau_i)$ from formula (21) in both cases of the aliveness status in order to consider all information available. As the data augmentation has no effect on the heterogeneous likelihood, we need to perform the same slice sampling technique for $\{r, \alpha, s, \beta\}$ as in the algorithm introduced

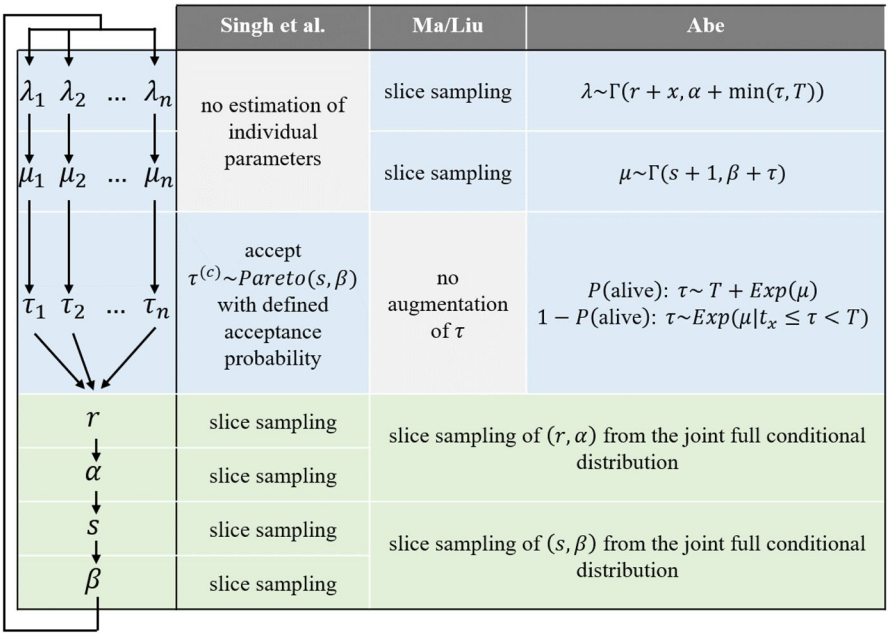

Fig. 1 Comparison of the different MCMC procedures

by Ma and Liu. Abe's algorithm in the described form is also coded in the BTYDplus package (Platzer 2016).

2.2.3 Singh et al. (2009)

Singh et al. (2009) use a divergent approach for their MCMC procedure. The major difference to the other algorithms is that they waive the individual parameters $\{\lambda_i, \mu_i\}$. They separately regard the NBD-distributed counting process on the one hand side and the Pareto-distributed dropout process on the other. The full conditional distributions in A2 require values for the dropout times $\{\tau_i\}$. The candidates for this individual lifetime are drawn from a truncated Pareto-II-distribution with parameters s and β. A candidate $\tau_i^{(c)}$is being accepted with probability (30) from A3.3. The heterogeneity parameters r, α, s, and β are drawn from (26)–(29) using a slice sampling routine.

2.2.4 Similarities and differences

Figure 1 shows a schematic overview of similarities and differences of the three MCMC algorithms.

Even though Singh et al. (2009) and Abe (2009) both use a data augmentation procedure, Singh et al.'s (2009) algorithm reveals some weaknesses in the direct comparison. The most obvious one is that they do not provide estimates for $\{\hat{\lambda}_i, \hat{\mu}_i\}$, which makes certain evaluations impossible on the individual level. Additionally, their data augmentation technique allows the estimated dropout times to stay constant over many iterations. In contrast, Abe's augmented lifetimes inevitably change in every step, which leads to a more granular distribution. Lastly, their algorithm underperforms in terms of computational cost. Using data augmentation simplifies the posterior distributions and thus generally speeds up the estimation. The absence of the individual parameter level though causes higher data input requirements in the slice sampling step, which is very time-consuming.

Ma and Liu (2007) use the very laborious slice sampling routine for the estimation of the individual parameters $\{\hat{\lambda}_i, \hat{\mu}_i\}$, which impedes the determination of lifetime estimates. These are two major disadvantages compared to the data augmentation procedures.

At this point, we can conclude that the conception of Abe's algorithm is superior to the others as it unifies all advantages concerning (1) individual parameter estimates, (2) augmented dropout dates, and (3) calculation time while showing no weaknesses compared to the other methods.

When we think of the likelihood as a distribution, there are three measures of central tendency that we can use as point estimates, namely the mode, the mean, and the median. MLE is defined as the likelihood mode, whereas MCMC procedures also provide the mean and median of the posterior parameter distributions. We additionally examine which of these measures is the best point estimate regarding parameter recovery in our simulation study.

All MCMC algorithms in this study are performed using four chains with random initial values and 5,000 draws each, where the first 2,000 steps are defined as the burn-in. The MLE comparison values are determined by a routine from the BTYD package in R (Dziurzynski et al. 2014) which is based on the optim() function of the stats package (R Core Team 2020).

2.3 Performance Metrics

Let $\theta_i = \left(r, \alpha, s, \beta, E(\lambda) = \frac{r}{\alpha}, E(\mu) = \frac{s}{\beta}\right)_i$ denote the vector of the true parameter values and $\hat{\theta}_{i,k}$ the parameter point estimate (mean, median, or MLE) for customer i ($i = 1, ..., N$) of data set k ($k = 1, ..., K$). We analyse the recovery of the heterogeneity parameter estimates $(\hat{r}, \hat{\alpha}, \hat{s}, \hat{\beta})$ of our simulated data sets by using the mean percentage error (MPE) and the mean absolute percentage error (MAPE), which are defined as:

$$MPE\left(\hat{\theta}\right) = \frac{100}{K \cdot N} \cdot \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\hat{\theta}_{i,k} - \theta_{i,k}}{\theta_{i,k}}, \tag{8}$$

$$MAPE\left(\hat{\theta}\right) = \frac{100}{K \cdot N} \cdot \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\left|\hat{\theta}_{i,k} - \theta_{i,k}\right|}{\theta_{i,k}}. \tag{9}$$

For the performance comparison of the recovery of the individual parameters $\{\lambda_i\}$ and $\{\mu_i\}$, the MPE and MAPE are unsuitable measures as the parameter values may be very close to zero and therefore produce infinite or undefined M(A)PE values. Hence, we draw on the concept of the mean arctangent absolute percentage error (MAAPE) instead (Kim and Kim 2016). The MAAPE is defined as follows:

$$MAAPE\left(\hat{\theta}\right) = \frac{100}{K \cdot N} \cdot \sum_{k=1}^{K} \sum_{i=1}^{N} arctan\left(\frac{\left|\hat{\theta}_{i,k} - \theta_{i,k}\right|}{\theta_{i,k}}\right). \tag{10}$$

The $arctan(x)$ function is bound to $\left[0, \frac{\pi}{2}\right]$, which makes the MAAPE robust against very small parameter values as well as estimate outliers.

Abe and Singh et al. provide augmented values for the individual dropout times $\{\hat{\tau}_i\}$, which we use to derive a different type of information on the activity status of an individual customer than $P(alive)$ gives us. We define the survival accuracy SA as

$$SA = \frac{100}{K \cdot N} \cdot \sum_{k=1}^{K} \sum_{i=1}^{N} I_{\{(T_i - \tau_i) \cdot (T_i - \hat{\tau}_i) > 0\}}, \tag{11}$$

with

$$I_{\{(T_i-\tau_i)\cdot(T_i-\hat{\tau}_i)>0\}} = \begin{cases} 1, & (T_i-\tau_i)(T_i-\hat{\tau}_i) > 0 \\ 0, & (T_i-\tau_i)(T_i-\hat{\tau}_i) \leq 0 \end{cases}, \tag{12}$$

where the indicator function is one if and only if the median estimate $\{\hat{\tau}_i\}$ states correctly whether a customer has defected.

The analysis of the future number of individual purchases x_i^* requires the mean absolute error (MAE) because the basic value may be zero and leads to errors in MPE, MAPE, and MAAPE. It is defined as

$$MAE\left(\hat{x^*}\right) = \frac{1}{K \cdot N} \cdot \sum_{k=1}^{K} \sum_{i=1}^{N} \left| \hat{x^*}_{i,k} - x^*_{i,k} \right|. \tag{13}$$

3 Study 1: Comparison of the MCMC procedures

3.1 Data set generation for study 1

Our parameter space for creating the synthetic data sets of study 1 is based on Fader et al. (2005) and uses $r, s \in \{0.25, 0.5, 0.75\}$ and $\alpha, \beta \in \{5, 10, 15\}$, allowing $3^4 = 81$ combinations of these parameter values. To reduce the computational cost, we choose a random sample of 500 with replacement from these 81 combinations. This implies an average number of $\frac{500}{3} = 166.7$ data sets for the analysis of a single underlying heterogeneity parameter value. As $E(\lambda) = \frac{r}{\alpha}$ and $E(\mu) = \frac{s}{\beta}$ rely on two parameters each, their average number of data sets reduces to $\frac{500}{3^2} = 55.6$. To simulate the individual purchase and dropout rate of the customers for each of these 500 data sets, we use the integrated random number generator in R and draw a sample of 1,500 individual values for $\{\lambda_i\}$ and $\{\mu_i\}$ from $\Gamma(r, \alpha)$ and $\Gamma(s, \beta)$ respectively. For each data set and customer, we now draw one realisation from $Exp(\mu_i)$ to receive the lifetime $\{\tau_i\}$ of the individual customers by using the integrated random number generator for the exponential distribution. Lastly, we simulate an initial purchase time for each customer as a fraction of the first time unit and generate successive inter-purchase times by drawing random values from $Exp(\lambda_i)$ until the individual lifetime $\{\tau_i\}$ is exceeded. We simulate these transaction data for a total period of 104 time units, of which 78 are used as the calibration period and the remaining 26 as the forecast period, corresponding to 1.5 years and 6 months on a weekly basis. To reduce the computational effort, we keep these values as well as the cohort size of 1,500 fixed throughout study 1.

3.2 Comparison of the hyper parameters

The application of MCMC procedures requires the choice of a prior distribution for the heterogeneity parameters (r, α, s, β). This means that we need to choose a distribution type and specify its parameters ("hyper parameters"). We perform the

MCMC algorithms with different hyper parameters to learn about the sensitivity of the parameter recovery. Using gamma distributions $\Gamma(h_1, h_2)$ as the conjugate prior, we need to specify their expected values and CV given by $\frac{h_1}{h_2}$ and $h_1^{-0.5}$, respectively. As each of the heterogeneity parameters (r, α, s, β) can take three values, we choose the middle one as the expected value of the prior distribution, i.e. $E(r) = E(s) = \frac{h_1}{h_2} = 0.5$ and $E(\alpha) = E(\beta) = \frac{h_1}{h_2} = 10$. This enables us to analyse the effect of the expectation of $\Gamma(h_1, h_2)$ being set too large (for $r, s = 0.25$ and $\alpha, \beta = 5$), too small (for $r, s = 0.75$ and $\alpha, \beta = 15$), or accurate (for $r, s = 0.5$ and $\alpha, \beta = 10$). We simultaneously vary the coefficient of variation (CV) of $\Gamma(h_1, h_2)$ between 0.01 and 2.

Figure 2 shows the MAPE of all four heterogeneity parameters for all three MCMC procedures based on their median draw, both averaged over the 500 data sets (bold red line) and separated by their underlying parameter value. Most plots for the total MAPE show a slight U-shape. The numeric values suggest

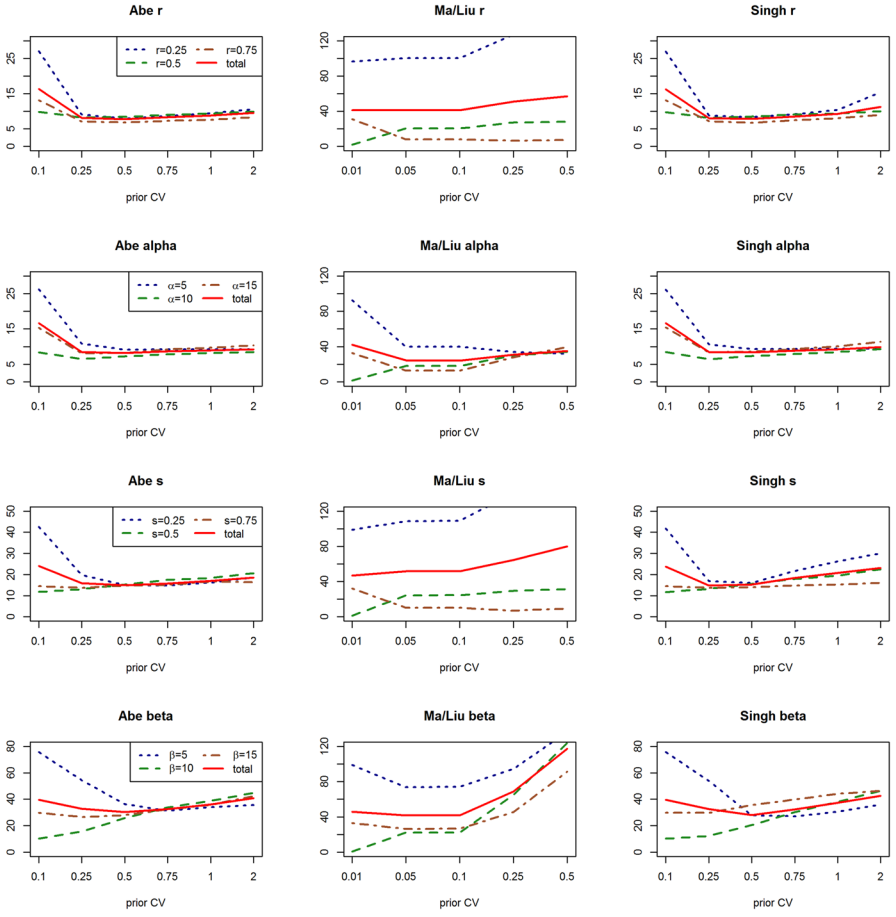

Fig. 2 MAPE of the heterogeneity parameters for the different hyper distributions of study 1

an optimal CV of 0.5 for the algorithms of Abe and Singh et al. for the given parameter space. Their courses of the MAPE for the low (blue dotted line) and high (green dashed line) parameter values illustrate that an overrated prior mean causes higher deviations than an underrated, in particular when being combined with a small CV. Thus, in case of uncertainty about the choice of hyper parameters when employing the algorithms of Abe or Singh et al., it is recommended to combine a smaller expected value with a larger dispersion. The estimates of Ma and Liu though show a high sensitivity towards both the mean and the CV of the hyper distribution. Considering the different scaling of their plots on both axes because of their high MAPE values, they only generate comparatively good estimates when using highly informative prior distributions. For the further comparison with the other procedures, we use $CV = 0.1$ when estimating parameters for Ma and Liu's algorithm, as this value provides the smallest MAPE for all four heterogeneity parameters.

In addition to the conjugate prior distribution, we also apply uninformative hyper prior distributions based on Jeffreys (1946). However, as they turn out to be less accurate than a gamma distribution, they will not be considered in the further analysis. To decide between the mean and median draw as the best fitting point estimate, we compare their MPE and MAPE values. The median outperforms the mean in all respects, especially for larger CV values as these promote outliers. For the further analysis, we will therefore restrict to using the median draw whenever a point estimate is required and compare the performance metrics for the different MCMC procedures with their respective optimal hyper parameters.

3.3 Comparison of the MCMC procedures

3.3.1 Recovery of the heterogeneity parameters

Table 1 reports the MPE and MAPE values for θ using the optimal hyper distribution of each algorithm as derived above. As Fig. 2 already suggested, Ma and Liu's procedure generates highly overrated estimates because the given parameter

Table 1 MPE and MAPE for θ

	MPE				MAPE			
	Abe	Ma/Liu	Singh	MLE	Abe	Ma/Liu	Singh	MLE
r	2.2	37.0	1.6	3.2	7.7	41.0	7.8	9.8
α	3.7	18.8	3.4	5.4	8.2	24.2	8.3	9.9
s	0.2	47.8	−7.5	15.5	14.8	51.7	15.3	32.8
β	4.0	20.1	−7.4	45.7	30.2	41.7	28.0	77.3
$E(\lambda) = \frac{r}{\alpha}$	−0.9	15.0	−1.3	−1.9	5.0	19.4	5.0	5.4
$E(\mu) = \frac{s}{\beta}$	4.1	38.6	7.1	6.8	18.7	50.5	18.9	30.2

Abe and Singh et al. with hyper CV 0.5, Ma and Liu with hyper CV 0.1

space is too broad to satisfy the need for very informative priors. Singh et al.'s results are very similar to Abe's in the absolute deviation but show an underestimation in the dropout process parameters s and β. The MLE routine performs well for the purchase process but produces high deviations in the dropout process. We can therefore conclude that Abe's algorithm is clearly preferred over all other estimation procedures.

3.3.2 Recovery of the individual parameters

As Singh et al. and MLE do not provide any individual parameter information, we can only compare the parameter recovery of the individual parameters $\{\lambda_i, \mu_i\}$ for Abe's and Ma and Liu's algorithms. Table 2 reports the MAAPE values which were calculated in total as well as separated by the number of repurchases in the calibration period. We should consider that very small parameter values promote higher relative deviations and therefore increase the MAAPE. The results show that the MAAPE for $\{\lambda_i\}$ decreases with an increasing number of purchases because a higher number of repurchases gives us more information on the individual λ and also implies larger values for λ. Opposed to this, the estimation of μ deals with two contrary effects. A higher number of repurchases gives more indirect information on the lifetime. Simultaneously, it also suggests a later dropout and thus a lower value of μ, which promotes larger values for the MAAPE. These effects cause decreasing MAAPE values for $\{\lambda_i\}$ and a U-shaped curve for $\{\mu_i\}$. As the M(A)PE values for $E(\lambda)$ and $E(\mu)$ in Table 1 already suggested, Abe's individual estimates outperform Ma and Liu's in the MAAPE. This holds in particular for two hardly traceable domains, namely the purchase parameter $\{\lambda_i\}$ for the cohort of single buyers and the complete dropout process, represented by $\{\mu_i\}$.

3.3.3 Survival accuracy

Abe and Singh et al. both provide augmented values for the individual lifetimes $\{\tau_i\}$. The distribution of the survival accuracy (11) over the 500 data sets is box-plotted in Fig. 3 and shows a significantly higher rate for Abe's algorithm than for Singh et al. In Fig. 4, we compare the actual activity status with the median draw of $\{\tau_i\}$ for all customers of all data sets. According to the plotted figures, Abe' algorithm performs particularly better at correctly classifying the

Table 2 MAAPE for study 1, Abe with hyper CV 0.5, Ma and Liu with hyper CV 0.1

		Number of repurchases							Total
		0	**1**	**2**	**3**	**4**	**5**	**> 5**	
λ	Abe	0.84	0.60	0.49	0.42	0.37	0.34	0.26	0.68
	Ma and Liu	0.91	0.62	0.50	0.43	0.38	0.34	0.26	0.72
μ	Abe	0.80	0.78	0.78	0.74	0.78	0.79	0.81	0.80
	Ma and Liu	0.85	0.83	0.83	0.84	0.84	0.84	0.86	0.85

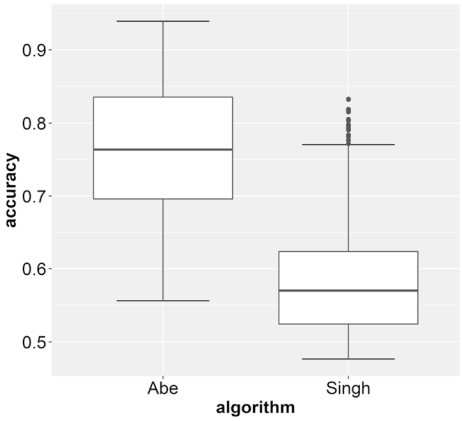

Fig. 3 Survival accuracy

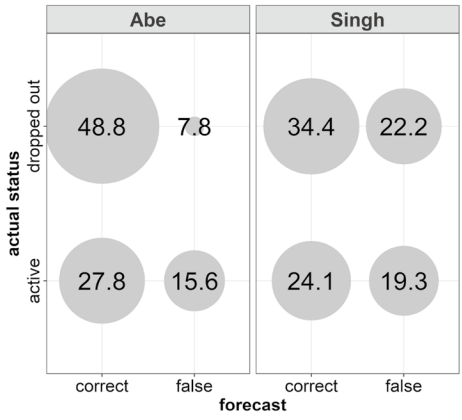

Fig. 4 Survival cross tables (in %)

churned customers. The reason for this can be attributed to the different simulation routines for $\{\tau_i\}$, which are described in A3.1 and A3.3. As Singh et al. use an acceptance probability, the values for $\{\tau_i\}$ tend to remain unchanged for many simulation steps, leading to a less accurate distribution and thus to a smaller survival accuracy.

3.3.4 Forecast accuracy

In our final analysis of study 1, we compare the forecast accuracy of the four algorithms. Corresponding to a length of 6 months on a weekly basis, we calculate $E(x^*)$ for our forecast period of 26 time units using three different methods. These methods depend on the available parameters and are described in A4. Method 1 solely requires the heterogeneity parameters (r, α, s, β) and can hence be applied to all procedures. Method 2 is the individual expectation conditional on the customer still being active. As it requires values for $\{\lambda_i\}$, $\{\mu_i\}$, and $\{\tau_i\}$, it can only be applied to Abe's algorithm. In method 3, the above condition is removed

by multiplying the individual expectations with $P(alive)$ and thus it only requires values for $\{\lambda_i\}$ and $\{\mu_i\}$. Therefore, it can be applied to the procedures of Abe and Ma and Liu. We conjecture that considering individual level parameters in methods 2 and 3 should lead to better forecasts for an individual customer, whereas the use of the heterogeneous estimates in method 1 should yield a better forecast on the aggregated level.

Table 3 reports different measures for the forecast accuracy. The individual MAE is defined in (13) and shows the mean deviation over all customers and data sets on the individual level. The aggregated MPE and MAPE refer to the total sum of future purchases of all customers in a data set. These forecast metrics reveal three effects. Firstly, small individual MAE values do not necessarily imply a small M(A)PE on the accumulated level. As zero buyers cannot be underrated, a general underestimation leads to a small MAE on the individual level for zero buyers but to a higher deviation in the accumulated purchases. This effect is very distinct here as 80.3% of the customers over all data sets made no purchase in the forecast period. Secondly, Ma and Liu's overestimation of $E(\mu)$ (see Table 1) implies shorter lifetimes and hence less future purchases leading to the previously described effect of small individual MAE values but large M(A)PE values. Thirdly, we can observe the expected effect that individual level parameters perform well when considering the individual purchase behaviour whereas heterogeneous estimates do better on the aggregated level. The reason for this can be seen in a "smoothing" effect which impedes extremely small values for x^*. Exemplarily for Abe's algorithm, 54.3% of all customers have a forecast of less than 0.01 purchases when using the individual method 3, whereas method 1 predicts less than 0.01 purchases for only 1.4% of all customers. We can conclude that the choice of the best forecast method depends on its purpose. Forecasts on the cohort level should be made by using method 1 either with Abe's MCMC algorithm or MLE. Since an individual method should be used to receive the most accurate individual forecast, the application of MCMC is required. In this case, Abe slightly outperforms Ma and Liu.

Table 3 Purchase forecast accuracy for the different procedures and formulas

		Estimates	Individual MAE			Aggregated	
		used	$x^* = 0$	$x^* > 0$	Total	MPE	MAPE
Method 1	Abe	r, α, s, β	0.23	1.53	0.48	1.4	5.4
	Singh	r, α, s, β	0.23	1.52	0.48	2.2	5.6
	Ma/Liu	r, α, s, β	0.20	1.56	0.47	5.8	8.7
	MLE	r, α, s, β	0.23	1.53	0.48	0.0	6.0
Method 2	Abe	$\{\lambda_i, \mu_i, \tau_i\}$	0.14	1.60	0.42	17.2	17.3
Method 3	Abe	$\{\lambda_i, \mu_i\}$	0.16	1.59	0.43	15.4	15.5
	Ma/Liu	$\{\lambda_i, \mu_i\}$	0.14	1.62	0.43	21.5	21.5

Table 4 Behavioural characteristics and derived parameter values

Behavioural characteristic		Numerical value		
		Low	Medium	High
Purchase frequency	$E(\lambda) =$	1/6	1/2	1/1
Dropout rate	$DR(\mu) =$	50/1000	275/1000	500/1000
Purchase process heterogeneity	$CV(\lambda) =$	1/2	3/4	3/2
Propout process heterogeneity	$CV(\mu) =$	3/4	1/1	2/1

Table 5 Derived parameter space

Parameter space of study 2
$r \in \{0.44, 1.78, 4\}$
$\alpha \in \{0.44, 0.89, 1.78, 2.67, 3.56, 4, 8, 10.67, 24\}$
$s \in \{0.25, 1, 1.78\}$
$\beta \in \{0.36, 0.78, 1.44, 2.56, 3.11, 4.87, 5.53.19.50, 34.66\}$

3.4 Summary of study 1

The inferences we can draw from study 1 regarding the best Pareto/NBD procedure are very clear. Abe's algorithm is the only one that provides the full range of estimate values and can hence be applied to all covered analyses. From all the performance metrics examined, it is at least equal to the other procedures but mostly outperforms them. In addition, it requires considerably less computing time than the other two algorithms. Furthermore, we showed that the augmented values of $\{\tau_i\}$ are a valuable addition to the parameter estimates. Concerning the sensitivity analysis, we have shown that the choice of the hyper parameters has a crucial influence on the parameter estimation.

To investigate the data set requirements for applying the Pareto/NBD model in study 2, we will limit ourselves to the superior algorithm of Abe and compare it with MLE as a benchmark.

4 Study 2: Data Set requirements

In study 2, we replicate the simulation study of Hoppe and Wagner (2010) and examine whether the minimal data sets requirements they derived for MLE can be relaxed when using Abe's MCMC procedure.

4.1 Data set generation for study 2

The simulation framework developed by Hoppe and Wagner (2010) is based on behavioural characteristics. As these cannot directly be mapped by the heterogeneity parameters, the authors define the average purchase frequency, the average dropout

rate, and their dispersions within the cohort and translate them back into values for (r, α, s, β) using the formulas (3), (5), and (7). Table 4 shows the behavioural characteristics used for study 2. The $3^4 = 81$ combinations of their values result in the heterogeneity parameter space as noted in Table 5.

For each of the 81 scenarios, Hoppe and Wagner (2010) created 100 synthetic data sets for cohort sizes of $N \in \{250, 500, 750, 1000, 1250, 1500\}$ and calibration periods of $T_{cal} \in \{12, 18, 24, 30\}$ time units. Since the parameter estimation with a MCMC procedure requires a multiple of time and storage space compared to MLE, we reduce the sample size of the synthetic data sets to 30 replications for each scenario and perform both MLE and Abe's MCMC algorithm.

Similar to study 1, we randomly generate 1,500 individual parameters $\{\lambda_i, \mu_i\}$ for each data set of each combination of (r, α, s, β) and simulate the individual lifetimes and purchase profiles. For smaller cohort sizes $N < 1,500$, we only consider the first n customers and vary the cut-off dates for shorter calibration periods of $T_{cal} < 30$.

4.2 Technical limitations

The number of data sets that cannot not be estimated with the MLE routine from the BTYD package (Dziurzynski et al. 2014) increases with a decreasing number of customers and calibration period length. In total, 17.6% of all data sets can either not be estimated or contains parameter estimates equal to their upper boundary defined in the optimisation routine and are therefore regarded as illegitimate estimates. Hence, we additionally apply the solnl() optimiser from the NlcOptim package (Chen and Yin 2019). As the optim() function tends to struggle with small values of $E(\mu)$ and solnl() with large values of $E(\mu)$, we are able to reduce the ratio of missing or illegitimate estimates to 7.6%. In cases where both routines produced results, we choose the one with the higher log-likelihood value. Hoppe and Wagner (2010) reported similar estimation problems in their study but gave no detailed information on their fail ratio.

4.3 Hyper parameter sensitivity

For the analysis of the hyper parameter sensitivity of Abe's MCMC algorithm, we use the mean of the parameter space in Table 5 as the prior mean and vary the CV of the heterogeneity parameters between 0.5 and 5. For each of the 81 parameter combinations, we apply these values to a sample of three data sets. Since the influence of the hyper parameters increases with a decreasing size of the data set (Edwards et al. 1963), we use reasonably small data set restrictions based on Hoppe and Wagner (2010) by specifying $N = 750$ and $T_{cal} = 18$ time units to reduce the computational effort of this analysis.

Figure 5 shows the MAPE of $\hat{\theta}$ with respect to the different hyper parameters. The numeric values of the MAPE as well as the MPE reveal the smallest accumulated error for a CV of the heterogeneity parameters of 1 which we

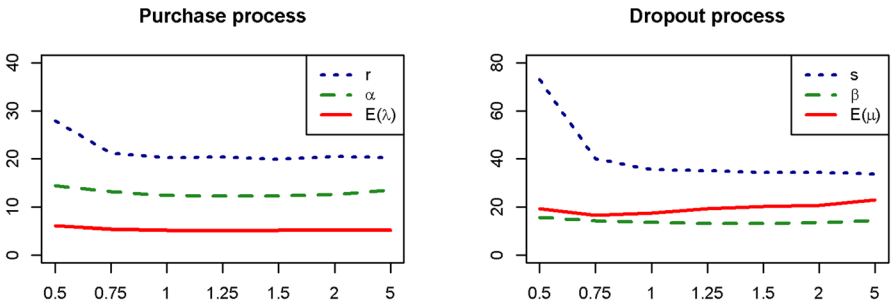

Fig. 5 MAPE for the different hyper CVs of study 2

therefore apply to the whole data set framework. The parameter space in study 2 has a much wider range than in study 1, explaining this higher value of the hyper CV. However, this enhanced range of parameters also yields another effect. As we use the same hyper parameters for each parameter combination regardless of their true underlying values, we should be aware that the performance metrics we receive in study 2 will be more conservative than in study 1.

4.4 Parameter Recovery

Table 6 compares the marginal MAPE values (i.e. the MAPE averaged over all parameter sets and data sets) of Abe's algorithm with MLE (in *italic*) and shows their progression over the different numbers of customers and observation

Table 6 Marginal MAPE for the study 2 data sets

	N						T			
	250	500	750	1, 000	1, 250	1, 500	12	18	24	30
r	20.6	15.5	13.2	11.7	10.6	9.7	15.0	13.8	13.0	12.5
	31.3	*20.0*	*17.1*	*14.7*	*12.7*	*11.3*	*20.7*	*18.2*	*16.4*	*16.0*
α	20.5	15.7	13.4	12.2	11.3	10.6	15.1	14.1	13.5	13.1
	31.2	*20.4*	*17.4*	*15.2*	*13.6*	*12.4*	*20.6*	*18.7*	*17.2*	*16.9*
s	26.4	22.1	20	18.5	17.6	16.6	28.3	21.2	16.8	14.5
	47.9	*35.0*	*28.9*	*24.6*	*22.9*	*21.9*	*47.1*	*31.3*	*23.3*	*19.1*
β	47.3	41.5	37.3	34.0	31.9	30.3	51.3	38.0	31.2	27.8
	98.1	*65.7*	*51.5*	*42.9*	*38.8*	*37.0*	*87.9*	*54.9*	*43.2*	*36.4*
$E(\lambda) = \frac{r}{\alpha}$	8.2	6.1	5.4	4.9	4.5	4.2	6.1	5.6	5.4	5.4
	8.3	*6.2*	*5.4*	*4.9*	*4.5*	*4.2*	*6.1*	*5.6*	*5.4*	*5.3*
$E(\mu) = \frac{s}{\beta}$	35.8	22.3	17.6	15.3	13.9	12.7	24.7	20.1	17.6	16.0
	116.1	*37.5*	*21.7*	*19.1*	*16.1*	*14.2*	*67.0*	*46.0*	*19.7*	*16.9*

MLE values are written in *italic*

periods. MCMC outperforms the MLE in the recovery of the true parameters in every respect, in particular in $E(\mu)$ for very small data sets with $N \leq 500$ and $T_{cal} \leq 18$ time units. This is mainly driven by the fact that the ML strives exclusively for the mode as the single desirable point of the likelihood, which is challenging when dealing with very uninformative data sets. In contrast to this, MCMC considers the entire distribution and is thus less prone to extreme outliers in case of very flat likelihood functions.

Further, Table 6 shows that the recovery of the purchase process parameters is more sensitive to the cohort size N than to the length of the calibration period T_{cal}. Opposed to this and in line with the results obtaines by Hoppe and Wagner (2010), the estimates of the dropout process are highly dependant on both N and T_{cal}.

4.5 Minimal data set requirements

Hoppe and Wagner (2010) derived the minimal data set requirements solely based on a criterion for $E(\lambda)$, disregarding the dropout process. Due to the low sensitivity of the purchase process to T_{cal}, they could restrict their limit recommendations to specifications of N. They determined the 95% quantile of the MAPE for $E(\lambda)$ for each combination of purchase frequency, dropout rate (as defined in Table 4), N, and T_{cal}.

We cannot fully replicate the minimal requirements of Hoppe and Wagner (2010) when applying the thresholds to our MLE values. Though the MAPE of the MCMC estimates outperforms MLE, the difference is not large enough to allow a relaxation of their minimal requirements on the cohort size. To the contrary, we receive even more restrictive data set limitations for medium and high purchase rates. Still, we enhance their data set specifications with conditions of $E(\mu)$ using the same method as for $E(\lambda)$. Lewis (1982) defines thresholds for the interpretation of the MAPE. The limit of 10% used by Hoppe and Wagner (2010) for $E(\lambda)$ corresponds to a highly accurate estimation. As the parameters of the dropout process are considerably more difficult to estimate, we use the MAPE threshold of 50% for a reasonably accurate estimation for $E(\mu)$. We apply these twofold thresholds to $E(\lambda)$ and $E(\mu)$

Table 7 Enhanced data set requirements

		Dropout rate		
		Low	Medium	High
Purchase frequency	Low	$N \geq 1,000$	$N > 1,500$	$N > 1,500$
		$T_{cal} \geq 24$	$T_{cal} \geq 18$	$T_{cal} \geq 24$
	Medium	$N \geq 1,000$	$N \geq 1,250$	$N > 1,500$
		$T_{cal} \geq 12$	$T_{cal} \geq 12$	$T_{cal} \geq 12$
	High	$N \geq 750$	$N \geq 1500$	$N > 1,500$
		$T_{cal} \geq 18$	$T_{cal} \geq 12$	$T_{cal} \geq 12$

with a type-I error of 5%. T_{cal} serves as a second dimension in the minimal requirements because of the high sensitivity of $E(\mu)$ estimate to the length of the calibration period.

Table 7 shows the new, combined data set requirements that result from applying the $E(\lambda)$ and $E(\mu)$ criteria to our MCMC estimates. These are based on the purchase and dropout rate defined in Table 4. The combinations of N and T_{cal} which are stated here use the lowest possible value of the cohort size. Hence, the requirements related to T_{cal} can be relaxed when N is increased.

5 Discussion and implications

Using MCMC algorithms for the Pareto/NBD model is a powerful tool and undoubtedly worth the additional implementation effort. However, the results from study 1 emphasise the necessity of a sensitivity analysis for the hyper parameters. The prior distributions that we derived in our studies might give a tentative idea of how in particular the CV of the heterogeneity parameters could be chosen. Still, our results do not replace the acquisition of prior information on the specific data sets. Regarding the choice of the hyper parameters, available information from different sources like experts, theories, or other data sets should be considered (Rossi and Allenby 2003).

Despite the application of two MLE routines, we received no results for 7.6% of the data sets, whereas the parameters of all data sets could be estimated with MCMC while simultaneously reducing the risk of extreme outliers. Therefore, using MCMC algorithms for CBA model parameter estimations in practice is less defective but still manageable concerning calculation time when being applied to a single data set only.

We contribute to extant literature by demonstrating that the parameter recovery and forecasting accuracy of Abe's algorithm is superior to other MCMC methods and to MLE. Moreover, it is the only procedure that generates the complete posterior distribution of the parameters on the individual as well as on the aggregate level and augments the unobserved individual dropout times $\{\tau_i\}$. Thus, it provides the entire range of available information. In particular compared to MLE, the use of Abe's algorithm herewith enriches the toolbox available to marketing management. Study 1 shows that using the individual parameters $\{\lambda_i, \mu_i\}$ rather than $\{r, \alpha, s, \beta\}$ on the aggregate level increases the accuracy of individual purchase forecasts. In addition, the provision of the full posterior distributions allows to explicitly account for (forecasting) uncertainty through confidence intervals or other distributional measures.

The individual level parameter values $\{\lambda_i, \mu_i\}$ enable managers to identify customers with a high purchase rate and a short estimated lifetime (i.e. large λ_i and μ_i) for individually targeted marketing activities like discount coupons or win-back campaigns. Moreover, the posterior distribution of these individual parameters can be used to calculate the distribution of the next customer purchase time, allowing management to use predefined quantiles for activity timing. In cases where point estimates are required, study 1 shows that the median draw of the posterior distribution outperforms the ML estimators. On the aggregate level, the

posterior distributions can be used to enrich the calculation of the CLV by considering uncertainty through confidence intervals.

Regardless of whether the parameters are estimated using MLE or MCMC, the size of the available data set plays a decisive role for the practical application of the Pareto/NBD model. In study 2, we amend the extant minimal data set requirements of Hoppe and Wagner (2010) which are restricted to the required cohort size by additionally considering a threshold based on the dropout process. Further, we expand the minimal requirements with a second dimension, including limitations for the minimal length of the calibration period.

Like any simulation-based research, there are limitations to the results of these studies. We were not able to fully replicate Hoppe and Wagner's (2010) results and found generally stricter cohort size requirements for medium and high purchase frequencies. This may, to a certain extent, be caused by the smaller number of replications (30 instead of 100) that we used to reduce the computational effort. In addition, these requirements refer to perfectly Pareto/NBD distributed data sets of a predefined parameter space. The examination of their validity for data sets which violate these prerequisites is left for future research. Within our simulation studies, we applied identical prior distributions to all data sets irrespective of the true heterogeneity parameter values, which lead to more conservative values of the performance metrics. When employing an MCMC procedure to a single real data set, we can assume that the prior distribution may fit better. We thus propose replicating study 2 with data set specific hyper parameters in future research. This might allow relaxing the minimal data sets requirements in case of more precise prior information. Furthermore, the influence of the hyper parameters should also be examined in the context of a decreasing data set size. Future research can also test the MCMC procedure not only against MLE but may use alternative parameter estimation approaches like quantile-based procedures.

A Technical appendix

This technical appendix is organised as follows. Chapter A1 contains different versions of the Pareto/NBD likelihood depending on the information available as well as $P(alive)$ which is defined as the probability that a customer is still active in $\{T_i\}$. In A2, we derive the (hyper) posterior distribution for the different MCMC procedures and describe their single steps in A3. In A4, we present the different methods for the future purchase forecast.

A1 General formulas

For an individual customer i, Abe (2009) presents some intermediate results for the Pareto/NBD model, which we use to explain the distributions we use for the different MCMC procedures.

If the customer is still active in T, the likelihood of the given purchase pattern is given by

$$L(\lambda, \mu | x, t_x, \tau > T) = \frac{\lambda^x t_x^{x-1}}{\Gamma(x)} e^{-(\lambda+\mu)T}. \tag{14}$$

If the customer has become inactive at a time $\tau \in (t_x, T]$, the likelihood is given by

$$L(\lambda, \mu | x, t_x, t_x \leq \tau < T) = \frac{\lambda^x t_x^{x-1}}{\Gamma(x)} \mu e^{-(\lambda+\mu)\tau}. \tag{15}$$

As Ma and Liu (2007) do not augment the dropout times $\{\tau_i\}$, they use the individual likelihood that Fader and Hardie (2005) noted as

$$L(\lambda, \mu | x, t_x, T) = \frac{\lambda^x}{\lambda + \mu} \left(\mu e^{-(\lambda+\mu)t_x} + \lambda e^{-(\lambda+\mu)T} \right). \tag{16}$$

The probability of a customer still being active in T was derived by Schmittlein et al. (1987) as

$$P\left(\tau > T | \lambda, \mu, x, t_x, T\right) = \frac{L(\lambda | x, t_x, \tau > T) \cdot P(\tau > T | \mu)}{L(\lambda, \mu | x, t_x, T)}. \tag{17}$$

A2 Posterior distributions

Bayes' theorem states that the posterior distributions π we draw from in our MCMC procedures are proportional to the product of the corresponding likelihood and the prior distribution. Depending on the information we have on τ, we receive the following formulas for the posterior distributions

- if $\tau > T$:

$$\pi(\lambda | x, t_x, \tau > T, \mu, r, \alpha) \propto L(\lambda, \mu | x, t_x, \tau > T) \cdot g(\lambda | r, \alpha)$$

$$\propto \lambda^{r+x-1} e^{-\lambda(\alpha+T)} \propto \Gamma(x + r, \alpha + T), \tag{18}$$

$$\pi(\mu | x, t_x, \tau > T, \lambda, r, \alpha) \propto L(\lambda, \mu | x, t_x, \tau > T) \cdot g(\mu | s, \beta)$$

$$\propto \mu^{s-1} e^{-\mu(\beta+T)} \propto \Gamma(s, \beta + T). \tag{19}$$

- if $t_x < \tau \leq T$:

$$\pi(\lambda | x, t_x, t_x < \tau \leq T, \mu, r, \alpha) \propto L(\lambda, \mu | x, t_x, t_x \leq \tau < T) \cdot g(\lambda | r, \alpha)$$

$$\propto \lambda^{r+x-1} e^{-\lambda(\alpha+\tau)} \propto \Gamma(x + r, \alpha + \tau), \tag{20}$$

$$\pi(\mu | x, t_x, t_x < \tau \leq T, \lambda, r, \alpha) \propto L(\lambda, \mu | x, t_x, t_x \leq \tau < T) \cdot g(\mu | s, \beta)$$

$$\propto \mu^{s} e^{-\mu(\beta+\tau)} \propto \Gamma(s + 1, \beta + \tau). \tag{21}$$

- if we have no information on τ:

$$\pi(\lambda | x, t_x, T, \mu, r, \alpha) \propto L(\lambda, \mu | x, t_x, T) \cdot g(\lambda | r, \alpha)$$

$$\propto \frac{\lambda^{x+r-1}e^{-\lambda\alpha}}{\lambda+\mu}\left(\mu e^{-(\lambda+\mu)t_x}+\lambda e^{-(\lambda+\mu)T}\right), \tag{22}$$

$$\pi(\mu|x,t_x,T,\lambda,r,\alpha) \propto L(\lambda,\mu|x,t_x,T)\cdot g(\mu|s,\beta)$$

$$\propto \frac{\lambda^{x+s-1}e^{-\mu\beta}}{\lambda+\mu}\left(\mu e^{-(\lambda+\mu)t_x}+\lambda e^{-(\lambda+\mu)T}\right). \tag{23}$$

The posterior distributions in (18) to (21) are proportional to gamma distributions. Therefore, if τ is known, we can draw values for λ and μ straight from these distributions without having to use sampling algorithms like Metropolis–Hastings or slice sampling which induce large computational cost when being applied for each customer.

To derive the posterior heterogeneity distributions, we assume that r, α, s, and β are themselves gamma distributed with $r \sim \Gamma(h_1,h_2), \alpha \sim \Gamma(h_3,h_4), s \sim \Gamma(h_5,h_6)$, and $\beta \sim \Gamma(h_7,h_8)$. We then receive the following joint distributions with $x = \{x_i\}, t_x = \{t_{x_i}\}$, and $T = \{T_i\}$ for the purchase and the dropout process, respectively: $\pi\left(r,\alpha|,x,t_x,T,\lambda,\mu,h_1,h_2,h_3,h_4\right)$

$$= \prod_{i=1}^{N}\left[L(\lambda_i,\mu_i|x,t_x,T)\cdot\frac{\alpha^r\lambda_i^{r-1}e^{-\lambda_i\alpha}}{\Gamma(r)}\right]\frac{h_2^{h_1}r^{h_1-1}e^{-rh_2}}{\Gamma(h_1)}\cdot\frac{h_4^{h_3}\alpha^{h_3-1}e^{-\alpha h_4}}{\Gamma(h_3)}$$

$$\propto r^{h_1-1}e^{-rh_2}\cdot\alpha^{h_3-1}e^{-\alpha h_4}\prod_{i=1}^{n}\frac{\alpha^r\lambda_i^{r-1}e^{-\lambda_i\alpha}}{\Gamma(r)}, \tag{24}$$

and $\pi\left(s,\beta|x,t_x,T,\lambda,\mu,h_1,h_2,h_3,h_4\right)$

$$= \prod_{i}^{N}\left[L(\lambda_i,\mu_i|x,t_x,T)\cdot\frac{\beta^s\mu_i^{s-1}e^{-\mu_i\beta}}{\Gamma(s)}\right]\frac{h_6^{h_5}s^{h_5-1}e^{-sh_6}}{\Gamma(h_5)}\cdot\frac{h_8^{h_7}\beta^{h_7-1}e^{-\beta h_8}}{\Gamma(h_7)}$$

$$\propto s^{h_5-1}e^{-sh_6}\beta^{h_7-1}e^{-\beta h_8}\prod_{i=1}^{N}\frac{\beta^s\mu_i^{s-1}e^{-\mu_i\beta}}{\Gamma(s)}. \tag{25}$$

We draw the parameter values from (24) and (25) by using a slice sampling routine.

Singh et al. (2009) use the NBD and Pareto distributions separately. For the purchase process with r and α, the full conditional distributions are given by $\Pi(r|x,t_x,T,\tau,\alpha,h_1,h_2)$

$$= \frac{h_2^{h_1}r^{h_1-1}e^{-rh_2}}{\Gamma(h_1)}\prod_{i=1}^{N}\frac{\Gamma(r+x_i)}{\Gamma(r)\Gamma(x_i+1)}\left(\frac{\alpha}{\alpha+min(\tau_i,T_i)}\right)^r\left(\frac{min(\tau_i,T_i)}{\alpha+min(\tau_i,T_i)}\right)^{x_i}, \tag{26}$$

$\Pi(\alpha|x,t_x,T,\tau,r,h_3,h_4)$

$$= \frac{h_4^{h_3}\alpha^{h_3-1}e^{-\alpha h_4}}{\Gamma(h_3)}\prod_{i=1}^{N}\frac{\Gamma(r+x_i)}{\Gamma(r)\Gamma(x_i)}\left(\frac{\alpha}{\alpha+min(\tau_i,T_i)}\right)^r\left(\frac{min(\tau_i,T_i)}{\alpha+min(\tau_i,T_i)}\right)^{x_i}. \tag{27}$$

For the dropout process with s and β, Singh et al. (2009) use the Pareto distribution. Therefore, the full conditional distributions are given by

$$\Pi(s|x,t_x,T,\tau,\beta,h_1,h_2) = \prod_{i=1}^{N} \frac{s\beta^s}{(\beta+\tau)^{s+1}} \frac{h_6^{h_5} s^{h_5-1} e^{-sh_6}}{\Gamma(h_5)} \propto \frac{s^{n+h_5-1}\beta^{sn} e^{-sh_6}}{\prod_{i=1}^{N}(\beta+\tau_i)^{s+1}}, \tag{28}$$

$$\Pi(\beta|x,t_x,T,\tau,\beta,h_1,h_2) = \prod_{i=1}^{N} \frac{s\beta^s}{(\beta+\tau_i)^{s+1}} \cdot \frac{h_8^{h_7} \beta^{h_7-1} e^{-\beta h_8}}{\Gamma(h_7)} \propto \frac{\beta^{ns+h_7-1} e^{-\beta h_8}}{\prod_{i=1}^{N}(\beta+\tau_i)^{s+1}}. \tag{29}$$

A3 MCMC procedure steps

A3.1 Abe

1. Initialise parameters $\{\lambda_i\}, \{\mu_i\}, \{\tau_i\}, r, \alpha, s, \beta$.
2. Using (18) for $\tau_i > T_i$ and from (20) for $t_x < \tau_i \leq T_i$, we can draw $\{\lambda_i\}$ straight from the combined gamma distribution
   $\lambda_i \sim \Gamma\big(x_i + r, \alpha + min\big(\tau_i, T_i\big)\big)$.
3. As outlined in (), we always use the specific value of $\{\tau_i\}$ and thus draw $\{\mu_i\}$ straight from the gamma distribution given in (21):
   $\mu_i \sim \Gamma(s+1, \beta+\tau_i)$.
4. Draw the aliveness vector $\{z_i\}$ using (17) with $P(z_i = 1) = P\big(\tau > T|\lambda_i, \mu_i, x_i, t_{x_i}, T_i\big) = \dfrac{L(\lambda_i|x_i, t_{x_i}, \tau_i > T_i) \cdot P(\tau_i > T_i|\mu_i)}{L(\lambda_i, \mu_i|x_i, t_{x_i}, T_i)}$ and $P(z_i = 0) = 1 - P(z_i = 1)$.
5. Draw $\tau = \{\tau_i\}$ using an exponential distribution for $z = 1$ and a double truncated exponential distribution for $z = 0$:
   $$\tau_i \sim \begin{cases} T_i + rexp(\mu_i), & z_i = 1 \\ -\dfrac{ln\Big[(1-runif)\cdot e^{-(\lambda_i+\mu_i)t_{x_i}} + unif \cdot e^{-(\lambda_i+\mu_i)T_i}\Big]}{\lambda_i+\mu_i}, & z_i = 0 \end{cases},$$
   where $runif \in [0,1]$ is a random number.
6. Draw r and α simultaneously from (24) using slice sampling:
   $(r,\alpha) \sim r^{h_1-1} e^{-rh_2} \cdot \alpha^{h_3-1} e^{-\alpha h_4} \prod_i \frac{\alpha^r \lambda_i^{r-1} e^{-\lambda_i \alpha}}{\Gamma(r)}$
7. Draw s and β simultaneously from (25) using slice sampling:
   $(s,\beta) \sim s^{h_5-1} e^{-sh_6} \beta^{h_7-1} e^{-\beta h_8} \prod_i \frac{\beta^s \mu_i^{s-1} e^{-\mu_i \beta}}{\Gamma(s)}$
8. Repeat from step 2 with updates parameter values.

A3.2 Ma and Liu

1. Initialise parameters $\{\lambda_i\}, \{\mu_i\}, r, \alpha, s, \beta$.
2. As Ma and Liu do not augment $\{\tau_i\}$, we draw $\{\lambda_i\}$ from (22) using a slice sampling routine:
$$\lambda_i \sim \frac{\lambda_i^{x_i+r-1} e^{-\lambda_i \alpha}}{\lambda_i+\mu_i}\left(\mu_i e^{-(\lambda_i+\mu_i)t_{x_i}} + \lambda_i e^{-(\lambda_i+\mu_i)T_i}\right)$$
3. Draw $\{\mu_i\}$ from (23) for the same reason using slice sampling:
$$\mu_i \sim \frac{\lambda_i^{x_i+s-1} e^{-\mu_i \beta}}{\lambda_i+\mu_i}\left(\mu_i e^{-(\lambda_i+\mu_i)t_{x_i}} + \lambda_i e^{-(\lambda_i+\mu_i)T_i}\right)$$
4. Draw r and α simultaneously from (24) using slice sampling:
$$(r, \alpha) \sim r^{h_1-1} e^{-rh_2} \cdot \alpha^{h_3-1} e^{-\alpha h_4} \prod_i^N \frac{\alpha^r \lambda_i^{r-1} e^{-\lambda_i \alpha}}{\Gamma(r)}$$
5. Draw s and β simultaneously from (25) using slice sampling:
$$(s, \beta) \sim s^{h_5-1} e^{-sh_6} \beta^{h_7-1} e^{-\beta h_8} \prod_i^N \frac{\beta^s \mu_i^{s-1} e^{-\mu_i \beta}}{\Gamma(s)}$$
6. Repeat from step 2 with updates parameter values.

A3.3 Singh et al.

1. Initialise parameters $\{\tau_i\}, r, \alpha, s, \beta$.
2. Draw a candidate vector $\left\{\tau_i^{(c)}\right\} > \{t_{x_i}\}$ from the truncated Pareto-II-distribution with parameters (s, β). This is given by
$$F_{\text{trunc}}(\tau_i) = \frac{F(\tau_i) - F(t_{x_i})}{1 - F(t_{x_i})} = \frac{\left[1 - \left(1 + \frac{\tau_i}{\beta}\right)^s\right] - \left[1 - \left(1 + \frac{t_{x_i}}{\beta}\right)^s\right]}{1 - \left[1 - \left(1 + \frac{\tau_i}{\beta}\right)^s\right]} = \left(\frac{t_{x_i} + \beta}{\tau_i + \beta}\right)^s - 1 \text{ and}$$
therefore inversion gives $\tau_i = (t_{x_i} + \beta)(1 - unif)^{-\frac{1}{s}} - \beta$.
3. Calculate the likelihoods for the old and new $\{\tau_i\}$-vector and accept the entry $\tau_i^{(c)}$ with $p_i = \dfrac{L(\tau_i^{(c)}|r, \alpha)}{L(\tau_i^{(c)}|r, \alpha) + L(\tau_i|r, \alpha)}$.
The acceptance probability can be reduced to
$$p_i = \frac{1}{1 + \left(\dfrac{\alpha + min(\tau_i^{(c)}, T_i)}{\alpha + min(\tau_i, T_i)}\right)^{r+x_i}}. \tag{30}$$
4. Draw r from (26) using slice sampling:
$$r \sim \frac{h_2^{h_1} r^{h_1-1} e^{-rh_2}}{\Gamma(h_1)} \prod_{i=1}^n \frac{\Gamma(r + x_i)}{\Gamma(r)\Gamma(x_i + 1)} \left(\frac{\alpha}{\alpha + min(\tau_i, T_i)}\right)^r \left(\frac{min(\tau_i, T_i)}{\alpha + min(\tau_i, T_i)}\right)^{x_i}.$$
5. Draw α from (27) using slice sampling:
$$\alpha \sim \frac{h_4^{h_3} \alpha^{h_3-1} e^{-\alpha h_4}}{\Gamma(h_3)} \prod_{i=1}^n \frac{\Gamma(r + x_i)}{\Gamma(r)\Gamma(x_i)} \left(\frac{\alpha}{\alpha + min(\tau_i, T_i)}\right)^r \left(\frac{min(\tau_i, T_i)}{\alpha + min(\tau_i, T_i)}\right)^{x_i}$$
6. Draw s from (28) using slice sampling:
$$s \sim \frac{\beta^{ns+h_7-1} e^{-\beta h_8}}{\prod_{i=1}^n (\beta + \tau_i)^{s+1}}$$

7. Draw β from (29) using slice sampling:
$$\beta \sim \frac{\beta^{ns+h_7-1}e^{-\beta h_8}}{\prod_{i=1}^{n}(\beta+\tau_i)^{s+1}}$$
8. Repeat from step 2 with updates parameter values.

A4 Determination of the purchase forecast x^*

The conditional expectation of the future purchases can be determined in different ways depending on the available parameters. Method 1 (Schmittlein et al. 1987) only uses the heterogeneity parameters $\{r, \alpha, s, \beta\}$ and can hence be used for all algorithms. It is provided in the BTYD package in R (Dziurzynski et al. 2014) and is given by Fader and Hardie (2005) as

$$E\left(x_i^* \middle| r, \alpha, s, \beta, x_i, t_{x_i}, T_i, T^*\right) = E\left(x_i^* \middle| r+x, \alpha+T_i, s, \beta+T, T^*\right) \cdot P\left(\tau_i > T_i \middle| r, \alpha, s, \beta, x_i, t_{x_i}, T_i\right). \quad (31)$$

Method 2 can be applied to Abe's algorithm only and is based on the individual form (Fader et al. 2005a) that requires values for $\{\lambda_i, \mu_i, z_i\}$ where $z_i = 1$ for $\tau_i > T_i$ and $z_i = 0$ for $\tau_i \leq T_i$:

$$E\left(x_i^* \middle| \lambda_i, \mu_i, \tau_i > T_i, T^*\right) = E\left(x_i^* \middle| \lambda_i, \mu_i, z = 1, T^*\right) = \frac{\lambda_i}{\mu_i} \cdot \left(1 - e^{-\mu_i T_i^*}\right). \quad (32)$$

Since Abe's procedure does not only provide values for z_i but distinct estimates for τ_i, we can modify (32) to

$$E\left(x_i^* \middle| \lambda_i, \mu_i, T_i, T^*\right) = \frac{\lambda_i}{\mu_i} \cdot \left(1 - e^{-\mu_i \cdot min(\tau_i, T_i^*)}\right). \quad (33)$$

Method 3 requires values for $\{\lambda_i, \mu_i\}$ only and can thus be applied to Abe's and Ma and Liu's procedure. We have no information on the status of the individual customer and therefore need to multiply (33) with $P(alive)$:

$$E\left(x_i^* \middle| \lambda_i, \mu_i, T^*\right) = E\left(x_i^* \middle| \lambda_i, \mu_i, \tau_i > T_i, T^*\right) \cdot P\left(\tau_i > T_i \middle| \lambda_i, \mu_i, t_{x_i}, T_i\right) = \frac{\frac{\lambda_i}{\mu_i}(1 - e^{\mu_i})}{1 + \frac{\mu_i}{\lambda_i+\mu_i}\left(e^{(\lambda_i+\mu_i)(T_i - t_{x_i})} - 1\right)}. \quad (34)$$

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abe M (2009) "Counting your customers" one by one: a hierarchical bayes extension to the Pareto/NBD model. Mark Sci 28:541–553

Batislam EP, Denizel M, Filiztekin A (2007) Empirical validation and comparison of models for customer base analysis. Int J Res Mark 24:201–209

Bemmaor AC, Glady N (2012) Modeling purchasing behavior with sudden 'death': a flexible customer lifetime model. Manag Sci 58:1012–1021

Chen X, Yin X (2019) NlcOptim: solve nonlinear optimization with nonlinear constraints. R package version 6. https://CRAN.R-project.org/package=NlcOptim. Accessed 12 Aug 2021

Dziurzynski L, Wadsworth E, McCarthy D (2014) BTYD: implementing buy til you die models. R package version 2:4. https://cran.r-project.org/web/packages/BTYD/. Accessed 12 Aug 2021

Eddelbuettel D (2013) Seamless R and C++ integration with Rcpp. Springer, New York

Eddelbuettel D, François R (2011) Rcpp : seamless R and C++ integration. J Stat Softw 40:1–18

Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. Psychol Rev 70:193

Fader PS, Hardie BG (2005) A note on deriving the Pareto/NBD model and related expressions. http://brucehardie.com/notes/009/. Accessed 15 Apr 2020

Fader PS, Hardie BG, Lee KL (2005a) "Counting your customers" the easy way: an alternative to the Pareto/NBD model. Mark Sci 24:275–284

Fader PS, Hardie BG, Lee KL (2005b) RFM and CLV: using iso-value curves for customer base analysis. J Mark Res 42(4):415–430

Glady N, Baesens B, Croux C (2009) A modified Pareto/NBD approach for predicting customer lifetime value. Expert Syst Appl 36(2):2062–2071

Gupta S, Hanssens D, Hardie B, Kahn W, Kumar V, Lin N, Sriram S (2006) Modeling customer lifetime value. J Serv Res 9:139–155

Hoppe D, Wagner U (2010) Small sample properties of the Pareto/Negative binomial distribution model. Mark ZFP 32:39–50

Jain D, Singh SS (2002) Customer value research in marketing: a review and future directions. J Interact Mark 16:34

Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc R Soc Lond Ser A Math Phys Sci 186:453–461

Jerath K, Fader PS, Hardie BG (2011) New perspectives on customer "death" using a generalization of the Pareto/NBD model. Mark Sci 30:866

Kim S, Kim H (2016) A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast 32:669–679

Lewis CD (1982) Industrial and business forecasting methods. Butterworths, London

Ma S, Büschken J (2011) Counting your customers from an "always a share" perspective. Mark Lett 22:243–257

Ma SH, Liu JL (2007) The MCMC approach for solving the Pareto/NBD model and possible extensions. Third Int Conf Nat Comput (ICNC 2007) 2:505–512

McCarthy DM, Fader PS (2018) Customer-based corporate valuation for publicly traded noncontractual firms. J Mark Res 55(5):617–635

Neal R (2003) Slice sampling. Ann Stat 31:705–767

Paap R (2002) What are the advantages of MCMC based inference in latent variable models? Stat Neerl 56:2–22

Platzer M (2016) BTYDplus: Probabilistic models for assessing and predicting your customer base. R package version 1:1. https://CRAN.R-project.org/package=BTYDplus. Accessed 12 Aug 2021

Platzer M, Reutterer T (2016) Ticking away the moments: timing regularity helps to better predict customer activity. Mark Sci 35:779–799

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/. Accessed 12 Aug 2021

Robert C (2007) The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer, New York
Rossi PE, Allenby GM (2003) Bayesian statistics and marketing. Mark Sci 22:304–328
Schmittlein DC, Morrison DG, Colombo R (1987) Counting your customers: who are they and what will they do next? Manag Sci 33:1–24
Schmittlein DC, Peterson RA (1994) Customer base analysis: an industrial purchase process application. Mark Sci 13:41–67
Schweidel DA, Park YH, Jamal Z (2014) A multiactivity latent attrition model for customer base analysis. Mark Sci 33:273–286
Singh SS, Borle S, Jain DC (2009) A generalized framework for estimating customer lifetime value when customer lifetimes are not observed. Quant Mark Econ 7:181–205
Tanner M, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–540