sBG-MCMC

Aaron D. Ramsey

January 2024

1 Introduction

In contractual settings, understanding customer attrition is crucial. Arguably, it is the duty of the manager to harness the full potential of attrition rates, given the privilege of its observation. Supplementing attrition is its counterpart, retention, commonly employed in calculating customer tenure and lifetime value. These metrics, vital for businesses, must be completely owned and understood.

Commonly, retention (the proportion of customers retained between successive periods) is assumed to be constant across a cohort. This is differentiated from what we will call the survival rate—the proportion of the total population alive in a given period. This *good-enough* approach, however, leads to a gross underestimation of tenure and value. At first glance, it may seem logical, but a more thorough examination reveals its weaknesses. A constant rate of retention requires a uniform attrition rate among a cohort's customers, irrespective of the observation period. This, however, defies intuition. It seems implausible that a customer who has resubscribed to Netflix for the 30th time would have the same likelihood of leaving in the 31st period as a new subscriber in their first. Empirically, it is even shown retention rates increase over time.

Accordingly, some would argue that these customers are *getting better*, but this contradicts the nature of human behavior. People don't aspire to become better consumers; they are simply guided by their own, seemingly random, desires. From our perspective, the *good customers* had always been good; identifying them, however, required observation. Customers are drawn to try new products; and, if in agreement with their preferences, they continue that relationship with the business. Such preferences remain unchanged (assuming stationarity) by the mere act of trying a product. Just as a coin landing on tails to infinity will never happen, all customers will churn, particularly if an alternative product better aligns with their preferences or current necessities.

How can we ascertain what actions our customers will take, or more aptly, how do we discern the assortment of customers we possess? Each customer, acting independently, obscures the underlying factors influencing their behavior. Yet, we endeavor to do our best. Fortunately, we have a proxy for determining customers with similar behavior patterns—the duration of our relationship.

To harness this relationship/churn data and our beliefs of customer behavior, we will employ the shifted beta-geometric (sBG) distribution (Fader and Hardie, 2007). The sBG has a hierarchical structure. On the individual level—the lowest—we model customer tenure

using the geometric distribution, particularly the shifted variant. To briefly summarize, geometric distribution aligns with our understanding of customer attrition. It is parameterized by θ (churn propensity), which encodes the various unobservable influences impacting a customer's decision to churn—product preferences, personality traits, occupational hazards, etc. Given we know an individual's true θ, we can calculate the probability of that customer surviving through period t: $p(T > t|\theta)$.

One level higher, the customer's theta is itself characterized by a beta distribution. This distribution accounts for our intuition regarding customer heterogeneity. It represents the spectrum of θ values among our customers and is characterized by two hyperparameters, α and β. The versatile shapes of this distribution, and the corresponding retention curves they imply, can be examined through the interactive graph here[1].

A result of modeling fixed churn rates are increasing retention rates. It's an artifact of the drop-out effect of bad customers. Accordingly, it would be ill-advised to only consider retention rates when analyzing one's customer base. This is evident in the survival curve, which reveals that customer cohorts with high churn rates rapidly vanish, despite achieving significant retention in later stages. In contrast, cohorts with a lower churn propensity retain a substantial portion of their initial base while still attaining high retention. Typically, companies excessively inclined towards the higher-churn segment of the distribution either cease to exist or adapt, attracting more high-value customers.

2 Methodology

The motivation behind this analysis is straightforward. Having previously implemented the sBG employing Empirical Bayes, a compelling desire to explore a more comprehensive approach emerged. My methodology features the fully Bayesian interpretation of the model; and, to benchmark, I will include results from the Empirical Bayes method (one whose simplicity allows for implementation in Excel).

2.1 Bayesian Interpretations

Understanding the implication of a Bayesian interpretation is important when designing the architecture of a model. Generally, we consider all unknown parameters as random variables with assigned probabilities for their possible values. This structure is then integrated with observed data to produce an updated estimate of our *belief* in model parameters—the posterior distribution. Choosing prior distributions that are also parametric creates a hierarchical structure (such as our sBG), which complicates the posterior distribution. Such complications lead to complicated posterior distributions with no standard form, requiring excessive resources to compute. In order to effectively sample these distributions, algorithms such as Random-Walk Metropolis (RWM) and, more recently, Hamiltonian Monte Carlo (HMC) are used to generate samples from complex posterior distributions. Alternatively, relaxed Bayesian assumptions also provide an excellent alternative without requiring intricate algorithms. Empirical Bayes is one such result of partially relaxing our Bayesian assumptions.

[1]For printed versions, the url is https://aarondaelramsey.com/academics/statistics/sBG/#betaDist

Under a Bayesian interpretation, any unknown parameters are considered random variables. Parallel with observed data, we can choose to imbue these parameters with our subjective beliefs, codified by prior distributions. The result is an updated estimation of our belief in the parameters' values, as given by a joint posterior distribution. For example, the joint distribution for an individual can be calculated using Bayes' rule:

$$\begin{aligned} p(\theta, \alpha, \beta|y) &= \frac{p(y|\theta, \alpha, \beta) \cdot p(\theta, \alpha, \beta)}{p(y)} \\ &= \frac{p(y|\theta, \alpha, \beta) \cdot p(\theta|\alpha, \beta) \cdot p(\alpha, \beta)}{p(y)} \\ &= \frac{p(y|\theta) \cdot p(\theta|\alpha, \beta) \cdot p(\alpha, \beta)}{p(y)} \end{aligned}$$

In fact, this joint distribution is exactly what the fully Bayesian approach uses. Here we also show the expanded form of the evidence—$p(y)$.

$$p(\theta, \alpha, \beta|y) = \frac{p(y|\theta) \cdot p(\theta|\alpha, \beta) \cdot p(\alpha, \beta)}{\int\int\int p(y|\theta) \cdot p(\theta|\alpha, \beta) \cdot p(\alpha, \beta)\, d\alpha\, d\beta\, d\theta}$$

The integral(s) in the denominator is not always tractable; and, unfortunately, given our model's structure, we will not be able to find a closed-form solution. Luckily the desired result, $p(\theta, \alpha, \beta|y)$, depends only on the values of θ, α, and β. Therefore, removing any terms without dependencies on the parameters keeps the result proportional to the original distribution. Having already integrated out these parameters, the bottom term of the posterior (the evidence) depends only on y and can be removed. The result is a distribution proportional to the original up to a normalizing constant.

$$p(\theta, \alpha, \beta|y) \propto p(y|\theta) \cdot p(\theta|\alpha, \beta) \cdot p(\alpha, \beta)$$

If we, for instance, were only inerested in estimating the thetas of our model (individual churn rates), we could integrate our proportional posterior over the population parameters α and β:

$$\begin{aligned} \int\int p(\theta, \alpha, \beta|y)\, d\alpha\, d\beta &\propto \int\int p(y|\theta) \cdot p(\theta|\alpha, \beta) \cdot p(\alpha, \beta)\, d\alpha\, d\beta \\ p(\theta|y) &\propto p(y|\theta) \cdot \int\int p(\theta|\alpha, \beta) \cdot p(\alpha, \beta)\, d\alpha\, d\beta \end{aligned}$$

Unfortunately, we are again in the same position from before; there is no closed form solution to this integral. A common solution for such issues is the Gibbs sampler. The Gibbs sampler is an iterative algorithm relying on the conditional distributions of the model parameters. It is in the family of Markov Chain Monte Carlo samplers and easy to both understand and implement. Unfortunately, for our specific model, the conditional distribution of α and β do not have a standard form, a requirement for the Gibbs sampler. The most popular extension addressing such concessions, Random-Walk Metropolis, proposes samples

for parameters lacking standard conditional distributions from standard ones. The Gaussian is a common choice given its symmetry. The proposed samples are then accepted or rejected according to an acceptance criterion. Another class of algorithms, subset of MCMC, is Hamiltonian Monte Carlo. Both algorithms will be discussed and implemented later in this paper.

A less complicated approach is Empirical Bayes. It's a procedure for performing statistical analysis/inference on data where the prior distribution is determined by the observed data—thus the term empirical. This is in contrast to the standard fully Bayesian approach where the prior beliefs are fixed before any data are observed. It is considered a good-enough approximation to the fully Bayesian treatment of our hierarchical model.

Instead of allowing the highest level model parameters to vary, they are set to their most likely values. To find these values, one needs to maximize the marinal likelihood $p(y|\alpha, \beta)$ with respect to the population-level parameters. This distinction contributes to Empirical Bayes' other name—maximum marginal likelihood. This structure greatly reduces computational complexity allowing analysis to be easily performed in Excel. With the advent of efficient and effective computational techniques, the fully Bayesian approach has mostly replaced Empirical Bayes methods; however, its ease of use, interpretability, and predictive power can neither be overstated nor overlooked.

Once found, the values of α and β can be used to find the conditional posterior distribution of θ. The parameters α and β are fixed points. From a Bayesian perspective, this is the same as conditioning on the parameters desired to be fixed.

$$\begin{aligned} p(\theta|y, \alpha, \beta) &= \frac{p(y|\theta) \cdot p(\theta|\alpha, \beta)}{p(y|\alpha, \beta)} \\ &= \frac{p(y|\theta) \cdot p(\theta|\alpha, \beta)}{\int p(y|\theta) \cdot p(\theta|\alpha, \beta) d\theta} \end{aligned}$$

2.2 Data and Model

The example data consist of 1,000 customers and their corresponding churn periods, which have been censored at the 7^{th} period. That is, it is unknown when each customer surviving through the 7^{th} period churned. Each customer's observed churn period, denoted y_i, is assumed to follow a Geometric distribution with a corresponding parameter representing churn propensity, θ_i. This structure implies an independent, constant retention rate for each customer. Each θ_i is assumed to be drawn from the same Beta distribution, described by parameters, α and β. Such structure can be represented by the following formulas:

$$\begin{aligned} y_i &\sim \text{Geom}(\theta_i) \\ \theta_i &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

A helpful analogy for internalizing this structure is to imagine each customer within a cohort (in our example the original 1000 customers acquired in the same period) reaching into a bag of coins and selecting one. The probability of this coin landing on heads is their propensity to churn (θ_i). The mix of coins within the bag follows a Beta distribution; and,

coin selection is only done once, at the beginning of the customer relationship. At each period, the customers flip their coins, and those whose coin lands on heads, churn. These constraints again underscore the two main assumptions of the model: each period, a customer decides to leave the firm with a personal, constant probability θ_i (stationarity) where each θ_i was generated by a Beta distribution at the cohorst inception. The Beta distribution is a convenient and intuitive choice as it exhibits conjugacy with the Geometric distribution and is bounded between 0 and 1.

In order for us to estimate these parameters, we need to incorporate the data with our model. Normally, it would involve calculating the product of likelihoods and prior distributions; however, recall that our data are right-censored at the 7th period. This structure necessitates the use of two different likelihood formulas: one based on the geometric's pdf; the other, its survival function:

$$\begin{aligned}
p(\boldsymbol{\theta}, \alpha, \beta | \boldsymbol{y}) &\propto p(\boldsymbol{y} | \boldsymbol{\theta}, \alpha, \beta) \cdot p(\boldsymbol{\theta}, \alpha, \beta) \\
&= p(\boldsymbol{y} | \boldsymbol{\theta}, \alpha, \beta) \cdot p(\boldsymbol{\theta} | \alpha, \beta) \cdot p(\alpha, \beta) \quad \text{where } p(\alpha, \beta) \propto 1 \\
&= \left(\prod_{i=1}^{1000} \left(p(y_i | \theta_i, \alpha, \beta) p(\theta_i | \alpha, \beta) \right) \right) \cdot p(\alpha, \beta) \\
&\propto \prod_{i=1}^{1000} \left(p(y_i | \theta_i) p(\theta_i | \alpha, \beta) \right) \\
&= \prod_{i=1}^{759} \left(\theta_i (1-\theta_i)^{y_i - 1} \frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{B(\alpha, \beta)} \right) \prod_{i=760}^{1000} \left((1-\theta_i)^7 \frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{B(\alpha, \beta)} \right) \\
&= \left(\frac{1}{B(\alpha, \beta)} \right)^{1000} \left(\prod_{i=1}^{759} \left(\theta_i^{\alpha} (1-\theta_i)^{\beta + y_i - 2} \right) \prod_{i=760}^{1000} \left(\theta_i^{\alpha-1} (1-\theta_i)^{\beta+6} \right) \right)
\end{aligned}$$

It should be noted that I have chosen α and β to have a flat prior for simplicity. The result is $p(\alpha, \beta) \propto 1$. Additionally, it's important to note that I have hard-coded y_i for the censored cell at $t = 7$ in the second product. This is specific to this dataset, and I want to make it clear that this value doesnt change in the second product. Generally, the the survival function for the geometric is $p(T > t) = (1-\theta_i)^t$; however, this may cause confusion for some users. One can reparameterize the survival function to be one more than the last observed date (8 in our case) to make implementation in Python more straightforward. Additionally, the indices of the products are also hardcoded to values specific to our dataset. To generalize the posterior, a data-agnostic formula can be given by:

$$p(\boldsymbol{\theta}, \alpha, \beta | \boldsymbol{y}) \propto \left(\frac{1}{B(\alpha, \beta)} \right)^{N} \left(\prod_{i=1}^{N_u} \left(\theta_i^{\alpha} (1-\theta_i)^{\beta + y_i - 2} \right) \prod_{i=N_u+1}^{N} \left(\theta_i^{\alpha-1} (1-\theta_i)^{\beta + y_i - 1} \right) \right)$$

$$p(\boldsymbol{\theta}, \alpha, \beta | \boldsymbol{y}) \propto \left(\frac{1}{B(\alpha, \beta)} \right)^{N} \left(\prod_{i=1}^{N_u} \left(\theta_i^{\alpha} (1-\theta_i)^{\beta + y_i - 2} \right) \prod_{i=N_u+1}^{N} \left(\theta_i^{\alpha-1} (1-\theta_i)^{\beta + y_i - 2} \right) \right)$$

Where N represents the total number of customers in the data, N_u represents the number of customers existing in the uncensored cells, and the top and bottom equations represent

respectively the two previously mentioned interpretations of the survival function. For clarity, I will be using the former definition of the function in all subsequent formulae.

2.3 Random-Walk Metropolis

We now have the form of our joint posterior, so we need to sample from it. Unfortunately, due to the complications previously discussed, such a task is not so straigtforward. This particular solution will employ the Metropolis algorithm, a Markov chain Monte Carlo method. This algorithm allows one to find samples of each parameter by conditioning on the current samples of the others. It also allows for some parameters to possess non-standard conditional distributions.

The initial step involves determining the conditional posteriors of all our parameters. This is achieved by disregarding terms that include the parameters on which we are conditioning. Consequently, the resulting distribution is proportional to its true value. For example, the parameters $\boldsymbol{\theta}$ have conditional posterior distributions proportional to the following equations:

$$p(\theta_i|\theta_{j\neq i},\alpha,\beta,\boldsymbol{y}) \propto \begin{cases} \theta_i^{\alpha}(1-\theta_i)^{\beta+y_i-2} & \text{for } 1 \leq i < 760 \\ \theta_i^{\alpha-1}(1-\theta_i)^{\beta+6} & \text{for } i \geq 760 \end{cases}$$

We recognize these as the functional forms of the Beta distribution. We will see these again later when discussing empirical Bayes:

$$\theta_i|\theta_{j\neq i},\alpha,\beta,\boldsymbol{y} \sim \begin{cases} \text{Beta}(\alpha+1,\beta+y_i-1) & \text{for } 1 \leq i < 760 \\ \text{Beta}(\alpha,\beta+7) & \text{for } i \geq 760 \end{cases}$$

Similarly, by ignoring elements of the joint distribution not reliant on α and β, a joint conditional distribution can be calculated. Unfortunately, this results in a non-standard distribution, which will require the Metropolis step of our algorithm to find samples from:

$$p(\alpha,\beta|\boldsymbol{y},\boldsymbol{\theta}) \propto \prod_{i=1}^{1000}\left(\frac{\theta_i^{\alpha}(1-\theta_i)^{\beta}}{B(\alpha,\beta)}\right)$$

In an iteration of the Metropolis algorithm, samples are obtained from each conditional distribution using the current samples (or previous samples, depending on one's perspective) Each time a sample for a parameter is approved, it becomes the current sample for that parameter, which is then used in subsequent samplings. The scheme below illustrates the process of sampling using this algorithm.

1. Initialize the parameters $\boldsymbol{\theta}$, α, and β, the starting points in the state space, and choose a symmetric proposal distribution $q(x^*|x^{t-1})$ for your parameters generated by non-standard conditional distributions.

2. For each iteration $t = 1, 2, \ldots, T$:

   (a) Sample the parameters $\boldsymbol{\theta}$ using
   $$\theta_i|\theta_{j\neq i},\alpha,\beta,y_i \sim \begin{cases} \text{Beta}(\alpha+1,\beta+y_i-1) & \text{for } 1 \leq i < 760, \\ \text{Beta}(\alpha,\beta+7) & \text{for } i \geq 760 \end{cases}$$

(b) Generate a proposal x^* from the proposal distribution $q(x^*|x^{t-1})$ for both α and β. Where σ_α and σ_β are tuning parameters for the algorithm. (See Appendix A.1 for information about non-symmetric proposal distributions)

i. $\alpha^* \sim \text{Normal}(\alpha^{t-1}, \sigma_\alpha)$

ii. $\beta^* \sim \text{Normal}(\beta^{t-1}, \sigma_\beta)$

(c) Input the proposed values into their non-standard joint distribution and compare to the current values using a ratio r.

i. $r = \frac{p(\alpha^*, \beta^*|\boldsymbol{y}, \boldsymbol{\theta})}{p(\alpha^{t-1}, \beta^{t-1}|\boldsymbol{y}, \boldsymbol{\theta})}$

(d) We accept the proposed values with probability r. (i.e. when r is greater than u where u $\sim$ Uniform$(0, 1)$)

(e) If accepted, our current sample values are updated to the proposed values. Otherwise, the old samples remain.

Sometimes, it is computationally useful to work with the logarithms of our distribution. Utilizing the properties of logarithms, r can be represented as a difference of logarithms and compared to $\log(u)$:

$$
\begin{aligned}
p(\alpha, \beta|\boldsymbol{y}, \boldsymbol{\theta}) &\propto \prod_{i=1}^{1000} \left(\frac{\theta_i^\alpha (1-\theta_i)^\beta}{B(\alpha, \beta)} \right) \\
&= B(\alpha, \beta)^{-1000} \cdot \prod_{i=1}^{1000} \left(\theta_i^\alpha (1-\theta_i)^\beta \right) \\
\log(p(\alpha, \beta|\boldsymbol{y}, \boldsymbol{\theta})) &\propto \log(B(\alpha, \beta)^{-1000}) + \log\left(\prod_{i=1}^{1000} \theta_i^\alpha (1-\theta_i)^\beta \right) \\
&= -1000 \log(B(\alpha, \beta)) + \log(\theta_1^\alpha \cdot \ldots \cdot \theta_{1000}^\alpha \cdot (1-\theta_1)^\beta \cdot \ldots \cdot (1-\theta_{1000})^\beta) \\
&= -1000 \log(B(\alpha, \beta)) + \log(\theta_1^\alpha) + \ldots + \log(\theta_{1000}^\alpha) \\
&\quad + \log((1-\theta_1)^\beta) + \ldots + \log((1-\theta_{1000})^\beta) \\
&= -1000 \log(B(\alpha, \beta)) + \alpha \log(\theta_1) + \ldots + \alpha \log(\theta_{1000}) \\
&\quad + \beta \log(1-\theta_1) + \ldots + \beta \log(1-\theta_{1000}) \\
&= -1000 \log(B(\alpha, \beta)) + \alpha \sum_{i=1}^{1000} (\log(\theta_i)) + \beta \sum_{i=1}^{1000} (\log(1-\theta_i)) \\
&= \alpha \sum_{i=1}^{1000} (\log(\theta_i)) + \beta \sum_{i=1}^{1000} (\log(1-\theta_i)) - 1000 \log(B(\alpha, \beta))
\end{aligned}
$$

$$
\log(r) = \log(p(\alpha^*, \beta^*|\boldsymbol{y}, \boldsymbol{\theta})) - \log(p(\alpha^{t-1}, \beta^{t-1}|\boldsymbol{y}, \boldsymbol{\theta}))
$$

2.4 Hamiltonian Monte Carlo

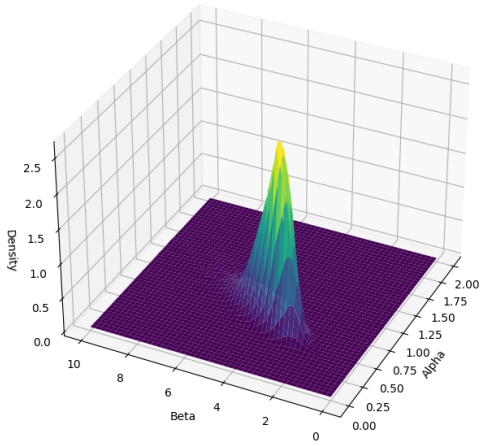
In most cases, using RWM to generate samples from a posterior works well; however, it isn't without its issues. Namely, RWM can and will make disproportionately many proposals outside of the typical set compared to how likely those points are to exist. Fortunately, the accept/reject criteon in RWM ensures that we don't stray from the typical set in a way that alters the characteristics of the posterior. Higher rejection, however, decreases the effectiveness of the algorithm to explore the typical set. In multiple dimensions, the tuning parameter is generally set at a value yielding a 23% acceptance rate. We see this first hand in the sBG. The algorithm struggles to diffuse through the posterior space; and, in fact, the auto-correlation is so high, roughly 800,000 samples were needed across our four chains to achieve an effective sample size of 1000. This is neither reasonable nor practical. Therefore, we must look towards a more reasonable solution—Hamiltonian Monte Carlo.

Hamiltonian Monte Carlo (HMC) is a class of algorithms which uses Hamiltonian trajectories to generate samples from a posterior distribution. One of earliest implementations, Hybrid Monte Carlo, generated samples using Hamiltonian trajectories integrated for some constant time ($L * \epsilon$) to generate the metropolis proposals instead of using a proposal distribution. While Hybrid Monte Carlo outperformed RWM, there were still many inefficiencies and nuisances within its implementation. Among those were the tuning of model parameters L (number of leapfrog steps) and ϵ (step size), providing gradients for trajectory calculations, and the waste of information contained within the trajectory's intermediate points. Without significant HMC experience, tuning these parameters optimally was a non-trivial task. Many posterior distributions can exhibit pathological behaviors with ill-set tuning parameters, especially those of high dimension—a feature which ironically is the appeal of the HMC algorithm.

In Gelman and Hoffman (2012) a new HMC algorithm named the No U-Turn Sampler (NUTS) was proposed. It featured automatically controlled tuning parameters, included information from trajectories' intermediate points, and offered a solution for the provision of gradients required by the algorithm. The specifics of implementing NUTS will be discussed later, so for now lets try to understand how this class of algorithms works.

2.4.1 Intuition

Intuition for NUTS/HMC can be more easily distilled by the construction of a physical system. In this system, assume a position in $\mathbb{R}^3$ to be represented by a small particle (or ball if it helps one's visualization) on a frictionless surface whose height is represented by our posterior distribution. Accordingly, the projection of this ball onto the xz plane represents the position in parameter space–known as the configuration space mechanics literature. This construction can sometimes affectionately be called Mount Likelihood.



In standard machine learning algorithms, the goal is to descend/ascend using a function's gradients to guide the algorithm to some maximum or minimum. That, however, is not our goal. We want to incorporate the posterior's gradient information to more efficiently explore all of the typical set not only a mode of the distribution. In simple terms, we want to integrate physics into our system—thus the introduction as a physical system.

As it stands, our physical construction does not afford us any conveniences. Placing our ball on the constructed surface would have the ball slide (no rolling without friction) away, having never explored regions of higher density. A solution could be to imbue the ball with some momentum in the direction of higher density; however, there are two considerations. Firstly, there would need to be a determined direction in which to give the momentum. Secondly, assuming the ball were sent towards a high density region, it would not remain there for a proportional length of time. Similar to before and without one's intervention, it would slide down the slope to disparate regions outside the typical set. As a result a stopping criterion of some sophistication would need to be applied. In either case, signicant bias would be introduced to the system.

The solution, at a high level, is to invert the posterior; so, instead of looking at Mount Likelihood, we observe the Ravine of Rareness. The more likely a set of parameters, the more negative the height. This ensures, at any time, the particle tends toward a local minimum. There need not be any concern of the particle sliding to distant points in space. So the problem is solved, and we need only to place a particle anywhere on the curve and it will tend towards high-density regions. We just need to solve Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \quad \text{and} \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}$$

$$\text{where} \quad H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{p}^T M^{-1} \boldsymbol{p}$$

The daunting task of finding a solution to these equations must be achieved through numerical integration. That aside, we currently have only a naive understanding of what

physics we are inducing on the system. As a result of introducing Hamiltonian dynamics, we have implicitly augmented our space with D additional parameters. For this example, there are now four—two for position ($\boldsymbol{q}$) and two for their corresponding momenta ($\boldsymbol{p}$). The bottom equation, known as the Hamiltonian, incorporates all four parameters in the augmented space. It is the sum of the potential and kinetic energies contained within the system. The potential energy can be written as the height of the particle, which is generally given as the inverted log of the posterior. The choice of kinetic enegry, is somewhat up to us, though care must be taken in its choice as not all kinetic energies are valid or useful. The standard choice is known as the Euclidean-Gaussian kinetic energy.

To further describe the physical implications of HMC, the only forces acting upon the particle are gravity and the reacation from the posterior's surface—again no friction. Accordingly, the total energy of the system remains constant along the path defined by these equations. Viewed from the perspective of our augmented system, these paths of constant energy define level sets of H:

$$H^{-1}(E) = \{\boldsymbol{q}, \boldsymbol{p} \mid H(\boldsymbol{q}, \boldsymbol{p}) = E\}$$

Put into plain English this represents the set of all inputs of the ordered pair $(\boldsymbol{q}, \boldsymbol{p})$ (points in position-momentum phase space) that when used to evaluate the Hamiltonian, results in the same constant energy E. A result of the particle's movements being confined to a specific level set is its tendency to return to its initial position—think of a swinging pendulum. Such behavior would offer poor exploration of the posterior; and, more importantly, without updating the level set the particle exists on, it will only continue traveling through the same path each time we calculate its trajectory. We need a procedure to transition the particle between level sets. Fortunately, our choice of kinetic energy is a result of constructing a probability distribution over the momentum. For a more in-depth exploration of these topics, Betancourt (2017) provides incredible insights and explanations for anyone interested in a more rigorous introduction to HMC. For our purposes, this choice of kinetic energy allows us to resample momenta from a Gaussian distrbibution following each sample of the model parameters.

To summarize the resultant structure, we have first augmented our model with auxiliary parameters representing the momentum, $\boldsymbol{p}$, of the particle in each dimension of the parameter space. We then separate the sampling process of the algorithm into two distinct steps. The first is a deterministic exploration of an energy level using Hamilton's equations. The second is a stochastic resampling/updating of our momenta which places the particle on a new level set to explore. Instead of diffusing inefficiently through parameter space alone, we now only diffuse through the various level sets. After each sampling of momenta, the algorithm relies on a predetermined path to make large jumps from the initial point. The efficiency of the deterministic step can be controlled by tuning the L and ϵ parameters. For example, the NUTS algorithm determines L as the trajectory is being built. It uses a stopping criteon that halts the numerical integrator once the trajectory begins to turn around on itself—hence, the No U-Turn element of its name. Secondly, it uses a dual averaging algorithm (Nesterov, 2009) allowing for an optimal ϵ to be set during the warm-up period.

2.4.2 Implementation

Talk about using torch for autodiff. Discuss that I took a log transformation of the parameters to help with autocorr and ensuring alpha, beta > 0. Talk about how I used and implemented an optimal metric.

Implementing HMC of any sort can be arduous. Before implementing, whether you are using bespoke code, STAN, pymc, etc. I would recommend in addition to the Betancourt paper cited earlier as well as the original NUTS paper reading Neal (2011). The code contained within this projects repository on github contains my implementation. It deviates from the NUTS implementation and more closely resembles the scheme laid out by Betancourt—what is currently implemented by STAN.

2.5 Empirical Bayes

My coverage of Empirical Bayes will be brief. I have distilled the vital components for the purpose of comparison. Fader and Hardie (2007) offers an in-depth derivation and handling of the Empirical Bayes implementation of the sBG within Excel.

Our goal is to find point estimates of α and β that maximize the likelihood of the data given the model. In our case, the model is the marginal likelhood function of the data—$p(y|\alpha, \beta)$.

$$
\begin{aligned}
p(\theta|y, \alpha, \beta) &= \frac{p(y|\theta) \cdot p(\theta|\alpha, \beta)}{p(y|\alpha, \beta)} \\
&= \frac{p(y|\theta) \cdot p(\theta|\alpha, \beta)}{\int p(y|\theta) \cdot p(\theta|\alpha, \beta) d\theta} \\
\int_0^1 p(Y = y|\theta) \cdot p(\theta|\alpha, \beta) d\theta &= \int_0^1 \theta(1-\theta)^{y-1} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} d\theta \\
p(Y = y|\alpha, \beta) &= \frac{1}{B(\alpha, \beta)} \cdot \int_0^1 \theta(1-\theta)^{\beta+y-2} d\theta \\
&= \frac{B(\alpha+1, \beta+y-1)}{B(\alpha, \beta)}
\end{aligned}
$$

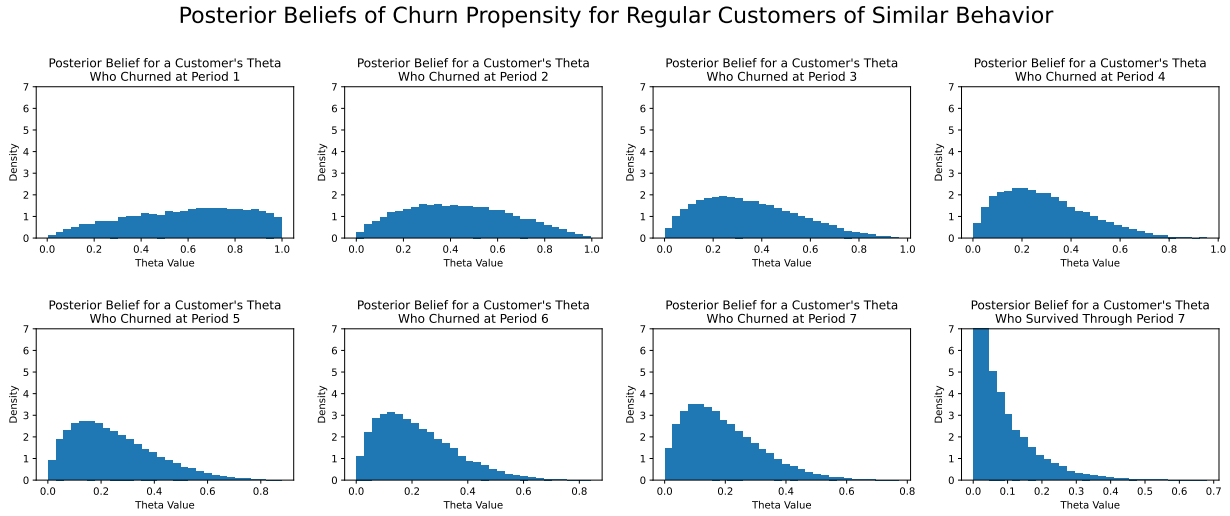
A similar process can be used to find $p(Y > y|\alpha, \beta)$

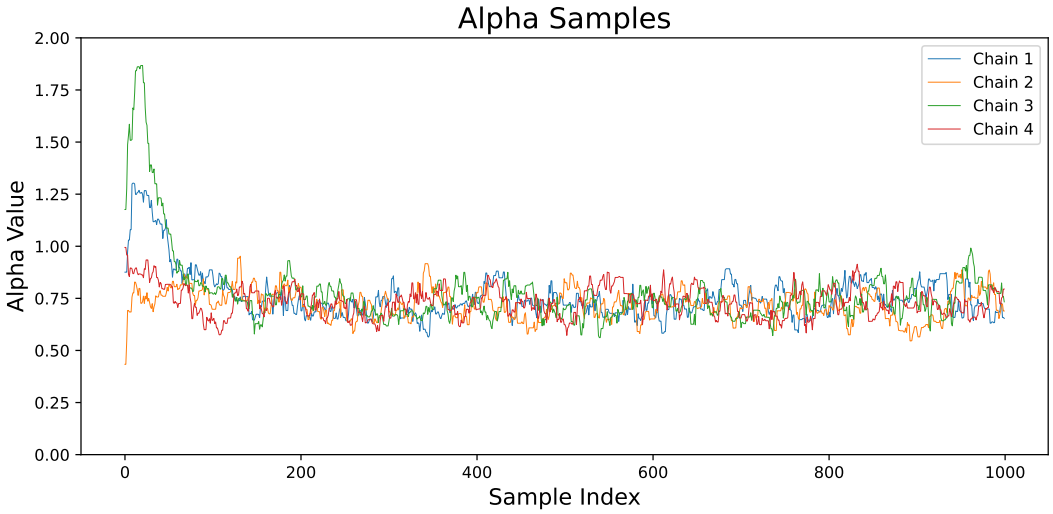
$$
p(Y > y|\alpha, \beta) = \frac{B(\alpha, \beta+y)}{B(\alpha, \beta)}
$$

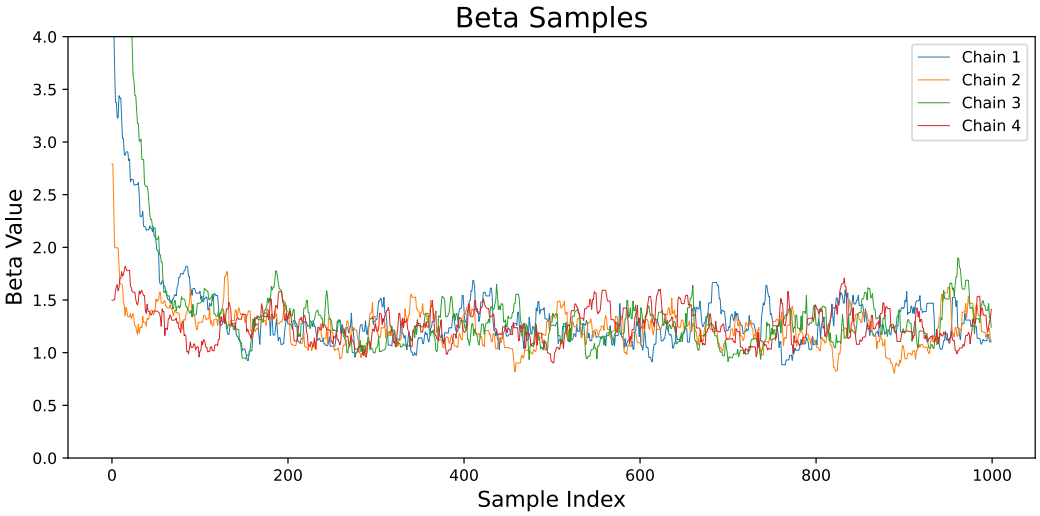
These two formulas applied to the data will give us a resultant likelihood which we can maximize using the solver in excel. Specifically, we can apply $p(Y = y|\alpha, \beta)$ to the uncensored cells and $p(Y > y|\alpha, \beta)$ to the censored cell. Traditionally, the likelihood is a product of these probabilities across the data. To ensure the lack of precision our computers have doesn't affect the calculations, it is common to take a log of the likelihood. This results in a sum of the log probabilities multiplied by the number of corresponding customers associated with it.

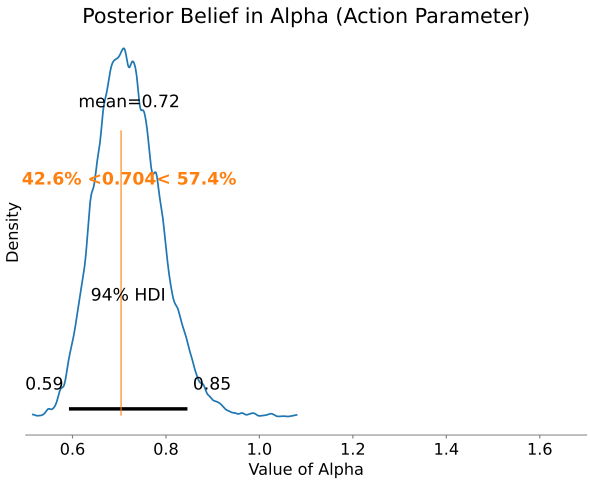
3 Results and Analysis

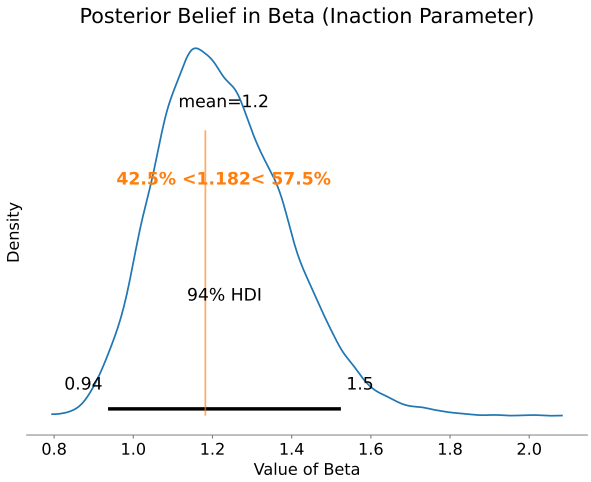
Theta posertiors are cool but mostly useless for churned customers. Lets see if typing more words allows me to rectify this error.

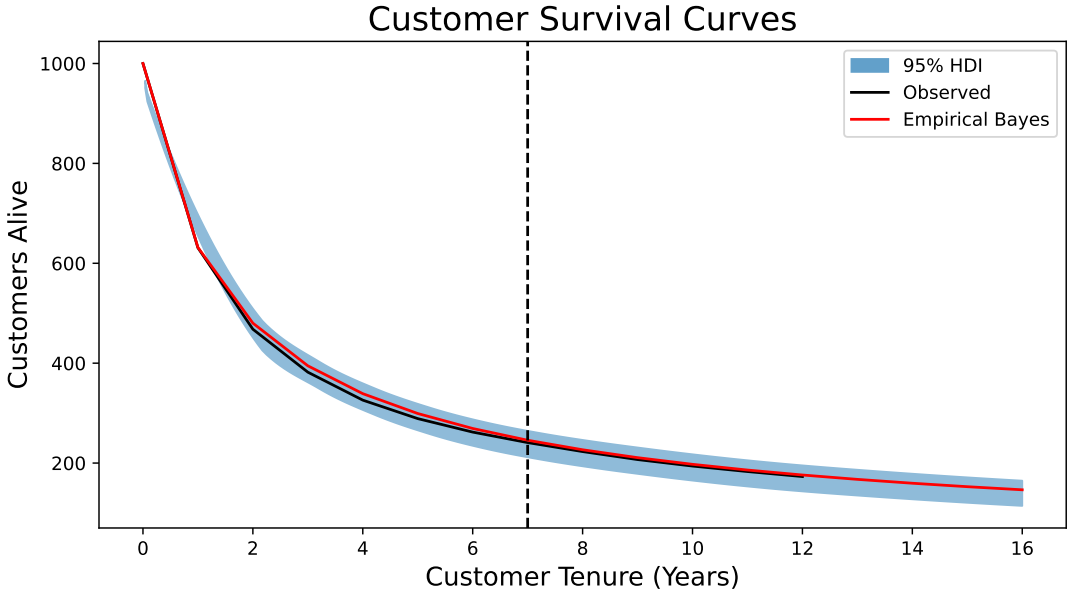


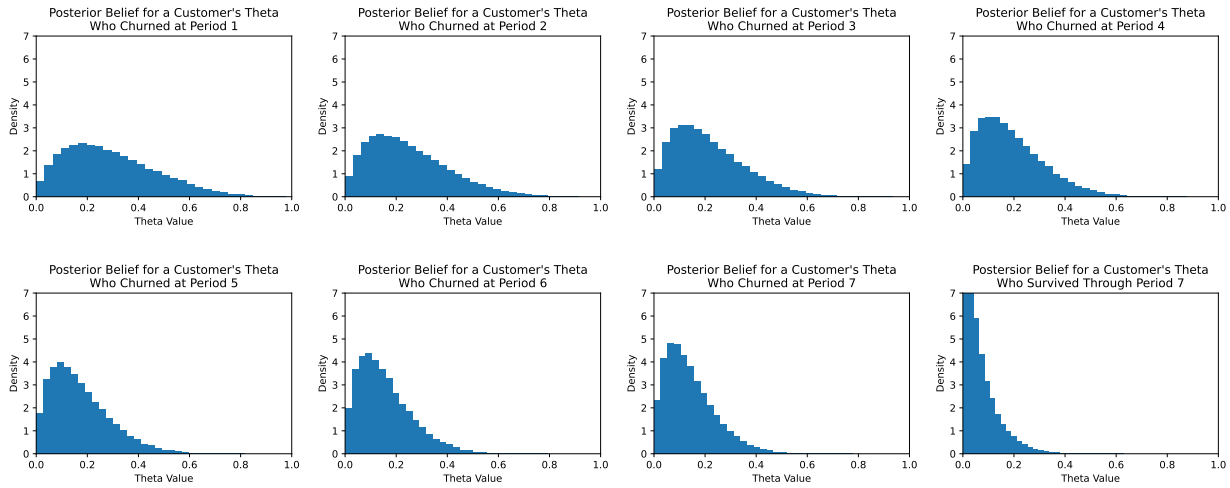


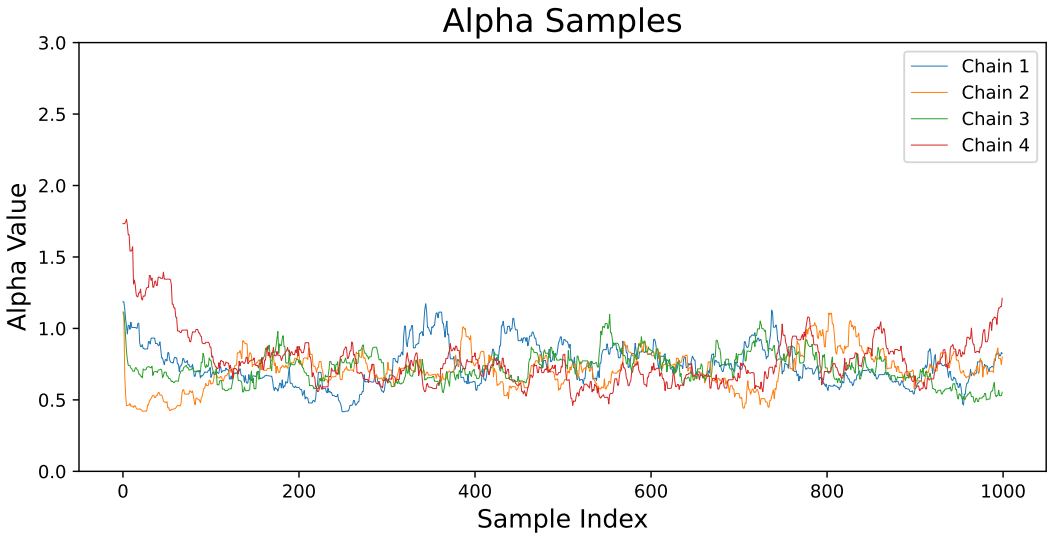

Beta Samples
Chain 1
Chain 2
Chain 3
Chain 4
Beta Value
Sample Index
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
200
400
600
800
1000

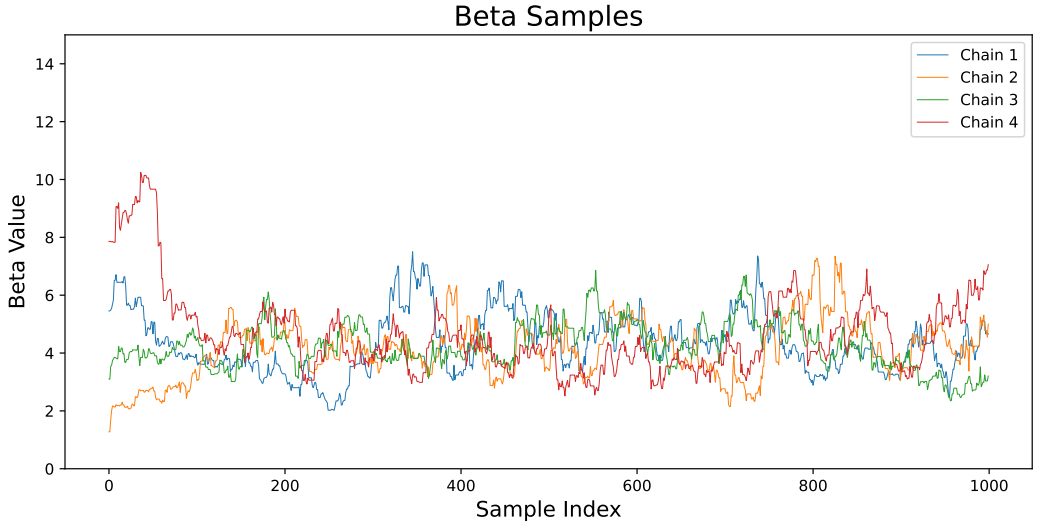

Posterior Belief in Alpha (Action Parameter)
mean=0.72
42.6% <0.704< 57.4%
94% HDI
0.59
0.85
Density
Value of Alpha
0.6
0.8
1.0
1.2
1.4
1.6

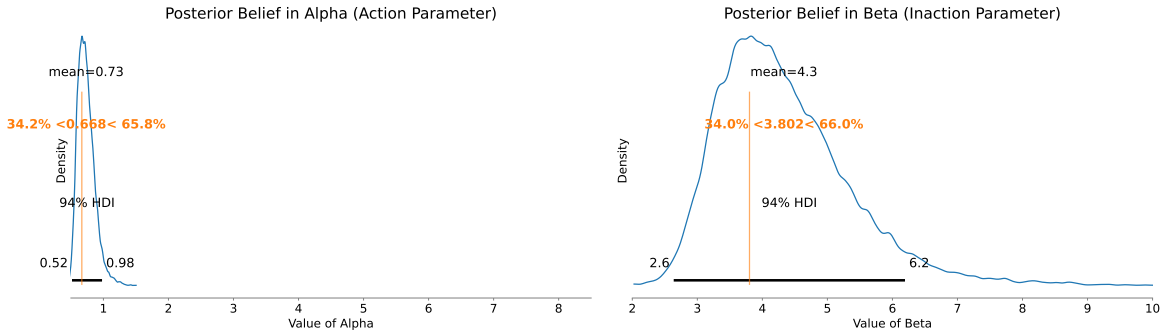

Posterior Belief in Beta (Inaction Parameter)
mean=1.2
42.5% <1.182< 57.5%
94% HDI
0.94
1.5
Density
Value of Beta
0.8
1.0
1.2
1.4
1.6
1.8
2.0

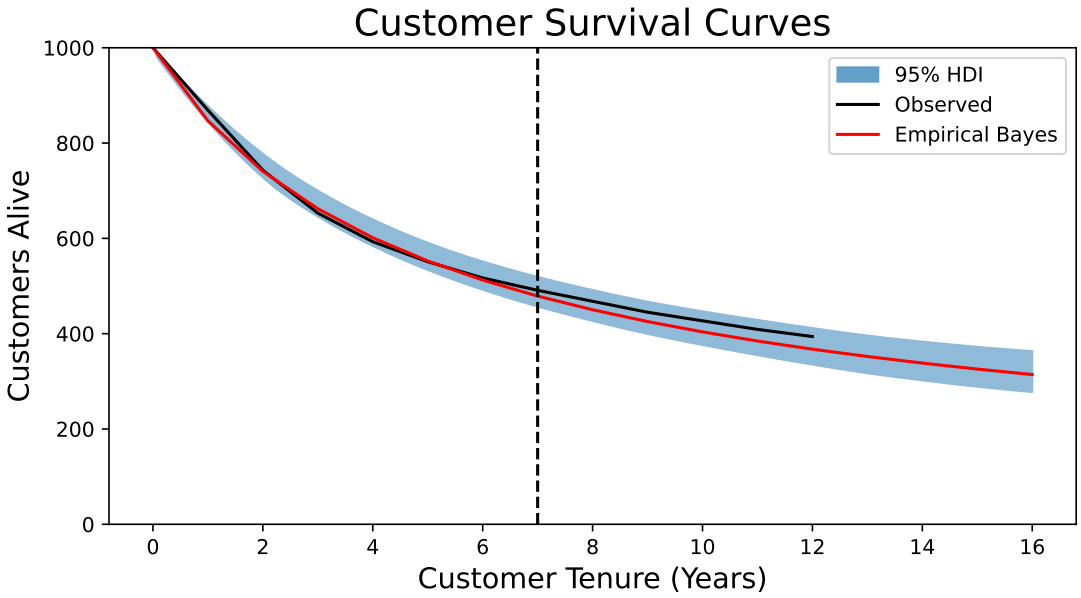

Customer Survival Curves
95% HDI
Observed
Empirical Bayes
Customers Alive
Customer Tenure (Years)
200
400
600
800
1000
0
2
4
6
8
10
12
14
16


Posterior Beliefs of Churn Propensity for Regular Customers of Similar Behavior
Posterior Belief for a Customer's Theta Who Churned at Period 1
Posterior Belief for a Customer's Theta Who Churned at Period 2
Posterior Belief for a Customer's Theta Who Churned at Period 3
Posterior Belief for a Customer's Theta Who Churned at Period 4
Posterior Belief for a Customer's Theta Who Churned at Period 5
Posterior Belief for a Customer's Theta Who Churned at Period 6
Posterior Belief for a Customer's Theta Who Churned at Period 7
Postersior Belief for a Customer's Theta Who Survived Through Period 7
Density
Theta Value


Alpha Samples
Chain 1
Chain 2
Chain 3
Chain 4
Alpha Value
Sample Index


Beta Samples
Chain 1
Chain 2
Chain 3
Chain 4
Beta Value
Sample Index





4 Conclusion and Considerations

I want to eventually implement a generalized version of the stopping criterion as well as a riemannian-gaussian version of the model.

Empirical Bayes

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)} = \frac{p(y|\theta) \cdot p(\theta)}{\int (p(y|\theta)p(\theta))d\theta}$$

Fully Bayesian (Our Approach)

$$p(\theta|y) = \frac{p(y|\theta) \cdot \int\int p(\theta|\alpha,\beta)p(\alpha,\beta)\,d\alpha\,d\beta}{p(y)} = \frac{p(y|\theta) \cdot \int\int p(\theta|\alpha,\beta)p(\alpha,\beta)\,d\alpha\,d\beta}{\int\int\int p(y|\theta)p(\theta|\alpha,\beta)p(\alpha,\beta)\,d\alpha\,d\beta\,d\theta}$$

A Appendix

There are many techniques and variants of MCMC methods. As we have seen, HMC can be especially useful when dealing with problematic posteriors; however, there are still times when additional techniques must be utilized. This appendix holds the few that I have

implemented for this specific use case. It is not an exhaustive list, or even necessary, but those on similar paths to mine will find them well-suited for their needs.

A.1 Metropolis-Hastings Ratio

A.2 Log of Log-Normal Distribution

A.3 Re-parameterization

A.4 Mass Matrix/Metric