Bayesian Hierarchical Modeling for Predicting Small Business Sales

Matthew Holcomb
mwholcomb@ucdavis.edu | matthew.w.holcomb@gmail.com

Source: drewmaddenart.com

An Undergraduate Research Project — Supervised by Professor Peter Kramlinger

Abstract

This paper reviews fundamental Bayesian methodologies and demonstrates their application to an intricate, real-world data set. With an overall goal of predicting sticker sales from a small business, a collection of candidate models are developed using modern hierarchical modeling approaches, including the Metropolis-Hastings algorithm: a Markov-Chain Monte-Carlo approach for approximating unknown distributions. The predictive capabilities of each model are evaluated using Leave-One-Out Cross-Validation and compared against a standard non-Bayesian approach: maximum likelihood estimation (MLE). Results show improved performance by two of the proposed hierarchies compared to MLE, suggesting that the inclusion of prior knowledge is beneficial in this setting. Possible implementations of the models with regards to a business setting are also discussed.

I Introduction

Over the summer of 2023, I worked with a close friend, Drew Madden, to organize and summarize invoice data he had through his sticker business, Rivers to Sea Stickers[1]. While I enjoyed the process of translating sales info into simple reports on top customers, I was also interested in trying to come up with a way to predict future orders using the data that was already present. This eventually led to the beginning of a project where, under the supervision and guidance of Prof. Peter Kramlinger, the goal was to do just that: construct and compare a range of Bayesian hierarchical models which could be used to predict future orders, which could then be used by Drew as a tool to determine the best timing to reach out to customers regarding placing new orders. This report is the culmination of that work, and is structured to be an extended and informative application of a few modern Bayesian processes, using Drew's sticker sales as a reference and motivation. *Note: Any invoice/sales data referenced in this report has been made anonymous, and is not available for public use. However, the evaluation of the models was done with the actual data, and the results of those evaluations can be found in later sections.*

Using [5] and [4], Bayesian hierarchies were constructed to model sticker sales by customers with at least eight total orders. The first two hierarchies utilize conjugate priors, resulting in analytical forms for the posterior distributions and thus a closed form for the Bayes estimator under squared error loss. The third hierarchy does not have an analytical form for the posterior - it is instead approximated using Markov-Chain Monte-Carlo methods (Metropolis-Hastings), where the Bayes estimator is computed as the empirical mean of the resulting sampled distribution. Evaluations of these three hierarchies using Leave-One-Out Cross-Validation ([2], Chapter 5) was used to compare their predictive capabilities. The implementation of these models into a business setting - i.e., how one might use them to make predictions - is introduced, followed by a brief discussion regarding future improvements and extensions.

II The Data

Invoices for all orders filled by River to Sea Stickers between January 1, 2020 and March 1, 2024 were compiled to form the dataset used in this project. Each invoice contained the customer's name, the date the order was placed, and the amount spent on the order. The focus was placed primarily on *when orders were placed*, and the order amounts were not considered (potential extensions utilizing the order amounts is considered further in Section IV). Orders were grouped by customer for further analysis - in total, there were roughly 1,000 invoices split by the $N = 229$ unique customers (note that the amount of orders placed per customer varied).

Initially, the order data for each customer was in the form of binary values corresponding to whether an order was placed on a specific date. However, evaluating the data in this form was difficult. The main issue revolved around the fact that the binary data did not appear independent; it is reasonable to assume that the probability of an order is influenced by how recently the last order was placed. Thus, the original form of the data could not be considered independent and identically distributed (i.i.d.), which would lead to problems in the hierarchical Bayes setting.

[1]Please note that this has **no relation** to the slogan "from the river to the sea" - it is derived from Drew's connection to river and ocean sports. More information about Drew's stickers can be found on his website.

II.1 Transforming the Data

A transformation of the data was utilized to combat the issue stemming from the lack of independence within the binary data. Rather than using the dates at which orders were placed, the temporal component of the data was changed to be a numerical value representing **the time since the previous order**. Essentially, for each order, the time in months would be calculated since the last order was placed (another way to think of it is as the time *between* orders). Using a single customer as an example, we can see the transformation directly in Figures II.1 and II.2.

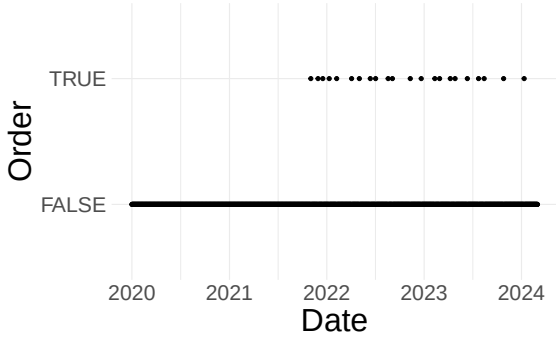

Figure II.1: Binary Data

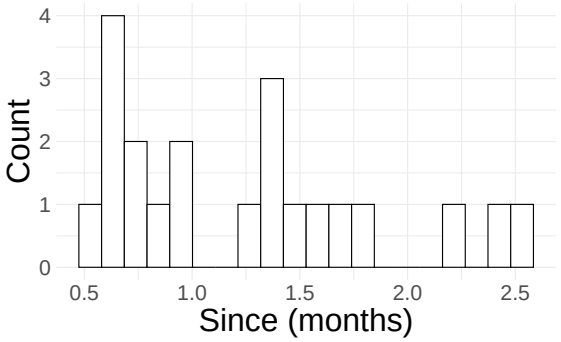

Figure II.2: Transformed Data

Figure II.1 shows the binary values for each day within our date range, where "TRUE" values indicate days in which an order was placed. As discussed, it is difficult to approach the modeling of this type of data, as each variable is dependent on the previous order(s) - we would expect the probability of an order to change depending on how long it has been since the previous order. However, if we consider the spacing of the orders instead, we end up with data plotted in Figure II.2. After transforming the data, the assumption of independence seems much more reasonable.

Note: The first order by each customer has a 'since' value which is computed based on the time since January 1, 2020 (because it is the first order placed by the customer and thus has no preceding orders). As a result of this, the first order placed by each customer is omitted from the transformed data, as it does not correspond to the time since the previous order, and is typically very (uncharacteristically) large.

The use of this transformation allows for the data to be interpreted as independent and identically distributed (i.i.d.) random variables, which is necessary for the use of Bayesian hierarchies - we will explore this further in Section III.

II.2 Notation

The following notation will be used throughout the remainder of this report in reference to the transformed data:

- N: The number of unique customers.
- $i \in \{1, \ldots, N\}$: The customer index.
- n_i: The total number of orders placed by store i minus one (excluding the first order).
- $j \in \{1, \ldots, n_i\}$: The order index.
- X_{ij}: The time in months it had been since the previous order.
- $\boldsymbol{x}_i = \{x_{ij}\}_{j \in \{1,\ldots,n_i\}}$: All orders placed by store i (again, excluding the first order[1]).

III Methods

Bayesian statistics is based on the concept of considering parameters to be random variables, where one can use prior knowledge alongside new data to compute a posterior distribution for which that parameter is thought to lie. Consider flipping a brand new coin many times, with the goal of estimating the probability of getting 'Heads'. One might assign a prior distribution (the distribution of a parameter before observing any data), and then compute a posterior distribution conditional on the observed coin flips to then obtain the conditional expectation to estimate the parameter. Furthermore, the prior can be assigned based on previous knowledge - in this case, one might choose a prior distribution which has a mean of 0.5, as it is known that a typical coin has a 50% chance of landing on heads. This is in contrast to a frequentist approach, in which one would estimate a 'true' parameter directly from the data, based only off of the proportion of observed 'Heads'.

For example, if the coin was flipped 10 times and 3 heads were observed, a frequentist approach would estimate the true parameter to be 0.3. However, the use of a prior which has mean 0.5 combined with the small sample of 10 flips would result in a distribution of the parameter which is centered closer to 0.5. If the coin was in fact 'normal,' then it is reasonable to conclude that the Bayesian approach with regards to this example would produce a better estimate of the parameter. This is a key motivation for the use of the Bayesian framework to model the sticker data available, as many customers have a small amount of orders, and may be benefited by an appropriately chosen prior.

This section outlines the construction of three different Bayesian hierarchical models, with an overall goal of identifying the best option for predicting future sticker orders based on current invoice data. First, the theoretical approaches necessary for finding the form of the Bayes estimator for each hierarchy will be discussed. Then, the application of these methods will be demonstrated on the sales data from Rivers to Sea Stickers, and model performance will be evaluated. Lastly, the 'best' model as determined by the evaluation method will be used to establish a process for predicting future orders.

III.1 Posterior Computation

As mentioned, Bayesian hierarchical modeling focuses on computing the posterior distribution for a parameter of interest, using a prior distribution which is chosen before seeing the data and a likelihood function which is based on the observed data itself. Finding this posterior distribution is the main goal - after doing so, it is possible to compute the expected value of both the parameter and new observations, which is relevant to this project's goal of predicting future orders.

Consider data in the form $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ which has been sampled/collected from a distribution for which the p.d.f. (probability density function) is $f(x \mid \theta)$. The posterior distribution of θ (our parameter of interest) is denoted $\pi(\theta \mid \boldsymbol{x})$, and can be computed using Bayes' theorem ([5], Theorem 7.2.1).

Posterior Equation (Bayes' Theorem)

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(x_1 \mid \theta) \ldots f(x_n \mid \theta)\pi(\theta)}{g_n(\boldsymbol{x})}$$

The denominator, $g_n(\boldsymbol{x})$, is often difficult to compute directly. Fortunately, there are ways to compute the posterior without ever doing so.

Consider the p.d.f. of the Normal (Gaussian) distribution:

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

If μ is the parameter of interest, and σ is treated as a constant value, one can re-write this form such that the values which are constant with respect to μ are simply written as some constant value c:

$$f(x \mid \mu, \sigma) = c \times e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

This can be taken one step further by writing the p.d.f. without the constant value. Consider some arbitrary equality $A = c \times B$, where c is some constant with respect to A. One can re-write this as $A \propto B$, which gives that A is directly proportional to B. This can be applied to the form of the Normal distribution:

$$f(x \mid \mu, \sigma) \propto_{\mu} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Returning to the form of the posterior - since $g_n(\boldsymbol{x})$ is a constant with respect to the parameter θ, one can write

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(x_1 \mid \theta) \dots f(x_n \mid \theta)\pi(\theta)}{g_n(\boldsymbol{x})} \quad \propto_{\theta} f(x_1 \mid \theta) \dots f(x_n \mid \theta)\pi(\theta)$$

Since it is known that the posterior is *some distribution*, if the form of the numerator is that of a known distribution, the denominator can be inferred - giving the distribution of the posterior.

In the cases where the numerator is not in the form of a known distribution and the denominator is not obtainable analytically, Markov-Chain-Monte-Carlo methods (such as Metropolis-Hastings) can be used to approximate the distribution of the posterior instead. This will be discussed in more detail later on.

III.2 Bayes Estimator

The choice of estimator for the parameter θ is subjective. A justification for the estimation procedure that is used in the subsequent sections is outlined below ([1], Chapters 1 and 4).

In order to evaluate an estimator $\hat{\theta}$ for some parameter θ, one can use its *loss* $L(\theta, \hat{\theta})$. The loss of an estimator can be interpreted as the amount lost if the parameter equals θ and the estimate equals $\hat{\theta}$ ([5], Definition 7.4.2). It is natural to choose an estimator $\hat{\theta}$ which minimizes the expected loss; which is called the *risk*:

$$R(\theta, \hat{\theta}) = \mathbb{E}\left[L(\theta, \hat{\theta})\right]$$

In the Bayesian setting, the parameter $\theta \in \Theta$ has some prior distribution $\pi(\theta)$, which supposedly reflects which θ's are "likely" to occur. Thus, one can compute the *average risk* by taking the expected "weight" of the risk $R(\theta, \hat{\theta})$:

$$r(\pi, \hat{\theta}) = \mathbb{E}^{\pi}\left[R(\theta, \hat{\theta})\right]$$

Then, the best estimator is one which minimizes the average risk $r(\pi, \hat{\theta})$ ([1], Section 1.5).

Once the data $\boldsymbol{X} = \boldsymbol{x}$ has been observed, the posterior $\pi(\theta, \boldsymbol{x})$ can be used in place of the prior in the steps above. Note that since the estimator will now depend on the data, it can be denoted $\hat{\theta}(\boldsymbol{x})$. The goal remains the same; to minimize the expected risk, for which it can be shown ([1], Equation 4.1):

$$\arg\min_{\hat{\theta}(\boldsymbol{x})} \mathbb{E}\left[R(\theta, \hat{\theta}(\boldsymbol{x}))\right] = \arg\min_{\hat{\theta}(\boldsymbol{x})} \mathbb{E}\left[L(\theta, \hat{\theta}(\boldsymbol{x})) \mid \boldsymbol{x}\right]$$

Thus the same 'best' estimator which minimizes the average risk also minimizes the posterior risk, and is known as the *Bayes estimator*.

III.2.1 Bayes Estimator Under Squared Error Loss

Squared error loss, defined as $L(\theta, a) = (\theta - a)^2$, is the most commonly used loss function in estimation problems, and will be what is used for the entirety of this project. The Bayes estimator under squared error loss can thus be written as

$$\arg\min_{\hat{\theta}(\boldsymbol{x})} \mathbb{E}[L(\theta, \hat{\theta}(\boldsymbol{x})) \mid \boldsymbol{x}] = \arg\min_{\hat{\theta}(\boldsymbol{x})} \mathbb{E}[(\theta - \hat{\theta}(\boldsymbol{x}))^2 \mid \boldsymbol{x}].$$

The estimator $\hat{\theta}(\boldsymbol{x})$ which minimizes this equation is $\hat{\theta}^*(\boldsymbol{x}) = \mathbb{E}(\theta \mid \boldsymbol{x})$, which is the posterior mean ([1], Section 4.4.2). Note that, going forward, the notation $\hat{\theta}$ will be used to indicate the Bayes estimator for the parameter θ, and is equivalent to $\hat{\theta}^*(\boldsymbol{x})$ in this context.

For instance, consider a customer which placed n orders for which the times (in months) between orders, $\boldsymbol{x}$, have mean $\bar{\boldsymbol{x}}_n = 1$. After choosing a distribution to model the data $f(X \mid \theta)$ as well as a prior distribution for the parameter θ, suppose one was able to use the processes described in Section III.1 to find that the form of the posterior is $\pi(\theta \mid \boldsymbol{X}) \sim \text{Gamma}(\alpha = 2, \beta = \bar{\boldsymbol{x}}_n)$. Then, one could compute the Bayes estimator $\hat{\theta}$ using the mean of the Gamma distribution (see Figure III.1):

$$\hat{\theta} = \mathbb{E}(\theta \mid \boldsymbol{X}) = \alpha\beta = (2)(\bar{\boldsymbol{x}}_n) = 2$$

However, if the mean of the posterior is not known, one might still be able to generate a sample from it. Suppose this was the case - one could instead draw a sufficiently large (e.g. $S = 1000$ values) sample from the posterior and compute the mean of those values through computer simulation.

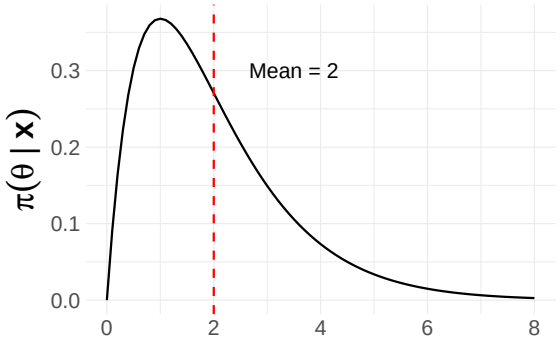

Figure III.1: Theoretical Gamma Distribution

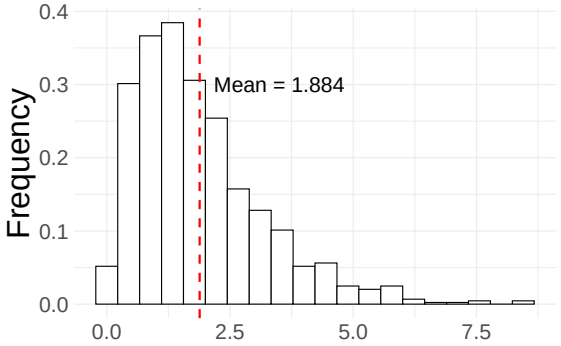

Figure III.2: Empirical Gamma Distribution

Figures III.1 and III.2 show the theoretical and empirical densities of a Gamma distribution with parameters $\alpha = 2$, $\beta = 1$, where the theoretical distribution has a mean of exactly 2 and the empirical distribution has a mean of 1.9. Either of these two values would then be chosen as a Bayes estimator for the parameter θ based on the method used.

III.3 Choosing a Prior

A benefit of choosing Bayesian hierarchical modeling in this setting is the incorporation of prior knowledge into the modeling of customer-specific sales. By definition, the prior distribution of a parameter θ must be a probability distribution over the parameter space Ω. However, before the data has been collected or observed, one's past experience and knowledge might lead to the belief that θ is more likely to lie in certain regions of Ω than in others ([5], Chapter 7). Furthermore, it is common to pick a distribution which is computationally usable, such as a conjugate prior. The usefulness of an appropriate choice can be seen later in Sections III.4.2 and III.5.2, in which the posterior distribution is analytically obtainable.

Consider the transformed data which represents the time between orders placed by different customers. It is reasonable to assume that the probability of an order would be higher a few months after the previous order as opposed to immediately after (i.e. a few days) or much longer after (i.e. a year or two). While the model must allow for the possibility of orders being placed in an appropriate range (technically, an order *could* be placed many years after the previous one), the belief that certain events are more likely than others can be incorporated into the model through the prior. Thus, one can pick a prior distribution which represents that knowledge.

For example, suppose that the Gamma distribution is chosen to model the time between orders with regards to Drew's sticker business. The shape parameter is set to be $\alpha = 2$, and the scale parameter β is considered to be random. One might choose a prior $\pi(\beta)$ for which the mean of the prior corresponds to a reasonable value for β to take. Consider the distributions of the Gamma density with shape $\alpha = 2$ and scale β for different values of β seen in Figure III.3.

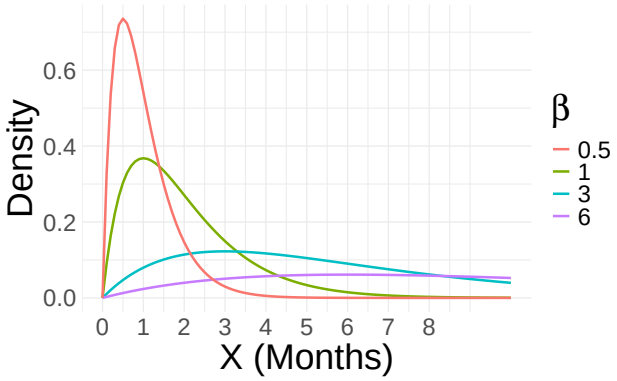

Figure III.3: Distribution for $X \sim \text{Gamma}(2, \beta)$

One can see the changes in the Gamma distribution that results from choosing different scale parameters. Looking at the four p.d.f.s and comparing them to what is known about sticker orders, it would be reasonable to infer values of β which may be most appropriate. For example, it can be seen that for $\beta = 1$, the distribution puts the largest weight on values of X between roughly 1 and 3. If it is generally known that most orders are placed a few months after one another, the choice of a prior with a mean of 1 would make sense. In contrast, for $\beta = 6$, the probabilities of an order being placed a few months versus nearly a year after the most recent order appear quite similar - this makes less sense in regards to the sticker data.

For the remainder of the project, the choice hyper-parameters set on the prior distributions for each of the three hierarchies will not be discussed in depth. However, they are chosen with regards to this idea; where the mean of the prior is reasonable with regards to the distribution chosen for the data. Furthermore, it is important to note that this is only one of many ways to approach selecting a prior in the Bayesian framework.

III.4 Hierarchy 1: Conjugate Poisson-Gamma Hierarchy

One might consider the discrete Poisson distribution to model the transformed data, as its form is similar to the shape the time between orders takes for most customers. In this case, one would model time between orders as a Poisson process - that is, $X_{ij} \sim \text{Poisson}(\lambda_i)$, where λ_i is the *rate* parameter for each customer (i). The probability mass function (p.m.f.) of the Poisson distribution is:

$$P(X = x \mid \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Note: Since the Poisson distribution is a discrete process, the time between orders must similarly be discrete $(X_{ij} \in \{0, 1, 2, \dots\})$

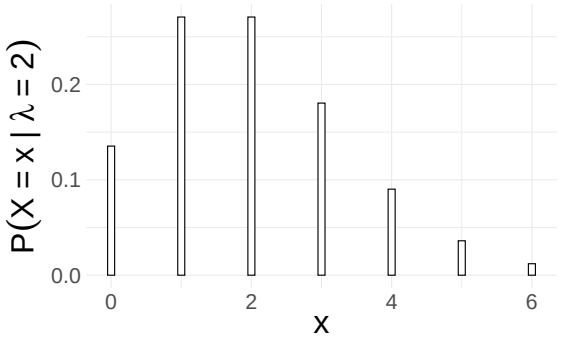

Figure III.4: Poisson Distribution, $\lambda = 2$

III.4.1 Hierarchy

If choosing the Poisson distribution to model the data, one can consider a conjugate Gamma prior - where $\lambda_i \sim \text{Gamma}(\alpha, \beta)$. The hyper-parameters α and β can be chosen by the user, while λ_i is treated as a customer-specific random variable. This allows one to construct the conjugate Poisson-Gamma hierarchy:

$$\begin{aligned} X_{ij} \mid \lambda_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

The Gamma distribution is considered a conjugate prior because the form of the posterior distribution, in this case, is also a Gamma distribution. This can be seen by computing the posterior distribution.

III.4.2 Posterior Computation

Recalling the form of the posterior, one can substitute the parameters from the hierarchy and write:

$$\pi(\lambda_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{f(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \lambda_i)\pi(\lambda_i)}{g_n(\boldsymbol{x}_i)}$$

Since the denominator of this function does not depend on the parameter λ_i:

$$\pi(\lambda_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) \propto_{\lambda_i} f(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \lambda_i)\pi(\lambda_i)$$

Recall the form of the Gamma distribution's p.d.f.:

$$f(X = x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \quad \text{for } X = x \mid \alpha, \beta \sim \text{Gamma}(\alpha, \beta)$$

One can plug the forms of the likelihood (Poisson) and prior (Gamma) into this equation - combining terms with respect to the parameter λ_i eventually leads to a familiar form:

$$\begin{aligned}
\pi(\lambda_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) &\propto_{\lambda_i} f(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \lambda_i)\pi(\lambda_i) \\
&\propto_{\lambda_i} \left(\prod_{j=1}^{n_i} \frac{\lambda_i^{x_{ij}}}{x_{ij}!} e^{-\lambda_i} \right) \left(\frac{1}{\Gamma(\alpha)\beta^\alpha} \lambda_i^{\alpha-1} e^{-\lambda_i \beta^{-1}} \right) \\
&\propto_{\lambda_i} \left(\lambda_i^{\sum x_{ij}} e^{-n_i \lambda_i} \right) \left(\lambda_i^{\alpha-1} e^{-\lambda_i \beta^{-1}} \right) \\
&\propto_{\lambda_i} \lambda_i^{\sum x_{ij} + \alpha - 1} e^{-\lambda_i (n_i + \beta^{-1})} \\
&\propto_{\lambda_i} \lambda_i^{n_i \bar{\boldsymbol{x}}_i + \alpha - 1} e^{-\lambda_i / (n_i + \beta^{-1})^{-1}} \text{ , where } \bar{\boldsymbol{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}
\end{aligned}$$

This is Gamma distributed with parameters $n_i \bar{\boldsymbol{x}}_i + \alpha$ and $(n_i + \beta^{-1})^{-1}$, giving:

$$\lambda_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i \sim \text{Gamma}(n_i \bar{\boldsymbol{x}}_i + \alpha, (n_i + \beta^{-1})^{-1})$$

III.4.3 Bayes Estimator

Under squared error loss, the Bayes estimator is the expected value (mean) of the posterior. Since the posterior distribution of λ_i is Gamma, this is computed as:

$$\hat{\lambda}_i = E(\lambda_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = (n_i \bar{\boldsymbol{x}}_i + \alpha) \left(\frac{1}{n_i + \beta^{-1}} \right) = \frac{n_i \bar{\boldsymbol{x}}_i + \alpha}{n_i + \beta^{-1}}$$

This gives us a closed form for our Bayes estimator of the parameter λ_i. One can plug in the data $\boldsymbol{X}_i = \boldsymbol{x}_i$ as well as our parameters α and β to this form to estimate λ_i.

The Bayes estimator is useful for making future predictions. Specifically, it gives a form for the posterior distribution, which can be used to model the expected behavior of future observations. In this hierarchy, it is defined that the time between orders, X_{ij}, are Poisson distributed with parameter λ_i. Thus, the posterior distribution models the probability of how much time there will be between future orders, which can be used to predict how long it will take for future orders to be placed. For instance, one could use the posterior to compute the probability of an order being placed within the next month, or the time interval for which there is an estimated 75% chance of an order occurring within.

For example - suppose $\hat{\lambda}_i = 2$ was computed. One could compute the probability that the next order will be between c and d months in the future using the estimate for the parameter λ_i:

$$P(c \leq X_{ij+1} \leq d \mid \lambda_i) \approx P(c \leq X_{ij+1} \leq d \mid \hat{\lambda}_i)$$

This use of the Bayes estimator will be important later when we wish to utilize our model to make predictions, as well as when evaluating the performance of each of our hierarchies, and is discussed further in Section III.9.

A downside of using the Poisson distribution is the fact that it is a discrete distribution. This means that we lose out on information - for instance, if there are orders placed on January 25th and then February 9th, a discrete model would use that the orders were either 1 or 0 months apart, while a continuous model would be able to use fractions of a month to more precisely capture the time between the orders. This brings us to Hierarchy 2, which instead utilizes the continuous Gamma distribution to model the time between orders.

III.5 Hierarchy 2: Conjugate Gamma-InverseGamma Hierarchy

Consider using the continuous Gamma distribution to model the time between orders. The probability density function (p.d.f.) of the Gamma distribution is

$$f(X = x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}.$$

Since the Gamma distribution uses two parameters, α and β, it is customary to only choose one to model as a random variable while fixing the other. Typically β is chosen to be the random variable (this is due to some computation issues if using α), giving the notation $X_{ij} \sim \text{Gamma}(\alpha, \beta_i)$.

Figure III.5: Gamma Density, $\alpha = 2, \beta = 0.7$

III.5.1 Hierarchy

One might consider using an Inverse-Gamma distribution to model the prior, as it is conjugate to the Gamma likelihood. Once again, hyper-parameters are chosen by the user, but since α and β are being used for the likelihood, one might use a and b instead. This allows for the construction of the conjugate Gamma-Inverse-Gamma hierarchy:

$$\begin{aligned} X_{ij} \mid \beta_i &\sim \text{Gamma}(\alpha, \beta_i) \\ \beta_i &\sim \text{Inverse-Gamma}(a, b) \end{aligned}$$

The p.d.f. of the Inverse-Gamma can be used to write out the form for the prior, and can be written, for $X \sim \text{Inverse-Gamma}(a, b)$, as

$$f(X = x) = \frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x},$$

where $\mathbb{E}(X) = \frac{b}{a-1}$.

Then, for $\beta_i \sim \text{Inverse-Gamma}(a, b)$:

$$\pi(\beta_i) = \frac{b^a}{\Gamma(a)} \beta_i^{-a-1} e^{-b/\beta_i}$$

As seen with the Poisson-Gamma hierarchy, computation of the posterior is straight-forward due to the choice of a conjugate prior distribution.

III.5.2 Posterior Computation

One can plug the forms of the likelihood and prior into the formula for the posterior:

$$\begin{aligned}
\pi(\beta_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) &= \frac{f(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \beta_i)\pi(\beta_i)}{g_n(\boldsymbol{x}_i)} \\
&\propto_{\beta_i} f(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \beta_i)\pi(\beta_i) \\
&\propto_{\beta_i} \left(\prod_{j=1}^{n_i} \frac{1}{\Gamma(\alpha)\beta_i^{\alpha}} x_{ij}^{\alpha-1} e^{-x_{ij}/\beta_i} \right) \left(\frac{b^a}{\Gamma(a)} \beta_i^{-a-1} e^{-b/\beta_i} \right) \\
&\propto_{\beta_i} \left(\frac{1}{\beta_i^{n_i\alpha}} e^{-\sum x_{ij}/\beta_i} \right) \left(\frac{1}{\beta_i^{a+1}} e^{-b/\beta_i} \right) \\
&\propto_{\beta_i} \beta_i^{-(n_i\alpha+a)-1} e^{-(\sum x_{ij}+b)/\beta_i}
\end{aligned}$$

This is the form of the Inverse-Gamma distribution with parameters $(n_i\alpha + a)$ and $(\sum_{j=1}^{n_i} x_{ij} + b)$, giving the posterior distribution of β_i:

$$\beta_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i \sim \text{Inverse-Gamma}(n_i\alpha + a, n_i\bar{\boldsymbol{x}}_i + b), \text{ where } \bar{\boldsymbol{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

III.5.3 Bayes Estimator

Under squared error loss, the Bayes estimator is the expected value (mean) of the posterior. Since the posterior distribution of β_i is Inverse-Gamma, this is computed as:

$$\hat{\beta}_i = E(\beta_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{n_i\bar{\boldsymbol{x}}_i + b}{n_i\alpha + a - 1}$$

Once again, the use of a conjugate prior leaves one with a form for the Bayes estimator $\hat{\beta}_i$, which can then be used to make inferences about the data, or to predict the values of future observations.

III.6 Posterior Approximation Using MCMC

So far, methods for obtaining the posterior distribution and Bayes estimator for a parameter of interest have been discussed for cases in which a conjugate prior is chosen to form a Bayesian hierarchy. However, in the case where a prior is chosen which does not result in an analytical form for the posterior, a different approach must be considered. While the Bayes estimator under squared error loss is still the posterior mean, there is now no closed-form solution for that quantity.

The most common way to work around this issue is to sample from the posterior distribution and approximate the quantities of interest (i.e., the mean). Although it is not possible to sample from an unknown posterior, a stochastic process can be constructed which has the posterior as a limiting distribution. Such sampling procedures are available (e.g., Metropolis-Hastings and Gibbs sampling) and are members of a family of algorithms known as Markov-Chain Monte-Carlo (MCMC) Methods.

This section introduces the Independent Metropolis-Hastings algorithm which will be used for Hierarchy 3, which has an unknown posterior distribution (Section III.7).

III.6.1 Markov-Chain Monte-Carlo Methods: A Brief Introduction

The term 'Monte-Carlo' refers to the general concept of using computer simulation to approximate probability distributions. For example, if one was interested in the probability of having a 'full-house' in poker, it could be approximated by simulating many hands of poker and observing the proportion in which that occurred. 'Markov chains,' on the other hand, refer to the modeling of random-chance events which depend solely on

the previous event. Consider flipping a fair coin multiple times, keeping track of the total number of heads flipped. The probability of having ten heads after the next throw depends on the number of heads before the throw, but the actual sequence of heads and tails does not have to be memorized.

'Markov-Chain Monte-Carlo' algorithms are the combination of these two concepts - simulating each step from some random distribution, but stringing those steps together to form a 'random walk,' where the 'direction' at each step/state is random, while the starting point is simply wherever the previous step left off. This project will introduce one of such methods - the Independent Metropolis-Hastings algorithm - but many more MCMC applications can be found in [4].

III.6.2 Independent Metropolis-Hastings

Note: This section is adapted directly from Section 7.4 of [4], and focuses on introducing and defining the Independent Metropolis-Hastings algorithm with regards specifically to a Bayesian posterior target density.

The Metropolis-Hastings algorithm is useful for generating a sample from some *target* density f without directly sampling from it. The algorithm is given (where the density q is called the *proposal density*) as:

Metropolis-Hastings. Given $X_n = x_n$:

(1) Generate $Y_n \sim q(y \mid x_n)$

(2) Set $X_{n+1} = \begin{cases} Y_n \text{ with probability } \rho(x_n, Y_n) \\ x_n \text{ with probability } 1 - \rho(x_n, Y_n) \end{cases}$, where $\rho(x_n, Y_n) = \min\left\{ \dfrac{f(y)}{f(x)} \dfrac{q(x \mid y)}{q(y \mid x)}, 1 \right\}$

Consider a Bayesian setting in which one wishes to model the posterior distribution for some parameter θ by approximation. One can set the posterior $\pi(\theta \mid \boldsymbol{x})$ to be the target density, and since the data is independent, the proposal density can be chosen to be the prior $\pi(\theta)$. This alteration, where $q(x \mid y)$ is taken to be equivalent to $q(x)$, is known as the *Independent Metropolis-Hastings*, and is particularly useful in this setting.

Substituting the proposal and target density into the Metropolis-Hastings algorithm, one has:

$$\rho(\theta, v) = \min\left\{ \frac{\pi(v \mid \boldsymbol{x})}{\pi(\theta \mid \boldsymbol{x})} \frac{\pi(\theta)}{\pi(v)}, 1 \right\} = \min\left\{ \frac{f(\boldsymbol{x} \mid v)}{f(\boldsymbol{x} \mid \theta)} \frac{\pi(v)}{\pi(\theta)} \frac{f(x)}{f(x)} \frac{\pi(\theta)}{\pi(v)}, 1 \right\} = \min\left\{ \frac{f(\boldsymbol{x} \mid v)}{f(\boldsymbol{x} \mid \theta)}, 1 \right\}$$

This form is useful - using Independent Metropolis Hastings, one only needs to know the form of the likelihood and prior to draw from the posterior distribution. Drawing a large enough amount of samples will thus allow for the approximation of the posterior distribution without additional computation.

III.7 Hierarchy 3: Gamma-Beta Hierarchy

So far, the hierarchies that have been discussed have utilized conjugate prior distributions which lead to straightforward posterior computations. Once again, consider using the Gamma distribution with a fixed rate parameter α and a random scale β_i to model the data. However, rather than setting a conjugate Inverse-Gamma prior, one might choose to use the following:

$$\frac{\beta_i}{1+\beta_i} \sim \text{Beta}(a, b)$$

While the Beta distribution is bounded by 0 and 1 and wouldn't be appropriate to model a parameter which can have values outside of that range, this ratio is an alternative which fits our data.

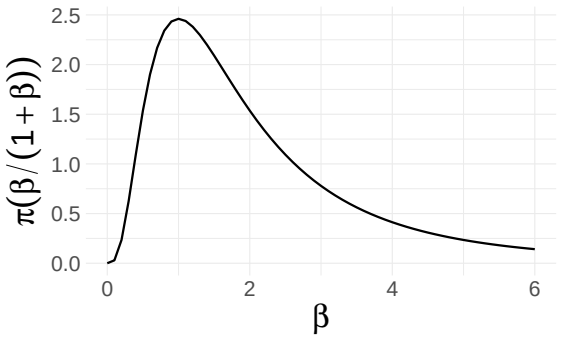

Figure III.6: Beta Prior for Hierarchy 3

III.7.1 Hierarchy

Putting the prior and likelihood together as before, we can write out the Gamma-Beta hierarchy:

$$X_{ij} \mid \beta_i \sim \text{Gamma}(\alpha, \beta_i)$$
$$\frac{\beta_i}{1+\beta_i} \sim \text{Beta}(a, b)$$

Finding an analytical form of the posterior distribution as with the previous hierarchies is very difficult (or potentially impossible) - the steps to show this are not included in this report, but it becomes quickly apparent by attempting the derivation directly. Thus, the approximation of the posterior using MCMC is necessary.

III.7.2 Using Metropolis-Hastings to Estimate the Posterior

One can re-write the Metropolis-Hastings algorithm in terms of this hierarchy for clarity:

Independent Metropolis-Hastings Algorithm for Hierarchy 3

Consider a store-specific example (dropping the i index) with orders $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. Given a starting value β_1, the steps for $t \in \{2, \ldots, S\}$ will be:

(1) Generate $\beta_{new} \sim \pi(\beta)$

(2) Set $\beta_t = \begin{cases} \beta_{new} \text{ with probability } \rho(\beta_{new}, \beta_{t-1}) \\ \beta_{t-1} \text{ with probability } 1 - \rho(\beta_{new}, \beta_{t-1}) \end{cases}$, $\rho(\beta_{new}, \beta_{t-1}) = \min\left\{\frac{f(\boldsymbol{x} \mid \beta_{new})}{f(\boldsymbol{x} \mid \beta_{t-1})}, 1\right\}$

The forms of the prior distribution as well as the ratio within $\rho(\beta_{new}, \beta_{t-1})$ are both known. Thus, the computational steps for this algorithm would be:

1. Draw β_{new} from the prior distribution by first drawing a value v from the Beta distribution with parameters a, b and then using the following equality:

$$\frac{\beta}{1+\beta} = v \iff \beta = \frac{v}{1-v}$$

2. Compute $\rho(\beta_{new}, \beta_{t-1})$ using the drawn value of β_{new} and the previous value β_{t-1}. To do so, one must find the form of $f(\boldsymbol{x} \mid \beta)$ (*Note that constant terms with respect to β can be dropped, as they will cancel*

out due to the ratio format of ρ):

$$f(\boldsymbol{x} \mid \beta) = \prod_{j=1}^{n} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x_j/\beta} \propto_{\beta} \frac{1}{\beta^{n\alpha}} \exp\left\{ -\frac{\sum_{j=1}^{n} x_j}{\beta} \right\}$$

One can see that the ratio $\frac{f(\boldsymbol{x}|\beta_{new})}{f(\boldsymbol{x}|\beta_{t-1})}$ can be found by plugging in β_{new} and β_{t-1} to $f(\boldsymbol{x} \mid \beta)$ respectively. This ratio can be compared to the value of 1 to give $\rho(\beta_{new}, \beta_{t-1})$.

3. Set β_t. This can be done by drawing a value U from the Uniform distribution between 0 and 1, and accepting the new value β_{new} if $\rho(\beta_{new}, \beta_{t-1}) > U$.

4. Repeat this process many times (i.e., $S = 10,000$), keeping track of the number of times a new value is rejected, and storing β_t for $t \in \{2, \ldots, S\}$. The number of times a new value is rejected can indicate whether the chosen stationary distribution is reasonable, while the values of β_t give an empirical distribution which approximates the posterior.

III.7.3 Bayes Estimator

Since the Bayes estimator under squared error loss is defined as the posterior mean, one simply computes the average of the β_t values accepted by the Independent Metropolis-Hastings algorithm (this is the empirical mean of the posterior distribution).

III.7.4 Metropolis-Hastings Example

While the first two hierarchies discussed have a closed-form solution for the Bayes estimator, this third hierarchy requires the use of MCMC methods to approximate the posterior as has been discussed. The following is an example of the computation of the Bayes estimator under this third hierarchy.

Consider an individual store i with $n_i = 21$ total orders and mean time between orders $\bar{\boldsymbol{x}}_i = 1.25$. Setting parameters $\alpha = 2, a = 5, b = 5$, one can follow the hierarchy and steps as described in Section III.6.2 to run the Independent Metropolis-Hastings algorithm (the R code for this algorithm can be found in the Appendix). The resulting values can be seen plotted below:

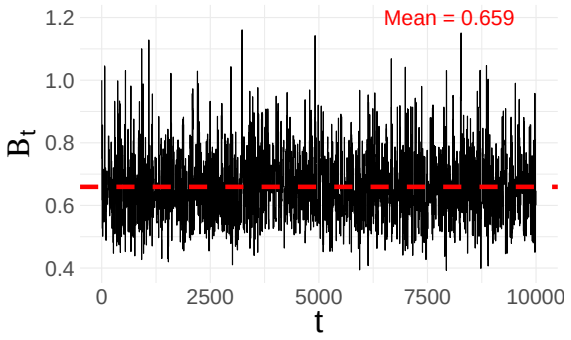

Figure III.7: Trace Plot of MCMC Draws

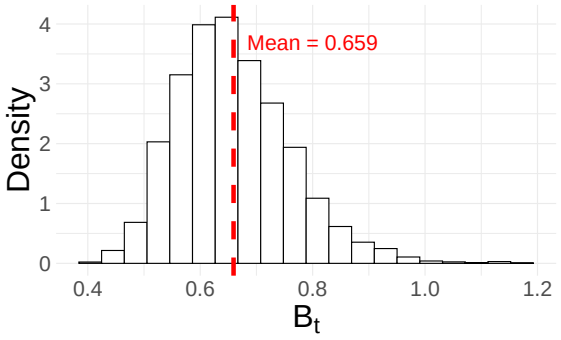

Figure III.8: Empirical Posterior Distribution

Figure III.7 depicts what is called a 'trace plot' of the MCMC sample made up of the $S = 10,000$ draws using the algorithm. The 'chain' (the pattern made by the line) seems to explore the state/sample space appropriately - that is, the sample seems to be relatively consistent, and evenly dispersed across all the draws (for more background/elaboration on MCMC diagnostics, see [3]). Figure III.8 shows the empirical distribution of the draws, which appears similar to what we would expect - the highest density areas are centered around the mean, as opposed to the distribution being much more spread out or irregular.

As mentioned previously, the Bayes estimator in this case is simply the sample mean of the empirical distribution; $\hat{\beta} = 0.659$. Recall from the hierarchy (for a store-specific example), $X_j \mid \beta \sim \text{Gamma}(\alpha, \beta)$, where

$\alpha = 2$ is set and β is treated as a random variable. One can substitute the Bayes estimator, giving:

$$\mathbb{E}[X \mid \hat{\beta}] = \alpha\hat{\beta} = (2)(0.659) = 1.318$$

This is quite close to the mean time between orders, $\bar{\boldsymbol{x}}_i = 1.25$, and is an indication that the approximation resulted in an appropriate value.

Furthermore, the Bayes estimator can be used to plot the Gamma distribution over the densities of the actual data (for 21 total orders), which can be seen to the right. For the store which was used in this example, one can see that the Gamma distribution with 'rate' parameter $\alpha = 2$ and 'scale' parameter $\hat{\beta} = 0.659$ follows a similar pattern to the distribution of the data.

Note: This is not unique to this hierarchy - one can compare the theoretical distributions from each of the hierarchies to the store data and will see similar patterns.

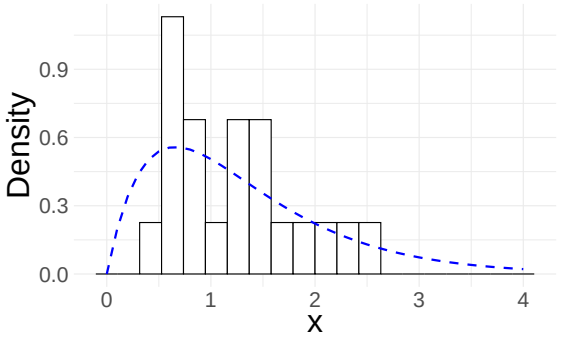

Figure III.9: Customer Data vs Gamma Fit

III.8 Evaluating and Comparing the Hierarchies

In order to compare hierarchies, one must choose from a variety of options. Leave-One-Out Cross-Validation seemed most appropriate in this case, where for each order by each store, each hierarchical model can be fit using the remaining orders, and the expected value for the model can be compared to the order that was left out. The steps for doing this can be seen in more detail below, and were adapted from Chapter 5 of [2].

III.8.1 Steps for Leave-One-Out Cross-Validation (LOOCV)

The Leave-One-Out Cross-Validation Error 'CV' can be computed using the following steps. Depending on which hierarchy one chooses to use, parameters which are fixed remain consistent for all stores (as they are *prior* knowledge and not meant to be store-specific), but parameters may be different across the different hierarchies. *Note that references to 'orders' is in regards to the transformed variables which represent the time between orders (X_{ij}).*

The goal of LOOCV is to, for each data point, fit the model using the remaining data, estimate the missing data point, and then compare the difference between the estimated and true values. Since there are many different customers, this will be done for each customer individually first, and then the mean squared error (MSE) for each customer will be averaged to produce the overall Cross-Validation error (CV).

Cross-Validation Algorithm for the Sticker Data:

1. For each customer $i \in \{1, \ldots, N\}$,
   (a) For each order $j \in \{1, \ldots, n_i\}$,
      i. Define the data without the j-th observation as $\boldsymbol{x}_i^{-j} = \{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{n_i}\}$.
      ii. Compute the Bayes estimator based on the shortened data; $\hat{\theta}_i(\boldsymbol{x}_i^{-j})$.
      iii. Define $\hat{\theta}_i = \hat{\theta}_i(\boldsymbol{x}_i^{-j})$, and compute $\hat{x}_{ij} = \mathbb{E}(X_{ij} \mid \hat{\theta}_i)$.
      iv. Compute the squared error by comparing the true value to the estimate: $\text{SE}_j = (x_{ij} - \hat{x}_{ij})^2$.
   (b) Compute the customer-specific mean squared error: $\text{MSE}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \text{SE}_j$.
2. Compute the overall Cross-Validation error as: $\text{CV} = \frac{1}{N}\sum_{i=1}^{N} \text{MSE}_i$.

This overall estimate, CV, gives an idea of the accuracy of each model, and can be used to compare the hierarchies directly with a single value - where the model with the lowest CV can be considered to be the 'best' option. It is important to note that picking different fixed parameters (i.e. α, a, b) will change these errors. Furthermore, one might choose to use LOOCV to evaluate the choice of parameters.

III.8.2 LOOCV Results for Each Hierarchy

Leave-One-Out Cross-Validation errors were computed for each of the three hierarchies using the steps outlined in III.7.1. This was done using five different cutoffs of 4, 6, 8, 10, and 12 total orders (only including customers with at least that many orders). Additionally, the sample mean was used in place of a fourth hierarchy, where predictions were computed using the average time between orders from the customer instead of using a Bayesian hierarchy ($\hat{x}_{ij} = \bar{\boldsymbol{x}}_i^{-j}$). This was chosen because the sample mean is the maximum likelihood estimator (MLE) for the parameters λ and β in the Poisson and Gamma distributions respectively, and provides a non-Bayesian estimation procedure for comparison.

Cutoff	Hierarchy 1	Hierarchy 2	Hierarchy 3	Mean
4	10.8570	10.1503	10.0009	10.9654
6	7.7472	7.1724	7.043	7.5490
8	5.2086	4.8192	4.7541	4.9989
10	3.1002	2.7986	2.7808	2.8691
12	1.7925	1.5748	1.5781	1.5956

Table III.1: LOOCV Results

It can be seen that Hierarchies 2 and 3 performed better than Hierarchy 1 and the baseline approach of using the sample mean. While Hierarchy 1's predictive ability was likely impaired by the use of discrete as opposed to continuous values to represent the time between orders, these results support the decision to use a Gamma as opposed to a Poisson process to model the transformed 'time between orders' invoice data from Drew's customers.

Furthermore, while the continuous data was used for Hierarchies 2 and 3 as well as the mean, it can be seen that the two Bayesian hierarchies resulted in a lower cross-validation error for all three cutoffs. Additionally, the difference between the two Gamma hierarchies and the mean increased as the cutoff decreased, supporting the idea that the Bayesian approach of incorporating prior information is beneficial for smaller samples.

III.9 Implementation of the Models

Recall the purpose of this project; to create a model which can be used to determine when Drew should reach out to a customer about placing a new order. In this section, two approaches for this are proposed, and are both based on the following:

Consider the setting in which all invoices up to the current day have been collected. In other words, for each customer, the transformed data representing the time between orders ($\boldsymbol{X}_i = \boldsymbol{x}_i$) is available. Furthermore, for each customer, the time between the most recent order and the current date has been computed in months, which will be denoted c_i. For instance, if customer i most recently placed an order exactly one month earlier, then $c_i = 1$. One can follow the steps outlined in Sections III.1-3 to construct a hierarchy to model this data and find Bayes estimators for each store.

III.9.1 Predicting Orders Within an Interval

If the interest is in predicting when a future order will be placed, one can choose an interval of time and then compute the probability of an order being placed within that interval. Furthermore, the time since the most recent order can be used as additional information. For instance, if it is known that no order has been placed in the last c_i months, then the probability of an order being placed in the next m months can be conditioned on this information.

Suppose that it has been c_i months since the most recent order for store i, and the interest is in computing the probability of an order being placed within the next m months. This can be computed as:

$$P(c_i \leq X \leq c_i + m \mid X \geq c_i)$$

This is equivalent to the ratio of the region in Figure III.10 which is colored in green (for $c_i = 1$ and $m = 1$) compared to the rest of the non-red area under the curve.

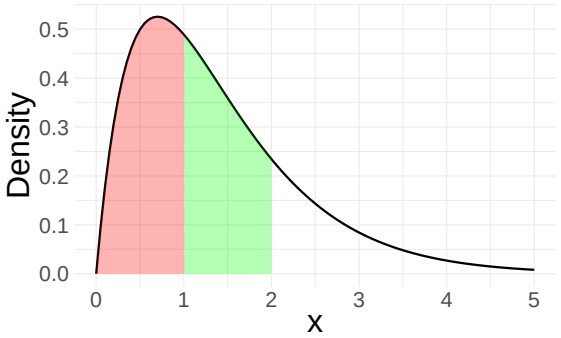

Figure III.10: The Interval, Visualized

Furthermore, one can show

$$P(c_i \leq X \leq c_i + m \mid X \geq c_i) = \frac{P(c_i \leq X \leq c_i + m)}{P(X \geq c_i)},$$

or more generally:

$$P(a \leq X \leq b \mid X \geq c) = \frac{P(a \leq X \leq b)}{P(X \geq c)}, \text{ where } b \geq a \geq c \geq 0.$$

Note that this form can be used to find the probability of an order within any desired interval, as long as the bounds of the interval are set appropriately.

Then, using the resulting probability, one can decide whether it is worth it to reach out during a given time period.

III.9.2 Picking When to Reach Out

Rather than computing the probability of an order falling within some interval, one might instead wish to identify a threshold which could be used to determine when to reach out to a customer. For instance, if the model assigns a probability of 75% to the chance of a customer placing an order within the first two months after the previous one, it would be reasonable to reach out if they hadn't done so by that time.

In this context, it makes sense to can use the cumulative density function (c.d.f.) to represent the probability of an order being made by at least some time period c, which can be written as:

$$F_X(c) = P(X \leq c) = \int_0^c f(t)\ dt,$$

where $f(x)$ is the p.d.f. of X.

It is straightforward to use the c.d.f. with the c_i values computed for each customer to find the probability of an order being placed by the current date. Then, if the model suggests a probability of an order within said interval that exceeds the set threshold, it is reasonable to reach out to see if that customer would want to place an order.

For example, suppose an individual customer for which the Bayes estimator for Hierarchy 2 was computed to be $\hat{\beta} = 0.7$ for $\alpha = 2$. Furthermore, suppose that the time since the most recent order is exactly 2 months. Then, the probability of an order being placed within the time since the last order could be computed as $P(X \leq 2)$. This can be visualized in Figures III.11 and III.12, where the red region and point both indicate the probability $P(X \leq 2)$ for $X \sim \text{Gamma}(\alpha = 2, \hat{\beta} = 0.7)$.

One could set a threshold to use as an indication for when it would be best to reach out to a customer - a decision to be made at the discretion of the user. In this case, if the threshold was chosen to be 75%, this method would indicate that the customer should be contacted, since the probability of them having placed

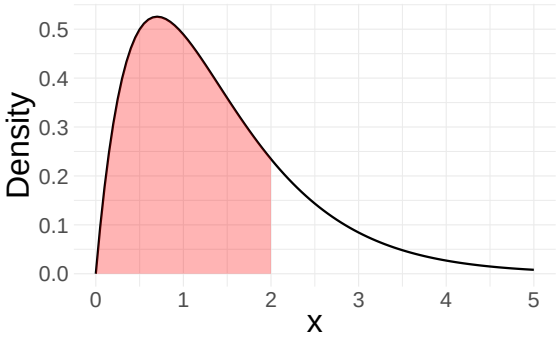

Figure III.11: Interval Visualized (p.d.f.)

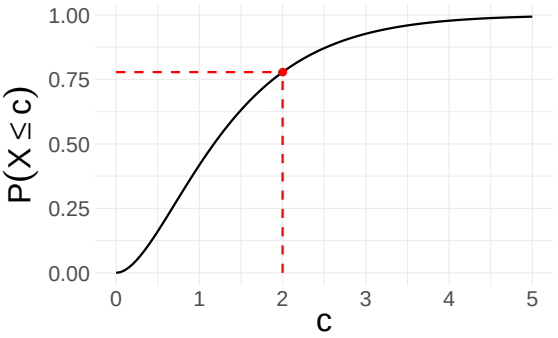

Figure III.12: Interval Visualized (c.d.f.)

an order within the last two months exceeds that threshold. Further analysis could be done to optimize this choice, but will not be covered in this project.

Overall, a prediction process might look like the following (using Hierarchy 2 for reference):

1. Compute the Bayes estimator $\hat{\beta}_i$ for the current customer. This will be used for the computation of the c.d.f.
2. Get the date of the most recent order, as well as the current date. Compute the difference in months between the two dates, which can be denoted as c.
3. Compute the probability of an order having been made in that time period; that is $P(X \leq c)$.
4. If that probability is above the chosen threshold, reach out to the customer about placing a new order.
5. If that probability is not above the chosen threshold, one may consider computing the time until the threshold is reached; that is, finding the number of months m such that $P(X \leq m) > t$, where t is the chosen threshold. Then one could wait until it has been that long to reach out instead.

Again, this is not the only way to make predictions, but it seems the most straight forward of the options available for this setting.

IV Discussion

This paper has explored the development and testing of three different models meant to predict sticker sales using approaches such as Bayesian hierarchies and Metropolis-Hastings. The use of the Bayesian framework was motivated by the size of the data, where many customers had very few individual orders. The Bayesian approach of incorporating prior knowledge sought to combat this issue; where those with small sample sizes could be 'supported' by an appropriately assigned prior. While the processes used represent steps taken for the specific data provided by Drew Madden's sticker sales, it is my hope that they can be easily followed for future applications.

Leave-One-Out Cross-Validation results support the choice of Bayesian methods to work with this data, as the second and third hierarchies outperformed the baseline example which utilized the non-Bayesian Maximum Likelihood Estimator (MLE). While the first hierarchy had the largest cross-validation error for nearly all cutoffs, it is likely that this is due to information loss from treating the data as discrete. It would be interesting to compare the performance of the hierarchies if the time between orders was measured in days rather than months, as this would solve the issue of information loss. Furthermore, for smaller cutoffs in which customers with very few orders were included, LOOCV results showed a greater difference between the second and third hierarchies' performances compared to that of the MLE. Some research has suggested that the inclusion of an appropriate prior can be advantageous for smaller samples [6], which may be the case here.

A limitation of the data stems from the way that orders are placed through Drew's store. In some cases, customers will reach out directly about placing a new order, and in others it will not be until Drew reaches out himself that a customer does so. Thus, there may be cases where a customer is *ready* to place an order, but since they hadn't been contacted, the time at which they eventually do is a bit later on. This project was motivated by this issue; where the goal was to create a tool which could be used to improve the process of reaching out to customers at the best time. Future data collection would be benefited by including information about this, as well as maintaining records of when customers were contacted and chose not to place new orders.

A natural continuation would be to implement one of these hierarchies as a part of Drew's business and observe their effectiveness. Additionally, while the data used in this project revolves around the times at which orders are placed, there was no inclusion of the *size* of the orders. It is natural to presume that, for a single customer, larger orders should lead to an increased amount of time before the next one in comparison to a smaller order. This information could be added to the hierarchies discussed in some ways, and might improve their predictive capabilities.

References

[1] James O. Berger. *Statistical Decision Theory: Foundations, Concepts, and Methods.* Springer Series in Statistics, 1980.

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer Texts in Statistics, 2nd edition, 2013.

[3] Taboga, Marco (2021). "Markov Chain Monte Carlo (MCMC) diagnostics", Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. Online appendix.

[4] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods.* Springer Science+Business Media Inc., 2nd edition, 2004.

[5] Mark J. Schervish and Morris H. DeGroot. *Probability and Statistics.* Addison-Wesley, 4th edition, 2012.

[6] Rens van de Schoot, Joris J. Broere, Koen H. Perryck, Mariëlle Zondervan-Zwijnenburg, and Nancy E. van Loey. "analyzing small data sets using bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors". *European Journal of Psychotraumatology*, 2015.

Acknowledgements

A special thanks to Prof. Peter Kramlinger for his guidance and support throughout the entirety of this project. He spent many hours meeting with me to discuss the progress I was making, and many more reviewing my work on his own time to provide valuable insights and direction.

Similarly, I wanted to say thank you to Drew Madden for allowing me to use his sales data. If interested, Drew's stickers can be found online[1], as well as in many stores throughout California.

[1] In case the hyperlink is not working: **www.drewmaddenart.com**

Code Appendix

```r
## Metropolis-Hastings Algorithm for Section III.6.4 Example ##

# Example Data
X <- c(0.8869908, 0.5913272, 0.7884363, 0.9198423, 1.8396846, 0.9526938, 1.3469120, 0.6570302,
       1.5440210, 0.5256242, 2.2010512, 1.3469120, 1.6754271, 0.5913272, 1.3140604, 0.5913272,
       1.5111695, 1.3797635, 0.6898817, 2.3981603, 2.5295664)

# Initialize hyperparameters, compute n
a <- 5; b <- 5; alpha <- 2
n <- length(X)

f <- function(Beta) { # Likelihood
  return(1 / (Beta^(n*alpha)) * exp(-sum(X) / Beta))
}

rho <- function(Beta_new, Beta_prev) {
 return(min(f(Beta_new) / f(Beta_prev), 1))
}

S <- 10000 # Set number of repetitions

Beta_Values <- numeric(length = S)
Beta_Values[1] <- 1
num_rejects <- 0

for (t in 2:S) {
  # Step 1:
  v <- rbeta(n = 1, shape1 = a, shape2 = b)
  Beta_New <- v / (1 - v)

  # Step 2:
  RHO <- rho(Beta_New, Beta_Values[t-1])

  # Step 3:
  U <- runif(1, min = 0, max = 1)
  if (RHO > U) {
    Beta_Values[t] <- Beta_New
  } else {
    Beta_Values[t] <- Beta_Values[t-1]
    num_rejects <- num_rejects + 1
  }
}

# Plot of Beta Values
plot(Beta_Values, type = 'l', xlab = expression(t), ylab = expression(B[t]))

# Plot Example Data and Gamma Fit
hist(X, prob = TRUE, col = "lightblue", main = NULL, xlab = expression(X[i]),
     ylab = "Density", xlim = range(0,5), ylim = range(0,1), breaks = 10)

curve(dgamma(x, shape = alpha, scale = mean(Beta_Values)), add = TRUE,
             col = "darkblue", lwd = 2, lty = "dashed")
```