I^2SDS
The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2009-3
March 12, 2009

Bayesian Analysis of Abandonment in Call Center Operations

Tevfik Aktekin
Institute for Integrating Statistics in Decision Sciences
The George Washington University, USA

Refik Soyer
Department of Decision Sciences
The George Washington University, USA

Bayesian Analysis of Abandonment in Call Center Operations

Tevfik Aktekin
Department of Decision Sciences
The George Washington University

Refik Soyer
Department of Decision Sciences
The George Washington University

Abstract

In this paper we consider modeling abandonment behavior in call centers. We present several time to event modeling strategies and develop Bayesian inference for posterior and predictive analyses. Different family of distributions, piecewise time to abandonment models and mixture models are introduced and their posterior analysis is carried out using Markov chain Monte Carlo methods. We illustrate implementation of the proposed models using real call center data, present additional insights that can be obtained from the Bayesian analysis and discuss implications for different customer profiles.

1 Introduction and Overview

Call center operations typically consist of three fundamental processes of arrival, service and abandonment. One of the main challenges that modern day call center practitioners are faced with is to find a balance between efficiency of the call center operation and quality of service offered to the customers. Careful study of these processes may provide managerial insights about several operating characteristics of the underlying queuing system and could help practitioners to find such a balance.

The focus of this paper will be on the study of the abandonment process in call centers. Abandonment, or patience, is defined as the time a customer is willing to wait before abandoning the queue (see Mandelbaum and Shimkin (2000)). Call center customers expect fast and efficient service, and if their expectations are not met, abandonments will occur more frequently and might lessen the perceived quality of a call center. According to Mandelbaum and Shimkin (2000), AT&T studies indicate that a 15 seconds wait for an agent caused a 44% abandonment rate; for a 30 seconds wait that figure increased to 69%. Also, the Help Desk Institute, in its annual report, indicates that around 43% of all call centers have a targeted abandonment rate, and about 40% of the call centers observe an abandonment rate of over 10%. In toll free services, service providers are required to pay the holding times of their customers. Thus, study of the abandonment process in call centers plays a crucial role also from an economic point of view.

Most of the literature on abandonment processes in call centers is from a queuing theory perspective and emphasizes modeling and development of associated performance measures. There is a lack of research in statistical inference with the exception of Brown et al. (2005) who analyze an anonymous call center from a queuing science perspective by focusing on statistical aspects of the three fundamental processes. In this paper we will attempt to fill this gap by analyzing the abandonment behavior exhibited by different call center customer profiles and in so doing, we will take a Bayesian viewpoint.

The model that is most commonly used in call center analysis is the so called M/M/s (Erlang-C) model where customers are assumed to have infinite patience. Garnett et al. (2002) make an argument about why the Erlang-A model (an M/M/s+M queue where customers' time to abandon the queue follows an exponential distribution) is superior from an optimal staffing point of view and provide comparisons of performance measures for Erlang-C and Erlang-A models. From a practical

point of view, the only parameter that the call center managers have control over is the number of agents working for a call center during a specific period of time. Call center management consists of labor intensive operations, where staffing comprise 60-80% of the overall operating budget (Aksin et al. (2007)). Therefore, a detailed study of the abandonment processes in call centers and its effect on staffing will be of interest to call center practitioners. Garnett et al. (2002) point out that most of the literature and practice in call center operations ignore the effects of abandonment which leads to either over or under staffing. The Erlang-A model is based on the assumption of exponentially distributed abandonment times, an extension is discussed in Bacelli et al. (1984) who introduce a queuing system with general abandonment distribution, G and obtain certain operating characteristics. Brandt and Brandt (1997) provide an extension in the form of a birth and death process for the $M(n)/M(m)/s+G$ model where n is the number of callers in the system, and $m = min(n, s)$ is the number of busy servers. More recently, Zeltyn and Mandelbaum (2005) provide a summary of operating characteristics for the M/M/s+G with additional operating characteristics. Brown et al. (2005) provide non-parametric estimates of the hazard rate for the abandonment process and point out that abandonment does not exhibit exponential behavior in their data set.

An important issue associated with abandonment in call centers is the effect of announcements on system performance. Typically, one can think of two types of announcements. The first is where an announcement is made regarding the expected wait in the queue (or the exact position of the caller in the queue) during fixed intervals, the second is where the customer is informed that all lines are busy and asked to hold on the line until a server becomes available. In this paper, we will adopt the latter definition and will investigate whether different customer profiles change their abandonment behavior as they experience announcements. In order to improve system service along with customer satisfaction, call center management will be interested in predicting delays or the effects of informing customers of expected delays. Whitt (1999a), Whitt (1999b) and Aksin et al. (2008) investigate the issue of delays, informing customers of these delays and their effect on system performance. Zohar et al. (2002) make the argument that customer patience is a function of several covariates and study its behavior with respect to mean waiting time in the queue.

What would be of interest to call center practitioners is a detailed study of the abandonment distribution for different types of customers for call center design and staffing purposes. For

instance, new/potential customers might exhibit different abandonment behavior as opposed to regular ones. Another issue of interest is whether call center customers exhibit monotonic or non-monotonic abandonment behavior, namely whether customers change their abandonment behavior as they experience announcements or as they wait in the line. In this paper we introduce models to describe abandonment process in call centers for different customer profiles. We also develop Bayesian analysis of these models using Markov chain Monte Carlo methods. Duration models such as generalized gamma family of distributions as well as piecewise and mixture models are considered for describing customer abandonment behavior motivated by the behavior of their hazard rates (or abandonment rates). To the best of our knowledge, these models, that are capable of capturing non-monotonic abandonment rates, have not been considered in the call center modeling literature. Furthermore, Bayesian view point has not been previously implemented in analysis of call center abandonment data.

A synopsis of our paper is as follows. In section 2, we cover the preliminaries and the properties of our proposed models. The Bayesian analysis of our proposed work, predictive analysis and model comparison criteria will be summarized in Section 3. We will illustrate the proposed models using real call center abandonment data with different customer profiles in Section 4. Section 5 will conclude our study with further remarks and potential future work.

2 Modeling the Abandonment Rate in Call Centers

Palm (1953) is the first one who has pointed out the relationship between the impatience of a customer and the hazard rate. Brown et al. (2005) represent the hazard rate of the patience distribution and virtual waiting time via non-parametric estimates. Mandelbaum and Shimkin (2000) study how changes in the hazard rate for the abandonment distribution effects the rational decision of when to abandon the system. They provide optimal rules of abandonment for cases when the hazard rate is increasing, decreasing and increasing-decreasing.

Time to abandonment of a customer in a virtual queue can be considered as similar to the time to failure of an item in reliability/survival analysis. Reliability function (or survival function) is defined as $F(t|\theta) = P(T \leq t|\theta)$, where $t \geq 0$ and T is said to be the life length of an item. A closely

related concept is the model failure rate (or hazard rate) defined as

$$r(t) = \lim_{t\to\infty} \frac{P(t \leq T \leq t + dt | T \geq t)}{dt} = \frac{f(t)}{1 - F(t)}. \tag{2.1}$$

The failure rate can be thought as a measure of the risk that a failure will occur at t. For small dt, it can be interpreted as the probability that an item of age t will fail in $(t, t + dt)$; see Singpurwalla (2006). This provides a natural way of thinking about the abandonment rate of a customer in a virtual queue. In other words, following (2.1), the random variable T can be thought as the time a customer will wait before abandoning the queue. Therefore $r(t)$ can be referred to as the customer abandonment rate in a call center.

In what follows, we introduce different classes of time to event (duration) models to describe the customer abandonment behavior in call centers and discuss their properties. These models are motivated by the behavior of abandonment rates observed in actual call centers.

2.1 Generalized Gamma Family of Models

The generalized gamma distribution was first introduced by Stacey (1962) in the context of reliability. Its flexible structure makes it a good candidate for modeling time to event phenomenon, in reliability and survival analysis. Pham and Almahana (1995) discuss its properties and the behavior of its hazard rate for different values of its parameters. Dadpay et al. (2007) integrate it into the information theoretic literature, discuss additional properties and propose its Bayesian estimation. In what follows, we will discuss its properties in the context of abandonment in call centers.

Let T denote the time to abandonment of customers in a call center. The density function of the generalized gamma random variable T is

$$f(t|\alpha, \gamma, \lambda) = \frac{\gamma \lambda^{\alpha} t^{\alpha\gamma - 1}}{\Gamma(\alpha)} exp\{-\lambda t^{\gamma}\}, \tag{2.2}$$

where $t \geq 0$, $\alpha, \gamma, \lambda > 0$, α and γ are shape parameters and λ is the scale parameter.

As discussed in Pham and Almahana (1995) and Dadpay et al. (2007), several well known distributions can be obtained using the generalized gamma family parametrization. For instance, for $\gamma = \alpha = 1$ one can obtain the exponential distribution, for $\alpha = 1$ the Weibull distribution, for

$\gamma = 1$ the gamma distribution, for $\gamma = 2, \alpha = 1/2$ the half-normal distribution, for $\gamma = 2, \alpha = 1$ the Rayleigh distribution, for $\gamma = 2, \alpha = 3/2$ the Maxwell-Boltzmann distribution and for $\gamma = 2, \alpha = k/2, (k = 1, \cdots)$ the Chi distribution. In the limiting case, as $\alpha \to \infty$, the lognormal model can be obtained.

The cumulative distribution for the generalized gamma model is given by

$$F(t|\alpha, \gamma, \lambda) = \frac{\Gamma_{\lambda t^\gamma}(\alpha)}{\Gamma(\alpha)}, \tag{2.3}$$

where

$$\Gamma_{\lambda t^\gamma}(\alpha) = \int_0^{\lambda t^\gamma} x^{\alpha-1} exp(-x) dx. \tag{2.4}$$

Its power moment is given by

$$E(T^s|\alpha, \gamma, \lambda) = (\frac{1}{\lambda})^{\frac{s}{\gamma}} \frac{\Gamma(\alpha + s/\gamma)}{\Gamma(\alpha)}, \text{ for } s > 0. \tag{2.5}$$

Combining (2.3) and (2.4) the hazard rate can be obtained as

$$r(t|\alpha, \gamma, \lambda) = \frac{\gamma \lambda^\alpha t^{\alpha\gamma-1} exp\{-\lambda t^\gamma\}}{\Gamma(\alpha) - \Gamma_{\lambda t^\gamma}(\alpha)}. \tag{2.6}$$

An attractive feature of the generalized gamma family is the flexibility of its hazard rate function which can be used to represent non-monotonic unimodal or bathtub shaped hazard functions (different shapes of the hazard rate can be seen in Figure 1). This is desirable in abandonment process modeling, since as discussed in Brown et al. (2005), the hazard rate can exhibit non-monotonic behavior. It can be shown that for $\gamma \neq 1$, if $(1 - \alpha\gamma)/[\gamma(\gamma - 1)]$ is a strictly positive constant, then the hazard rate is bathtub shaped for $\gamma > 1$ and inverse bathtub shaped for $0 < \gamma < 1$. Otherwise, the hazard rate is increasing for $\gamma > 1$ and decreasing for $0 < \gamma < 1$.

As pointed out previously, in the limit when $\alpha \to \infty$, one can obtain the lognormal family of models as a special case of the generalized gamma model. The density function of the lognormal random variable is given by

$$f(t|\mu, \sigma) = (t\sqrt{2\pi}\sigma)^{-1} exp\{-\frac{1}{2}\frac{(log(t) - \mu)}{\sigma^2}\}, \; t > 0. \tag{2.7}$$

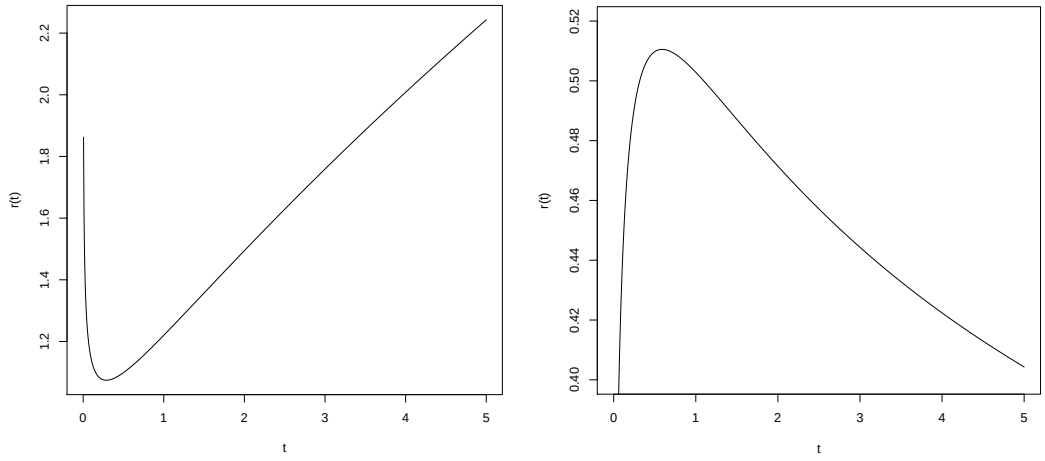

Figure 1: Hazard Rates for Generalized Gamma; $\alpha = 0.5, \lambda = 0.5, \gamma = 1.6$ (left) and $\alpha = 2, \lambda = 1.33, \gamma = 0.66$ (right)

The hazard rate exhibits non-monotonic behavior in t as seen in Figure 2. If the customers exhibit an increasing abandonment rate followed by a decrease than the lognormal density will be a proper candidate for abandonment behavior modeling. In other words, customers may be more patient when they join the queue but after waiting a certain amount of time they might start abandoning followed by a period of decreasing abandonment rate. This would indicate a lognormal type of abandonment behavior as shown in Figure 2. The hazard rate of the lognormal density can be obtained as

$$r(t|\mu,\sigma) = \frac{f(t|\mu,\sigma)}{1 - \Phi(log(T),\mu,\sigma)}, \tag{2.8}$$

where $f(t|\mu,\sigma)$ is given in (2.7) and $\Phi(log(T),\mu,\sigma)$ is the cumulative density function of the normal random variable $log(T)$ with parameters μ and σ.

2.2 Mixture Models

An alternate modeling strategy to describe abandonment behavior in call centers is use of finite mixture models. What makes the finite mixture models attractive from a modeling point of view is their flexibility in terms of the hazard rate function. Different combinations of family of distributions can capture several different shapes for the hazard rate along with bimodal type of behavior in the respective densities. As pointed out by Diebolt and Robert (1994), mixture models can be thought

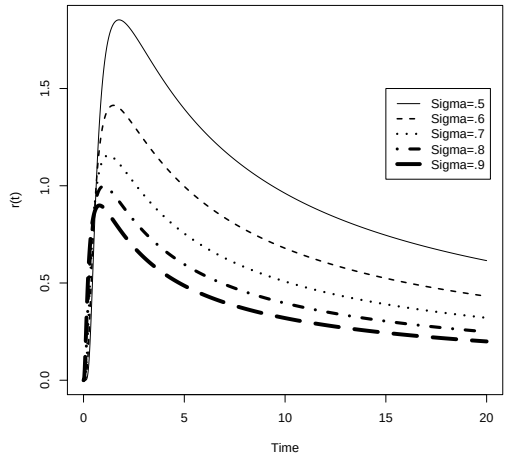

Figure 2: Hazard Rates for the Lognormal Distribution

of as an alternative to non-parametric models and they are less restrictive as opposed to standard parametric distributions.

Following Gilks et al. (1996), a finite mixture model for $f(t)$ can be defined as

$$f(t|\theta_1, \cdots, \theta_k) = \sum_{i=1}^{k} p_i f(t|\theta_i), \tag{2.9}$$

where $\sum_{i=1}^{k} p_i = 1$ (usually referred to as mixing weights), $k \geq 0$ is any integer, and $f(t|\theta_i)$ for $i = 1, \cdots, k$ represent the densities of k different components (also referred to as the components of the mixture). As pointed out by Gilks et al. (1996), depending on the context the mixture is being used for, the components of the mixture may or may not have any physical meaning. For instance in reliability, one can think of the individual components as items coming from different populations (such as items manufactured by different machines, factories, etc...). For the abandonment modeling purposes we will not attempt to assign a physical meaning to the individual components and will simply refer to them as components of the abandonment mixture.

Suppose that $k = 2$ in (2.9), then the mixture density can be written as

$$f(t|\theta_1, \theta_2) = pf(t|\theta_1) + (1-p)f(t|\theta_2), \tag{2.10}$$

and the mixture hazard rate is given by

$$r(t|\theta_1, \theta_2) = w(t)r(t|\theta_1) + [1 - w(t)]r(t|\theta_2), \tag{2.11}$$

where $0 \leq w(t) \leq 1$ and

$$w(t) = \frac{pF(t|\theta_1)}{pF(t|\theta_1) + (1-p)F(t|\theta_2)}. \tag{2.12}$$

It follows from (2.12) that

$$min\{r(t|\theta_1), r(t|\theta_2)\} \leq w(t) \leq max\{r(t|\theta_1), r(t|\theta_2)\}. \tag{2.13}$$

Wondmagegnehu et al. (2005) discuss the behavior of the hazard rate for a mixture density with two components for different distributions, discuss further properties in the limits and point out how bathtub type of hazard rate behavior can be obtained under different parametrization of the mixture.

Gilks et al. (1996) point out that maximum likelihood estimation of mixture models is not straightforward in most cases and it may not exist in others. Therefore using Bayesian methods for parameter estimation is the natural way to approach the problem. Gilks et al. (1996) discuss a general Bayesian estimation method for finite mixtures for the exponential family of distributions. A Bayesian estimation via Gibbs sampling and data augmentation can be implemented in an efficient manner as will be discussed in the sequel.

In our development, we will attempt to model T with k mixtures regardless of when the announcements are made. In doing so, we will consider lognormal and Weibull densities as components of the mixture. For instance, the density plots for mixture models with two Weibull and three lognormal components are shown in Figure 3 (with equal mixing probabilities). A similar behavior is observed in actual time to abandonment data shown in Figure 4, suggesting evidence for presence of mixtures.

2.3 A Piecewise Time to Abandonment Model

A preliminary analysis of call center abandonment data, as done in Brown et al. (2005), suggests that upon joining the queue abandonment rate is increasing, followed by a bathtub and then a fairly constant type of behavior. Such behavior can also be observed from the density plots of time to

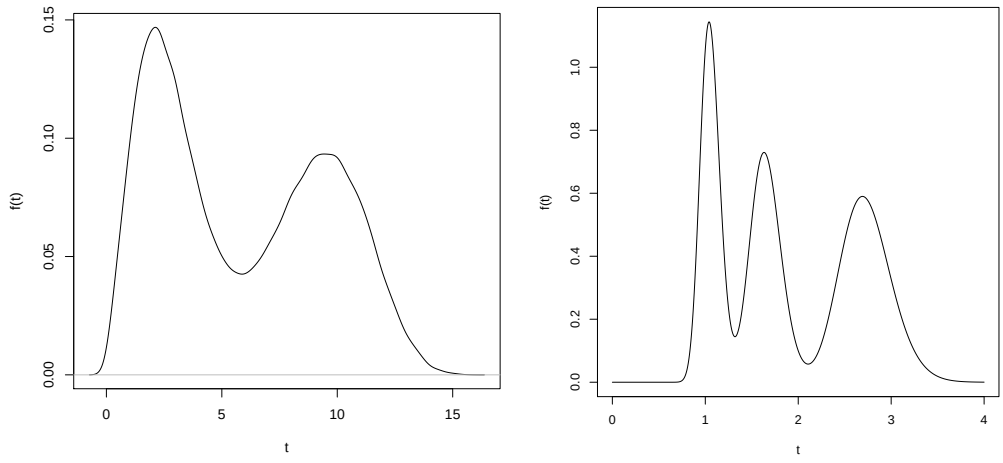

Figure 3: Mixture Density Plots with Weibull Components $k = 2$ (left) and Lognormal Components $k = 3$ (right)

abandonment given in Figure 4. Here we can see that there is a peak around 5-10 seconds and this is followed by a second peak at 60 seconds which coincides with the first announcement time. This suggests that the abandonment rate can exhibit different behavior before or after announcements. Thus, as an alternative strategy, a piecewise model with switching points defined by announcement times can be considered for describing the behavior of abandonment rate.

As before, let T denote the time to abandonment. Then, a piecewise time to abandonment density for T with switching points as announcements is defined by

$$f(t|\boldsymbol{\Theta}) = \begin{cases} Cf_0(t|\boldsymbol{\theta_0}), & \text{for} \quad 0 < t \leq \tau_1 \\ Cf_1(t|\boldsymbol{\theta_1}), & \text{for} \quad \tau_1 < t \leq \tau_2 \\ \qquad\qquad ... & \\ Cf_k(t|\boldsymbol{\theta_k}), & \text{for} \quad \tau_k < t < \infty \end{cases} \tag{2.14}$$

where

$$C\int_0^{\tau_1} f_0(t|\boldsymbol{\theta_0}) + \cdots + C\int_{\tau_k}^{\infty} f_k(t|\boldsymbol{\theta_k}) = 1, \tag{2.15}$$

and $\boldsymbol{\Theta} = \{\boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_k}, \tau_1, \cdots, \tau_k\}$. The switching point τ_i represents the time of the i^{th} announcement (for example, in our call center data every 60 seconds there is an announcement), $f_i(t|\boldsymbol{\theta_i})$ is the piecewise density of the abandonment rate before the $(i+1)^{th}$ announcement and $\boldsymbol{\theta_i}$ is the

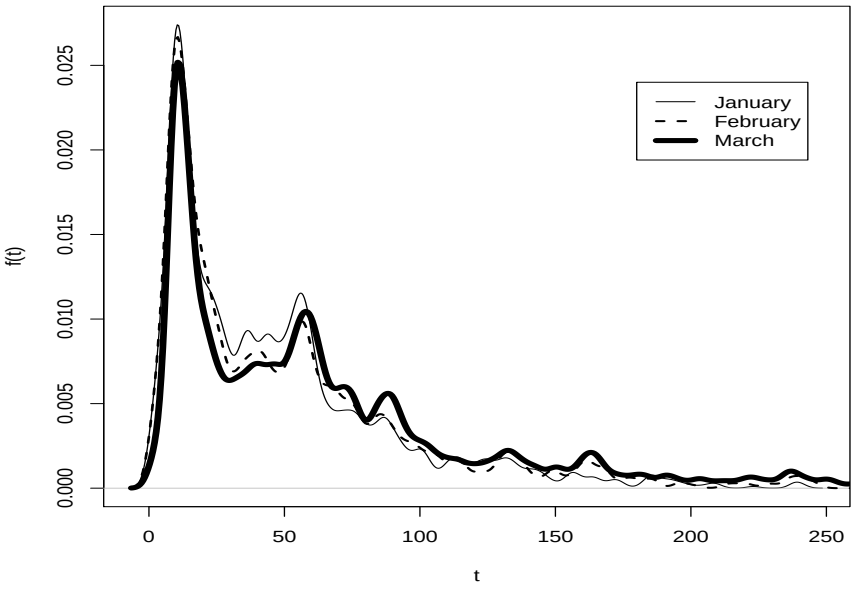

Figure 4: Time to Abandonment Densities for January, February, March of Regular Customers

corresponding vector of paramaters for $i = 1, \cdots, k$.

Different types of customer behavior before or after announcements can be captured in the above model where lognormal or Weibull models can be used as density functions.

3 Bayesian Inference for Abandonment Models

As previously discussed, the generalized gamma, the mixture and the piecewise models can capture the non-monotonic abandonment rate behavior typically observed in call center operations. In the sequel, given time to abandonment data on n customers, we discuss posterior and predictive Bayesian analyses of these models using Markov chain Monte Carlo methods. We let $D = \{t_1, \cdots, t_n\}$ denote the n observed abandonment times of a given customer profile.

3.1 Inference for the Generalized Gamma Family

For the generalized Gamma parametrization introduced in (2.2), if we assume a joint prior $p(\alpha, \gamma, \lambda)$, then the joint posterior distribution $p(\alpha, \gamma, \lambda|D)$ is obtained proportional to

$$p(\alpha, \gamma, \lambda) \prod_{i=1}^{n} \frac{\gamma \lambda^{\alpha} t_i^{\alpha\gamma - 1}}{\Gamma(\alpha)} exp\{-\lambda t_i^{\gamma}\}. \tag{3.1}$$

The posterior distribution $p(\alpha, \gamma, \lambda|D)$ can not be obtained analytically for any choice of the prior $p(\alpha, \gamma, \lambda)$ in (3.1). Therefore in order to obtain the joint posterior distributions of the model parameters we will use a Gibbs sampler along with the random walk Metropolis-Hasting algorithm whose proposal density is multivariate normal. In so doing, we assume independent gamma priors for α, γ, and λ. In implementation of the Gibbs sampler, the choice of a gamma prior for λ enables us to obtain the full conditional of the λ as a gamma distribution. More specifically, if we assume a gamma prior with parameters a and b, denoted as $\lambda \sim G(a, b)$, then the full conditional for λ is given by

$$p(\lambda|\alpha, \gamma, D) \propto \lambda^{\alpha n + a - 1} e^{-\lambda(b + \sum_{i=1}^{n} t_i^{\gamma})}, \tag{3.2}$$

implying that $(\lambda|\alpha, \gamma, D) \sim G(\alpha n + a, b + \sum_{i=1}^{n} t_i^{\gamma})$.

If we let $\boldsymbol{\theta} = \{\gamma, \alpha\}$, then following Chib and Greenberg (1995) the steps in the Metropolis-Hasting algorithm and the Gibbs sampler using (3.2) can be summarized as follows

1. Assume the starting points $\boldsymbol{\theta^{(0)}}$ at $t = 0$.

   Repeat for $t > 0$,

2. Generate $\lambda^{(t)}$ from $G(\alpha^{(t)} n + a, b + \sum_{i=1}^{n} t_i^{\gamma^{(t)}})$.

3. Generate $\boldsymbol{\theta^*}$ from $q(\boldsymbol{\theta}^*|\boldsymbol{\theta^{(t)}})$ and u from $U(0, 1)$.

4. If $u \leq \alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^*)$ then set $\boldsymbol{\theta^{(t)}} = \boldsymbol{\theta}^*$; else set $\boldsymbol{\theta^{(t)}} = \boldsymbol{\theta^{(t)}}$ and $t = t + 1$,

where

$$\alpha(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta}^*) = min \left\{ 1, \frac{\pi(\boldsymbol{\theta^*}|\lambda^{(t)}) q(\boldsymbol{\theta^{(t)}}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta^{(t)}}|\lambda^{(t)}) q(\boldsymbol{\theta}^*|\boldsymbol{\theta^{(t)}})} \right\}. \tag{3.3}$$

In (3.3), $q(.|.)$ is the multivariate normal proposal density and $\pi(.|.)$ is the full conditional that we need to generate samples from. If we repeat the above a large number of times then we obtain samples from the joint posterior distribution $p(\boldsymbol{\theta}, \lambda|D)$.

3.1.1 Special Case: The Lognormal Model

As $\alpha \to \infty$ in the generalized gamma model (2.2) we obtain the special case of the lognormal model (2.7). Let $y_i = log(t_i)$ for $i = 1, \cdots, n$ represent the log of n abandonment times and $\phi = 1/\sigma^2$ represent the precision parameter. Also assume that a priori, $\mu \sim N(m_0, C_0)$ and $\phi \sim G(a_0, b_0)$ where μ and ϕ are independent. In order to obtain the joint posterior distribution of the parameters, $p(\mu, \phi|D)$ where $D = \{y_1, \cdots, y_n\}$, it is possible to obtain a full Gibbs sampler. The full conditional for μ can be written as

$$(\mu|\phi, D) \sim N(m_1, C_1), \tag{3.4}$$

where $C_1 = (n\phi + 1/C_0)^{-1}$ and $m_1 = C_1(n\phi\bar{y} + m_0/C_0)$ where $\bar{y} = (\sum_i^n y_i/n)$. Also the full conditional of ϕ is given by

$$(\phi|\mu, D) \sim G(a_1, b_1), \tag{3.5}$$

where $a_1 = a_0 + n/2$ and $b_1 = b_0 + \sum_i^n [(y_i - \mu)^2/2]$. Therefore, given (3.4) and (3.5) we can easily implement the Gibbs sampler as follows

- Assume the starting points $(\mu^{(0)}, \phi^{(0)})$.
- Generate $\mu^{(1)}$ from $(\mu|\phi^{(0)}, D)$ and $\phi^{(1)}$ from $(\phi|\mu^{(0)}, D)$.
- $\cdots$
- Generate $\mu^{(j)}$ from $(\mu|\phi^{(j-1)}, D)$ and $\phi^{(j)}$ from $(\phi|\mu^{(j-1)}, D)$.

If we repeat the above a large j number of times then we obtain samples from $p(\mu, \phi|D)$.

3.2 Inference for the Mixture and Piecewise Models

In our discussion of Bayesian inference for the mixture model (2.9), we consider the lognormal and the Weibull distributions for the components of the mixture and discuss the implementation of the Markov chain Monte Carlo methods. More specifically, following Diebolt and Robert (1994), we will present a data augmentation step within the Gibbs sampler for both Weibull and lognormal k component mixtures.

Let $\boldsymbol{\Theta} = \{\boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_k}, p_1, \cdots, p_k\}$ where $\boldsymbol{\theta_j}$ for $\{j = 1, \cdots, k\}$ are the parameters of the respective components, p_j for $\{j = 1, \cdots, k\}$ are the mixing weights. Furthermore assume that z_{ij} is a

latent variable indicating which component the i^{th} observation t_i belongs to, that is, $z_{ij} \in \{0, 1\}$ and $\sum_{j=1}^{k} z_{ij} = 1$. In the data augmentation method $\boldsymbol{z} = \{\boldsymbol{z_1}, \cdots, \boldsymbol{z_n}\}$ is referred to as the incomplete data where $\boldsymbol{z_i} = \{z_{i1}, \cdots, z_{ik}\}$ for $i = 1, \cdots, n$. Given $\boldsymbol{z}$, the density of the i^{th} observation from (2.9) can be rewritten as

$$f(t_i|\boldsymbol{\Theta}, \boldsymbol{z}) = \prod_{j=1}^{k} p_j^{z_{ij}} f(t_i|\boldsymbol{\theta_j})^{z_{ij}}, \text{ for } i = 1, \cdots, n. \tag{3.6}$$

Therefore given a sample size of n, let $D = \{t_1, \cdots, t_n\}$ the likelihood term with the missing data structure can be written as

$$L(\boldsymbol{\Theta}; D, \boldsymbol{z}) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_j^{z_{ij}} f(t_i|\boldsymbol{\theta_j})^{z_{ij}}. \tag{3.7}$$

We assume that the mixing weights, $(p_1, \cdots, p_k)$ follow a Dirichlet prior with paremeters $(\alpha_1, \cdots, \alpha_k)$. Therefore the Gibbs sampler with the data augmentation step for a general parameter vector $\boldsymbol{\Theta}$ can be summarized as follows

- Assume the starting points $\boldsymbol{\Theta^{(0)}}$.
- Generate $\boldsymbol{z^{(0)}}$ from $p(\boldsymbol{z}|D, \boldsymbol{\Theta^{(0)}})$ and $\boldsymbol{\Theta^{(1)}}$ from $p(\boldsymbol{\Theta}|D, \boldsymbol{z^{(0)}})$.
- $\cdots$
- Generate $\boldsymbol{z^{(m-1)}}$ from $p(\boldsymbol{z}|D, \boldsymbol{\Theta^{(m-1)}})$ and $\boldsymbol{\Theta^{(m)}}$ from $p(\boldsymbol{\Theta}|D, \boldsymbol{z^{(m-1)}})$.

If we repeat the above a large m number of times then we obtain samples from $p(\boldsymbol{\Theta}, \boldsymbol{z}|D)$. In order to implement this algorithm one needs to obtain $p(\boldsymbol{z}|\boldsymbol{\Theta}, D)$ and $p(\boldsymbol{\Theta}|\boldsymbol{z}, D)$. The full conditional of $\boldsymbol{z}$ is given by

$$p(\boldsymbol{z_i}|D, \boldsymbol{\Theta}) \sim Multinomial(\pi_{i1}, \cdots, \pi_{ik}), \tag{3.8}$$

where

$$\pi_{ij} = \frac{p_j f(t_i|\boldsymbol{\theta_j})}{\sum_{l=1}^{k} p_l f(t_i|\boldsymbol{\theta_l})}, \tag{3.9}$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, k$.

Furthermore, we can show that $p(\boldsymbol{\Theta}|D,\boldsymbol{z}) = p(p_1,\cdots,p_k|\boldsymbol{z},D)p(\boldsymbol{\theta_1}|\boldsymbol{z},D)\cdots p(\boldsymbol{\theta_k}|\boldsymbol{z},D)$ where

$$(p_1,\cdots,p_k|\boldsymbol{z},D) \sim Dir(\alpha_1+\sum_{i=1}^{n} z_{i1},\cdots,\alpha_k+\sum_{i=1}^{n} z_{ik}), \tag{3.10}$$

that is, a Dirichlet posterior for the mixing weights. Note that under the lognormal model, $p(\boldsymbol{\theta_1}|\boldsymbol{z},D),\cdots,p(\boldsymbol{\theta_k}|\boldsymbol{z},D)$ can be obtained using (3.4) and (3.5) with $\boldsymbol{\theta_j} = (\phi_j,\mu_j)$ for $j = 1,\cdots,k$.

In the case of Weibull components with scale parameters λ_j and shape parameters γ_j, for $j = 1,\cdots,k$, we have $\boldsymbol{\theta_j} = (\lambda_j,\gamma_j)$. In order to obtain samples from the joint posterior distribution of $p(\boldsymbol{\theta_j}|\boldsymbol{z},D)$ for $j = 1,\cdots,k$, the Gibbs sampler with the Metropolis-Hastings step can be used as discussed in the generalized gamma model. If we assume independent gamma priors for λ_j's as $\lambda_j \sim G(a_j,b_j)$, then for any particular form of priors on γ_j's we can obtain the full conditional of λ_j's as gamma distributions given by

$$(\lambda_j|\boldsymbol{z},\gamma_j,D) \sim G(a_j+\sum_{i}^{n} z_{ij}, b_j+\sum_{i}^{n} z_{ij}t_i^{\gamma_j}). \tag{3.11}$$

The full conditional distributions for γ_j's can not be obtained analytically for any choice of a prior form. Thus, we can use the Metropolis-Hastings algorithm to generate samples from $p(\gamma_j|\boldsymbol{z},\lambda_j,D)$ in implementing the Gibbs sampler.

3.2.1 Inference for the Piecewise Model

Similarly posterior samples for the piecewise abandonment rate models introduced in (2.14) can be obtained using the random walk Metropolis-Hastings algorithm as summarized for the generalized gamma model (without the full conditional generation step for λ). In our development, we have considered the Weibull and the lognormal distributions as the components of the piecewise models with a switching point coinciding with the first announcement (i.e. $\tau_0 = 60s$). Therefore, the relevant joint posterior is $p(\boldsymbol{\Theta}|D)$ and $\boldsymbol{\Theta} = \{\boldsymbol{\theta_0},\boldsymbol{\theta_1}\}$ where $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$ represent the parameters of the first and the second components of the piecewise model respectively.

3.3 Predictive Inference and Model Comparison

Once the samples from the joint posterior densities are obtained, it is possible to obtain the predictive posterior densities and predictive failure rates for different customer profiles. These enable

us to compare different abandonment rate behavior in call centers implied by different models and assess the model fit to data. Based on S posterior realizations from the joint posterior distribution of the parameter vector $\boldsymbol{\theta}$, the predictive posterior density can be obtained as

$$p(t|D) = \frac{1}{S}\sum_{i=1}^{S} f(t|\boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)}|D). \tag{3.12}$$

As pointed out by Lynn and Singpurwalla (1997) the predictive abandonment rate cannot be calculated as $\int r(t|\boldsymbol{\theta})p(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$. Given S set of samples the predictive abandonment rate can be obtained as

$$r(t|D) = \frac{\frac{1}{S}\sum_{i=1}^{S} f(t|\boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)}|D)}{1 - \frac{1}{S}\sum_{i=1}^{S} F(t|\boldsymbol{\theta}^{(j)})p(\boldsymbol{\theta}^{(j)}|D)}. \tag{3.13}$$

Once posterior samples are available from the distributions of model parameters, model comparison can be made using sampling based methods as suggested by Gelfand (1996). In what follows, we summarize the comparison criteria discussed by Gelfand (1996).

Let, as before, D represent a sample of observed times of abandonment, $\{t_i; i = 1, \cdots, n\}$. The cross validation predictive density for observation i is defined as $f(t_i|D_{(-i)})$, where $D_{(-i)}$ represents data D except for t_i. Note that $f(t_i|D_{(-i)})$ represents values of t_i which are supported by the model constructed using $D_{(-i)}$ and is given by

$$f(t_i|D_{(-i)}) = \int f(t_i|\boldsymbol{\theta}, D_{(-i)})f(\boldsymbol{\theta}|D_{(-i)})d\boldsymbol{\theta}. \tag{3.14}$$

Since the observations are conditionally independent given $\boldsymbol{\theta}$, $f(t_i|\boldsymbol{\theta}, D_{(-i)}) = f(t_i|\boldsymbol{\theta})$. Therefore, using (3.14) the pseudo-Bayes factor (PBF) is defined as

$$\prod_{i=1}^{n} \frac{f(t_i|D_{(-i)}, M_1)}{f(t_i|D_{(-i)}, M_2)}, \tag{3.15}$$

where M_1 and M_2 are any of the two proposed models. In addition, $C_i = \frac{f(t_i|D_{(-i)},M_1)}{f(t_i|D_{(-i)},M_2)}$ is referred to as the conditional predictive ordinate (CPO) ratio for the i^{th} observation. In order to be able to compute (3.15) directly, we need to obtain $f(t_i|D_{(-i)}, M_1)$ and $f(t_i|D_{(-i)}, M_2)$ for all $i = \{1, \cdots, n\}$ which are not available in closed form. However, using the samples of $\boldsymbol{\theta}$ generated via Markov chain Monte Carlo methods for each proposed model, it is possible to estimate the cross validation

predictive density as follows

$$\begin{aligned} f(t_i|D_{(-i)}) &= \frac{f(D)}{f(D_{(-i)})} \\ &= \frac{1}{\int \frac{f(D_{(-i)},\boldsymbol{\theta})}{f(D,\boldsymbol{\theta})} f(\boldsymbol{\theta}|D)d\boldsymbol{\theta}} \\ &= \frac{1}{\int \frac{1}{f(t_i|D_{(-i)},\boldsymbol{\theta})} f(\boldsymbol{\theta}|D)d\boldsymbol{\theta}}, \end{aligned}$$

for $i = 1, \cdots, n$. Therefore, a Monte Carlo estimate of $f(t_i|D_{(-i)})$ can be obtained as

$$\hat{f}(t_i|D_{(-i)}) = \frac{1}{\frac{1}{S}\sum_{j=1}^{S} \frac{1}{f(t_i|D_{(-i)},\boldsymbol{\theta}^{(j)})}}, \tag{3.16}$$

where S is the number of samples generated and $\boldsymbol{\theta}^{(j)}$ is the j^{th} vector of the generated parameter samples. Since given $\boldsymbol{\theta}$, t_i's are independent, $f(t_i|D_{(-i)}, \boldsymbol{\theta}^{(j)}) = f(t_i|\boldsymbol{\theta}^{(j)})$ can be used in (3.16). Finally, once the cross validation predictive densities are estimated using (3.16), coupled with the pseudo-Bayes factor as introduced in (3.15), we can compare the proposed time to abandonment models. Alternatively, we will investigate the fit of the proposed models to abandonment data from a future month. For instance using the models constructed using the abandonment behavior in January, we can assess their fit to abandonment data in February and March.

4 Numerical Illustrations

In order to show the implementation of the Bayesian methods introduced in the previous section we have used real call center abandonment data from an anonymous bank operation. A detailed description of the data can be found at Data (2000). We have carried out Bayesian inference for the abandonment behavior observed in the month of January for two different customer profiles; regular customers and stock exchange customers. Next we discuss the implications of the models on call center customer abandonment behavior, effect of announcements on such behavior, change of behavior from one month to the next as well as for different customer profiles. In addition, we summarize issues such as model fit, predictive abandonment rate behavior and predictive performance.

4.1 Case 1: Regular Customers

First we discuss the model fit issue using the sampling based method introduced in (3.14). Let $log(\prod_{i=1}^{n} \hat{f}(t_i|D_{(-i)}, M_j))$ be called the global CPO of model j in the log scale and represent how much the model j supports the data. In other words, higher the global CPO better is the fit of the model.

Models	Global CPO
LogNormal	-7859
Generalized Gamma	-7822.48

Table 1: Global CPO in the Log Scale for the Generalized Gamma and Lognormal Models

Models	Global CPO
Piecewise Weibull	-7833.59
Piecewise LogNormal	-7818.48

Table 2: Global CPO in the Log Scale for the Piecewise Models

Models	Global CPO
Mixture Weibull (2)	-7776.11
Mixture Weibull (3)	-7763.62
Mixture LogNormal (2)	-7823.11
Mixture LogNormal (3)	-7760.74

Table 3: Global CPO in the Log Scale for the Mixture Models

Based on the global CPO values shown in Tables 1, 2 and 3 the best fit is obtained using the three component lognormal mixture model. Also we note that the lognormal does not seem to support the abandonment behavior of the regular customers. This might be due to the specific abandonment behavior shown by regular customers and is discussed in the sequel. Although the three component lognormal mixture model seems to provide the best fit, we have also investigated for each time to abandonment sample how much better it was outperforming its closest fit competitor. In order to be able to observe this, consider the difference, $log(\hat{f}(t_i|D_{(-i)}, M_1)) - log(\hat{f}(t_i|D_{(-i)}, M_2))$, for which positive values represent support for the first model for the i^{th} sample (and vice versa). The behavior of this log difference for different sample values is shown in Figure 5. Mostly this value oscillates around the value of zero, indicating a balance between the two models. However as can be observed from the high positive values in the graph, there are several occasions where the first

model (three component lognormal mixture) outperforms the second one (three component Weibull mixture).

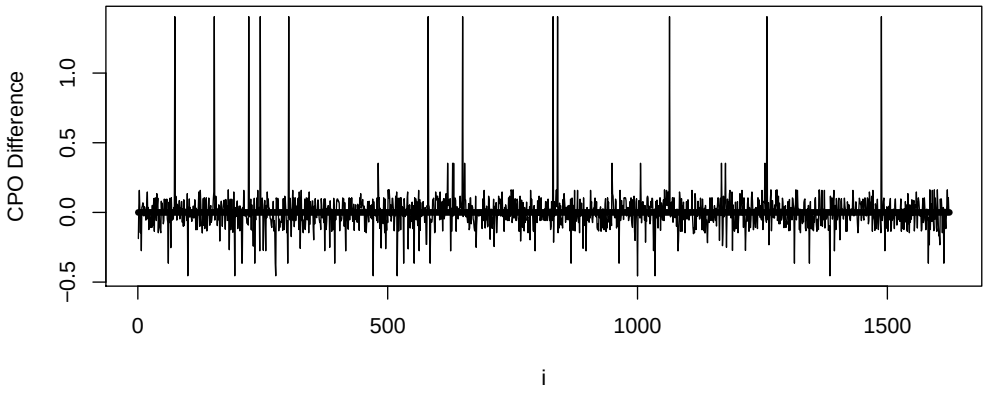

Figure 5: $log(\hat{f}(t_i|D_{(-i)}, M_1)) - log(\hat{f}(t_i|D_{(-i)}, M_2))$ vs Sample i

Another issue that we have investigated was the potential existence of higher order mixtures. All the higher order mixture models seemed to be indistinguishable from the three-component mixture models as suggested by similar parameter values. We have investigated the possibility of fourth,fifth and sixth order mixtures and found no evidence in favor of higher order mixtures for both the lognormal and Weibull models.

The mixture and the piecewise models seem to support the behavior of the regular customers as exhibited in the general pattern of Brown et al. (2005). This suggests that regular customers change their abandonment behavior. The customers tend to become more or less patient as they wait more in the line or as they experience announcements. What would be of interest to call center practitioners is the investigation of how regular customers exhibit different abandonment behavior over time or before/after announcements. This can be observed via the posterior predictive abandonment rates implied by different models and are obtained as discussed in (3.13). The posterior predictive abandonment rates of some of the models are shown in Figures 6 and 7 from which non-monotonic abandonment rate can be inferred. This supports the claim that regular customers exhibit different abandonment rate behavior as they wait in the line or experience announcements (non-monotonic). The non-monotonic behavior is also supported by the posterior results implied

by the generalized gamma distribution whose abandonment rate is the most flexible and is not a mixture nor a piecewise model and is among the best fitting models (see Figure 7). Furthermore all models implicate that as customers join the queue their abandonment rate is increasing at first, in other words they are less patient upon arrival. However as they wait more in the queue or experience announcements they become more patient and their abandonment rate decreases.

An attractive feature of the Bayesian modeling of the abandonment rate is that we can now formally test the hypothesis of non-monotonic behavior. Since the generalized gamma model can exhibit both monotonic and non-monotonic abandonment rate behavior you can use its posterior samples to test the hypothesis. Therefore, using the notation introduced in (2.2), formally the hypothesis testing can be setup as

$$H_0 : \{\frac{1-(\alpha\gamma)}{\gamma(\gamma-1)} > 0|D\} \text{ vs. } H_1 : \{\frac{1-(\alpha\gamma)}{\gamma(\gamma-1)} \leq 0|D\}$$

where $Pr(H_0 : \{\frac{1-(\alpha\gamma)}{\gamma(\gamma-1)} > 0|D\})$ was found to be equal to one. In other words, we can infer that regular customers exhibit non-monotonic behavior with probability one.

Another interesting finding is the one implied by the piecewise lognormal model. After the 60^{th} second which coincides with the first announcement (it is also the switching point for the piecewise model), the customers begin to exhibit a decreasing abandonment rate behavior, that is they become more patient as soon as they hear the announcement (see Figure 7). A similar behavior can also be inferred from the lognormal mixture models where customers begin to exhibit a decreasing abandonment rate behavior slightly after the 60^{th} second.

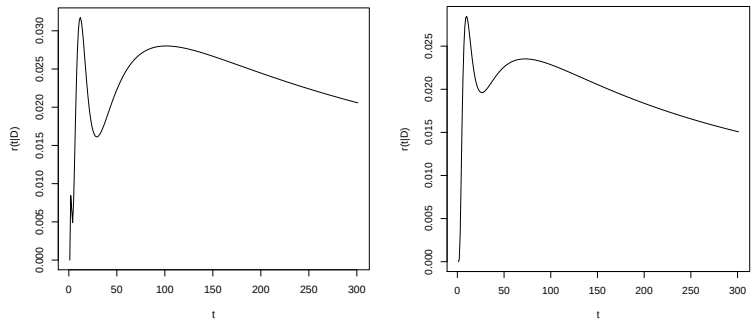

Figure 6: Predictive Abandonment Rates for Regular Customers: Three-Component Lognormal Mixture (left) and Two-Component Lognormal Mixture (right)

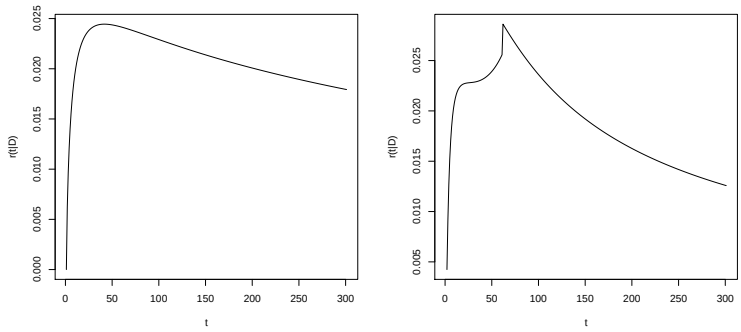

Figure 7: Predictive Abandonment Rates for Regular Customers: Generalized Gamma (left), Piecewise Lognormal (right)

In Tables 1-3, we presented comparison of the models based on abandonment data during the month of January. We are also interested in how these models perform in predicting abandonment behavior in other months. Thus, we next obtained one-month ahead fits using the best fitting models to investigate if there are clear differences between the months of January and February (see Figure 4). The results are very encouraging, the best fit rankings mostly stayed the same (except for the two component mixture lognormal model which seem to provide a slightly better fit than the piecewise lognormal and generalized gammma models) and there is strong evidence that the customer abandonment behavior does not change from month to month. This also suggests that the proposed models were able to adequately capture the abandonment behavior of regular customers. In order to assess the fits, we have used one-month ahead log-likelihoods as follows

$$log\{\prod_{i=1}^{n_f} \frac{1}{S} \sum_{j=1}^{S} f(t_i^f|\boldsymbol{\theta_j})\}, \tag{4.1}$$

where n_f is the sample size for the month of February, t_i^f for $i = 1, \cdots, n_f$ are the n_f time to abandonment samples for February customers and $\boldsymbol{\theta_j}$ for $j = 1, \cdots, S$ are S posterior samples obtained in the previous month. The log-likelihoods for the best fitting models are shown in Table 4.

Posterior predictive distributions which are obtained as in (3.12) for the top three best fitting models, that it the mixture models are shown in Figure 8 against the actual abandonment data. An interesting property that can be noted is that the lognormal mixture captures a small proportion

Model	Log Likelihood
Mixture LogNormal (3)	-8090.952
Mixture LogNormal (2)	-8142.75
Mixture Weibull (3)	-8099.588
Mixture Weibull (2)	-8123.107
Piecewise LogNormal	-8149.24
Generalized Gamma	-8154.642

Table 4: One-Month Ahead Fits

of abandonments that occur immediately right after customers join the queue. In other words, one of the components of the lognormal mixture captures behavior that regular customers exhibiting that is referred to as balking in the queuing literature.

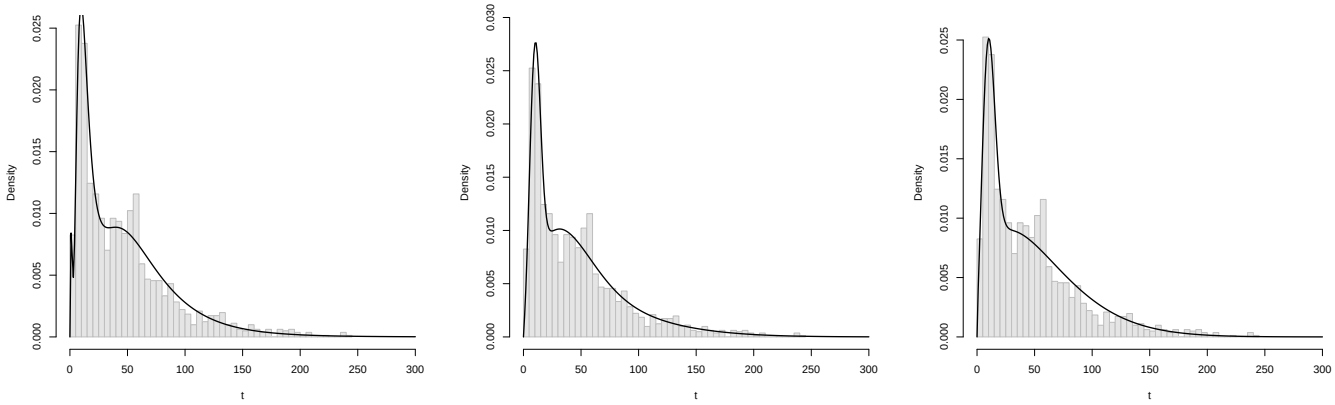

Figure 8: Posterior Predictive Fits vs. Actual Data

Table 5 shows a summary of statistics and Figure 9 shows the respective density plots of the posterior parameters of the best model fit, the lognormal mixture model with three components.

Parameters	μ_1	μ_2	μ_3	ϕ
Mean	0.5025	2.5560	4.1004	3.5609
St.Deviation	0.1494	0.0300	0.0226	0.1728
Parameters	p_1	p_2	p_3	
Mean	0.0180	0.3990	0.5830	
St.Deviation	0.0040	0.0154	0.0156	

Table 5: Posterior Parameters Summary Statistics for the Lognormal Mixture Model with Three Components

The results presented in this section are based on the Markov chain Monte Carlo sampling

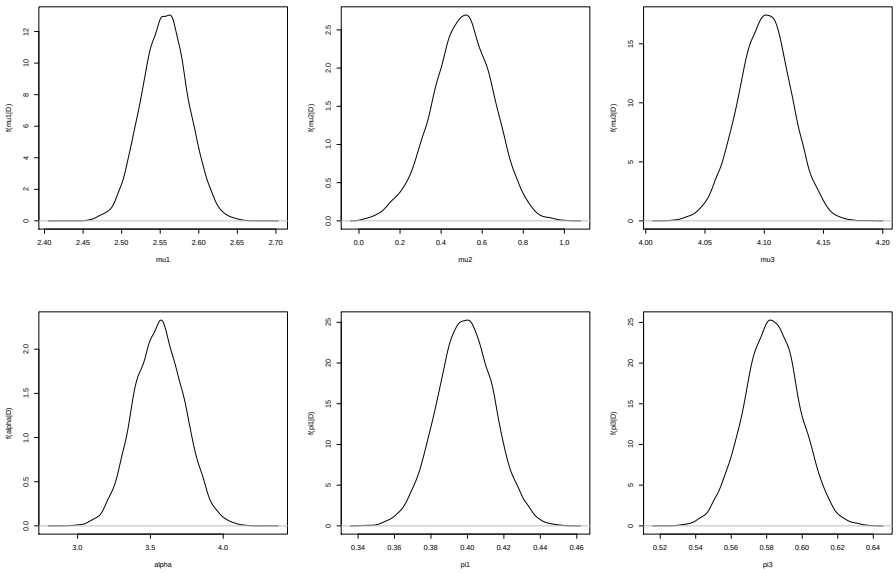

Figure 9: Posterior Parameters Density Plots for the Lognormal Mixture Model with Three Components

techniques discussed in the previous section. Therefore we briefly discuss issues of convergence and in doing so, we present the results obtained using the generalized gamma model. For the sake of preserving space we will omit a detailed discussion of convergence for the rest of the models for which similar results were obtained. An informal way of assessing convergence is to observe the behavior of the trace plots of the posterior parameters. As shown in Figure 10, where $\lambda = e^{\theta}$, all trace plots exhibit a fairly constant behavior around a specific value which is the first evidence of convergence.

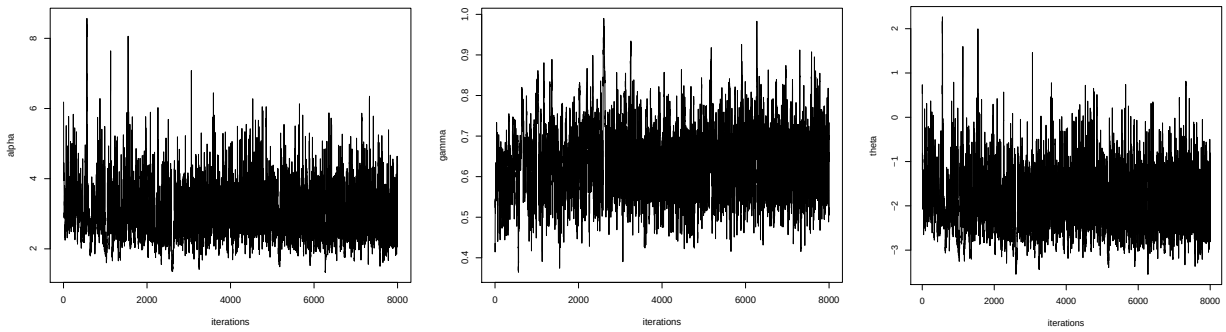

Figure 10: Trace Plots of α (left), γ (middle) and θ (right)

A more formal way of assessing convergence can be carried out using the Brooks and Gelman plots and the shrink factor (see Brooks and Gelman (1998) for a detailed discussion). In order to assess convergence using the methods discussed in Brooks and Gelman (1998) we have used three different chains with different starting points. According to Brooks and Gelman (1998), if the shrink factor also referred to as the scale reduction point estimate is around 1 then convergence is said to have been attained. The Brooks and Gelman plots are shown in Figure 11 where the shrink factor approaches 1 as the number of iterations increases.

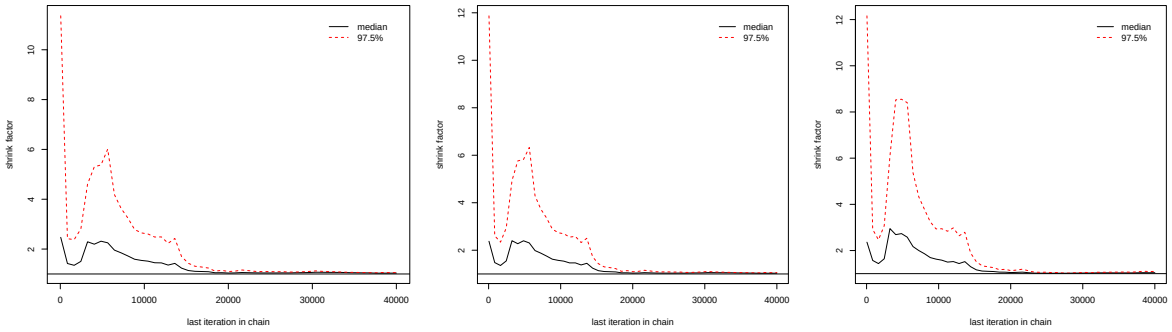

Figure 11: Brooks and Gelman Plots

The scale reduction point estimates for the parameters α and γ were 1.04 and 1.03 for θ which also supports what we were able to conclude by observing the trace plots and the Brooks and Gelman plots.

4.2 Case 2: Stock Exchange Customers

Another customer profile for which we have investigated abandonment behavior is the stock exchange customers. As shown in Figure 12, stock exchange customers exhibit a different time to abandonment behavior as opposed to the regular customers. Our first impression was that they did not seem to exhibit mixture behavior. We have investigated potential mixture behavior using Weibull components and found no evidence in its favor. The expected values of the shape parameters, as in (2.9), were very close to 1 and mixing probabilities were equal (~ 0.5) which shows evidence against mixing. We have also considered a Weibull model to be able to assess whether stock exchange customers were exhibiting increasing, decreasing or constant abandonment rate behavior. The expected value of the posterior shape parameter ($E(\gamma|D)$) was found to be

0.9606, a value very close to 1. These results led us to believe that stock exchange customers were exhibiting an abandonment rate behavior which can be captured by the exponential model. We have carried out a conjugate Bayesian analysis of the exponential model (see Gelman et al. (2003) for instance) and compared it against the Weibull model using the pseudo Bayes factor (3.15). The pseudo Bayes factor was ~ 3.815, slightly in favor of the exponential model. This suggests that the stock exchange customers, unlike the regular customers, do not change their abandonment behavior (monotonic abandonment rate) as they wait more in the line or as they experience announcements as implied by both the Weibull and the exponential models. This might be due to the fact that stock exchange service is time sensitive and do not offer the customer the luxury of redialing at a later time since abandoning might yield a lost investment opportunity for the customers. Whereas regular customers might redial at a later time when the lines are less busy.

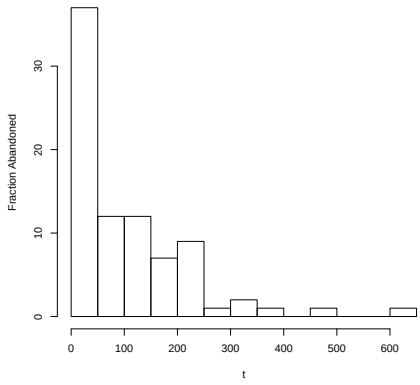

Figure 12: Histogram for Stock Exchange Customers in January

5 Concluding Remarks

In this paper we have introduced different strategies to model different customer abandonment behavior in call centers. Most of the research in call centers with abandonment (also referred to as impatience) has been conducted from a queuing perspective and to the best of our knowledge research from statistical inference has not been considered. This research aims to fill this gap and to bring concepts from survival and reliability analysis into the call center literature. In doing so we have introduced different family of distributions, mixture models and piecewise time to

abandonment models with their respective properties and their Bayesian analysis via Markov chain Monte Carlo methods. Furthermore we have introduced the notion of the abandonment rate by motivating its relationship to the hazard and failure rate concepts from survival and reliability analysis.

In order to show the implementation of the proposed models we have used real call center data for two customer profiles, regular and stock exchange customers. We were able to illustrate different customer abandonment rate behavior exhibited by different customer profiles and found evidence in favor of mixture models for regular customers due to their flexible hazard rate behavior and in favor of exponential behavior for stock exchange customers. Furthermore, the piecewise models showed that regular customers tend to change their abandonment behavior as they experience announcements, whereas the stock exchange customer do not tend to exhibit such behavior. In other words, regular customers exhibit non-monotonic abandonment rate and the stock exchange customers monotonic abandonment rate. We note here that the models introduced are general, that is they are applicable to call center operations with several different customer profiles. Combined with studies in queues with abandonment and their relationship to staffing (see Garnett et al. (2002) for instance), the proposed models can be used to infer operational regimes (as in Garnett et al. (2002)), level of service and operating characteristics (as in Zeltyn and Mandelbaum (2005) and Brandt and Brandt (1997)) implied by different abandonment behavior. We also note here that these studies can only be possible if other call center primitives such as the arrival rate and the service rate are given (or studied separately).

We believe that further research from a decision theoretic framework in call centers with abandonment is possible. Coupled with the proposed models, one can study the optimal announcement time which will maximize the utility of the call center operation from a maintenance problem point of view. Furthermore one can also study the effects of the optimal announcement times on call center operating characteristics implied by different customer profiles.

References

Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688.

Aksin, Z., Jouini, O., and Dallery, Y. (2008). Modeling call centers with delay information. *Submitted*.

Bacelli, F., Boyer, P., and Hebuterne, G. (1984). Single server queues with impatient customers. *Advances in Applied Probability*, 16(4):887–905.

Brandt, A. and Brandt, M. (1997). On the M(n)/M(m)/s queue with impatient calls. *Performance Evaluation*, 35:1–18.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434455.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hasting algorithm. *The American Statistician*, 49(4):327335.

Dadpay, A., Soofi, E. S., and Soyer, R. (2007). Information measures for generalized Gamma family. *Journal of Econometrics*, 138:568–585.

Data (2000). Technion, Israel Institute of Technology. Available at http://iew3.technion.ac.il /serveng /callcenterdata/.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56(2):363–375.

Garnett, O., Mandelbaum, A., and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management*, 4(3):208–227.

Gelfand, A. E. (1996). *Markov Chain Monte Carlo in Practice*, chapter 9.Model determination using sampling-based methods. Chapman & Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Lynn, N. J. and Singpurwalla, N. D. (1997). Comment: Burn-in makes us feel good. *Statistical Science*, 12(1):1319.

Mandelbaum, A. and Shimkin, N. (2000). A model for rational abondonments from invisible queues. *Queueing Systems*, pages 141–173.

Palm, C. (1953). Methods of judging the annoyance caused by congestion. *Tele*, (2):189–208.

Pham, T. and Almahana, J. (1995). The generalized Gamma distribution:its hazard rate and stress-strength model. *IEEE Transactitons on Reliability*, 44(3):568–585.

Singpurwalla, N. (2006). *Reliability and Risk A Bayesian Perspectiv*. John Wiley and Sons, Ltd.

Stacey, E. (1962). A generalization of the Gamma distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192.

Whitt, W. (1999a). Improving service by informing customers about anticipated delays. *Management Science*, 45(2):192–207.

Whitt, W. (1999b). Predicting queueing delays. *Management Science*, 45(6):870–888.

Wondmagegnehu, E. T., Navarro, J., and Hernandez, P. J. (2005). Bathtub shaped failure rates from mixtures: A practical point of view. *IEEE Transactitons on Reliability*, 54(2).

Zeltyn, S. and Mandelbaum, A. (2005). Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. Technical report, Technion, Israel Institute of Technology.

Zohar, E., Mandelbaum, A., and Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48(4):566–583.