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

Technical Report TR-2010-3
February 22, 2010

Bayesian Modeling of Queues in Call Centers

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

Refik Soyer
Department of Decision Sciences
The George Washington University

Bayesian Modeling of Queues in Call Centers

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

Refik Soyer
Department of Decision Sciences
The George Washington University

February 22, 2010

Abstract

Queuing models have been extensively used in call center analysis for obtaining performance measures and for developing staffing policies. However, almost all of this work have been from a pure probabilistic point of view and have not addressed issues of statistical inference. In this paper, we develop Bayesian analysis of call center queuing models by describing uncertainty about system primitives probabilistically. We consider models with impatient customers and develop details of Bayesian inference for queues with abandonment such as the M/M/s+M model. We illustrate the implementation of the Bayesian models using actual arrival, service, and abandonment data from call centers. We compare the M/M/s+M model results with those from the M/M/s queues and discuss implications of ignoring abandonment.

1 Introduction and Overview

A system where customers acquire service via phone and are served by contact centers can be defined as a call center. A call center queuing system consists of three main processes: arrival, service and abandonment. Given that one has a good understanding of these processes, relevant operating characteristics can be obtained using existing queuing theory (see Gross and Harris (1998) for a detailed coverage of queuing results). One of the main concerns of call center customers is fast

and efficient service. If such expectations are not met, abandonments will occur more frequently, which in turn will effect the perceived quality of a call center operation. The fraction of abandoned customers in the system can be considered as a proxy for the quality of service and therefore its assessment in the long run is of utmost importance to call center managers. Other relevant operating characteristics include number of customers in the system and in the queue, waiting time in the system and in the queue and the offered load which are all functions of the three fundamental processes.

The most commonly used queuing model in call center analysis is the M/M/s queue with s servers and having Markovian arrivals and service completions. This model assumes that customers do not abandon, that is, customers have infinite patience. The M/M/s model is also referred to as the Erlang-C in the queuing literature. Garnett et al. (2002) discuss the implications of ignoring abandonment and discuss its effects on certain operating characteristics and on staffing rules for different regimes. Another widely used model is the M/M/s/s model, referred to as the Erlang-B where no waiting is allowed. The customers are assumed to be turned away upon arrival if there are no available servers, that is, a system with no waiting space. Palm (1957) further extended the Erlang-B model by considering stochastic abandonment which he referred to as the patience time of a caller. Bacelli and Hebuterne (1981) discussed other properties of the model introduced in Palm (1957), and defined the M/M/s+M and M/M/s+G models, where in the last argument M stands for exponential distribution and G for a general distribution for time to abandonment (or patience time). The M/M/s+M model is called the Erlang-A model in the call center literature. Zeltyn and Mandelbaum (2005) and Mandelbaum and Zeltyn (2007) provide a good summary of the operating characteristics for M/M/s+M and M/M/s+G models along with their extensions. Another area of research in call centers with abandonment is the study of delay announcements on system performance. 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. Koole and Mandelbaum (2002) present a survey of recent research on queuing models in call centers.

Main focus of the call center queuing literature has been on probabilistic modeling and as a result, statistical issues have been neglected. Brown et al. (2005) point out the apparent need for

research emphasizing statistical analysis and refer to it as *queuing science* rather than queuing theory. The authors analyze an anonymous call center from a queuing science perspective by providing statistical analysis for the three fundamental processes of arrival, service and abandonment in call center operations. A similar view on the lack of statistical analysis has been expressed by Mandelbaum and Zeltyn (2007) who note that uncertainty in the system parameters is of concern to call center practitioners for staffing and sensitivity analysis purposes. In this paper, we intend to fill this gap in the literature by looking at the call center queuing models from a Bayesian perspective and in doing so, extend the literature in queuing science. Our objective is to develop a Bayesian analysis of the M/M/s+M queues to address issues such as

1. Modeling uncertainty and dependence in system primitives, that is, arrival, service and abandonment rates,
2. Assessment of operating characteristics such as number of customers in the system,
3. Implications of ignoring abandonment in the system,
4. Assessment of operational costs and optimal staffing.

To the best of our knowledge, this will be the first study in call center queuing literature from a Bayesian point of view. In the Bayesian approach uncertainty about any unknown quantity is handled via probability. Since the system primitives, such as the arrival, service and abandonment rates are unknown quantities, uncertainty about them will be described probabilistically as opposed to conventional call center queuing models where they are assumed to be unknown but fixed. Namely, instead of obtaining only point estimates for these parameters, the Bayesian approach will yield probability distributions of the system primitives which will be used to obtain the distributions of relevant operating characteristics. Thus, all inference questions will be addressed probabilistically using these distributions. Another attractive feature of the Bayesian approach is its ability to incorporate the call center management's opinions based on historical data and past experience into the model via the prior distributions of the system primitives. Furthermore, as pointed out by Lindley (1990) the Bayesian approach provides a coherent framework for making decisions and thus enables us to develop optimal strategies with regards to staffing of call centers.

The Bayesian approach has been considered in queuing literature. Some of the earlier works include McGrath et al. (1987a) and McGrath et al. (1987b) who address the fundamental concepts

of Bayesian queuing for M/M/1 type of queues. The first known full treatment of an M/M/1 queue from a Bayesian point of view is due to Armero and Bayarri (1994) where the focus is on the inference of the system primitives and the predictive distributions of main operating characteristics of the M/M/1 queue where independent priors are assumed for the arrival rate, λ, and the service rate, μ. Later on, Armero and Bayarri (1996) extend the results for M/M/1 queues to M/M/s queues. Bayesian analysis of Er/M/1 and Er/M/s queues where inter-arrival times are assumed to be Erlang and the service times are exponentially distributed are introduced by Wiper (1998) who provides estimates of the operating characteristics using Monte Carlo methods. However, analysis of queuing models with abandonment has not been previously considered in the Bayesian literature. Thus, in addition to being a contribution to the call center literature, our paper also represents a contribution to the Bayesian queuing analysis.

In this paper, we consider Bayesian modeling of queues for providing additional insights for management of call centers. We propose two different classes of models for system primitives. The first class assumes that the system primitives are independent and uses independent Gamma priors for the M/M/s+M (Erlang-A, where A stands for abandonment) queues. One of the advantages of the first model is its computational simplicity since the posterior distributions of the system primitives will be available in closed form. The second model relaxes the independence assumption and allows for dependent arrival, service and abandonment rates via a trivariate lognormal prior for the M/M/s+M model. The dependency assumption of the system primitives does not imply the dependence of the arrival, service and abandonment processes since given the system primitives they are still conditionally independent. To the best of our knowledge, this type of approach has not been considered in the call center literature previously. These models are compared with M/M/s queues where there is no abandonment. For each model posterior/prior implications and estimation of the operating characteristics are discussed, staffing and service level implications for the M/M/s+M models are introduced.

A quick summary of our paper is as follows. In section 2 we provide a summary of operating characteristics for the M/M/s+M (also referred to as Erlang-A) queues. Section 3 will present Bayesian modeling with independent and dependent priors for M/M/s+M queues. In Section 4, we discuss how to carry out Bayesian inference for the proposed models and develop techniques for obtaining posterior operating characteristics. This will be followed by a discussion of the effects

of the proposed models on staffing decisions and service level in Section 5 where comparisons are presented with M/M/s models using actual call center data with abandonment. Concluding remarks are given in Section 6.

2 M/M/s+M Queues and their Properties

Consider a queuing system, where inter-arrival times are exponential with rate λ, there are s identical servers with service times distributed according to an exponential distribution with rate μ and customers abandon the queue according to an exponential distribution with a constant rate of θ. This system is referred to as the Erlang-A model (where A stands for abandonment). Bacelli and Hebuterne (1981) discuss further properties of the model and call it the M/M/s+M queue. Mandelbaum and Zeltyn (2007) provide a detailed discussion of the operating characteristics for the Erlang-A model, discuss relevant computational issues, and note that the M/M/s+M model always reaches steady state due to its birth and death process properties. Therefore unlike the M/M/1 and M/M/s models, steady state conditions need not be investigated. In what follows we summarize the operating characteristics results from Mandelbaum and Zeltyn (2007) in call centers.

We first define the function called $A(x, y)$ as

$$A(x,y) = \frac{xe^y}{y^x}\gamma(x,y), \tag{2.1}$$

where $x, y > 0$ and $\gamma(x, y)$ is the incomplete Gamma function given by

$$\gamma(x,y) = \int_0^y t^{x-1}e^{-t}dt. \tag{2.2}$$

Given that the system is in steady state, the probability of having n customers in the system, conditioned on λ, μ and θ, can be obtained via

$$P_n = Pr(N = n|\lambda,\mu,\theta) = \begin{cases} P_s\frac{s!}{n!r^{s-n}} & \text{for} \quad 0 \le n \le s, \\ P_s\frac{(\frac{\lambda}{\theta})^{n-s}}{\prod_{k=1}^{n-s}(\frac{s\mu}{\theta}+k)} & \text{for} \quad n \ge s+1, \end{cases} \tag{2.3}$$

where

$$P_s = \frac{E_{1,s}}{1 + [A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta}) - 1]E_{1,s}}, \tag{2.4}$$

$$E_{1,s} = \frac{\frac{r^s}{s!}}{\sum_{j=0}^{s} \frac{r^j}{j!}}, \tag{2.5}$$

and $r = \frac{\lambda}{\mu}$. The term $E_{1,s}$ is also called the blocking probability in the M/M/s/s queue (Erlang-B). The probability that a customer will wait in the queue, $Pr(T_q > 0|\lambda, \mu, \theta)$ is given by

$$P_{T_q} = Pr(T_q > 0|\lambda, \mu, \theta) = \sum_{n=s}^{\infty} Pr(N = n|\lambda, \mu, \theta), \tag{2.6}$$

where $Pr(N = n|\lambda, \mu, \theta)$ for $n = s, \cdots, \infty$ can be obtained from (2.3). Garnett et al. (2002) discuss three call center operational regimes for limiting cases of the delay probability $Pr(T_q > 0|\lambda, \mu, \theta)$. For example, if $Pr(T_q > 0|\lambda, \mu, \theta) \to 0$ then the operational regime is called the *quality driven regime*. On the other hand, the case of $Pr(T_q > 0|\lambda, \mu, \theta) \to 1$ implies an *efficiency driven regime*.

The conditional probability of abandonment, given that all servers are busy and there are j customers in the queue upon arrival, can be obtained as

$$Pr_j(Ab|\mu, \theta) = \frac{(j+1)\theta}{s\mu + (j+1)\theta},\ j \geq 0. \tag{2.7}$$

The probability to abandon given that a customer is not served upon arrival is

$$P_{Ab|T_q} = Pr(Ab|T_q > 0, \lambda, \mu, \theta) = \frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}, \tag{2.8}$$

where $\rho = \frac{\lambda}{s\mu}$. The operating characteristics, (2.7) and (2.8), are directly related to the perceived quality of the call center operation from a customer perspective. Consequently, the steady state probability of abandonment can be calculated as the product of (2.8) and (2.6) as

$$P_{Ab} = Pr(Ab|\lambda, \mu, \theta) = (\frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}) \sum_{n=s}^{\infty} Pr(N = n|\lambda, \mu, \theta). \tag{2.9}$$

The steady state probability (2.9) can be used for determining an optimal staffing policy motivated by the trade-off between the server cost and the cost of abandoning (can be interpreted as a lost opportunity cost). Following Mandelbaum and Zeltyn (2007), the average operational cost (per

unit of time) can be defined as

$$C(s, \lambda, \mu, \theta) = cs + a\lambda Pr(Ab|\lambda, \mu, \theta), \tag{2.10}$$

where s is the number of servers, c is the staffing cost, a is the abandonment cost and $Pr(Ab|\lambda, \mu, \theta)$ is given by (2.9). Note that given λ, μ and θ, the average cost (2.10) can be used as an objective function to choose the optimal number of servers s. In other words, if λ, μ and θ are known, then we can obtain the value of s which will minimize the cost function (2.10). Since λ, μ and θ are treated as random quantities in the Bayesian framework, we need to minimize the expected value of (2.10) where expectation is taken with respect to the joint distribution of λ, μ and θ as

$$E\{C(s, \lambda, \mu, \theta)|D\} = \int_{\lambda} \int_{\mu} \int_{\theta} \{cs + a\lambda Pr(Ab|\lambda, \mu, \theta)\} p(\lambda, \mu, \theta|D) d\lambda \, d\mu \, d\theta. \tag{2.11}$$

Thus, (2.10) serves the role of a loss function in the Bayesian decision theoretic set up. The implications of the proposed models on the operating characteristics and on optimal staffing will be addressed in Section 5.

3 Bayesian Models for System Primitives

The rates of the arrival, service and abandonment, λ, μ and θ, respectively, are referred to as the system primitives. In call center queuing models, these rates are assumed to be fixed but unknown and are estimated from historical data. Since these rates are unknown quantities, by taking a Bayesian point of view, we define our uncertainty about them via probability distributions. In this section we introduce different modeling strategies for the system primitives in M/M/s+M queues in call centers and compare them with Bayesian models for M/M/s queues.

One can specify either dependent or independent priors for system primitives of M/M/s+M queues. Previous work in Bayesian analysis of M/M/s queues assume independent Gamma densities as priors for λ and μ; see Armero and Bayarri (1994). Similarly, we can use independent Gamma priors for all the three system primitives in M/M/s+M queues. In other words, we can assume that the abandonment rate, θ also follows a Gamma prior. The choice of Gamma density is attractive due to its conjugacy with the exponential likelihood function. Armero and Bayarri (1994) and Armero

and Bayarri (1996) discuss prior and posterior implications of using independent Gamma priors for λ and μ in M/M/s queues. These are also applicable in the posterior analysis of M/M/s+M queues.

An alternate modeling strategy is to assume a dependent prior distribution between two or more of the system primitives. If it is believed that the primitives are dependent on each other then the prior can be specified to reflect this belief. For instance, when the arrival rate is high, the abandonment rate might tend to go up due to the increased number of people in the queue or the service rate might go down (or up) due to the sudden increase of call arrivals. Another plausible scenario might be that as the number of arrivals increase, servers intentionally hang up on the customers as soon as the service starts, causing a decrease in the service times. In order to be able to capture this type of dependence structure in M/M/s+M queues, one can use a trivariate lognormal prior for λ, μ and θ with parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Phi} = [log(\lambda)\, log(\mu)\, log(\theta)]$. Therefore

$$p(\lambda, \mu, \theta) = (2\pi)^{-3/2}|\boldsymbol{\Sigma}|^{-3/2}(\frac{1}{\lambda\mu\theta})exp(-\frac{1}{2}[\boldsymbol{\Phi} - \boldsymbol{\nu}]'\boldsymbol{\Sigma}^{-1}[\boldsymbol{\Phi} - \boldsymbol{\nu}]), \tag{3.1}$$

where $\boldsymbol{\nu} = (\nu_1\ \nu_2\ \nu_3)'$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}. \tag{3.2}$$

The joint density (3.1) implies that $log(\lambda)$, $log(\mu)$ and $log(\theta)$ follow a multivariate normal with mean vector $\boldsymbol{\nu}$ and covariance matrix $\boldsymbol{\Sigma}$. We note that when the covariances in (3.2) ($\sigma_{i,j}$ where $i \neq j$) are all equal to zero, then λ, μ and θ are said to be independent a priori. This approach allows us to study the dependency structure in the light of new arrival, service and abandonment data via the posterior distributions.

Using the rules of probability we can rewrite $p(\lambda, \mu, \theta)$ as

$$p(\lambda, \mu, \theta) = p(\lambda|\mu, \theta)p(\mu|\theta)p(\theta), \tag{3.3}$$

where

$$\lambda|\mu, \theta \sim LN(\nu_1 + \begin{bmatrix} \sigma_{12} & \sigma_{13} \end{bmatrix} \begin{bmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} \mu - \nu_2 \\ \theta - \nu_3 \end{bmatrix}, \sigma_{11} - \begin{bmatrix} \sigma_{12} & \sigma_{13} \end{bmatrix} \begin{bmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{bmatrix} \begin{bmatrix} \sigma_{12} \\ \sigma_{13} \end{bmatrix}),$$

$$\mu|\theta \sim LN(\nu_2 + \frac{\sigma_{23}}{\sigma_{33}}[\theta - \nu_3], \sigma_{22} - \frac{\sigma_{23}^2}{\sigma_{33}}),$$

and

$$\theta \sim LN(\nu_3, \sigma_{33}).$$

Note that this prior can also be used for M/M/s queues, where abandonment rate θ does not apply. In this case (3.1) is a bivariate lognormal distribution of λ and μ. Furthermore, it can be shown that the distribution of λ/μ for the bivariate case is also lognormal, that is the prior density of ρ (or r in the multiple server queues) is given by

$$p(\rho) = (2\pi)^{-1/2}\beta^{-1}(\frac{1}{\rho})exp\{-\frac{1}{2}[(\frac{log(\rho - \alpha)}{\beta})^2]\}, \tag{3.4}$$

where $\alpha = \nu_1 - \nu_2$ and $\beta = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. This implies that depending on the prior parameters, there is a non zero probability that the system is not in steady state. Therefore by obtaining the posterior distribution of ρ, a formal steady state hypothesis for the system can be conducted and steady state behavior can be inferred using call center data.

If a certain ordering of the system primitives are known a priori, then an alternative dependent prior distribution can be specified. For example, if it is believed that $\lambda < s\mu$, then a bivariate gamma prior which reflects this ordering can be specified. Though such restrictions are not required in M/M/s+M queues, the ordering $\lambda < s\mu$ is necessary in the M/M/s models for steady state results to exist. The bivariate gamma prior for λ and μ is given by

$$p(\lambda, \mu) = \frac{s^{\alpha_1}\phi^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\lambda^{\alpha_2-1}(\mu - \frac{\lambda}{s})^{\alpha_1-1}e^{-\phi s\mu}, \tag{3.5}$$

where $s\mu > \lambda$ and $\phi, \alpha_1, \alpha_2 > 0$. The density (3.5) is also known as the McKay's bivariate Gamma distribution as discussed in Kotz et al. (2000). As a consequence of (3.5), the marginals of λ and μ are given by the following Gamma densities

$$\lambda \sim G(\alpha_2, \phi), \tag{3.6}$$

$$\mu \sim G(\alpha_1 + \alpha_2, s\phi). \tag{3.7}$$

Also, it is straightforward to show that the prior for $\rho = \frac{\lambda}{s\mu}$ is given by the beta density

$$\rho \sim B(\alpha_1, \alpha_2), \tag{3.8}$$

where $0 < \rho < 1$. Since the prior distribution for ρ is beta distributed, some of the predictive priors of the operating characteristics can be obtained in closed form for call center design purposes (prior to observing the data). The above implies that a priori there is always a positive correlation between λ and μ which is consistent with the steady state assumption. Other characteristics are given in Aktekin (2009).

4 Bayesian Analysis of M/M/s+M Queues

In this section, we introduce Bayesian inference techniques for the independent and dependent priors discussed previously and show how posterior predictive analysis for relevant operating characteristics can be carried out.

We assume that we have observed n_x inter-arrival times, n_y service times, n_z abandonment times and n_z' waiting times for customers who did not abandon. Therefore, let $(x_1, \cdots, x_{n_x})$ be a collection of n_x samples from the exponential inter-arrival times distribution with rate λ, $(y_1, \cdots, y_{n_y})$ be a collection of n_y samples from the exponential service times distribution with rate μ and $(z_1, \cdots, z_{n_z}, z_1', \cdots, z_{n_z'}')$ be a collection of $n_z + n_z'$ samples from the exponential abandonment times distribution with rate θ. We define $D = (x_1, \cdots, x_{n_x}, y_1, \cdots, y_{n_y}, z_1, \cdots, z_{n_z}, z_1', \cdots, z_{n_z'}')$ for the observed data in M/M/s+M queues. Note that this reduces to $D = (x_1, \cdots, x_{n_x}, y_1, \cdots, y_{n_y})$ for M/M/s queues that we will use for comparison purposes in Section 5.

Therefore given D, the likelihood function of λ, μ and θ for M/M/s+M queues is given by

$$L(\lambda, \mu, \theta; D) \propto \lambda^{n_x} e^{-\lambda \sum_{i=1}^{n_x} x_i} \mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i} \theta^{n_z} e^{-\theta \sum_{i=1}^{n_z} z_i} e^{-\theta \sum_{i=1}^{n_z'} z_i'}. \tag{4.1}$$

In (4.1) the last term in the likelihood is due to the survival function of the exponential distribution and represents the contribution of the customers who received service and thus did not abandon the queue. In other words, their actual abandonment time is censored. McGrath et al. (1987b) and Armero and Bayarri (1996) discuss different experimental designs for queuing systems and point out that regardless of the design scheme the likelihood function in (4.1) will be the same from a

Bayesian point of view. The call center data which will be used in the sequel is consistent with what is described previously, that is, there are n_x observed interarrivals, n_y service times and $n_z + n'_z$ abandonment times.

4.1 Posterior Analysis with Independent Gamma Priors

First we introduce a conjugate Bayesian analysis using independent Gamma priors for the system primitives. The joint prior density can be written as the product of the Gamma marginals

$$p(\lambda, \mu, \theta) = p(\lambda)p(\mu)p(\theta), \tag{4.2}$$

where $\lambda \sim G(a_0, b_0)$, $\mu \sim G(c_0, d_0)$ and $\theta \sim G(e_0, f_0)$. Note that in the case of M/M/s queues the density $p(\theta)$ will be unity in (4.2).

Given (4.1) and (4.2), the joint posterior of λ, μ and θ can be obtained as a product of independent gamma densities given by $\lambda|D \sim G(a_1, b_1)$, $\mu|D \sim G(c_1, d_1)$, $\theta|D \sim G(e_1, f_1)$ where $a_1 = a_0 + n_x$, $b_1 = b_0 + (\sum_{i=1}^{n_x} x_i)$, $c_1 = c_0 + n_y$, $d_1 = d_0 + (\sum_{i=1}^{n_y} y_i)$, $e_1 = e_0 + n_z$, $f_1 = f_0 + (\sum_{i=1}^{n_z} z_i) + (\sum_{i=1}^{n'_z} z'_i)$. We note that since conditional on the primitives λ, μ and θ, arrival, service and abandonment times are independent, use of independent priors resulted in independent posterior distributions.

The use of the conjugate Gamma priors allows us to obtain analytical results for certain quantities such as posterior means of the system primitives. For example, the posterior mean of the abandonment rate θ is given by

$$E(\theta|D) = \frac{e_0 + n_z}{f_0 + (\sum_{i=1}^{n_z} z_i) + (\sum_{i=1}^{n'_z} z'_i)}, \tag{4.3}$$

where an increase in the total time spent in the queue without abandonment through $(\sum_{i=1}^{n'_z} z'_i)$ indicates a decrease in the expected posterior abandonment rate. In other words as customers do not abandon the queue the overall abandonment rate goes down, as a result customers are said to have become more patient.

So far we have not made any assumptions about the system being in steady state or not. Posterior analysis of λ, μ and θ can be carried out regardless. As pointed out by Mandelbaum and Zeltyn (2007), the M/M/s+M model always reaches steady state due to its birth and death process

properties and therefore steady state tests will not be carried out for M/M/s+M queues. Steady state conditions for M/M/s queues can be easily tested using the posterior distributions. This is done by computing the posterior probability that $\rho < 1$ given the observed data.

4.2 Posterior Analysis with Trivariate Lognormal Prior

Unlike the independent Gamma prior model, when we use the trivariate lognormal prior, closed form results cannot be obtained for the posterior distributions of the system primitives, λ, μ and θ. But we can use Markov chain Monte Carlo methods to draw samples from the posterior distribution of the system parameters. Next we discuss the implementation of the Metropolis-Hastings algorithm for the trivariate lognormal prior model. Given the likelihood function (4.1) and the trivariate lognormal prior density (3.1), the joint posterior of λ, μ and θ can be obtained as

$$p(\lambda, \mu, \theta|D) \propto L(\lambda, \mu, \theta; D)p(\lambda, \mu, \theta). \tag{4.4}$$

In order to draw samples from (4.4) we can use the random walk Metropolis-Hastings algorithm with a multivariate normal proposal density. Let $\boldsymbol{\theta}$ be a vector of parameters (for the trivariate lognormal case $\boldsymbol{\theta} = \{log(\lambda), log(\mu), log(\theta)\}$, then following Chib and Greenberg (1995) the steps in the Metropolis-Hastings algorithm can be summarized as follows

- Assume the starting points $\boldsymbol{\theta^{(0)}}$ at $j = 0$.
  Repeat for $j > 0$,
- Generate $\boldsymbol{\theta^*}$ from $q(\boldsymbol{\theta}^*|\boldsymbol{\theta^{(j)}})$ and u from $U(0, 1)$.
- If $u \leq \alpha(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^*)$ then set $\boldsymbol{\theta^{(j)}} = \boldsymbol{\theta}^*$; else set $\boldsymbol{\theta^{(j)}} = \boldsymbol{\theta^{(j)}}$ and $j = j + 1$,

where

$$\alpha(\boldsymbol{\theta^{(j)}}, \boldsymbol{\theta}^*) = min\left\{1, \frac{\pi(\boldsymbol{\theta^*})q(\boldsymbol{\theta^{(j)}}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta^{(j)}})q(\boldsymbol{\theta}^*|\boldsymbol{\theta^{(j)}})}\right\}, \tag{4.5}$$

and $q(.|.)$ is the multivariate normal proposal density and $\pi(\boldsymbol{\theta}) = p(log(\lambda), log(\mu), log(\theta)|D)$ as in (4.4). If we repeat the above a large number of times then we obtain samples from $p(\boldsymbol{\theta}|D)$.

Posterior analysis using the bivariate gamma prior (3.5) again requires use of Markov chain Monte Carlo methods but in this case a Gibbs sampler with a rejection sampling step can be used for drawing samples from the joint posterior of λ and μ. The details are given in the Appendix.

4.3 Posterior Analysis of Operating Characteristics

Next we focus on the estimation of the posterior predictive distributions of the operating characteristics introduced in Section 2. Depending on the form of the priors, the posterior distributions of λ, μ and θ can either be obtained in closed form as in the case of independent Gamma priors, or by using Markov chain Monte Carlo methods which are discussed in the previous section.

Once the joint posteriors are obtained, the next step is to obtain the posterior predictive distributions of the operating characteristics. For instance, $Pr(Ab|D)$ can be computed as

$$Pr(Ab|D) = \int_{\lambda}\int_{\mu}\int_{\theta} Pr(Ab|\lambda,\mu,\theta)p(\lambda,\mu,\theta|D)d\lambda\, d\mu\, d\theta. \tag{4.6}$$

In order to evaluate (4.6), one can use a Monte Carlo approximation as

$$Pr(Ab|D) \approx \frac{1}{S}\sum_{j=1}^{S} Pr(Ab|\lambda^{(j)},\mu^{(j)},\theta^{(j)}), \tag{4.7}$$

where S represents the number of samples generated, and $\lambda^{(j)}$, $\mu^{(j)}$, $\theta^{(j)}$ are the samples generated from the joint posterior distribution of λ, μ and θ. Another measure of interest is the expected average cost given by (2.11) which can be used to determine optimal staffing. This can be approximated as

$$E\{C(s,\lambda,\mu,\theta)|D\} \approx \frac{1}{S}\sum_{j=1}^{S} C(s,\lambda^{(j)},\mu^{(j)},\theta^{(j)}), \tag{4.8}$$

Estimation of other measures of performance can be developed along the same lines. Illustrations of these along with their implications on optimal staffing will be addressed in the next section.

5 Numerical Illustrations using Call Center Data

In order to illustrate the implementation of the proposed models, we will use real call center data from an anonymous bank operation. A detailed discussion of the call center data can be found at Data (2000). For illustrative purposes we have used arrivals, service and abandonment times in the mornings of month February (between 7:00AM -10:30AM) for stock exchange customers whose abandonment times seem to exhibit exponential type of behavior during these intervals. In our analysis, both the independent Gamma and the trivariate lognormal prior models will be

considered for M/M/s+M queues and comparisons will be made with the M/M/s model.

5.1 Case 1: Independent Gamma Priors Case

In order to carry out the inference for the independent Gamma priors model, we have used flat but proper priors for λ, μ and θ. A posterior summary of the most commonly used operating characteristics of the system primitives for different number of servers is shown in Tables 1 and 2. A quick note about the summary statistics is that they are all conditioned on the uncertain system primitives, but for notational convenience dependence on the parameters λ, μ and θ is suppressed in all the tables. For instance, P_0 from Table 1 is in fact $Pr(N = 0|\lambda, \mu, \theta)$ as defined in (2.3) and the expected value of P_0 is $Pr(N = 0|D)$.

Summary of Operating Characteristics-I						
Number of Servers	Statistics	P_0	P_1	P_2	P_3	P_4
s=1	Expected Value	0.3167	0.2911	0.2000	0.1101	0.0509
	Standard Deviation	0.0281	0.0129	0.0076	0.0121	0.0103
s=2	Expected Value	0.3853	0.3543	0.1635	0.0646	0.0224
	Standard Deviation	0.0241	0.0019	0.0102	0.0078	0.0041
s=3	Expected Value	0.3955	0.3644	0.1685	0.0521	0.0140
	Standard Deviation	0.0230	0.0016	0.0110	0.0066	0.0027

Table 1: Summary of Operating Characteristics-I for Case 1

Summary of Operating Characteristics-II				
Number of Servers	Statistics	P_{T_q}	$P_{Ab\|T_q}$	P_{Ab}
s=1	Expected Value	0.6832	0.3794	0.2592
	Standard Deviation	0.0281	0.0309	0.0237
s=2	Expected Value	0.2602	0.1991	0.0519
	Standard Deviation	0.0243	0.0248	0.0089
s=3	Expected Value	0.0714	0.1304	0.0093
	Standard Deviation	0.0106	0.0187	0.0021

Table 2: Summary of Operating Characteristics-II for Case 1

As shown in Table 2, the posterior expected probability of not getting served upon arrival, $Pr(T_q > 0|\lambda, \mu, \theta)$ (denoted by P_{T_q} in the table), is decreasing as the number of servers increase. Similarly, the expected probability of abandoning, $Pr(Ab|\lambda, \mu, \theta)$ (that is, P_{Ab} in the table) is decreasing as there are more servers. Both measures provide insights for the call center management since they can be interpreted as measures of perceived service quality of the call center operation.

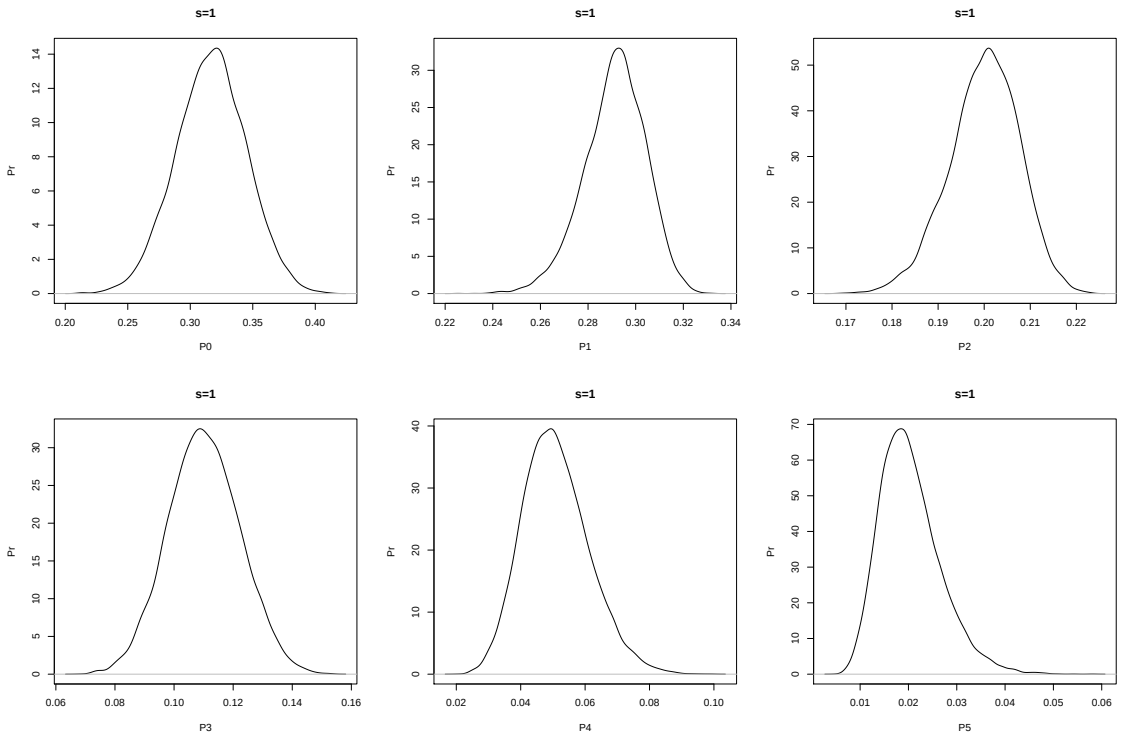

Figure 1: Posterior Distributions of $Pr(N = n|\lambda, \mu, \theta)$ for $n = 0, \cdots, 5$ when s=1

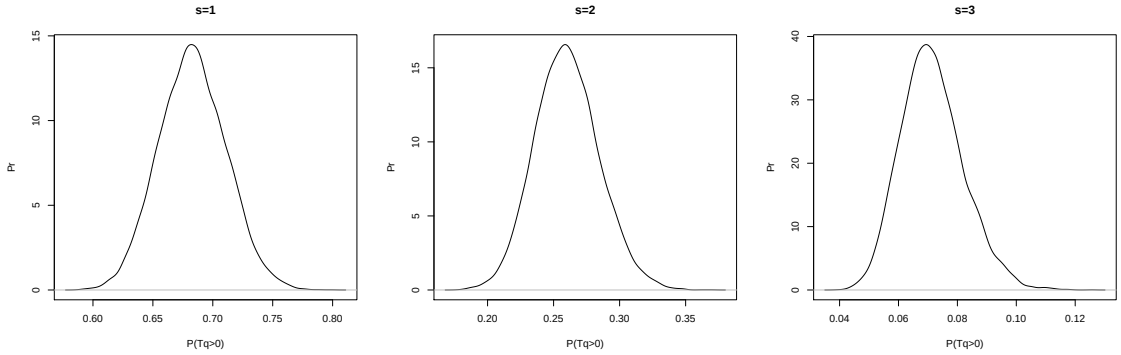

Figure 2: Posterior Distributions of $Pr(T_q > 0|\lambda, \mu, \theta)$ for s=1,s=2 and s=3

As a result of using the Bayesian approach, the system primitives as well as the operating characteristics will have their own distributions. Instead of solely using point estimates of the operating characteristics, the call center management can now construct credibility intervals and infer how sensitive a respective characteristic is to the changes in the system primitives. For instance as shown in Figures 1, 2 and 3, some of the density plots are not always symmetric and using the

mean as a point estimate may not be appropriate in all cases. The call center management can instead use the mode along with a 95 % credibility interval to make operational decisions.

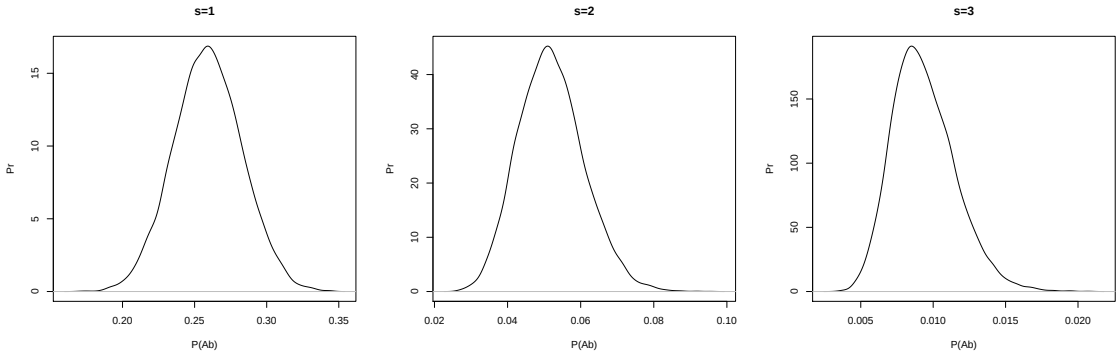

Figure 3: Posterior Distributions of $Pr(Ab|\lambda, \mu, \theta)$ for s=1,s=2 and s=3

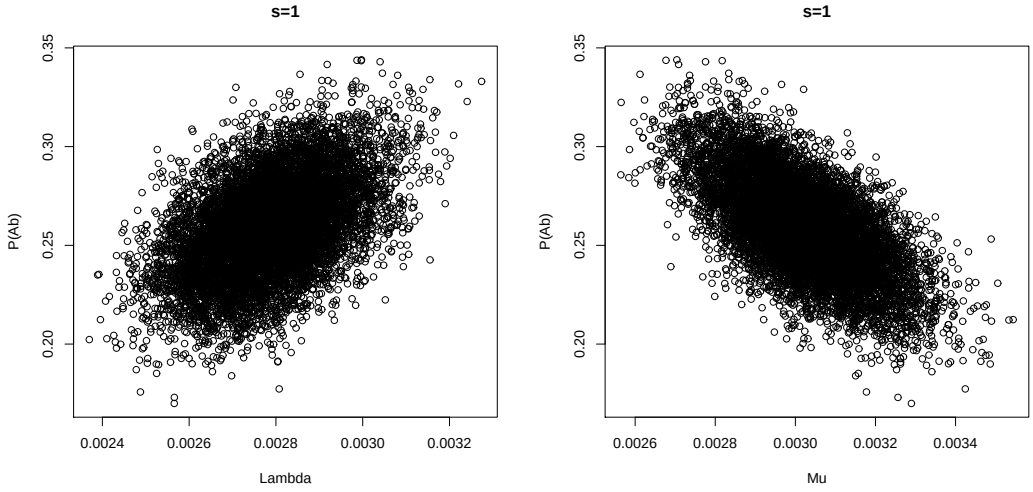

Figure 4: $Pr(Ab|\lambda, \mu, \theta)$ vs. λ (left) and $Pr(Ab|\lambda, \mu, \theta)$ vs. μ (right) for s=1

In Figure 4, we show the relationship between λ, μ and the probability of abandoning, $Pr(Ab|\lambda, \mu, \theta)$. We here note that $Pr(Ab|\lambda, \mu, \theta)$ is sensitive to the changes in the values of λ and μ regardless of the number of servers suggesting that use of point estimates may result in misleading conclusions. As expected, as arrivals become more frequent the fraction of customers that abandon the queue increases, whereas quicker service times lead to a decrease in the fraction of customers that abandon the queue.

5.2 Case 2: Trivariate Lognormal Prior Case

In Case 1, all system primitives were assumed to be independent from each other. Using the structure introduced for the trivariate lognormal prior model (Case 2), we can infer if the system primitives are in fact dependent or not. For example, if the arrivals were to become more frequent, the abandonment rate might tend to go up due to the increased number of people in the queue or the service rate might go down (or up) due to the sudden increase of call arrivals.

Similar to Case 1, we have used flat but proper priors in order to generate samples from (4.4). In our implementation of the Metropolis algorithm, 20,000 samples were obtained 5,000 of which were used as the burn-in sample. The trace plots for the posterior samples of λ, μ and θ are shown in Table 5 and the respective density plots in Table 6. Based on the trace plots we can fairly conclude that convergence has been attained.

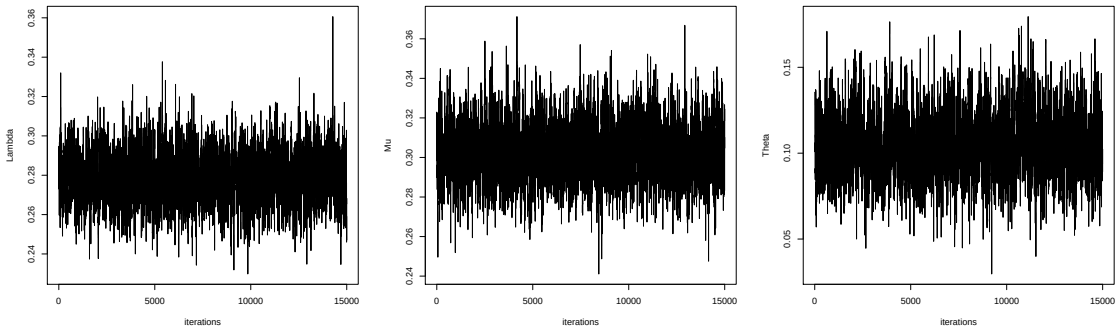

Figure 5: Trace Plots for λ (left), μ (middle) and θ (right)

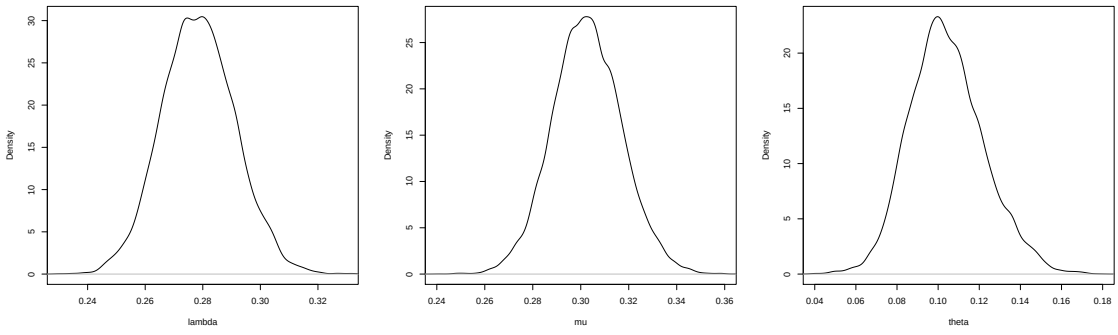

Figure 6: Posterior Density Plots for λ (left), μ (middle) and θ (right)

Next, similar to Case 1 we obtain the summary of the operating characteristics using the samples

generated from (4.4). As shown in Table 3 for $s = 1$, the summary statistics are fairly identical to those obtained for Case 1 (similar results were obtained for the summary statistics where $s = 2$ and $s = 3$). The density plots for the operating characteristics were identical to those obtained in Figures 1, 2 and 3. A boxplot of the steady state probabilities for $s = 1$ is shown in Figure 7.

Summary of Operating Characteristics-I						
Number of Servers	Statistics	P_0	P_1	P_2	P_3	P_4
s=1	Expected Value	0.3177	0.2915	0.1996	0.1096	0.0506
	Standard Deviation	0.0295	0.0139	0.0080	0.0126	0.0108

Table 3: Summary of Operating Characteristics-I for Case 2

Summary of Operating Characteristics-II				
Number of Servers	Statistics	P_{T_q}	$P_{Ab\|T_q}$	P_{Ab}
s=1	Expected Value	0.6882	0.3798	0.2591
	Standard Deviation	0.0295	0.0320	0.0250

Table 4: Summary of Operating Characteristics-II for Case 2

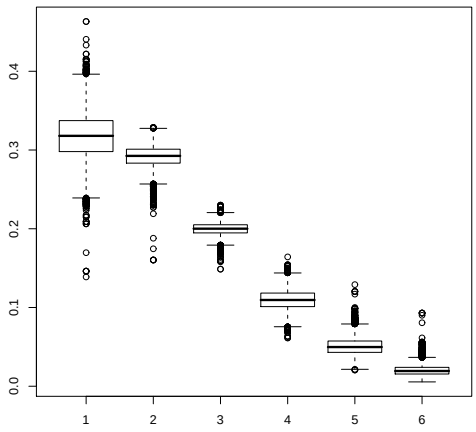

Figure 7: Boxplots for the Steady State Probabilities for Case 2 when s=1

The fact that the results are identical implies that the posterior distributions are very similar for Cases 1 and 2. The only difference between these two is that the latter allows for dependency between system primitives. As shown in Table 5, the posterior sample correlations for λ, μ and θ are all very close to zero. This shows evidence in favor of independent system primitives. Thus,

for this particular call center data, Case 1 seems to be the more appropriate model. However if there is a prior belief about potential dependence of the system primitives, then one can always test it using the lognormal model of Case 2. If evidence is found in favor of dependence then the operating characteristics should be obtained using samples generated from the trivariate lognormal prior model, failure to do so might lead to over or under estimation of the operating characteristics.

	λ	μ	θ
λ	1	0.0200	0.0014
μ	0.0200	1	-0.0004
θ	0.0014	-0.0004	1

Table 5: Posterior Sample Correlations for λ, μ and θ

5.3 Comparison with M/M/s Results

Classical queuing theory assumes steady state for analysis of M/M/s queues. An analogous scenario can be developed using the Bayesian framework. In other words, if a priori call center management believes that the system is in steady state then a bivariate Gamma prior can be used to model system primitives. In order to show the implications of ignoring abandonment, we have used the same arrival and service data as in Cases 1 and 2. Similar to the previous case, we have used flat but proper priors in order to generate samples from $p(\lambda, \mu|D)$, 20,000 samples were obtained 5,000 of which were used as the burn-in sample. Table 8 shows the trace plots and the density plots for λ and μ. Based on the trace plots we can fairly conclude that there are no convergence issues.

A summary of the most common operating characteristics for M/M/s queues as implied by the posterior samples of λ and μ for different number of servers is shown in Table 6 where the last column denotes the probability of $T_q = 0$ conditional on λ and μ. An important point about M/M/s queues is that they ignore the abandonment behavior of customers. For instance, as seen in Table 6 (for $s = 1$), the expected value of the probability that a customer will not wait upon arrival is about 0.1018 whereas it is estimated to be about 0.3168 for the M/M/s+M queues (as in Table 1). This under/over estimation in the operating characteristics emphasizes the importance of modeling the abandonment process in call center queuing applications. On the other hand, as the number of servers, s, go up the estimates for both M/M/s+M and M/M/s queues become similar.

Furthermore, the posterior distributions of the utilization, $p(\rho|D)$, for $s = 1, 2, 4$ are shown in Figure 9. An interesting finding is that for $s = 1$, there are several ρ values that are close to 1

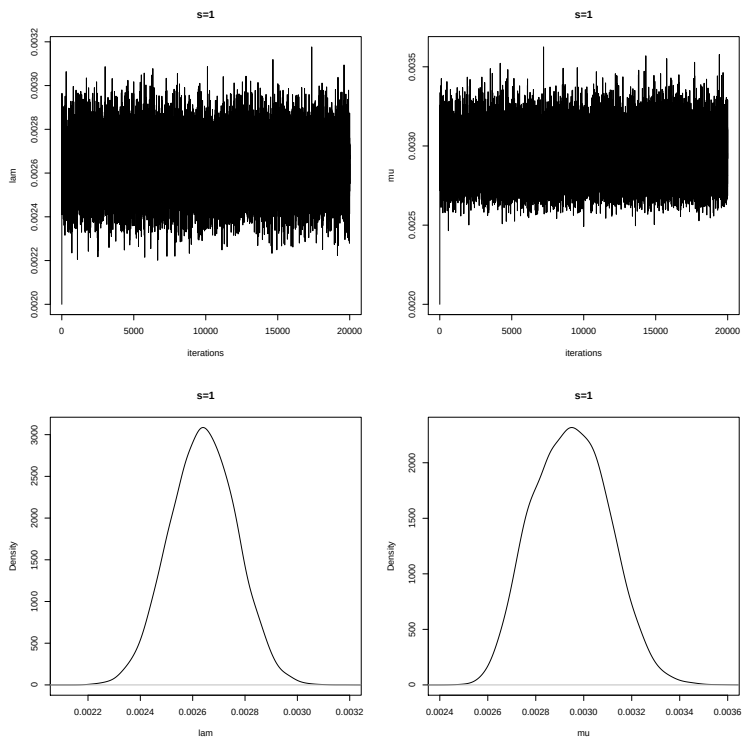

Figure 8: Trace Plots and Posterior Density Plots for λ (left) and μ (right) for M/M/s Model

Summary of Operating Characteristics						
Number of Servers	Statistics	P_0	P_1	P_2	P_3	$Pr(T_q = 0)$
s=1	Expected Value	0.1011	0.0860	0.0730	0.0620	0.1018
	Standard Deviation	0.0700	0.0559	0.0450	0.0366	0.0690
s=2	Expected Value	0.3806	0.3414	0.1525	0.0681	0.7204
	Standard Deviation	0.0334	0.0036	0.0104	0.0099	0.0366
s=3	Expected Value	0.4048	0.3618	0.1627	0.0490	0.9294
	Standard Deviation	0.0293	0.0024	0.0135	0.0078	0.0130

Table 6: Summary of Operating Characteristics for the M/M/s Model

implying a very high utilization of the system with a posterior expected value of 0.8988. The call center management can use this information and setup a tolerance level for utilization in order to adjust their staffing accordingly (or use it as a constraint in an optimal staffing problem). Since the utilization for M/M/s queues is assumed to be fixed in the conventional queuing models, being able to obtain the distribution of utilization is another important feature of the Bayesian approach.

Assessment of the steady state hypothesis given the data in M/M/s queues is an important contribution of the Bayesian analysis and it has no parallel in classical queuing theory. For instance

if we use the same data set as in Case 1 and assume that customers are patient (i.e. ignore abandonment), we can use the posterior distributions of λ and μ to test the hypothesis

$$H_0 : \{\rho > 1|D\} \text{ vs. } H_1 : \{\rho \leq 1|D\}.$$

In this case, we obtain $Pr(H_0 : \rho > 1|D) = 0.09786$ when $s = 1$. In other words there is almost a 10 % chance that the system is not in steady state. Whereas, in classical queuing theory the estimates for λ and μ would be 0.00278 and 0.00302 respectively. That is the system would have been assumed to be in steady state. When $s = 2$, $Pr(H_0 : \rho > 1|D)$ is approximately 0 indicating a sure steady state behavior. The inclusion of the steady state test here is for illustration purposes. We note here that the steady state hypothesis is not applicable to this particular example since there is abandonment in the system which already implies steady state behavior for M/M/s+M as discussed previously.

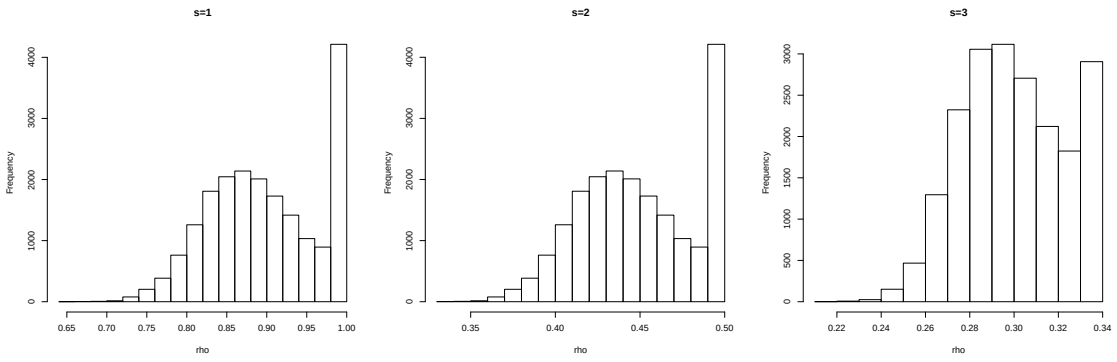

Figure 9: Posterior ρ for $s = 1, 2, 3$ in the M/M/s Model

5.4 Additional Insights from Bayesian Analysis

Implications of the proposed models on average operational cost (per unit of time), staffing and the level of service can provide additional managerial insights for call center practitioners. As pointed out previously, there is a trade-off between the server cost and the cost of abandoning (can be interpreted as a lost opportunity cost). Using the average operational cost (per unit of time) as introduced by (2.10), one can calculate the value of s which will minimize the cost function.

Since in our proposed models λ, μ and θ will have their own posterior distributions, we can

obtain the posterior distribution of the average operational cost as well as its expected value, i.e. $E\{C(s,\lambda,\mu,\theta)|D\}$ which is computed by (4.8). Thus, for a given ratio of c/a, we can obtain the value of s, say s^* which will minimize $E\{C(s,\lambda,\mu,\theta)|D\}$. Using the posterior samples obtained for Case 1 (results for Case 2 will be skipped since posterior distributions were identical), the cost functions were obtained for fixed c/a ratios for different number of servers as shown in Figures 10 and 11. Depending on the value of c/a, optimal staffing level changes. As c/a goes up, namely server cost becomes relatively more important, the optimal number of servers decreases. Whereas, as c/a goes down, namely abandonment cost becomes relatively more important, the optimal number of servers increases. In addition, the uncertainty associated with the total cost decreases as the number of servers increases as seen from boxplots given by Figures 10 and 11.

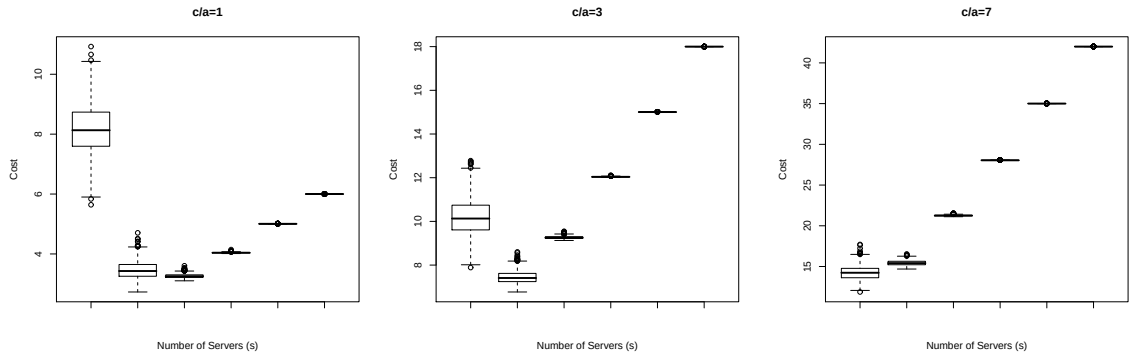

Figure 10: Cost Function vs. Number of Servers for fixed a/c

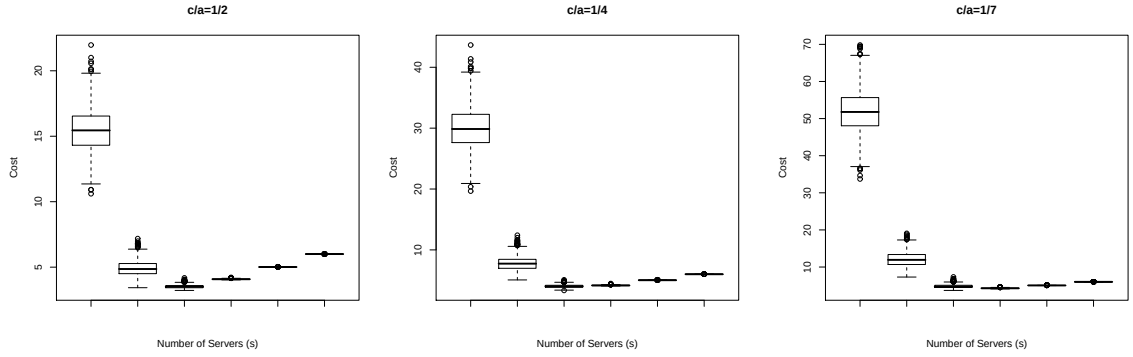

Figure 11: Cost Function vs. Number of Servers for fixed a/c

Since the total cost is a random variable when the system primitives are treated as uncertain quantities, alternatively one can calculate an optimal s^* value such that the probability of the total

cost being less than a fixed amount is at least, say 0.95. In other words choose s^* such that

$$Pr(C \leq c|\lambda, \mu, \theta, D) \geq \alpha_{\text{tolerance}}, \tag{5.1}$$

where c is the desired minimum cost and $\alpha_{\text{tolerance}}$ is a tolerance level such as 0.95 or 0.99.

Another commonly used approach for obtaining the appropriate level of staffing in call centers is the square root rule for safety staffing. Following Whitt (1992), let $r = \lambda/\mu$ denote the average offered load. Therefore, for M/M/s type of queues the square root rule for safety staffing is defined as follows,

$$s^* = r + \beta\sqrt{r}, \tag{5.2}$$

where s^* (an integer) represents the appropriate staffing level and β is a positive constant which is a function of level of service. As pointed out by Whitt (1992), (5.2) is appropriate for systems where the magnitude r, the offered load is moderate to large. Furthermore, β is obtained as a function of waiting time in the queue before service, $Pr(W > 0)$. The term $\beta\sqrt{r}$ is usually referred to as the excess capacity in the call center literature. Garnett et al. (2002) discuss the use of staffing rule (5.2), for M/M/s+M queues where the coefficient β is calculated as a function of both $Pr(W > 0)$ (probability of waiting before service) and θ (the abandonment rate). Furthermore, β is allowed to take negative values, namely the optimal staffing level can be less than the offered wait due to abandonment. For limiting values of the performance characteristics such as $Pr(W > 0)$ and $Pr(Ab)$, Garnett et al. (2002) discuss three different regimes for designing a call center; rationalized, quality-driven and efficiency-driven regimes. Based on the analysis, they conclude that using (5.2) along with the appropriate calculation of β, a call center manager can attain a balance between efficiency and service quality.

Alternatively, we can also study the effects of s^* obtained from (2.10) on the implied β index which can be obtained using (5.2). That is, for a given optimal number of servers we can infer the implied level of service, i.e. β which can be rewritten as $\frac{s^*-r}{\sqrt{r}}$. As a result, given the service level we can study the implications of the proposed models on the call center operational regimes which can be obtained based on β. For large values of β the operational regime is said to be Quality-Driven, for small values to be Efficiency-Driven and Quality-Efficiency-Driven if β takes values in (-1,2); see Garnett et al. (2002).

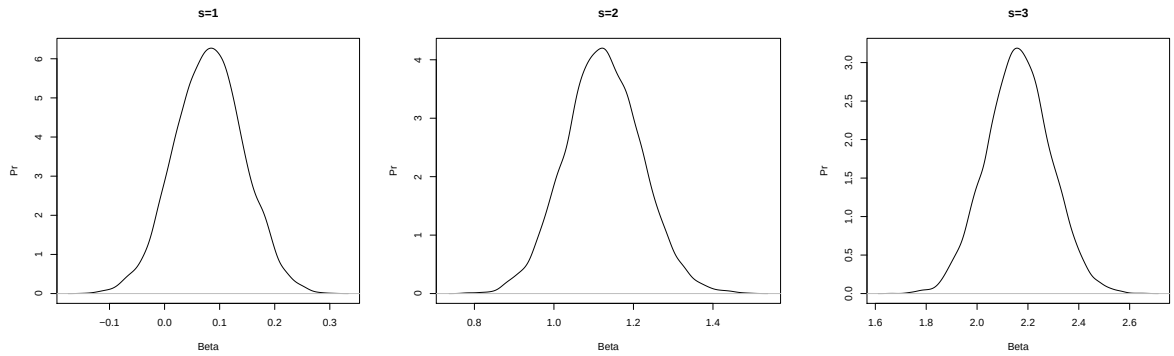

Figure 12: The Posterior Distributions of β for s=1,s=2 and s=3

Density plots for the implied level of service β for different optimal number of servers are shown in Figure 12 for the M/M/s+M queue. The posterior means of β, i.e. $E(\beta|D, s = s^*)$, for $s^* = 1, 2, 3$ were 0.0812, 1.124 and 2.163 respectively. In other words, the level of service is increasing that is the system is becoming more quality-driven as the optimal number of servers go up.

6 Concluding Remarks

In this paper we have introduced Bayesian queuing models for call center operations. We considered Bayesian models with independent Gamma priors and trivariate lognormal prior for M/M/s+M queues and compared them with M/M/s queues.

To the best of our knowledge this is the first study in the call center queuing literature from a Bayesian perspective where the system primitives are uncertain quantities and the first Bayesian treatment of queues with impatient customers. To be able to carry out inference for the system primitives, we have introduced Markov chain Monte Carlo methods such as Metropolis-Hastings and rejection sampling within Gibbs. In order to illustrate the implications and the implementation of the proposed models we have used real call center data from an anonymous bank operation. In doing so we have addressed the issues of interest as summarized in Section 1.

In the call center queuing literature, λ, μ and θ are assumed to be independent from each other. We have investigated if the independence assumption was adequate using real call center data. In doing so, we have developed the trivariate lognormal prior model to handle cases where dependent system primitives are suspected a priori. As a result we were able to conclude that the assumption was adequate, that is λ, μ and θ were found to be independent a posteriori as well in the light

of the data at hand. As for the steady state hypothesis, we have showed an example comparing the implications of our proposed models to those of the classical models. We have illustrated an example where the probability that the system was not in steady state was computed as 10% while the classical estimates of λ and μ were indicating a steady state behavior.

A natural result of treating λ, μ and θ probabilistically, are the distributions of operating characteristics, cost functions, and the level of service. We have specified a trade-off function, i.e. a total cost function, based on which we have developed a mechanism for finding the optimal number of servers. In doing so we were able to infer the level of service and the operational regime of the call center for various optimal number of servers.

Although the abandonment data at hand was exhibiting a fairly exponential type of behavior for the time period used which in turn validates the M/M/s+M model, further extensions where the abandonment process as a general distribution are possible and can be considered as a possible future extension. Furthermore, in typical call center operations, different agents might exhibit different service behavior for a given type of call. Therefore a hidden Markov model (or a mixture of exponential service times) for the service times can be investigated as a potential area for future work.

Appendix: Posterior Analysis with Bivariate Gamma Prior for $s = 1$

The joint posterior for λ and μ can be obtained as

$$p(\lambda, \mu|D) \propto \lambda^{n_x+\alpha_2-1} e^{-\lambda \sum_{i=1}^{n_x} x_i} \mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i+\phi} (\mu - \lambda)^{\alpha_1-1},$$

where $\mu > \lambda$. The above is not a known density form. Thus, we will use a Gibbs sampler to draw samples from the joint posterior. Since the modes from the conditional likelihoods are available, we can use a rejection sampling within Gibbs algorithm (see Smith and Gelman (1992)) First we need to obtain the full conditional densities, $p(\lambda|\mu, D)$ and $p(\mu|\lambda, D)$ as

$$p(\lambda|\mu, D) \propto \{\lambda^{n_x+\alpha_2-1}(\mu - \lambda)^{\alpha_1-1}\}\{e^{-\lambda \sum_{i=1}^{n_x} x_i}\} I(0, \mu),$$

and

$$p(\mu|\lambda, D) \propto \{\mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i+\phi}\}\{(\mu - \lambda)^{\alpha_1-1}\} I(\lambda, \infty).$$

Therefore the steps in rejection sampling within the Gibbs sampler can be summarized as follows

- Assume starting points $(\lambda^{(0)}, \mu^{(0)})$,
- Generate $\lambda^{(j)}$ from $(\lambda|\mu^{(j-1)}, D) \sim B(\alpha_1 + \alpha_2, \phi)I(0, \mu^{(j-1)})$ and u from $U(0,1)$,
- Accept if $\lambda^{(j)}$ if $u \leq \frac{L(\lambda^{(j)};\mu,D)}{L(\hat{\lambda};\mu,D)}$, else repeat the previous step,
- Generate $\mu^{(j)}$ from $(\mu|\lambda^{(j-1)}, D) \sim TrG(\alpha_1, \phi)I(\lambda^{(j-1)}, \infty)$ and u from $U(0,1)$,
- Accept if $\mu^{(j)}$ if $u \leq \frac{L(\mu^{(j)};\lambda,D)}{L(\hat{\mu};\lambda,D)}$, else repeat the previous step,

where the modes, $\hat{\lambda}$ and $\hat{\mu}$ can be obtained from the conditional likelihoods $L(\lambda; \mu, D)$ and $L(\mu; \lambda, D)$ as

$$\hat{\lambda} = \begin{cases} \mu, & \text{if} \quad \mu \leq \frac{n_x}{\sum_{i=1}^{n_x} x_i} \\ \frac{n_x}{\sum_{i=1}^{n_x} x_i}, & otherwise. \end{cases}$$

and

$$\hat{\mu} = \begin{cases} \lambda, & \text{if} \quad \lambda \geq \frac{n_y}{\sum_{i=1}^{n_y} y_i} \\ \frac{n_y}{\sum_{i=1}^{n_y} y_i}, & otherwise. \end{cases}$$

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

References

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

Aktekin, T. (2009). *Three Essays in Call Center Modeling, A Bayesian Perspective*. PhD thesis, The George Washington University.

Armero, C. and Bayarri, M. (1994). Bayesian prediction in M/M/1 queues. *Queuing Systems*, 15:401–417.

Armero, C. and Bayarri, M. (1996). *Bayesian Statistics 5*, chapter Bayesian Questions and Answers in Queues. Oxford University Press.

Bacelli, F. and Hebuterne, G. (1981). On queues with impatient customers. *Performance'81*, (2):159–179.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical analysis of a telephone call center: A queuing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hasting algorithm. *The American Statistician*, 49(4):327–335.

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

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

Gross, D. and Harris, C. M. (1998). *Fundamentals of Queuing Theory*. John Wiley & Sons.

Koole, G. and Mandelbaum, A. (2002). Queuing models of call centers: An introduction. *Annals of Operations Research*, (113):41–59.

Kotz, S., Johnson, N. L., and Balakrishnan, N. (2000). *Continuous Multivariate Distributions: Models and Applications*. John Wiley & Sons.

Lindley, D. V. (1990). The present position in Bayesian statistics. *Statistical Science*, 5(1):44–89.

Mandelbaum, A. and Zeltyn, S. (2007). *Advances in Service Innovations*, chapter Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers, pages 17–48. Springer.

McGrath, M. F., Gross, D., and Singpurwalla, N. D. (1987a). A subjective Bayesian approach to the theory of queues i - modeling. *Queuing Systems*, 1:317–333.

McGrath, M. F., Gross, D., and Singpurwalla, N. D. (1987b). A subjective Bayesian approach to the theory of queues ii - inference and information in M/M/1 queues. *Queuing Systems*, 1:335–353.

Palm, C. (1957). Research on telephone traffic carried by full availability groups. *Tele*, (1):107.

Smith, A. F. M. and Gelman, A. E. (1992). Bayesian statistics without tears: A Sampling perspective. *The American Statistician*, 46(2):84–88.

Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Management Science*, 38(5):708–723.

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

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

Wiper, M. (1998). Bayesian analysis of Er/M/1 and Er/M/c queues. *Journal of Statistical Planning and Inference*, 69:65–79.

Zeltyn, S. and Mandelbaum, A. (2005). Call centers with impatient customers: Many-server asymptotics of the M/M/n+G queue. *Queuing Systems*, (51):361–402.

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.