THREE ESSAYS IN CALL CENTER MODELING, A BAYESIAN PERSPECTIVE

By Tevfik Aktekin

B. S. in Mechanical Engineering, Yildiz Teknik University, Istanbul, 2002
M.B.A. in Management Decision Making, The George Washington University, 2004

A Dissertation submitted to

The Faculty of
The School of Business
The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 31, 2009

Dissertation Directed By
Refik Soyer
Professor of Decision Sciences and of Statistics

The School of Business of The George Washington University certifies that Tevfik Aktekin has passed the Final Examination for the degree of Doctor of Philosophy as of May 29, 2009. This is the final and approved form of the dissertation.

THREE ESSAYS IN CALL CENTER MODELING, A BAYESIAN PERSPECTIVE

Tevfik Aktekin

Dissertation Research Committee:

Refik Soyer, Professor of Decision Sciences and of Statistics, Dissertation Director

Srinivas Y. Prasad, Associate Professor of Decision Sciences, Committee Member

Jonathan R. Stroud, Assistant Professor of Statistics, Committee Member

Abstract of Dissertation

Three Essays in Call Center Modeling, a Bayesian Perspective

Our study aims to address three main issues in the area of call center modeling; within day and intra-week call arrival forecasting for staffing purposes, abandonment behavior of different call center customer profiles, and inference in measures of performance of the most commonly used call center queuing models. All three issues will be addressed from a Bayesian perspective implying that all uncertainties including those about model parameters will be described probabilistically. In our first essay we introduce a discrete time Bayesian state space model with Poisson measurements for intra-day call arrivals. We present the properties of our model and develop Bayesian inference. In so doing, we provide analytically tractable expressions for sequential updating for parameters, for smoothing and prediction of call arrivals and discuss how the model can be used for inter-weekly forecasts. In our second essay, we consider modeling abandonment behavior in call centers for different customer profiles. We present several time to event modeling strategies and develop Bayesian inference for posterior and predictive analysis. For the third essay we consider the following. 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 both patient and impatient customers and discuss their further extensions.

Table of Contents

Abstract of Dissertation …….iii

List of Figures ……..…….viii

List of Tables ……...…….xi

1 Introduction ……...1

2 Essay 1: Call Center Arrival Modeling, A Bayesian State Space Approach……..5

2.1 Summary ……..5

2.2 Introduction …….5

2.3 Call Center Arrival Modeling and Forecasting ……..9

2.4 A Discrete Time Poisson Model for Intra-day Call Arrivals ……..15

2.4.1 Long Run Behaviour of the Call Arrival Rate ……...20

2.4.2 Bayesian Learning about the Discount Parameter ……...22

2.4.3 Filtering Distribution and Retrospective Analysis ……...23

2.5 A Model for Inter-week Call Arrivals ……...25

2.6 Numerical Examples ……29

2.7 Concluding Remarks ……37

3 Essay 2: Bayesian Analysis of Abandonment in Call Center Operations ……...39

3.1 Summary ……39

3.2 Introduction and Overview …….39

3.3 Modeling the Abandonment Rate in Call Centers ……...44

3.3.1 Generalized Gamma and Exponentiated Weibull Models …….46

3.3.2 Mixture Models ……54

3.3.3 A Piecewise Time to Abandonment Model ……..58

3.4 Bayesian Inference ……60

3.4.1 Generalized Gamma and Exponentiated Weibull Inference ……..61

3.4.2 Piecewise Inference ……...64

3.4.3 Mixture Inference ……..64

3.4.4 Predictive Inference ……..68

3.4.5 Model Comparison Using Sampling Based Methods ……..69

3.5 Numerical Illustrations ……71

3.5.1 Case 1: Regular Customers ……..71

3.5.2 Case 2: Stock Exchange Customers ……79

3.6 Concluding Remarks ……82

4 Essay 3: Bayesian Modeling of Queues in Call Centers ……84

4.1 Summary ……84

4.2 Introduction ……...84

4.3 Queuing Models in Call Centers ……88

4.3.1 Some General Results for G/G/1 and G/G/s Queues ……...88

4.3.2 Operating Characteristics for M/M/1 Queues ……91

4.3.3 Operating Characteristics for M/M/s Queues ……93

4.3.4 Operating Characteristics for M/M/s+M Queues ……..94

4.4 Bayesian Queuing Models ……...98

4.5 Proposed Bayesian Queuing Models in Call Centers ……103

4.5.1 Independent Priors ……103

4.5.2 Dependent Priors ……104

4.6 Bayesian Analysis of Queues in Call Centers ……110

4.6.1 Posterior Analysis of Independent Gamma Priors ……111

4.6.2 Posterior Analysis of Trivariate Lognormal Prior ……..113

4.6.3 Posterior Analysis of Bivariate Gamma Prior ……115

4.6.4 Posterior Predictive Analysis ……117

4.7 Numerical Illustrations and Findings ……..119

4.7.1 Case 1: Independent Gamma Priors Case ……...119

4.7.2 Case 2: Trivariate Lognormal Prior Case ……123

4.7.3 Case 3: Bivariate Gamma Prior Case ……126

4.7.4 Additional Insights from Bayesian Analysis ……..128

4.8 Concluding Remarks ……..134

References ……137

List of Figures

For $\alpha_0 = 10$, $\beta_0 = .1$ (left) and $\alpha_0 = 100$, $\beta_0 = 1$ (right) ……...30

Posterior for different days of the week ……31

$p(N_{141} \mid N_{(140)})$ (left) and $p(N_{142} \mid N_{(141)})$ (right) ……33

Posterior means and the actual arrivals at different time intervals during Monday ……34

Posterior means $E(\theta_t \mid N_{(170)})$ versus time t for different days ……..34

Mean arrival rates during 7:00 to 9:30 AM(left) and 9:30 AM to 12: noon (right) 35

Mean arrival rates during 2:30 to 5:00 PM(left) and 5:00 to 9:05 PM (right) …….35

One week ahead forecasts for Monday (left) and Tuesday (right) versus actual call arrivals ……..37

Hazard Rates for Generalized Gamma; $\alpha = 3$, $\lambda = 2$, $\gamma = 1.6$ (left) and $\alpha = 0.9$, $\lambda = 0.5$, $\gamma = 0.9$ (right) ……49

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) ……50

Hazard Rates for the Lognormal Distribution …….. 51

Hazard Rates for Chen's Two Parameter Exponentiated Weibull …….54

Mixture Density Plots with Weibull Components k = 2 (left) and Lognormal Components k = 3 (right) ……..57

Time to Abandonment Densities for January, February, March of Regular Customers ……...59

Histograms for Regular Customers in January, February, March ……..60

$\log(\hat{f}(t_i \mid D_{(-i)},M_1)) - \log(\hat{f}(t_i \mid D_{(-i)},M_2))$ vs Sample i ……..73

Predictive Abandonment Rates for Regular Customers; Mixture Lognormal (3) (left), Mixture Lognormal (2) (right) …….. 75

Predictive Abandonment Rates for Regular Customers; Generalized Gamma (left), Piecewise Lognormal (right) ……..75

Posterior Predictive Fits vs. Actual Data ……78

Posterior Parameters Density Plots for the Lognormal Mixture Model with Three Components ……79

Trace Plots of α (left), γ (middle) and θ (right) ……..80

Brooks and Gelman Plots ……..80

Histogram for Stock Exchange Customers in January ……81

Posterior Distributions of $(P_i \mid D)$ for i = 1, … , 5 when s=1 ……121

$Pr(T_q > 0 \mid D)$ for s=1,s=2 and s=3 ……...121

$Pr(Ab \mid D)$ for s=1,s=2 and s=3 ……122

$Pr(Ab \mid D)$ vs. λ (left) and $Pr(Ab \mid D)$ vs. μ (right) for s=1 ……...123

Trace plots for λ (left), μ (middle) and θ (right) ……...124

Posterior density plots for λ (left), μ (middle) and θ (right) ……..124

Boxplots for the Steady State Probabilities for Case 2 when s=1 ……..125

Trace Plots and Posterior Density Plots for λ (left) and μ (right) for Case 3 …...127

Posterior ρ for s = 1, 2, 3 in Case 3 ……128

Cost Function vs. Number of Servers for fixed a/c ……131

Cost Function vs. Number of Servers for fixed a/c ……132

β vs. Optimal Number of Servers for Case 1 and 2 ……134

List of Tables

Posterior means and standard deviations for different days of the week……..32

One step ahead MAPE for different days ……...32

One week ahead MAPEs ……...36

Global CPO in the Log Scale for the Generalized Gamma and Exponentiated Weibull Models ……..72

Global CPO in the Log Scale for the Piecewise Models ……...72

Global CPO in the Log Scale for the Mixture Models ……..72

One-Month Ahead Fits ……..77

Posterior Parameters Summary Statistics for the Lognormal Mixture Model with Three Components ……78

Summary of Operating Characteristics-I for Case 1 ……...120

Summary of Operating Characteristics-II for Case 1 ……..120

Summary of Operating Characteristics-I for Case 2 ……...124

Summary of Operating Characteristics-II for Case 2 ……..125

Posterior Sample Correlations for λ, μ and θ ……...126

Summary of Operating Characteristics for Case 3 ……..127

1 Introduction

In the last decade call center operations became an inevitable part of many global businesses. Thanks to the technological advancement in the communications field, call centers grew to be the milieu through which firms interact with their potential and existing customers. The complexities involved in managing call center operations gave birth to managerial challenges worth investigating. Different stages of a call center operation contain managerial decision making scenarios where many processes and their components are unknown. Three fundamental uncertain processes in call center operations can be summarized as arrival, service and abandonment. Each offer challenges from both modeling and practical points of view. In general, the call center related literature consist of a wide range of fields such as forecasting, queuing, scheduling, customer behaviour and operations management.

Our study aims to address three main issues in the area of call center modeling; within day and intra-week call arrival forecasting for staffing purposes, abandonment behaviour of different call center customer profiles, and inference in measures of performance of the most commonly used call center queuing models. All three issues will be addressed from a Bayesian perspective implying that all uncertainties including those about model parameters will be described probabilistically. The Bayesian view point has been considered only very recently in call center literature by Weinberg et al. (2007) and Soyer and Tarimcilar (2008), both of which addressed the issue of forecasting call arrivals.

In our first essay we introduce a discrete time Bayesian state space model with Poisson measurements for intra-day call arrival forecasting and discuss how the model can be used for inter-weekly forecasts. We aim to illustrate the implementation of the model by using actual intra-day arrival data from a US commercial bank's call center. The novel contribution of this work is the proposed discrete time Poisson state space model which enables to obtain certain inferences analytically or by use of standard non-iterative Monte Carlo methods. From a practical point of view, the fact that state parameter updating and one-step-ahead call forecasting are analytically tractable makes the proposed model attractive. From a staffing point of view, parameter updating efficiency and adaptability of the proposed model are other features that might be of interest to call center managers. Furthermore, the proposed model incorporates managerial insights from call center practitioners via the prior distribution on the parameters.

In the second essay, 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. Time to event (or duration) models such as generalized gamma, exponentiated Weibull, piecewise and mixture densities are considered for describing customer abandonment behavior. The proposed models are illustrated using real call center data for different customer profiles. These models that are capable of capturing non-monotonic abandonment rates have not been considered in the call center modeling literature. Furthermore, the Bayesian view point

has not been implemented in analysis of abandonment data. Thus, the modeling aspect as well as the Bayesian analysis of the corresponding models represent one of the contributions of our second paper. Also, Bayesian inference for exponentiated Weibull models can be considered as a modest contribution to applied Bayesian statistics literature. In addition, issues which are investigated in this study can be summarized as whether call center customers exhibit monotonic or non-monotonic abandonment behavior, whether diverse customer profiles exhibit different abandonment rate behavior and finally whether customers change their abandonment behavior as they experience announcements or as they wait in the line.

In our last essay we develop Bayesian analysis of call center queuing models and in so doing, we consider models with both patient and impatient customers and discuss their further extensions. We develop details of Bayesian inference for queues with abandonment such as M/M/s+M models and present relevant operating characteristics and inference about them. We illustrate analysis of our models using actual call center arrival, service, and abandonment data and discuss how the Bayesian analysis can be used to develop staffing policies such as determining number of servers and to infer service levels. The contributions of the third paper is twofold. Firstly, Bayesian analysis and its use in determining staffing and service levels has not been considered before in call center queueing models. Secondly, analysis of queueing models with abandonment has not been considered in the Bayesian queuing literature. In addition, issues such as correlated system primitives, prior assumption for steady state queues,

posterior steady state hypothesis, prior and posterior implications of the operating characteristics implied by different models, application of Little's formula in posterior Bayesian analysis and staffing implications are investigated. Furthermore, some of the prior distributions that we introduce to describe the potential dependence among queuing system primitives and their implications on performance measures seem to be new and represent a modest contribution.

2 Essay 1: Call Center Arrival Modeling, A Bayesian State Space Approach

2.1 Summary

In this paper we introduce a discrete time Bayesian state space model with Poisson measurements for intra-day call arrivals. We present the properties of our model and develop Bayesian inference. In so doing, we provide analytically tractable expressions for sequential updating for parameters, for smoothing and prediction of call arrivals and discuss how the model can be used for inter-weekly forecasts. We illustrate the implementation of the model by using actual intra-day arrival data from a US commercial bank's call center.

2.2 Introduction

Call centers provide vital contact points between firms and their potential and existing customers. In general, customers who are looking for service are served either by an interactive voice response unit (IVR) or by agents of the call center. The focus of this paper will be on calls handled by agents. In recent years, academic research in call center modeling has shown a drastic increase in volume. In a survey paper by Gans et al. (2003), 164 call center related papers are referenced. A detailed overview of recent call center related literature can be found in Aksin et al. (2007), where different research areas such as forecasting, queueing, scheduling and behavioural

issues are considered. Call center management consists of technology and labour intensive operations, where staffing comprise 60-80% of the overall operating budget (Aksin et al. (2007)). Therefore modeling uncertainty in call center arrivals is an important issue for call center management. In this paper, we address the issue of modeling daily call volumes in an inbound call center with particular emphasis on intra-day forecasting and illustrate how inter-week forecasts can be developed from the proposed models.

Due to the increased availability of call center databases, forecasting related studies have been an active area for academic research. Some of the recent work in this area can be summarized as follows. Jongbloed and Koole (2001) discuss that the arrival process at a Dutch insurance company call center follow a non-homogeneous Poisson process where the arrival rate is not constant over time. Avramidis et al. (2004) develop a model based on the correlation structure of the intra-day arrivals at a Bell Canada call center using doubly stochastic Poisson processes. Recent work by Weinberg et al. (2007) and Soyer and Tarimcilar (2008) are the first two known attempts to model call volumes using a Bayesian point of view. Soyer and Tarimcilar (2008) model the arrival process at a consumer electronics producer call center via a modulated Poisson process whose rate is a function of advertising specific characteristics as well as time. Weinberg et al. (2007) consider the intra-day arrival behaviour of a US commercial bank's call center and introduce a non-homogeneous Poisson model whose arrival rates are changing both intra-day and inter-day. A sequential

Monte Carlo algorithm is developed in order to estimate both the latent states and the model parameters. In order to develop the MCMC algorithm, the arrival count data is transformed into a Gaussian time series using a square root transformation. As noted by Aksin et al. (2007) the approach proposed by Weinberg et al. (2007) is computationally intensive especially for intra-day forecast updating.

Intra-day forecasting is an important issue in call center staffing decisions. Whitt (1999a) mentions the importance of near term staffing needs of a call center in order to carry out dynamic staffing with the aim of answering all calls. Near term staffing is usually referred to as the staffing requirements in the next minute, five minutes or ten minutes. Conventional methods used in the industry for call center staffing problems require that mean call volume at any given time interval is forecasted so that the required staffing level is $c\sqrt{m}$, where c is a constant defined over $1 \leq c \leq 10$ and m is the predicted mean for a given time interval. Another approach proposed by Whitt (1999a) for dynamic staffing is as follows. Let $s(t)$ be the required number of servers at time t, then

$$s(t) = [E(D(t)) + z_\alpha \sqrt{Var(D(t))} + 0.5] \tag{2.1}$$

where $P(N(0,1) > z_\alpha) = \alpha$, $[x]$ is the least integer greater than x, $D(t)$ and $V(t)$ are the respective mean and the variance of the predicted demand for time t. Therefore, availability of intra-day updating call enables call center managers adjust their staffing dynamically on a given day. Whitt (1999a) also introduces the idea of flexible staffing

where the call center agents have alternative work on a given day. In other words if the call volume is predicted to be low in the morning they can be assigned to other tasks, as soon as intra-day forecasts indicate an increase in the call arrivals they can be reassigned to answering inbound calls. Thus, it is important to be able to update forecasts rather easily at any time period in a given day.

In this paper, we propose an alternative modeling strategy via modeling the original count data rather than its transformation and consider a discrete time Bayesian state space model with Poisson measurements. The attractive feature of our model is that, under certain conditions, it provides analytically tractable expressions for sequential updating for the model parameters, for smoothing and predictions of call arrivals. Thus, it can easily be used with decision models for updating intra-day staffing schedules. Furthermore, the proposed approach is quite valuable in inferring inter-day as well as intra-day differences in arrival patterns and it can be extended to obtain for inter-week forecasts.

A synopsis of our paper is as follows: In Section 3, we review the most recent studies in call center arrival forecasting and discuss their respective properties. In Section 4, we introduce a Bayesian state space model for the daily within day call arrivals, show how the sequential updating of the parameters is implemented and discuss the underlying properties of the proposed model. In section 5, we present an extension of the model to consider interweekly forecasts. In section 6, a numerical example is illustrated for the model. Finally, in section 7 we conclude with comments

and suggestions for future extensions of our work.

2.3 Call Center Arrival Modeling and Forecasting

Jongbloed and Koole (2001) point out that the arrival rate in call centers during fixed time intervals is not constant and propose to model it as a stochastic quantity. In order to account for the overdispersion in arrival count data, the following Poisson mixture structure is proposed

$$P(X = x) = \int_0^\infty \frac{\lambda^x}{x!} e^{-\lambda} dH(\lambda), \tag{2.2}$$

where X is the number of call arrivals during a fixed interval on a given day and $H(\lambda)$ represents the distribution of the call arrival rate. Since $H(\lambda)$ is unknown, it needs to be estimated. Jongbloed and Koole (2001) proposed two modeling techniques, one parametric and one non-parametric. The parametric approach assumes that $H(\lambda) \sim G(r, s)$, i.e. a gamma density with parameters r and s. As a result, it can be shown that $P(X = x)$ is a negative binomial density with success probability $s/(1+s)$ and the number of successes r. As pointed out by Jongbloed and Koole (2001), the gamma density is unimodal and may not account for heterogeneous nature of the call arrivals on different time intervals. As an alternative, non-parametric estimation using maximum likelihood estimator for Poisson mixtures is proposed. Finally, the effects on staffing of these two approaches are illustrated using call center data from a Dutch insurance company.

Avramidis et al. (2004) discuss the existence of intra-day correlations of the call arrivals on a given day and point out that the model proposed by Jongbloed and Koole (2001) does not account for potential dependency structures between successive call arrival intervals. They address this issue by introducing three models. The first model assumes that a sequence of call arrivals on a given day with k intervals follows a negative multinomial distribution, where call arrivals given the random arrival rate function, $\Lambda(t)$ is a non-homogeneous Poisson process. The second model assumes a compound negative multinomial distribution where the parameters of the model are random and jointly follow a Dirichlet distribution. A drawback for the first two models is that they both have positive correlation structures between individual call arrivals. Therefore, a third model is proposed where the vector of arrival ratios follow a Dirichlet distribution. This model is noted to have more flexibility in terms of the correlation structure with respect to the other two models discussed previously. Finally, all three models are used to model a Bell Canada call center and comparisons are presented.

The article by Weinberg et al. (2007) studies the within-day arrival behaviour of a US commercial bank's call center. The authors introduce an inhomogeneous Poisson model whose arrival rates are changing both intra-day and inter-day. A sequential Markov chain Monte Carlo algorithm is developed in order to estimate both the latent states and the model parameters. In order to develop the Markov chain Monte Carlo algorithm, the authors transform the arrival count data into a Gaussian time series

using a modified square root transformation. Furthermore, one step ahead forecasts are obtained and are compared with competing models. The authors conclude that their proposed model improves the predictions by 25%. In what follows, we briefly outline the model as presented in Weinberg et al. (2007).

According to standard queuing theory, in general arrival rates follow a Poisson process with rate λ. Researchers in call center modeling have argued that this may not be the case in most call center arrival data. The authors propose that arrival rates to the call center at hand follow a non-homogeneous Poisson process with a stochastic arrival rate $\lambda(t)$. Here, $\lambda(t)$ changes during the day which takes into account the over-dispersion in the data.

Let N_{jk} denote the number of arrivals to the call center on day j during the time interval $[t_{k-1}, t_k]$ where $j = 1, \cdots, J$ and $k = 1, \cdots, K$. Therefore, the proposed model is

$$N_{jk} \sim Poisson(\lambda_{jk}), \lambda_{jk} = R_{d_j}(t_k)v_j + \epsilon_{jk} \tag{2.3}$$

where d_j represents the weekday corresponding to day j, λ_{jk} is the arrival rate on day j during the time period $[t_{k-1}, t_k]$, $R_{d_j}(t_k)$ is the proportion of daily calls on day j during the period $[t_{k-1}, t_k)$, v_j is a proxy for the total call arrivals on day j and ϵ_{jk} is a random error term. Also, the proportion of daily calls on day j is treated as a density as follows

$$\sum_{k=1}^{K} R_{d_j}(t_k) = 1 \tag{2.4}$$

where $d_j = 1, \cdots, 5$. Furthermore, the authors use the variance stabilizing transfor-

mation of the Poisson count data in order to be able to use the appropriate Markov chain Monte Carlo algorithms.

Using (2.3) and the transformation $y_{jk} = \sqrt{N_{jk} + \frac{1}{4}}$,

$$y_{jk} = g_{d_j}(t_k)x_j + \epsilon_{jk} \tag{2.5}$$

where $\epsilon_{jk} \sim N(0, \sigma^2)$, $g_{d_j}(t_k) = \sqrt{R_{d_j}(t_k)}$ and $x_j = \sqrt{v_j}$. The authors also make a point about the validity of the approximation used in (2.5) by pointing out the large enough size of the arrival counts at hand.

Furthermore, in order to capture the day to day relationship of the arrival counts, the authors propose a first order autoregressive type of model on x_j's.

$$x_j - \alpha_{d_j} = \beta(x_{j-1} - \alpha_{d_{j-1}}) + \eta_j \tag{2.6}$$

where $\eta_j \sim N(0, \psi^2)$ and α_{d_j} represents the intercept for day d_j.

Based on (2.5), the authors also assume the following structure on $g_{d_j}(t_k)$'s

$$\frac{d^2 g_{d_j}(t_k)}{dt^2} = \tau_{d_j} \frac{dW_{d_j}(t_k)}{dt_k} \tag{2.7}$$

where $W_{d_j}(t)$ are independent Wiener processes with $W_{d_j}(0) = 0$ and $Var(W_{d_j}(t)) =$

t. Also following (2.4) and the transformation mentioned before,

$$\sum_{k=1}^{K} g_{d_j}(t_k)^2 = 1 \tag{2.8}$$

where $d_j = 1, \cdots, 5$. An extra step is required in order for the Markov chain Monte Carlo algorithm to be developed properly. The authors define $z_{d_j} = \{g_{d_j}(t_k\}, \frac{dg_{d_j}(t_k)}{dt_k})$ and using (2.5)

$$y_{jk} = h^{'} z_{d_j}(t_k) x_j + \epsilon_{jk}, \epsilon_{jk} \sim N(0, \sigma^2), \tag{2.9}$$

$$x_j - \alpha_{d_j} = \beta(x_{j-1} - \alpha_{d_{j-1}}) + \eta_j, \eta_j \sim N(0, \psi^2), \tag{2.10}$$

$$z_{d_j}(t_k) = F(\delta) z_{d_j}(t_{k-1}) + u_k, u_k \sim N(0, \tau_{d_j}^2 U(\delta)) \tag{2.11}$$

where ϵ_{jk}, η_j and u_k are assumed to be pairwise independent, $\delta = t_k - t_{k-1}$ and $h^{'} = [1, 0]$. Also,

$$F(\delta) = \begin{pmatrix} 1 & \delta \\ 0 & 1 \end{pmatrix}, U(\delta) = \begin{pmatrix} \delta^3/3 & \delta^2/2 \\ \delta^2/2 & \delta \end{pmatrix}. \tag{2.12}$$

Therefore, (2.9), (2.10) and (2.11) consist of a multiplicative model with latent states where the efficient forward filtering backward sampling algorithm can easily be implemented. In order to carry out the Bayesian inference on the above mentioned model prior selection is discussed.

What follows the above mentioned model, is the implementation using Markov chain Monte Carlo algorithms and a discussion of convergence measures follows the implementation. The authors point out that closed form predictive densities for

$p(N_j|Y_{1:j-1})$ cannot be obtained, therefore they develop a Markov chain Monte Carlo algorithm to approximate the results. Consequently, one day ahead predictions for the proposed model and two regression models with seasonal and random effects are obtained for comparison purposes. As a result of the predictions, root mean square error (RMSE) and average percent error (APE) for data based on the next 64 days is obtained and compared against the two competing models. Based on out of sample predictions, the model proposed by the authors outperforms the other two seasonal linear models.

Most recently, Shen and Huang (2008) model call arrivals as a vector time series which has both intra-day and inter-day time components. Shen and Huang (2008) also transform the count data using the square root transformation and discuss issues such as updating using penalized least squares method, reducing data dimensionality using singular value decomposition and inter-day forecasting using univariate forecasting methods. Due to the lack of underlying probability distribution assumptions on call center data, distributional forecasts are obtained using non-parametric bootstrapping method. Finally, out of sample forecasts of two call center datasets are obtained and compared against competing models.

2.4 A Discrete Time Poisson Model for Intra-day Call Arrivals

Queuing theory has been commonly used in the call center modelling literature. However, appropriateness of the standard assumptions of queuing models have been ques-

tioned by many researchers. For example, Jongbloed and Koole (2001) and Weinberg et al. (2007) note that the assumption of Poisson arrivals with a constant rate over time does not apply to call center arrivals. Therefore they propose a stochastic arrival rate, θ_t, which evolves over time. Similarly, the independent increments property of Poisson processes has been criticized by Avramidis et al. (2004) who propose a Poisson processes with stochastic arrival intensities. In this paper, we consider a discrete time Poisson process with a stochastic arrival rate that evolves over time according to a discrete time Markov process. Our model and its properties are discussed in the sequel.

Let N_t be the number of call arrivals during the time period $t-1$ to t in a given day and θ_t be the corresponding arrival rate during the same time period. Note that the length of the time interval may change depending on the call arrival data at hand. For example, N_t may represent the call arrivals during the tth minute, t^{th} hour or t^{th} day depending on the nature of the data, but the proposed model can still be applicable.

Given the arrival rate θ_t, we assume that the number of call arrivals during period t is described by a discrete time nonhomogeneous Poisson process with probability distribution

$$p(N_t|\theta_t) = \frac{\theta_t^{N_t} e^{-\theta_t}}{N_t!}. \tag{2.13}$$

It is assumed that given θ_t's, N_t's are independent. In other words, the model implies that the independent increments property holds only conditional on θ_t and

N_t's are correlated. In the above, (2.13) acts as a measurement equation and is defined over discrete space in time.

For time evolution of θ_t's, we assume a Markovian structure where the relationship between the arrival rates is described by

$$\theta_t = \frac{\theta_{t-1}}{\gamma}\epsilon_t \tag{2.14}$$

where $(\epsilon_t|N^{(t-1)}) \sim Beta(\gamma\alpha_{t-1}, (1-\gamma)\alpha_{t-1})$ with $\alpha_{t-1} > 0$, $0 < \gamma < 1$, and $N^{(t-1)} = (N_1, \cdots, N_{t-1})$. In (2.14), γ acts like a discounting term. It follows from the above that $\theta_t < \frac{\theta_{t-1}}{\gamma}$. We can show that $\theta_t|\theta_{t-1}, N^{(t-1)} \sim Beta(\gamma\alpha_{t-1}, (1-\gamma)\alpha_{t-1}, 0, \frac{\theta_{t-1}}{\gamma})$, that is, a scaled Beta density given by

$$p(\theta_t|\theta_{t-1}, N^{(t-1)}) = \frac{\Gamma(\alpha_{t-1})}{\Gamma(\gamma\alpha_{t-1})\Gamma((1-\gamma)\alpha_{t-1})}(\frac{\gamma}{\theta_{t-1}})^{\alpha_{t-1}-1}\theta_t^{\gamma\alpha_{t-1}-1}(\frac{\theta_{t-1}}{\gamma} - \theta_t)^{(1-\gamma)\alpha_{t-1}-1}. \tag{2.15}$$

The state equation (2.14) also implies that $E(\theta_t|\theta_{t-1}, N^{(t-1)}) = \theta_{t-1}$, that is, a random walk type of evolution for the arrival rates. We note that a modified version of the evolution model (2.14) was first introduced in Smith and Miller (1986) where the measurement equation was exponential and later was used by Morali and Soyer (2003) in the context of software reliability.

A Poisson measurement model has not been considered in the earlier work, but it is possible to develop an analytically tractable Bayesian analysis for the model if we assume that at time 0, $(\theta_0|N^{(0)})$ is a gamma distribution. In other words, before any

of the call arrivals have been observed, we define our prior uncertainty regarding the call arrival rate as

$$(\theta_0|N^{(0)}) \sim Gamma(\alpha_0, \beta_o). \tag{2.16}$$

Following Smith and Miller (1986) we assume that

$$(\theta_{t-1}|N^{(t-1)}) \sim Gamma(\alpha_{t-1}, \beta_{t-1}) \tag{2.17}$$

which we can show by induction.

Using (2.15) and (2.17), we can obtain the prior of θ_t given $N^{(t-1)}$ via

$$p(\theta_t|N^{(t-1)}) = \int_{\gamma\theta_t}^{\infty} p(\theta_t|\theta_{t-1}, N^{(t-1)})p(\theta_{t-1}|N^{(t-1)})d\theta_{t-1}, \tag{2.18}$$

which reduces to a gamma density as

$$(\theta_t|N^{(t-1)}) \sim Gamma(\gamma\alpha_{t-1}, \gamma\beta_{t-1}). \tag{2.19}$$

It follows from the above that $E(\theta_t|N^{(t-1)}) = E(\theta_{t-1}|N^{(t-1)})$, whereas $V(\theta_t|N^{(t-1)}) = \frac{V(\theta_{t-1}|N^{(t-1)})}{\gamma}$. In other words, the model implies that as we move forward in time, expected call arrival rate stays the same but our uncertainty about arrival rate increases as a function of the discount factor γ.

Given the prior (2.17) and the Poisson observation model (2.13) we obtain the

posterior distribution $\theta_t|N^{(t)}$ using the Bayes' Rule,

$$p(\theta_t|N^{(t)}) \propto p(N_t|\theta_t)p(\theta_t|N^{(t-1)}). \tag{2.20}$$

The above implies that

$$p(\theta_t|N^{(t)}) \propto \theta_t^{\gamma\alpha_{t-1}+N_t-1}e^{-(\gamma\beta_{t-1}+1)\theta_t},$$

that is, the posterior distribution of the call arrival rate at time t is a gamma density

$$\theta_t|N^{(t)} \sim Gamma(\alpha_t, \beta_t) \tag{2.21}$$

where $\alpha_t = \gamma\alpha_{t-1} + N_t$ and $\beta_t = \gamma\beta_{t-1} + 1$.

The one-step ahead predictive distribution of call arrivals at time t given $N^{(t-1)}$ can be obtained via

$$p(N_t|N^{(t-1)}) = \int_0^\infty p(N_t|\theta_t)p(\theta_t|N^{(t-1)})d\theta_t \tag{2.22}$$

where $N_t|\theta_t \sim Poisson(\theta_t)$ and $\theta_t|N^{(t-1)} \sim Gamma(\gamma\alpha_{t-1}, \gamma\beta_{t-1})$. Therefore,

$$p(N_t|N^{(t-1)}) = \binom{\gamma\alpha_{t-1}+N_t-1}{N_t}(1-\frac{1}{\gamma\beta_{t-1}+1})^{\gamma\alpha_{t-1}}(\frac{1}{\gamma\beta_{t-1}+1})^{N_t}. \tag{2.23}$$

which is a negative binomial model denoted as

$$N_t|N^{(t-1)} \sim Negbin(\gamma\alpha_{t-1}, \gamma\beta_{t-1}). \tag{2.24}$$

Availability of one-step ahead predictive density in closed form is an attractive feature of the model from a practical point of view. Given (2.24), one can carry out one step ahead predictions and credibility interval calculations in a straightforward manner. Further results regarding one step ahead predictive distributions will be discussed in our numerical example section.

Although the k-step ahead predictive density is not analytically available, the k-step ahead predictive means can be easily obtained. Using a standard conditional expectation argument one can obtain $E(N_{t+k}|N^{(t)})$ as follows

$$E(N_{t+k}|N^{(t)}) = E_{\theta_{t+k}}(E(N_{t+k}|\theta_{t+k}, N^{(t)})) = E(\theta_{t+k}|N^{(t)}). \tag{2.25}$$

Furthermore, using the state equation we have

$$E(\theta_{t+k}|N^{(t)}) = E(\theta_t|N^{(t)}) \prod_{n=t+1}^{t+k} \frac{E(\epsilon_n|N^{(t)})}{\gamma} = E(\theta_t|N^{(t)}) = \frac{\alpha_t}{\beta_t} \tag{2.26}$$

where $E(\epsilon_n|N^{(t)}) = \gamma$ for any n. Therefore, combining (2.25) and (2.26), we can write

$$E(N_{t+k}|N^{(t)}) = E(\theta_{t+k}|N^{(t)}) = \frac{\alpha_t}{\beta_t}. \tag{2.27}$$

Due to the random walk type of structure introduced in (2.15), the above result simply indicates that k-step ahead forecasts given that we have observed the call arrivals up to time t are equal to α_t/β_t.

2.4.1 Long Run Behaviour of the Call Arrival Rate

Long run behaviour of the call arrival rate is of interest to call center practitioners. Under the proposed model we can analyze the long run behaviour of the call arrival rate θ_t via a study of the behaviours of α_t and β_t. In other words, if the call center system has been in operation for a long time, a steady state type behaviour might have been reached for the arrival rate. We note that the long run behaviour of α_t and β_t would determine how the mean and the variance of the arrival rate θ_t would change as functions of γ.

We define the steady-state value of β_t as $\beta = \lim_{t\to\infty} \beta_t$. By taking the limit on both sides of $\beta_t = \gamma\beta_{t-1} + 1$, we obtain

$$\beta = \frac{1}{1-\gamma} \tag{2.28}$$

For the long run behavior of α_t, using $\alpha_t = \gamma\alpha_{t-1} + N_t$ we can show that for a large k

$$\alpha_k = N_k + \gamma N_{k-1} + \cdots + \gamma^{k-1} N_1 + \gamma^k \alpha_0 \tag{2.29}$$

Therefore given (2.28) and (2.29), in the long run $E(\theta_t|N^{(t)}) = \frac{\alpha_t}{\beta_t} = (1-\gamma)\alpha_k$.

That is

$$E(\theta_k|N^{(k)}) = (1-\gamma)N_k + (1-\gamma)\gamma N_{k-1} + \cdots + (1-\gamma)\gamma^{k-1}N_1 + (1-\gamma)\gamma^k\alpha_0. \quad (2.30)$$

We note from (2.30) that the posterior mean of θ_t is an exponentially weighted average of the past observed call arrivals up to time t. Similarly, from (2.24) $E(N_t|N^{t-1}) = \frac{\alpha_{t-1}}{\beta_{t-1}} = (1-\gamma)\alpha_{k-1}$.

$$E(N_k|N^{k-1}) = (1-\gamma)N_{k-1} + (1-\gamma)\gamma N_{k-2} + \cdots + (1-\gamma)\gamma^{k-1}N_1 + (1-\gamma)\gamma^k\alpha_0 \quad (2.31)$$

Therefore, we can also conclude that one step ahead predictive mean of the call center arrivals at time $k-1$ is a weighted average of the past arrivals.

It follows from the above that for $\gamma < 0.5$ both for $E(\theta_k|N^{(k)})$ and $E(N_k|N^{k-1})$, the model assigns a relatively larger weight on the most recent call arrival, whereas for $\gamma > 0.5$ the weight is relatively smaller. Following a similar logic, we can also obtain the variance of $\theta_t|N^{(t)}$ and $N_t|N^{(t-1)}$ in the long run as follows

$$V(\theta_k|N^{(k)}) = (1-\gamma)^2\alpha_k \quad (2.32)$$

and

$$V(N_k|N^{(k-1)}) = \frac{(1-\gamma)}{\gamma}\alpha_{k-1} \quad (2.33)$$

where α_k is defined as in (2.29) for a large k. A quick note about the initial values of

α_0 and β_0 is that if the call center system has been in operation for a long time their affect on both the mean and the variance in the long run becomes negligible.

2.4.2 Bayesian Learning about the Discount Parameter γ

In previous sections, the discount factor γ has been assumed to be known in order for us to obtain a tractable Bayesian analysis. However, we can also treat γ as an unknown quantity and describe our uncertainty about it via a prior distribution, say $p(\gamma)$. Given, $N^{(t)}$, that is, call arrivals up to time t, the likelihood function of γ is given by

$$L(\gamma; N^{(t)}) = \prod_{k=1}^{t} p(N_k | N^{(k-1)}, \gamma), \tag{2.34}$$

where $p(N_k|N^{(k-1)}, \gamma)$ is given (2.24). The posterior distribution of γ can then be obtained via the Bayes' rule as

$$p(\gamma|N^{(t)}) \propto \prod_{k=1}^{t} p(N_k|N^{(k-1)}, \gamma)p(\gamma). \tag{2.35}$$

For any choice of prior $p(\gamma)$ in (2.35) the posterior distribution can not be obtained analytically. However, we can always sample from the posterior distribution of γ using a Markov chain Monte Carlo (MCMC) method such as the Metropolis-Hastings algorithm. Alternatively, a discrete prior can be used for γ over $(0, 1)$. For example, a discrete uniform prior between 0.01 and 0.99 can be a reasonable choice and this will be considered in our examples.

2.4.3 Filtering Distribution and Retrospective Analysis

In the previous sections our focus has been on prediction of call arrivals in the future. For inference on mean arrival rates at different points in time one is more interested in a retrospective type of analysis. In other words, given that we have observed the data $N^{(t)}$ at time t, we will be interested in the distribution of $\theta_{t-k}|N^{(t)}$ for all $k \geq 1$.

We can write

$$p(\theta_{t-k}|N^{(t)}) = \int p(\theta_{t-k}|\theta_{t-k+1}, N^{(t)})p(\theta_{t-k+1}|N^{(t)})d\theta_{t-k+1} \tag{2.36}$$

where $p(\theta_{t-k}|\theta_{t-k+1}, N^{(t)})$ is obtained via the Bayes' rule as

$$\begin{aligned} p(\theta_{t-k}|\theta_{t-k+1}, N^{(t)}) &= \frac{p(\theta_{t-k}|\theta_{t-k+1}, N^{(t-k)})p(N^*|\theta_{t-k}, \theta_{t-k+1}, N^{(t-k)})}{p(N^*|\theta_{t-k+1}, N^{(t-k)})} \\ &= p(\theta_{t-k}|\theta_{t-k+1}, N^{(t-k)}) \end{aligned}$$

with $N^* = (N_{t-k+1}, \cdots, N_t)$. Here, given θ_{t-k+1}, N^* is independent of θ_{t-k}, i.e. $p(N^*|\theta_{t-k}, \theta_{t-k+1}, N^{(t-k)}) = p(N^*|\theta_{t-k+1}, N^{(t-k)})$. Thus, (2.36) reduces to

$$p(\theta_{t-k}|N^{(t)}) = \int p(\theta_{t-k}|\theta_{t-k+1}, N^{(t-k)})p(\theta_{t-k+1}|N^{(t)})d\theta_{t-k+1}. \tag{2.37}$$

We can not obtain (2.37) analytically, but it is possible to evaluate it using Monte Carlo methods. More specifically, we can draw samples from $p(\theta_{t-k}|N^{(t)})$. This requires us to develop an efficient algorithm which would lead us to sample from the

joint density, i.e $p(\theta_1, \cdots, \theta_t|\gamma, N^{(t)})$, and then collect the samples corresponding to $p(\theta_{t-k}|\gamma, N^{(t)})$ for all $k \geq 1$.

Due to the Markovian nature of the state parameters, we can rewrite $p(\theta_1, \cdots, \theta_t|\gamma, N^{(t)})$ as

$$p(\theta_t|\gamma, N^{(t)})p(\theta_{t-1}|\theta_t, \gamma, N^{(t-1)}) \cdots p(\theta_1|\theta_2, \gamma, N^{(1)}). \tag{2.38}$$

We note that $p(\theta_t|\gamma, N^{(t)})$ is available from (2.21) and $p(\theta_{n-1}|\theta_n, \gamma, N^{(n-1)})$ for any n can be obtained as follows

$$p(\theta_{n-1}|\theta_n, \gamma, N^{(n-1)}) \propto p(\theta_n|\theta_{n-1}, \gamma, N^{(n-1)})p(\theta_{n-1}|\gamma, N^{(n-1)}) \tag{2.39}$$

where the first term is available from (2.15) and the second term from (2.19). It would be straightforward to show that $\theta_{n-1}|\theta_n, \gamma, N^{(n-1)} \sim Gamma((1-\gamma)\alpha_{n-1}, \beta_{n-1}+1)$ where $\gamma\theta_n < \theta_{n-1} < \infty$.

Therefore, given (2.38) and the posterior distribution of γ from (2.35), we can sample from $p(\theta_1, \cdots, \theta_t|\gamma, N^{(t)})$ sequentially simulating the individual state parameters as follows:

1. Generate $\gamma^{(i)}$ from $p(\gamma|N^{(t)})$.

2. Using the generated $\gamma^{(i)}$, sample $\theta_t^{(i)}$ from $\theta_t|\gamma, N^{(t)}$.

3. Using the generated $\gamma^{(i)}$, for each $n = t-1, \cdots, 1$ to generate $\theta_n^{(i)}$ from $\theta_n|\theta_{n+1}^{(i)}, \gamma, N^{(n)}$ where $\theta_{n+1}^{(i)}$ is the value generated in the previous step.

If we repeat the above large number of times, then we obtain samples from

$p(\theta_1, \cdots, \theta_t|\gamma, N^{(t)})$ which allows us to obtain a density estimate for $p(\theta_{t-k}|\gamma, N^{(t)})$ for all $k \geq 1$. Further details about the filtering distributions are presented in our numerical example section.

2.5 A Model for Inter-week Call Arrivals

It is possible to extend the model discussed in the previous section to describe arrivals for each day of the week. In so doing, we consider each day of the week separately and assume that the behaviour of a given day is the same from one week to another. In other words we assume that call arrival process for Monday on any week will exhibit similar behaviour to any other Monday in another week. We consider such a behaviour by assuming the weekly call arrivals for any day are exchangeable over time, that is, from one week to another.

Such an extension is helpful in providing one week ahead forecasts for staffing decisions. These forecasts are of interest to call center managers who would like to be able to determine staff schedules in advance for different time intervals in a given day.

Let $N_{t,j,k}$ denote the number of call arrivals during the time interval $(t-1, t]$ on the k^{th} day of the j^{th} week. We will refer to the interval $(t-1, t]$ as the t^{th} within day interval. Also, let $\theta_{t,k}$ be the corresponding arrival rate. We assume that for each

different day of the week the arrival rates follow the Markovian evolution

$$\theta_{t,k} = \frac{\theta_{t-1,k}}{\gamma_k}\epsilon_{t,k} \tag{2.40}$$

where $\epsilon_{t,k}|D^{(j)}_{t-1,k} \sim Beta(\gamma_k\alpha_{t-1,k}, (1-\gamma_k)\alpha_{t-1,k})$ and $D^{(j)}_{t-1,k} = \{\mathbf{N_1^{(t-1,k)}}, \mathbf{N_2^{(t-1,k)}}, \cdots, \mathbf{N_j^{(t-1,k)}}\}$. In other words, $D_j^{(t,k)}$ denotes all call arrivals observed up to time period t for the last j weeks for day k and $\mathbf{N_j^{(t,k)}}$ represents the call arrivals observed up to time period t for the j^{th} week of day k. In the above, γ_k acts like a common discounting term for day k. Thus, there is a different Markov evolution for each day of the week with a different discounting term.

In what follows, we will suppress the index k for notational convenience and focus on a particular day of the week. Thus, we write (2.40) as $\theta_t = \frac{\theta_{t-1}}{\gamma}\epsilon_t$. Given the arrival rate θ_t the distribution for the number of call arrivals for the j^{th} week during the t^{th} interval for a given day is described via the discrete time Poisson model

$$p(N_{t,j}|\theta_t) = \frac{\theta_t^{N_{t,j}} e^{-\theta_t}}{N_{t,j}!} \tag{2.41}$$

where $t = 1, \cdots, T$ and $j = 1, \cdots, J$. Here T is the number if time intervals available in a given day and J is the number of weeks available in the whole data. We assume that given $\Theta^{(T)} = (\theta_1, \cdots, \theta_T)$, N_j^T's are independent of each other, with $N_j^T = (N_{1,j}, \cdots, N_{T,j})$. Thus, $N_j^T; j = 1, \cdots, J$ form an exchangeable sequence of random variables. Our objective is to obtain Bayesian updating for call arrival rates after

observing J weeks each with T time intervals and to use this information to predict the call arrivals in each interval of $(J+1)th$ week.

To develop Bayesian inference for the model, we will process data in a sequential manner. Similar to our development in Section 2, prior to observing any data we assume that

$$(\theta_0|N^{(0)}) \sim Gamma(\alpha_0, \beta_0). \tag{2.42}$$

Given $D_{t-1}^{(j)}$, that is, given arrival counts up to time $t-1$ for the last j weeks, we can show by induction that

$$(\theta_{t-1}|D_{t-1}^{(j)}) \sim Gamma(\alpha_{t-1}, \beta_{t-1}). \tag{2.43}$$

and by combining (2.40) and (2.43), we can obtain

$$(\theta_t|D_{t-1}^{(j)}) \sim Gamma(\gamma\alpha_{t-1}, \gamma\beta_{t-1}). \tag{2.44}$$

Furthermore, the posterior distribution $\theta^k|D_t^{(j)}$ can be obtained using the Bayes' Rule, that is, by combining the likelihood function (2.41) with the prior (2.44), as

$$(\theta_t|D_t^{(j)}) \sim Gamma(\alpha_t, \beta_t) \tag{2.45}$$

where $\alpha_t = \gamma\alpha_{t-1} + \sum_{i=1}^{j} N_{t,i}$ and $\beta_t = \gamma\beta_{t-1} + j$. In other words, the posterior distribution of the call arrival rate at time t given that we have observed the call

arrivals for the last j weeks up to time period t is again a gamma density.

An attractive feature of the extended model is that we can now obtain the probability distribution of the call arrivals at time t for the j^{th} week given observed data from previous $j-1$ weeks. The posterior predictive distribution of $N_{t,j}|D_t^{(j-1)}$ will be

$$N_{t,j}|D_t^{(j-1)} \sim Negbin(\gamma\alpha_t, \gamma\beta_t). \tag{2.46}$$

The predictive distribution (2.46)provides a mechanism to make forecasts for the same day of the future weeks. For example, given call arrivals data for Monday from the previous $j-1$ weeks, we can make predictions for all the time periods for the next Monday. Similarly if we have the call arrivals data for all other days, we can provide predictions for the whole next week. This is a helpful feature for a week-ahead staffing decisions of call center managers. Also, the within day posterior predictive distribution of $N_{t,j}|D_{t-1}^{(j)}$ can be obtained as

$$N^{t,j}|D_{t-1}^{(j)} \sim Negbin(\gamma\alpha_{t-1}, \gamma\beta_{t-1}). \tag{2.47}$$

Similar to our original model proposed, we can assume that the discount factor γ is unknown and learn about it based on arrival counts data on different weeks of the given day. We can define our uncertainty about γ by a prior distribution $p(\gamma)$ over

$(0, 1)$ and obtain the posterior distribution $p(\gamma|D^{(J)}_{(T)})$ via

$$p(\gamma|D^{(J)}_{(T)}) \propto \prod_{j=1}^{J}\prod_{t=1}^{T} p(N_{t,j}|D^{(j)}_{(t-1)}, \gamma)p(\gamma) \tag{2.48}$$

where $p(N_{t,j}|D^{(j)}_{(t-1)}, \gamma)$ for any t is given by (2.47). As before we can either use MCMC methods for drawing posterior samples from the above distribution or alternatively can specify a discrete prior distribution for γ and evaluate posterior accordingly. Once the posterior distribution is available, using (2.46) we can obtain one week ahead predictions.

2.6 Numerical Examples

In this section we use actual data to show the implementation of the proposed models. The available data is on call center arrivals during different intervals of a day from a US commercial bank. Each day consists of 170 time intervals each of which is 5 minutes duration. In a given day, the call center is operational between 7 AM and 9:05 PM. A detailed summary of the data is available in Weinberg et al. (2007).

In our analysis, we initially focus on the first five days of the first week in the data set. In so doing, we assume a discrete uniform prior over $(0.01, 0.99)$ for the discounting term γ and study its behaviour via the posterior distribution. Using the discrete prior, the posterior distribution of γ given $N^{(t)}$ can be obtained via

$$p(\gamma = i|N^{(t)}) = \frac{p(N_1, \cdots, N_t|\gamma = i)p(\gamma = i)}{\sum_{j=0.01}^{0.99} p(N_1, \cdots, N_t|\gamma = j)p(\gamma = j)} \tag{2.49}$$

for $i = 0.01, \cdots, 0.99$.

We investigate the effect of the gamma prior parameters, α_0 and β_0 on the posterior distribution of γ. As shown in Figure 1, posterior distribution of γ is not very sensitive to the initial prior values as long as the prior mean, (α_0/β_0) is within the same order of magnitude as the actual call arrival data.

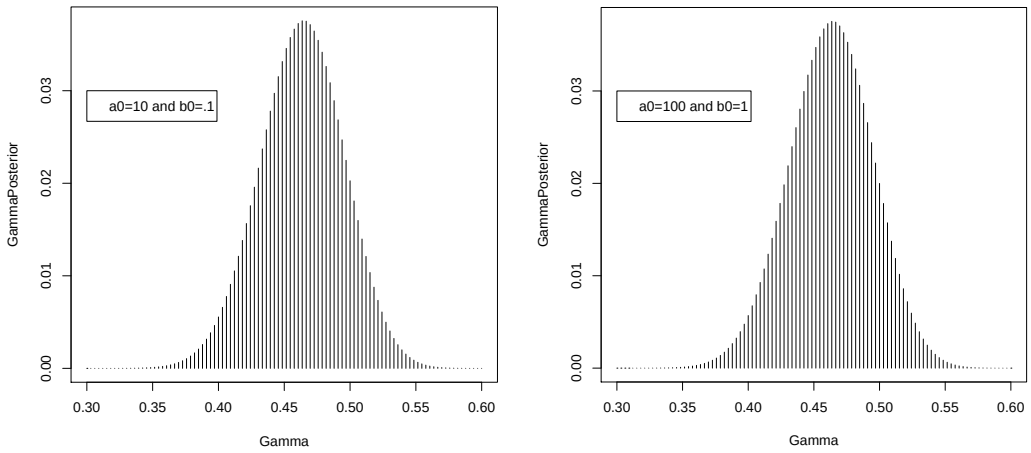

Figure 1: for $\alpha_0 = 10$, $\beta_0 = .1$ (left) and $\alpha_0 = 100$, $\beta_0 = 1$ (right)

In Figure 2 we present the posterior distributions of γ for different days of the week. We note that the discount term γ behaves more or less the same for different days of the week. Table 1 shows the posterior means and the standard deviations of γ's for different days of the week. Again, we note that there are no large differences between the means and the standard deviations for different days.

One step ahead forecasts for call arrivals can be obtained using the conditional posterior predictive distribution given by (2.24) with the posterior distribution of γ

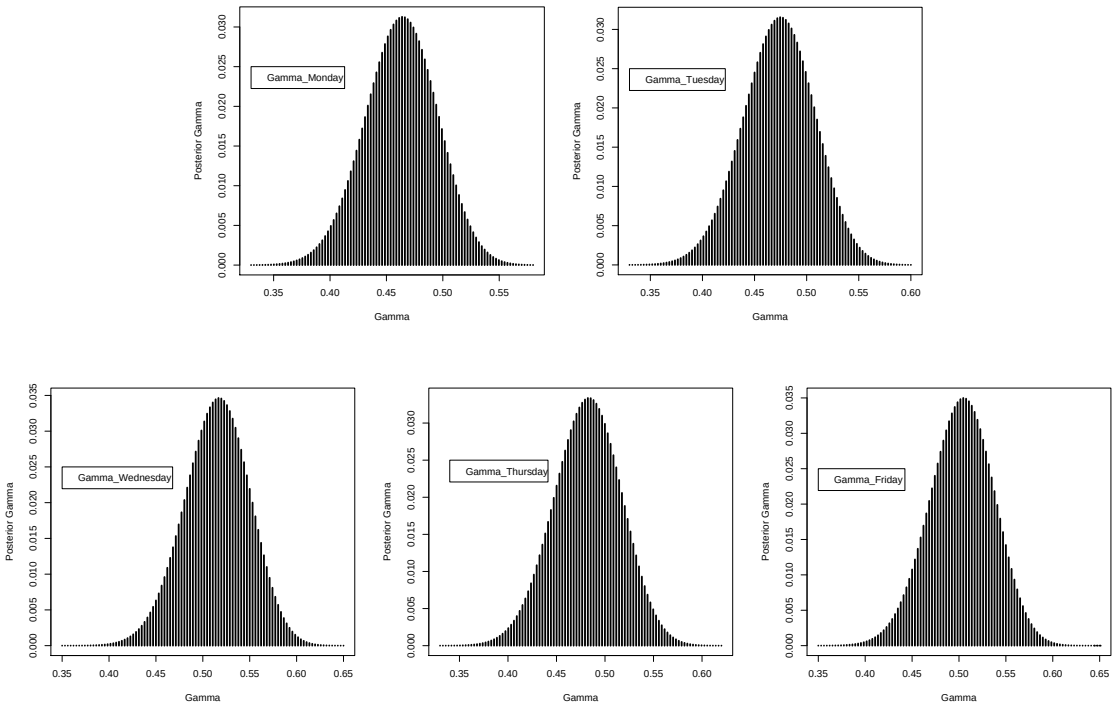

Figure 2: Posterior γ for different days of the week

given by (2.49). More specifically, to obtain $p(N_t|N^{(t-1)})$ for $t \geq 1$, we evaluate

$$p(N_t|N^{(t-1)}) = \sum_{i=0.01}^{0.99} p(N_t|N^{(t-1)}, \gamma = i)p(\gamma = i|N^{(t-1)}) \tag{2.50}$$

where $p(N_t|N^{(t-1)}, \gamma = i)$ is available from (2.24) and $p(\gamma = i|N^{(t-1)})$ is from (2.49). One step ahead posterior predictive distributions $p(N_{141}|N^{(140)})$ and $p(N_{142}|N^{(141)})$, using the first 140 and 141 observations, for Monday are given in Figure 3. Using (2.50) we have obtained one step ahead forecasts for different days of the week. In doing so, we have used the first 140 time periods of each day as historical data and sequentially obtained one step ahead forecasts. Based on the one step ahead forecasts

Days	Means	Std
Monday	.4628	0.03200
Tuesday	.4730	0.03425
Wednesday	.5147	0.03477
Thursday	.4817	0.03478
Friday	.5026	0.03441

Table 1: Posterior means and standard deviations for different days of the week

we have computed mean absolute percentage error for day d as

$$MAPE_d = \sum_{i=141}^{170} \frac{|N_i - \hat{N}|}{30N_i}, \tag{2.51}$$

where d =Monday, $\cdots$, Friday. We have also obtained mean absolute percentage errors for each day using the respective posterior means in 1 as the fixed values of γ. The results are identical as can be seen from the summary in Table 2.

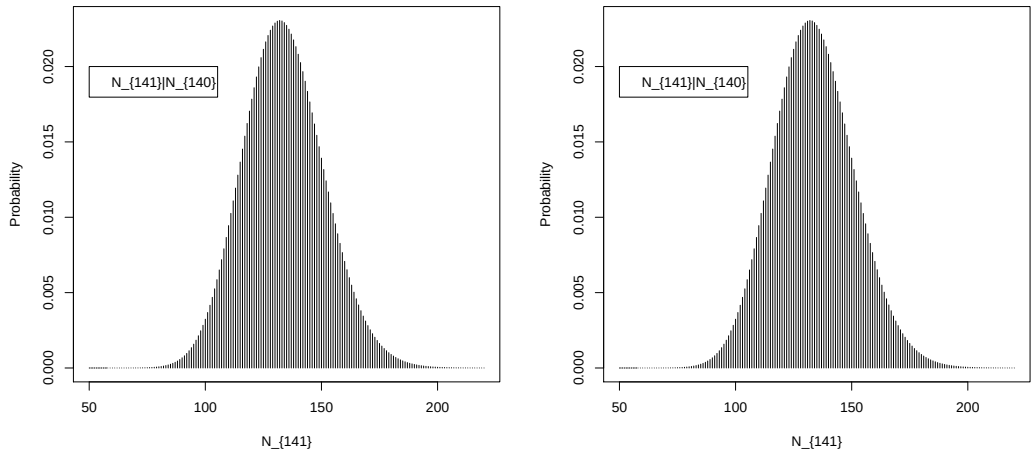

Figure 3: $p(N_{141}|N^{(140)})$ (left) and $p(N_{142}|N^{(141)})$ (right)

One way to infer how well the proposed model fits to the actual call arrival data on a given day is by a comparison of the posterior means, $E(\theta_t|N^{(t)})$ with the actual

Days	MAPE (γ unknown)	MAPE (γ fixed)
Monday	.081	.081
Tuesday	.093	.094
Wednesday	.137	.137
Thursday	.117	.117
Friday	.098	.099

Table 2: One step ahead MAPE for different days

values. Figure 4 shows the posterior means and the call arrivals on the same plot for the whole day on Monday and suggests that the model provides a reasonably good fit to the data.

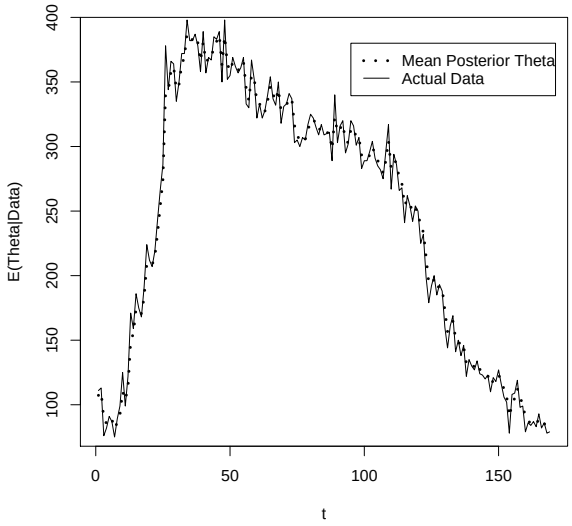

Figure 4: Posterior means and the actual arrivals at different time intervals during Monday.

A comparison of mean arrival arrival rates during different periods of each day can be made retrospectively by looking at the smoothed state means $E(\theta_t|N^{(170)})$'s based on all observed data in a given day. As discussed in section 2.3, the forward filtering backward sampling algorithm was implemented with 1,000 iterations for each

day. Figure 5 shows the behaviour of the filtered state means for different days of the week. We note that, retrospectively, arrival rates on Mondays are mostly higher than

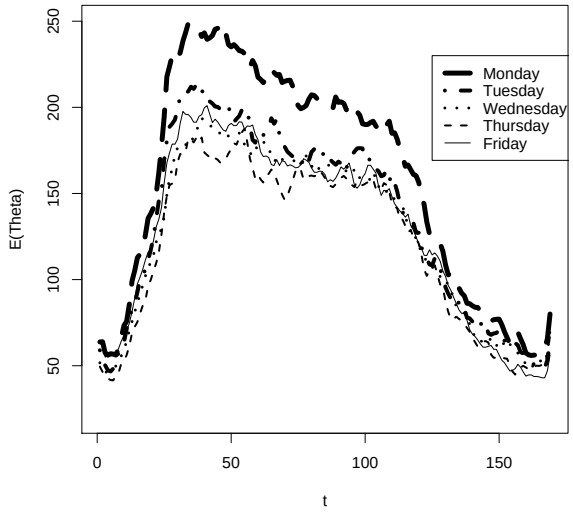

Figure 5: Posterior means $E(\theta_t|N^{(170)})$ versus time t for different days.

those on the other days. It may be of interest to look at the arrival patterns during different time periods in the day. As shown in Figure 6, between 7:00 AM and 12:00 noon, call arrival rates are higher on Monday, but we also note that Friday arrivals are almost as high as Monday in the morning whereas the lowest rate is observed on Thursday.

Similar insights can be obtained from Figure 7 which illustrates the arrival rate patterns between 2:30 and 9:05 PM.

Our analysis so far has focused on comparison of mean arrival rates during different days and intra-day call arrival predictions. As noted before, short term predictions are important for dynamic staffing. An equally important issue is the ability to predict

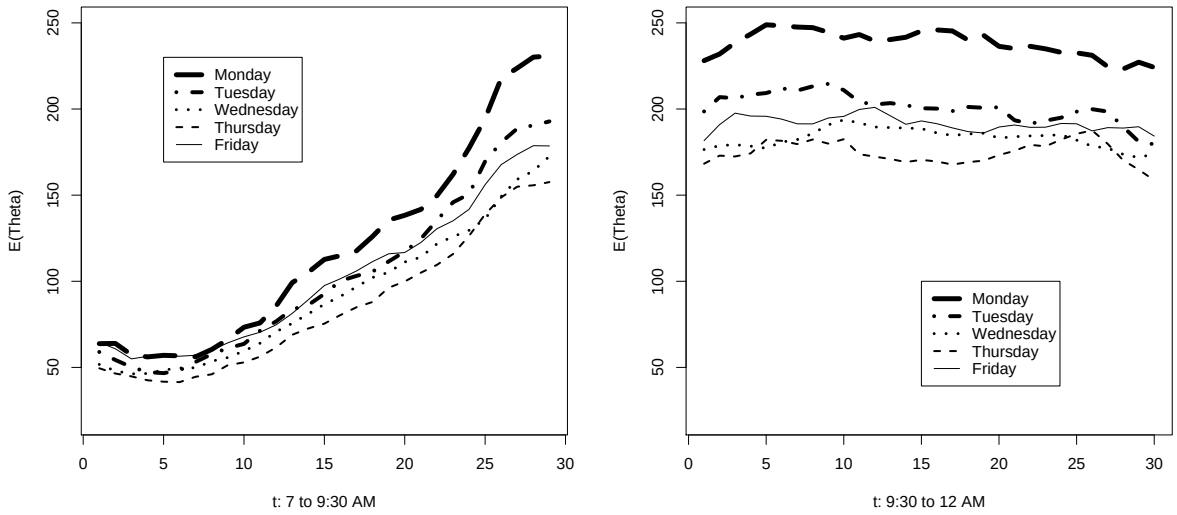

Figure 6: Mean arrival rates during 7:00 to 9:30 AM(left) and 9:30 AM to 12: noon (right)

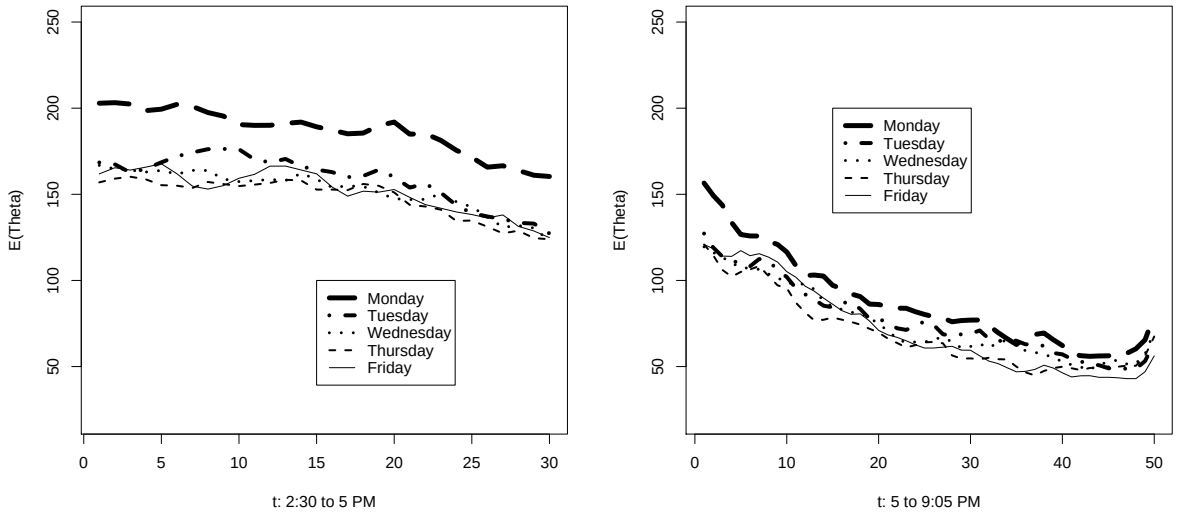

Figure 7: Mean arrival rates during 2:30 to 5:00 PM(left) and 5:00 to 9:05 PM (right)

the arrival pattern in future weeks for long term staffing decisions. As discussed in Section 3, our model enables us to obtain inter-weekly forecasts as well as the intra-

day ones. Using our development in Section 3 and more specifically using (2.46) we can obtain one week ahead predictions for each day of the week. For illustrative purposes we have used the first three weeks of call arrivals in order to predict the arrivals during all time periods of each day of the following week. In Figure 8 we present a comparison of one week ahead predictions for Monday and Tuesday arrivals with the actual call arrivals during the whole 170 time periods. In table 3 we present the mean absolute percentage errors (MAPEs) for each day of the week based on 170 time periods. We note that the daily error rate changes between 7.5 and 8.5 per cent which is comparable our intra-day predictions given in Table 2.

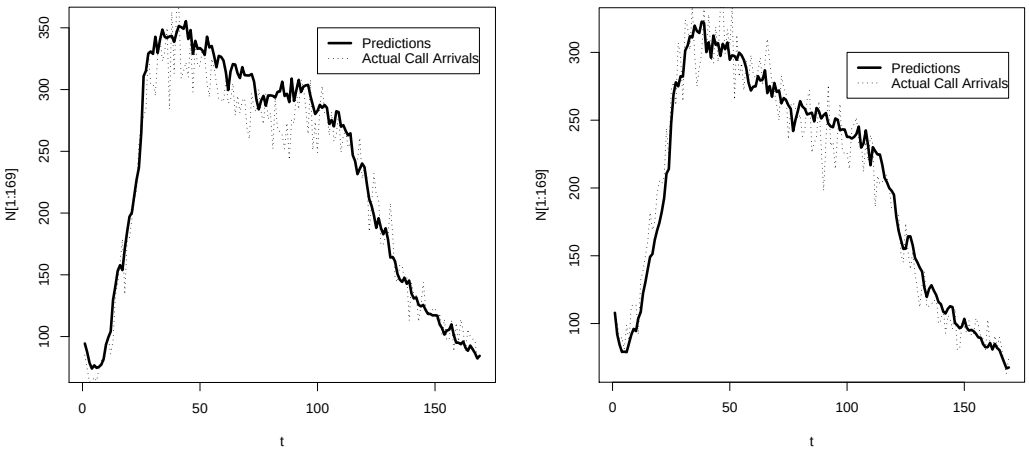

Figure 8: One week ahead forecasts for Monday (left) and Tuesday (right) versus actual call arrivals

2.7 Concluding Remarks

In this article we have developed a Bayesian state space model whose measurement equation follows a discrete time Poisson process and whose rates follow a discrete

Days	MAPE
Monday	0.0756
Tuesday	0.0759
Wednesday	0.074
Thursday	0.080
Friday	0.0855

Table 3: One week ahead MAPEs

time gamma process with beta distributed error terms. We provided closed form updating procedures for the model parameters under certain conditions, developed one step ahead predictions and introduced a forward filtering backward sampling algorithm in order to estimate smoothed arrival rates for different days of the week for inference purposes. Our model also allows us to obtain analytically tractable inter-week forecasts based on the previously observed call arrivals. We applied the proposed model to data from an anonymous U.S. commercial bank. Based on our comparison of the expected posterior means of the arrival rates and the actual arrival counts for each period of a given day we have shown that the proposed model provided a good fit to the data at hand. Furthermore, the model provided us with accurate one-week ahead forecasts.

We believe that modeling the call center arrivals using a Bayasian perspective provides additional insights for call center practitioners. For instance, instead of obtaining just one step ahead point estimates, now one can talk about the distributions of the estimates, the state parameters and their respective distributional properties such as the variance, mode and credibility intervals. Furthermore, the proposed model incorporates managerial insights from call center practitioners via the prior

distribution parameters. From a practical point of view, the fact that state parameter updating and one step ahead call forecasting is analytically tractable makes the proposed model attractive. From a staffing point of view, parameter updating efficiency and adaptability of the proposed model are other features that might be of interest to call center managers.

Our findings led us to believe that further extensions to the proposed model is possible if the discount term γ were to be treated as a common parameter for different days of the week which would introduce a dependence structure between the call arrivals of different days. Therefore, if one can develop a mechanism which can efficiently update the prior parameters of each day based on past data, the proposed model can be further extended. However we expect that this type of approach would lead us to lose the analytical tractability of our model and is currently being investigated as an extension to the problem at hand.

3 Essay 2: Bayesian Analysis of Abandonment in Call Center Operations

3.1 Summary

In this paper we consider modeling abandonment behavior in call centers for different customer profiles. We present several time to event modeling strategies and develop Bayesian inference for posterior and predictive analyses. Two different families 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 and discuss inferential implications for different call center customer profiles.

3.2 Introduction and Overview

Call center operations typically consist of three fundamental processes which can be summarized as 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 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 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 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 which is most probably due to the fact that call center abandonment data only recently became available. In this paper we will attempt to fill this gap by analyzing the abandonment behavior exhibited by different call center customer profiles from a Bayesian point of view.

Some of the most recent work in call center research with abandonment can be summarized as follows. Bacelli et al. (1984) consider a telecommunication network

where a customer leaves the system whenever the waiting time exceeds a random threshold. They describe the system by a $G/G/1+G$ queuing model where both inter-arrivals, service times and time to abandonment follow general distributions, G. 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. Garnett et al. (2002) analyze the design of call centers with impatient customers, where the abandonment is assumed to be exponentially distributed. They provide both a numerically tractable analysis and an asymptotic analysis for large call centers from which they provide rules of thumb for large call center design.

The model that is used most commonly 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 Erlang-A (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. 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. They assume a system in steady state and discuss its existence and uniqueness under certain conditions.

Brown et al. (2005) analyze an anonymous call center from a queuing science perspective by focusing on statistical analysis of three fundamental processes in call center operations; arrival, service and abandonment. They provide non-parametric estimates of the hazard rate for the abandonment process in question and develop an index for patience, which they refer to as the patience index that is calculated as the ratio of the number of customers served to the number that abandoned. The authors develop a solution assuming that time to abandonment is exponentially distributed (They also note that abandonment does not exhibit exponential behavior in their data set). A decision theoretic view is discussed by Mandelbaum and Shimkin (2000) where they propose a model with impatient customers who are assumed to act rationally without observing the queue length. An equilibrium point is obtained via perceived waiting costs and benefits of service based on individual utility. For different forms of the hazard rate function of the abandonment process the authors provide optimal rules for when to abandon for a given type of customer. Similarly, Hassin and Haviv (1995) provide equilibrium strategies for queues with impatient customers, and show

that a unique Nash equilibrium exists for joining or leaving a queue. Another area in call centers with abandonment which has been investigated is the effect of delays on call center operations. 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 (1999b), Whitt (1999c) and Aksin et al. (2008) investigate the issue of delays, informing customers of these delays and their effect on system performance.

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. 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. Time to event (or duration) models such as generalized gamma, exponentiated Weibull, piecewise and mixture densities 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 implemented in analysis of call center abandonment data. The issues which will be investigated in this study and would be of interest to call center practitioners can be summarized as whether

call center customers exhibit monotonic or non-monotonic abandonment behavior, whether diverse customer profiles exhibit different abandonment rate behavior and finally whether customers change their abandonment behavior as they experience announcements or as they wait in the line.

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.

3.3 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. They estimated the hazard via (*number of events during* $(t, t+\delta]$)/(*number at risk at* $t \times \delta$), where δ is defined as the interval length. Also 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.

There is a relationship between the time to abandonment of a customer and time to failure of an item in reliability modeling. The 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 discussed by Singpurwalla (2006) is the model failure rate (or hazard rate) defined as

$$r(t) = \lim_{t \to \infty} \frac{P(t \leq T \leq t + dt | T \geq t)}{dt} = \frac{f(t)}{1 - F(t)}. \tag{3.1}$$

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 $(t, t + dt)$. This provides a natural way of thinking about the abandonment rate of a customer in a virtual queue. In other words, following (3.1), 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 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.

3.3.1 Generalized Gamma and Exponentiated Weibull 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 type of phenomenon, such as reliability, survival and duration models. 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 a 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

$$f(t|\alpha,\gamma,\lambda) = \frac{\gamma\lambda^{\alpha}t^{\alpha\gamma-1}}{\Gamma(\alpha)}exp\{-\lambda t^{\gamma}\}, \tag{3.2}$$

where $t \geq 0$, $\alpha,\gamma,\lambda > 0$, α and γ are shape parameters and λ is the scale parameter. Using a different parametrization the density can also be written as

$$f(t|\alpha,\gamma,\lambda) = \frac{\gamma\lambda^{\alpha\gamma}t^{\alpha\gamma-1}}{\Gamma(\alpha)}exp\{-(\lambda t)^{\gamma}\}, \tag{3.3}$$

where $t \geq 0$ and $\alpha,\gamma,\lambda > 0$. In what follows we will use the parametrization presented in (3.2).

As discussed in Pham and Almahana (1995) and Dadpay et al. (2007), sev-

eral 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. Some properties of the generalized gamma family can be summarized as follows. Its power moment is given by

$$E(T^s|\alpha, \gamma, \lambda) = \{\frac{1}{\lambda}\}^{\frac{s}{\gamma}} \frac{\Gamma(\alpha + s/\gamma)}{\Gamma(\alpha)}, \tag{3.4}$$

where $s > 0$. Therefore, following (3.4) the mean and the variance of the generalized gamma distribution can be obtained as

$$E(T|\alpha, \gamma, \lambda) = \{\frac{1}{\lambda}\}^{\frac{1}{\gamma}} \frac{\Gamma(\alpha + 1/\gamma)}{\Gamma(\alpha)}, \tag{3.5}$$

and

$$V(T|\alpha, \gamma, \lambda) = \frac{(1/\lambda)^{2/\gamma}}{\Gamma(\alpha)} \{\Gamma(\alpha + 2/\gamma) - \frac{(\Gamma(\alpha + 1/\gamma))^2}{\Gamma(\alpha)}\}. \tag{3.6}$$

Another attractive feature of the generalized gamma family is that its hazard rate function is very flexible. One can obtain non-monotonic single peaked or bathtub shaped hazard functions (different shapes of the hazard rate can be seen in Figures 9 and 10). This is desirable in abandonment process modeling, since as discussed in

Brown et al. (2005), the hazard rate for the customers of the call center at hand (that will be analyzed in our numerical example section) exhibits increasing, decreasing, bathtub or constant type of behavior before or after announcements. The flexibility provided by the generalized gamma family makes it a good candidate for modeling the abandonment behavior of different customer profiles. The conditions for different hazard rate behavior can be summarized as follows. Using the parametrization introduced in (3.3), for $\gamma \neq 1$, if $(1 - \alpha\gamma)/[\gamma(\gamma - 1)]$ is a strictly positive constant, then the hazard rate is

- bathtub shaped for $\gamma > 1$,
- inverse bathtub shaped for $0 < \gamma < 1$.

Otherwise, the hazard rate is

- increasing for $\gamma > 1$,
- decreasing for $0 < \gamma < 1$.

The hazard rate for the generalized gamma family is obtained via

$$r(t|\alpha, \gamma, \lambda) = \frac{f(t|\alpha, \gamma, \lambda)}{1 - F(t|\alpha, \gamma, \lambda)}, \tag{3.7}$$

where $F(t|\alpha, \gamma, \lambda)$ is the cumulative distribution of T given by

$$F(t|\alpha, \gamma, \lambda) = \frac{\Gamma_{\lambda t^{\gamma}}(\alpha)}{\Gamma(\alpha)}, \tag{3.8}$$

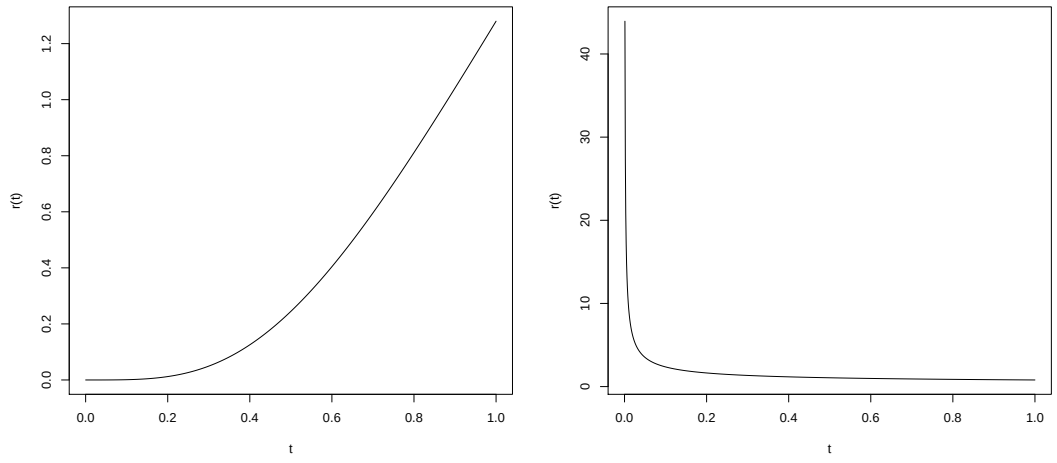

Figure 9: Hazard Rates for Generalized Gamma; $\alpha = 3, \lambda = 2, \gamma = 1.6$ (left) and $\alpha = 0.9, \lambda = 0.5, \gamma = 0.9$ (right)

where

$$\Gamma_{\lambda t^{\gamma}}(\alpha) = \int_0^{\lambda t^{\gamma}} x^{\alpha-1} exp(-x) dx. \tag{3.9}$$

Therefore, combining (3.8) and (3.9) as

$$F(t|\alpha, \gamma, \lambda) = \frac{1}{\Gamma(\alpha)} \int_0^{\lambda t^{\gamma}} x^{\alpha-1} exp(-x) dx, \tag{3.10}$$

and the hazard rate can be obtained as

$$r(t|\alpha, \gamma, \lambda) = \frac{\gamma \lambda^{\alpha} t^{\alpha\gamma-1} exp\{-\lambda t^{\gamma}\}}{\Gamma(\alpha) - \Gamma_{\lambda t^{\gamma}}(\alpha)}. \tag{3.11}$$

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. Next we briefly discuss the lognormal family of models and their respective properties. Let T denote

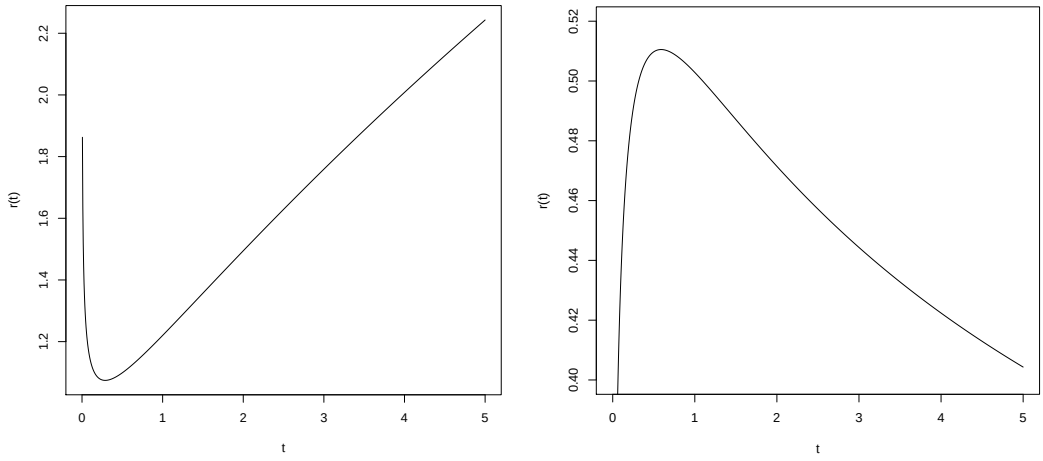

Figure 10: Hazard Rates for Generalized Gamma; $\alpha = 0.5, \lambda = 0.5, \gamma = 1.6$ (left) and $\alpha = 2, \lambda = 1.33, \gamma = 0.66$ (right)

the time to abandonment of customers in a call center. Therefore, following Johnson et al. (1994) the density function of the lognormal random variable is given by

$$f(t|\mu, \sigma) = (t\sqrt{2\pi}\sigma)^{-1} exp\{-\frac{1}{2}\frac{(log(t) - \mu)}{\sigma^2}\},\ t > 0. \tag{3.12}$$

An attractive feature of the lognormal density is that its hazard rate exhibits non-monotonic behavior in t as seen in Figure 11. If the customers exhibit an increasing abandonment rate followed by a decreasing abandonment rate behavior 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 11. The hazard rate of the lognormal density can be

obtained as

$$r(t|\mu,\sigma) = \frac{f(t|\mu,\sigma)}{1-\Phi(log(T),\mu,\sigma)}, \tag{3.13}$$

where $f(t|\mu,\sigma)$ is given in (3.12) and $\Phi(log(T),\mu,\sigma)$ is the cumulative density function of the normal random variable $log(T)$ with parameters μ and σ.

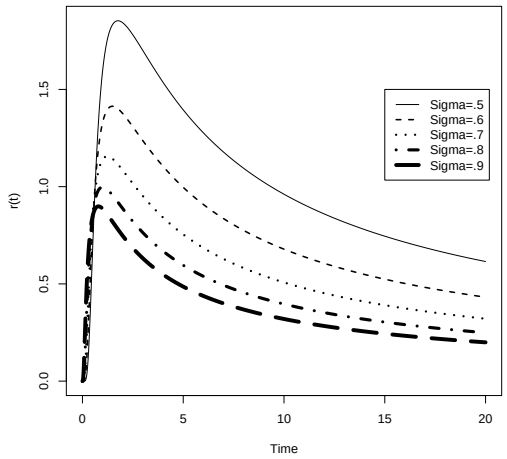

Figure 11: Hazard Rates for the Lognormal Distribution

Another family of distributions worth mentioning is the so called exponentiated Weibull family introduced to the literature by Mudholker and Srivastava (1993). Similar to the family of distributions discussed previously, the hazard rate of the exponentiated Weibull family is very flexible in the sense that it can exhibit increasing, decreasing, bathtub and constant types of behavior (which makes it a family of distributions to model the time to abandonment of customers in call centers worth studying). Further properties of the exponentiated Weibull family is discussed first in Chen (2000) and then more recently in Nassar and Eissa (2003). A Bayesian estimation for the exponentiated Weibull family is discussed in Nassar and Eissa (2004)

using both conjugate priors and the squared error loss form. Comparisons between these estimates and the maximum likelihood estimates is reported.

Let T denote the time to abandonment of customers in a call center, therefore following Mudholker and Srivastava (1993), the density function for T from the exponentiated Weibull family will be

$$f(t|\alpha,\theta,\sigma) = (\frac{\alpha\theta}{\sigma^{\alpha}})[1 - exp(-(t/\sigma)^{\alpha})]^{\theta-1}exp(-(t/\sigma)^{\alpha})(t/\sigma)^{\alpha-1}, \tag{3.14}$$

where $t \geq 0,\ \alpha,\theta,\sigma > 0,\ \alpha$ and θ are shape parameters and σ is the scale parameter.Also the cumulative probability distribution of T is

$$F(t|\alpha,\theta,\sigma) = [1 - exp(-(t/\sigma))^{\alpha}]^{\theta}. \tag{3.15}$$

Therefore, combining (3.14) and (3.15), the hazard rate function can be obtained as

$$r(t|\alpha,\theta,\sigma) = (\frac{\alpha\theta}{\sigma^{\alpha}})[1-exp(-(t/\sigma)^{\alpha})]^{\theta-1}exp(-(t/\sigma)^{\alpha})(t/\sigma)^{\alpha-1}[1-[1-exp(-(t/\sigma)^{\alpha})]^{\theta}]^{-1}. \tag{3.16}$$

Based on different parametrization different distributions and different hazard rates can be obtained; for instance, when $\alpha = 1, \theta = 1$, one can obtain the exponential distribution with constant hazard rate, when $\theta = 1$ the Weibull distribution, when $\alpha < 1,\ \theta < 1$ the hazard rate exhibits a decreasing behavior, when $\alpha > 1,\ \theta > 1$ an

increasing behavior, when $\alpha > 1$, $\theta < 1$ a bathtub behavior and when $\alpha < 1$, $\theta > 1$ an inverse bathtub (unimodal) behavior.

Nassar and Eissa (2003) discuss further statistical properties of the exponentiated Weibull family, such as its r^{th} moments, coefficient of variation, skewness, kurtosis, quantiles, median and mode. Furthermore, Nassar and Eissa (2004) develop a Bayesian estimation mechanism for the case when $\sigma = 1$ via different loss functions such as squared loss and LINEX loss.

Due to the complex three parameter nature of the exponentiated Weibull distribution, Chen (2000) considers a similar family of distributions with only two parameters whose hazard rate can also exhibit a bathtub behavior. In this paper we will refer to this distribution as the two parameter exponentiated Weibull whose density function is given by

$$f(t|\lambda,\beta) = \lambda\beta t^{\beta-1} exp(t^{\beta}) exp\{\lambda(1 - exp(t^{\beta}))\}, \tag{3.17}$$

where $t \geq 0,\ \lambda, \beta > 0$. Its cumulative distribution function is given by

$$F(t|\lambda,\beta) = 1 - exp\{\lambda(1 - exp(t^{\beta}))\}, \tag{3.18}$$

where $t \geq 0,\ \lambda, \beta > 0$ and its hazard rate function is given as

$$r(t|\lambda,\beta) = \lambda\beta t^{\beta-1} exp(t^{\beta}). \tag{3.19}$$

As indicated by (3.19), the hazard rate function of the distribution proposed by

Chen (2000) is fairly simple and has only two parameters, therefore it might be a good candidate for time to abandonment modeling for estimation and inference purposes as opposed to generalized gamma and exponentiated Weibull families of distributions (as distributions who can capture increasing, decreasing and bathtub hazard rate behavior). Different forms of its hazard rate function is shown in Figure 12.

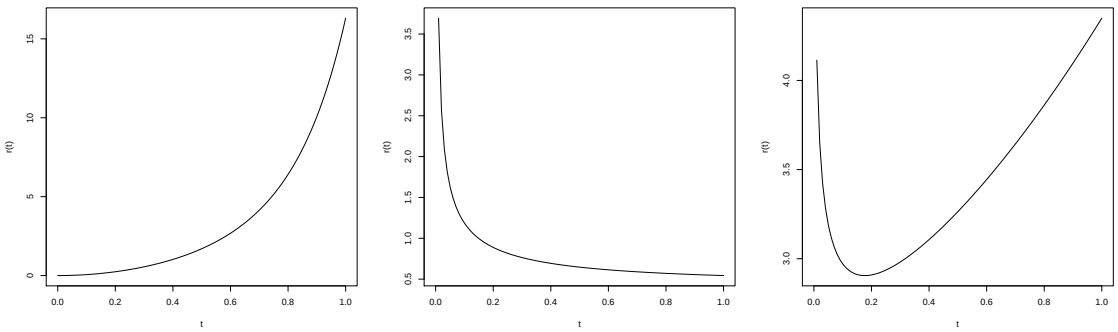

Figure 12: Hazard Rates for Chen's Two Parameter Exponentiated Weibull

3.3.2 Mixture Models

Another modeling approach that one can implement for time to abandonment data for call centers is the 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 of as an alternative to non-parametric models and they are less restrictive as opposed to

standard parametric distributions. Therefore, from an inference point of view using a mixture model can be preferred. As mentioned before, we will attempt to show how mixture models can be used to model time to abandonment data in call centers.

Let T denote the time to abandonment of customers in a call center, therefore following Gilks et al. (1996), a finite mixture model for $f(t)$ can be defined as

$$f(t) = \sum_{i=1}^{k} p_i f(t|\theta_i), \tag{3.20}$$

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, the components of the mixture do not necessarily have 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. Next, we discuss a mixture model with two components and show how its hazard rate is obtained.

Suppose that $k = 2$ in (3.20), therefore the mixture density can be written as

$$f(t) = pf(t|\theta_1) + (1-p)f(t|\theta_2), \tag{3.21}$$

and the mixture cumulative probability distribution is

$$F(t) = pF(t|\theta_1) + (1-p)F(t|\theta_2). \tag{3.22}$$

Therefore, the mixture hazard rate can be obtained as

$$r(t) = w(t)r(t|\theta_1) + (1-w(t))r(t|\theta_2), \tag{3.23}$$

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

$$w(t) = \frac{pF(t|\theta_1)}{pF(t|\theta_1) + (1-p)F(t|\theta_2)}. \tag{3.24}$$

Therefore, (3.24) also indicates that

$$min\{r(t|\theta_1), r(t|\theta_2)\} \leq w(t) \leq max\{r(t|\theta_1), r(t|\theta_2)\}. \tag{3.25}$$

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 (details of the Gibbs sampling and the data augmentation method can found in the sequel).

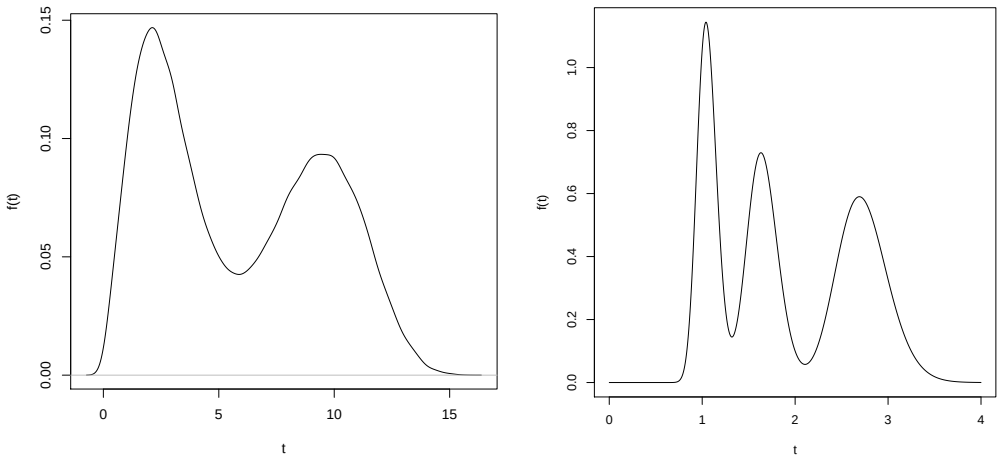

Figure 13: Mixture Density Plots with Weibull Components $k = 2$ (left) and Lognormal Components $k = 3$ (right)

In terms of modeling the time to abandonment in call centers using mixture models, we will attempt to model T as a whole with k mixtures. In other words regardless of when the announcements are made, we will assume that there are k different underlying components of T. 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 13 (with equal mixing probabilities). As evidence of a mixture structure, a similar behavior is observed in our time to abandonment data as seen in Figure 14.

3.3.3 A Piecewise Time to Abandonment Model

After preliminary analysis of our call center data and based on the non-parametric hazard rate estimates discussed in Brown et al. (2005), we have observed that upon joining the queue abandonment rate is increasing, followed by a bathtub type of behavior and a fairly constant behavior. For regular customers a similar behavior can be observed in Figure 15 where the y-axis represents the fraction of number of abandonment's and the x-axis the time intervals of 5 seconds in length (for January, February and March). Here it can be observed that after the second interval (exactly after 60 seconds which coincides with the first announcement time) there is a huge drop in the abandonment rate. Furthermore, the same abandonment behavior among regular customers is observed over a period of three months. Therefore, if one can model the abandonment behavior for a given month it can also represent the bahavior for the subsequent months.

As shown in Figure 15, the abandonment rate can exhibit different behavior before or after announcements. Therefore a piecewise abandonment rate model with switching points as announcements can be considered. Let T_i denote the time to abandonment of a type i customer in a call center (for $i = 1, \cdots, n$; and n is the number of different types of customers), for notational convenience we will suppress the index i, i.e. the type of customer. Therefore a piecewise time to abandonment

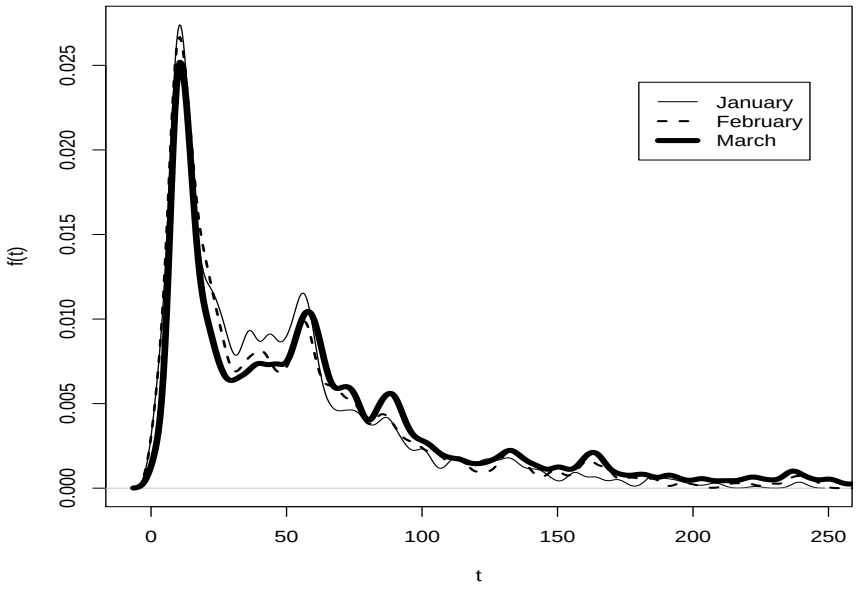

Figure 14: Time to Abandonment Densities for January, February, March of Regular Customers

density with switching points as announcements is defined as

$$A(t|\boldsymbol{\Theta}) = \begin{cases} Cf_0(t|\boldsymbol{\theta_0}), & \text{for} \quad 0 < t \leq \tau_1 \\ Cf_1(t|\boldsymbol{\theta_1}), & \text{for} \quad \tau_1 < t \leq \tau_2 \\ \qquad\qquad ... & \\ Cf_k(t|\boldsymbol{\theta_k}), & \text{for} \quad \tau_k < t < \infty \end{cases} \tag{3.26}$$

where

$$C\int_0^{\tau_1} f_0(t|\boldsymbol{\theta_0}) + \cdots + C\int_{\tau_k}^{\infty} f_k(t|\boldsymbol{\theta_k}) = 1, \tag{3.27}$$

i.e.

$$C = \frac{1}{\int_0^{\tau_1} f_0(t|\boldsymbol{\theta_0}) + \cdots + \int_{\tau_k}^{\infty} f_k(t|\boldsymbol{\theta_k})}, \tag{3.28}$$

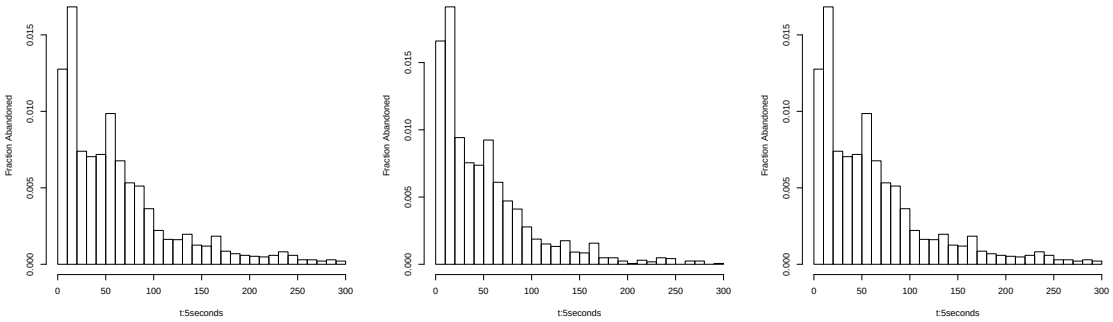

Figure 15: Histograms for Regular Customers in January, February, March

and $\boldsymbol{\Theta} = \{\boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta_k}, \tau_1, \cdots, \tau_k\}$, τ_i represents the time of the i^{th} announcement (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 corresponding vector of paramaters (for $i = 1, \cdots, k$).

We believe that different types of customer behavior before or after announcements can be captured in the above model where lognormal, Weibull or two parameter exponentiated Weibull models can be used as density functions. Density plots for regular customers for the first three months of our call center data are shown in Figure 14.

3.4 Bayesian Inference

As discussed previously, different parametrizations of the generalized gamma, exponentiated Weibull, the mixture and the piecewise models can capture the monotonic or non-monotonic abandonment rate behavior typically observed in call center operations. From a Bayesian point of view the joint posterior distribution of all the

parameters, that is $p(\boldsymbol{\theta}|D)$ can not be obtained analytically except for the lognormal case and thus requires development of Markov chain Monte Carlo methods. In the sequel we discuss posterior and predictive Bayesian analyses of the generalized gamma, lognormal, two parameter exponentiated Weibull, piecewise and mixture models with Weibull and lognormal components. In the sequel, we use the following notation for all the models. Let t_i for $i = 1, \cdots, n$ represent n abandonment times of a given customer profile and let $D = \{t_1, \cdots, t_n\}$.

3.4.1 Generalized Gamma and Exponentiated Weibull Inference

For the generalized Gamma parametrization introduced in (3.2) we have assumed that a priori, $\gamma \sim G(a, b)$, $\alpha \sim G(c, d)$ and $\lambda \sim G(e, f)$, where γ, α and λ are independent.

For the generalized gamma model there is no well known conjugate result. Therefore in order to obtain the joint posterior distributions of the model parameters we have used the Gibbs sampler along with the random walk Metropolis-Hasting algorithm whose proposal density is multivariate normal. Using (3.2) and the gamma prior for λ, the full conditional for the λ parameter can be written as follows

$$p(\lambda|\alpha, \gamma, D) \propto \lambda^{\alpha n+e-1} e^{-\lambda(f+\sum_{i=1}^{n} t_i^{\gamma})}, \tag{3.29}$$

that is $(\lambda|\alpha, \gamma, D) \sim G(\alpha n + e, f + \sum_i^n t_i^{\gamma})$. Also let $\boldsymbol{\theta}$ be a vector of the rest of the parameters ($\boldsymbol{\theta} = \{\gamma, \alpha\}$), then following Chib and Greenberg (1995) the steps in the Metropolis-Hastings algorithm and the Gibbs sampler using (3.29) 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 $G(\alpha^{(t)}n + e, f + \sum_{i=1}^{n} t_i^{\gamma^{(t)}})$.

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

$$\alpha(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta}^*) = min\left\{1, \frac{\pi(\boldsymbol{\theta^*}|\lambda^{(t)})q(\boldsymbol{\theta^{(t)}}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta^{(t)}}|\lambda^{(t)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta^{(t)}})}\right\}. \tag{3.30}$$

In (3.30), $q(.|.)$ 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 $p(\boldsymbol{\theta}, \lambda|D)$.

Next we discuss the special case of the generalized gamma family as $\alpha \to \infty$, that is the lognormal model. Following the model introduced in (3.12), let $y_i = log(t_i)$ for $i = 1, \cdots, n$ represent the log of n abandonment times of a given customer profile. Let ϕ represent the precision and let $1/\sigma^2 = \phi$. 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 using the following well known normal full

conditionals. The full conditional for μ can be written as

$$(\mu|\phi, D) \sim N(m_1, C_1), \tag{3.31}$$

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

$$(\phi|\mu, D) \sim G(a_1, b_1), \tag{3.32}$$

where $a_1 = a_0 + n/2$ and $b_1 = b_0 + \sum_i^n [(y_i - \mu)^2/2]$. Therefore, given (3.31) and (3.32) 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 $p(\mu, \phi|D)$.

For the two parameter exponentiated Weibull model introduced in (3.17), we have asssumed that a priori, $\lambda \sim LN(a, b)$ and $\beta \sim G(c, d)$ where λ and β are independent. Let $\boldsymbol{\theta}$ be the vector of parameters ($\boldsymbol{\theta} = \{\lambda, \beta\}$). There are no known full conditionals, therefore we can use the random walk Metropolis-Hastings algorithm as discussed

previously for the generalized gamma case (without the full conditional generation step for λ) in order to generate sample from $\boldsymbol{\theta}|D$ for the two parameter exponentiated Weibull model.

3.4.2 Piecewise Inference

Similarly posterior samples for the piecewise abandonment rate models introduced in (3.26) 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). We have considered three families of distributions as the components of the piecewise models; the Weibull, the lognormal and the two-parameter exponentiated Weibull distributions with a switching point coinciding with the first announcement (i.e. $\tau_0 = 60$). 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.

3.4.3 Mixture Inference

The last model for which Bayesian inference will be discussed is the mixture model introduced in (3.20). In the sequel we consider two different distributions for the components of the mixture, the lognormal and the Weibull distributions and discuss the implementation of their Markov chain Monte Carlo estimation.

Following Diebolt and Robert (1994) a data augmentation method for Gibbs sampling can be developed for both Weibull and lognormal k component mixtures as

follows. 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 and $\boldsymbol{z_i} = \{z_{i1}, \cdots, z_{ik}\}$ for $i = 1, \cdots, n$. Given $\boldsymbol{z}$, the density of the i^{th} observation from (3.20) can be rewritten as

$$f(t_i|\boldsymbol{\Theta}, \boldsymbol{z}) = \prod_{j=1}^{k} p_j^{z_{ij}} f(t_i|\boldsymbol{\theta_j})^{z_{ij}}, \text{ for } i = 1, \cdots, n. \tag{3.33}$$

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

$$L(\boldsymbol{\Theta}; D, \boldsymbol{z}) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_j^{z_{ij}} f(t_i|\boldsymbol{\theta_j})^{z_{ij}}. \tag{3.34}$$

Furthermore we assume that the mixing weights, $(p_1, \cdots, p_k)$ follow a Dirichlet prior with parameters $(\alpha_1, \cdots, \alpha_k)$ and the individual component priors follow the same format discussed in the lognormal inference section and the Weibull inference from the sequel. Therefore the Gibbs sampler algorithm with the data augmentation method for both cases 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 $\boldsymbol{z}^{(\boldsymbol{m-1})}$ from $p(\boldsymbol{z}|D, \boldsymbol{\Theta}^{(\boldsymbol{m-1})})$ and $\boldsymbol{\Theta}^{(\boldsymbol{m})}$ from $p(\boldsymbol{\Theta}|D, \boldsymbol{z}^{(\boldsymbol{m-1})})$.

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

$$p(\boldsymbol{z_i}|D, \boldsymbol{\Theta}) \sim Multinomial(\pi_{i1}, \cdots, \pi_{ik}), \tag{3.35}$$

where

$$\pi_{ij} = \frac{p_j f(t_i|\boldsymbol{\theta_j})}{\sum_{l=1}^{k} p_l f(t_i|\boldsymbol{\theta_l})}, \tag{3.36}$$

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

Furthermore using the conditional independence property of the model parameters we can write, $p(\boldsymbol{\Theta}|D, \boldsymbol{z}) = p(p_1, \cdots, p_k|\boldsymbol{z}, D)p(\boldsymbol{\theta_1}|\boldsymbol{z}, D) \cdots p(\boldsymbol{\theta_k}|\boldsymbol{z}, D)$ where

$$p(p_1, \cdots, p_k|\boldsymbol{z}, D) \sim D(\alpha_1 + \sum_{i=1}^{n} z_{i1}, \cdots, \alpha_k + \sum_{i=1}^{n} z_{ik}), \tag{3.37}$$

i.e. a Dirichlet posterior for the mixing weights. In addition $p(\boldsymbol{\theta_1}|\boldsymbol{z}, D), \cdots, p(\boldsymbol{\theta_k}|\boldsymbol{z}, D)$ can be obtained using the sampling method discussed previously for the lognormal models, for the Weibull sampling one can implement the following.

Let T follow a Weibull density with a scale parameter λ and a shape parameter γ. Let t_i for $i = 1, \cdots, n$ represent n abandonment times of a given customer profile.

Also assume that a priori, $\gamma \sim U(a, b)$ and $\lambda \sim G(c, d)$ where γ and λ are independent. In order to obtain the joint posterior distribution of the parameters, $p(\gamma, \lambda|D)$, we can use the Gibbs sampler with the Metropolis-Hastings algorithm (similar to the generalized gamma case). The full conditional for λ will be available as follows

$$p(\lambda|\gamma, D) \propto \lambda^{n+c-1} e^{-\lambda(d+\sum_{i=1}^{n} t_i^{\gamma})}, \tag{3.38}$$

that is $(\lambda|\gamma, D) \sim G(n + c, d + \sum_i^n t_i^{\gamma})$. However the full conditional distribution for γ is not a well known distribution. Therefore we can use the Metropolis-Hastings algorithm as in the case of the generalized gamma model to generate samples from $p(\gamma|\lambda, D)$. Then proceed with the Gibbs sampler as follows

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

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

In our numerical example section we discuss issues such as implementation of the previously introduced estimation methods, convergence and higher order component estimation for the mixture models using real call center abandonment data for different customer profiles.

3.4.4 Predictive Inference

Once the samples from the joint posterior densities are obtained using the Bayesian inference methods presented previously, it is possible to obtain the predictive posterior densities and predictive failure rates for different customer profiles. In doing so we can compare different abandonment rate behavior in call centers implied by different models and assess the model fit to data. The predictive posterior density can be obtained as

$$p(t|D) = \frac{1}{M}\sum_{i=1}^{M} f(t|\boldsymbol{\theta}^{(M)})p(\boldsymbol{\theta}^{(M)}|D). \tag{3.39}$$

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 M set of samples generated from any joint posterior $p(\boldsymbol{\theta}^{(M)}|D)$ the predictive abandonment rate can be obtained as

$$r(t|D) = \frac{\frac{1}{M}\sum_{i=1}^{M} f(t|\boldsymbol{\theta}^{(M)})p(\boldsymbol{\theta}^{(M)}|D)}{1 - \frac{1}{M}\sum_{i=1}^{M} F(t|\boldsymbol{\theta}^{(M)})p(\boldsymbol{\theta}^{(M)}|D)}. \tag{3.40}$$

In our numerical example we will obtain (3.40) for different customer profiles implied by different models which we believe will provide additional managerial insights regarding the abandonment bahavior of call center customers. Next, we discuss how the model comparison will be carried out.

3.4.5 Model Comparison Using Sampling Based Methods

As mentioned previously, closed form results for posterior parameters of the proposed models does not seem to be a possibility given the structure of the problem at hand. Therefore, for model comparison and selection purposes we will be using sampling based methods, more specifically we will implement the selection criteria introduced by Gelfand (1996). In what follows, we summarize the approach discussed by Gelfand (1996).

Let T, as before, represent the time to abandonment of customers in a call center and D represent a set of observations,$\{t_i; i = 1, \cdots, n\}$, of size n from T. Therefore the cross validation predictive density for any i is defined as

$$f(t_i|D_{(-i)}), \tag{3.41}$$

where $i = 1, \cdots, n$ and $D_{(-i)}$ represents the whole set of D except for t_i. In other words, (3.41) represents what values of t_i are likely when the model is constructed using the whole set of observations except t_i. Typically (3.41) can be computed as follows,

$$f(t_i|D_{(-i)}) = \int f(t_i|\boldsymbol{\theta}, D_{(-i)}) f(\boldsymbol{\theta}|D_{(-i)}) d\boldsymbol{\theta}, \tag{3.42}$$

where $\boldsymbol{\theta}$ is the set of parameters from any proposal time to abandonment density as introduced previously. Since the observations are conditionally independent given $\boldsymbol{\theta}$, it can be shown that $f(t_i|\boldsymbol{\theta}, D_{(-i)}) = f(t_i|\boldsymbol{\theta})$. Therefore, using (3.42) the pseudo-

Bayes factor (PBF) is defined as

$$\prod_{i=1}^{n} \frac{f(t_i|D_{(-i)}, M_1)}{f(t_i|D_{(-i)}, M_2)}, \tag{3.43}$$

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 (3.43) 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

$$\hat{f}(t_i|D_{(-i)}) = \frac{1}{\frac{1}{M}\sum_{j=1}^{M} \frac{1}{f(t_i|D_{(-i)},\hat{\boldsymbol{\theta}})}}, \tag{3.44}$$

where M is the number of samples generated and $\hat{\boldsymbol{\theta}}$ is the vector of the generated parameter samples. Since given $\boldsymbol{\theta}$, t_i's are independent, $f(t_i|D_{(-i)}, \hat{\boldsymbol{\theta}}) = f(t_i|\hat{\boldsymbol{\theta}})$ can

be used in (3.44). Finally, once the cross validation predictive densities are estimated using (3.44), coupled with the pseudo-Bayes factor as introduced in (3.43), 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 will assess their fit to abandonment data in February and March.

3.5 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, provide additional managerial insights and summarize issues such as model fit, predictive abandonment rate behavior and Markov chain Monte Carlo convergence.

3.5.1 Case 1: Regular Customers

First we discuss the model fit issue using the sampling based method introduced in (3.42). Let $log(\prod_{i=1}^{n} \hat{f}(t_i|D_{(-i)}, M_j))$ called the global CPO of model j in the log scale and represent how much the model j supports the data. In other words, the higher

the global CPO, the better the fit of the model.

Models	Global CPO
LogNormal	-7859
Expon.Weibull	-7955.11
Generalized Gamma	-7822.48

Table 4: Global CPO in the Log Scale for the Generalized Gamma and Exponentiated Weibull Models

Models	Global CPO
Piecewise Weibull	-7833.59
Piecewise LogNormal	-7818.48
Piecewise Expon.Weibull	-7859.2

Table 5: 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 6: Global CPO in the Log Scale for the Mixture Models

Based on the global CPO values shown in Tables 4, 5 and 6 the best fit is obtained using the three component lognormal mixture model. Also we note here that models such as the exponentiated Weibull and the lognormal do 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 seem to result with 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, let $log(\hat{f}(t_i|D_{(-i)}, M_1)) - log(\hat{f}(t_i|D_{(-i)}, M_2))$ for which positive values represent

how much better the first model is outperforming the second one for the i^{th} sample (and vice versa) and its behavior over time is shown in Figure 16. Mostly this value oscillates around the value of zero, indicating a balance between the two models. However as can be observed from the values on the upper levels of the graph (around the values of 1.5 on the y-axis) there are several occasions where the first model (three component lognormal mixture) extremely outperforms the second one (three component Weibull mixture).

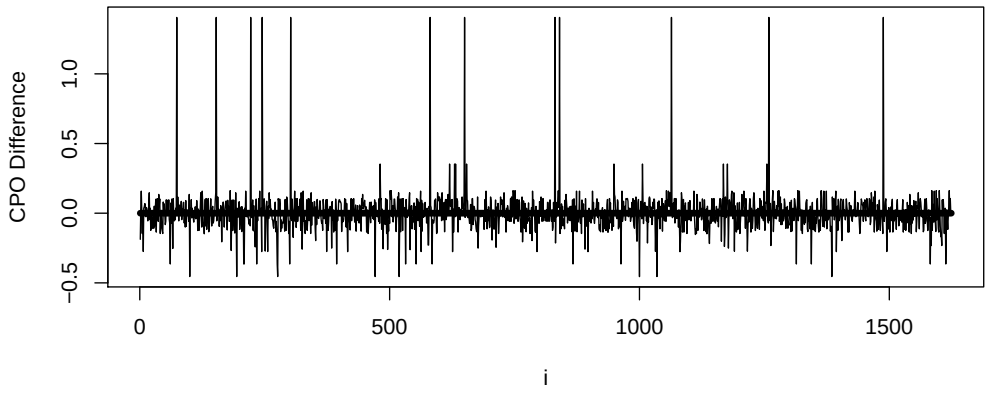

Figure 16: $log(\hat{f}(t_i|D_{(-i)}, M_1)) - log(\hat{f}(t_i|D_{(-i)}, M_2))$ vs Sample i

Another issue that we have investigated was the potential existence of higher order mixtures. For all the higher order mixture models we have encountered convergence issues. Models with more than three components seem to be indistinguishable due to the fact that higher order model parameters were very close to the previously obtained values for lower order mixtures. 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 (3.40). The posterior predictive abandonment rates of some of the models are shown in Figures 17 and 18 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 18). 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. Another interesting finding is the one implied by the piecewise lognormal model. After the 60^{th} second which co-

incides 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 17). 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.

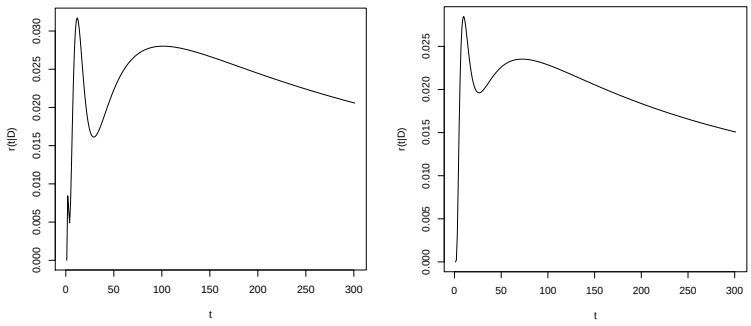

Figure 17: Predictive Abandonment Rates for Regular Customers; Mixture Lognormal (3) (left), Mixture Lognormal (2 (right)

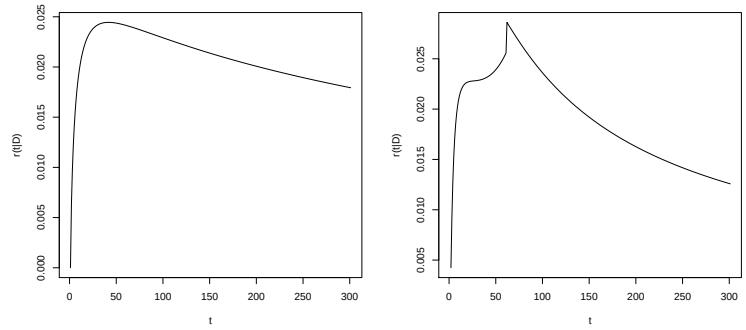

Figure 18: Predictive Abandonment Rates for Regular Customers; Generalized Gamma (left), Piecewise Lognormal (right)

Another 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 (3.2), 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.

Next we have 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 15 for the histogram of February). 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

$$log\{\prod_{i=1}^{n_f} \frac{1}{M} \sum_{j=1}^{M} f(t_i^f|\boldsymbol{\theta_j})\}, \tag{3.45}$$

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 $j = 1, \cdots, M$ are M posterior samples obtained in the previous month. The log-likelihoods for the best fitting models are shown in Table 7.

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 7: One-Month Ahead Fits

Posterior predictive distributions which are obtained as in (3.39) for the top three best fitting models, namely the mixture models are shown in Figure 19 against the actual abandonment data. An interesting property that can be noted is that the log-normal mixture captures a small proportion 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, a behavior referred to as balking in the queuing literature.

Table 8 shows a summary of statistics and Figure 20 shows the respective density plots of the posterior parameters of the best model fit, the lognormal mixture model with three components.

The results presented in this section are based on the Markov chain Monte Carlo sampling techniques discussed in the previous section. Therefore we need to discuss issues of convergence, in doing so we present the results obtained using the generalized

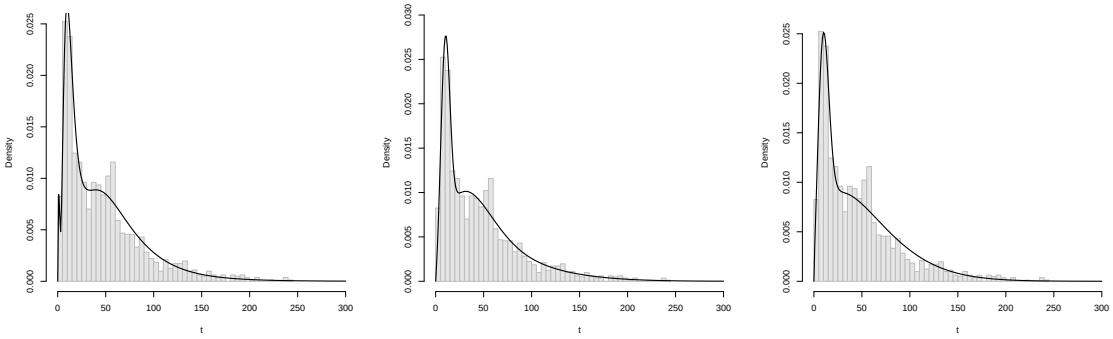

Figure 19: Posterior Predictive Fits vs. Actual Data

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	π_1	π_2	π_3	
Mean	0.0180	0.3990	0.5830	
St.Deviation	0.0040	0.0154	0.0156	

Table 8: Posterior Parameters Summary Statistics for the Lognormal Mixture Model with Three Components

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. We have used the parametrization as presented in (3.3) where $\lambda = e^{\theta}$. As shown in Figure 21, all trace plots exhibit a fairly constant behavior around a specific value which is the first evidence of convergence.

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.

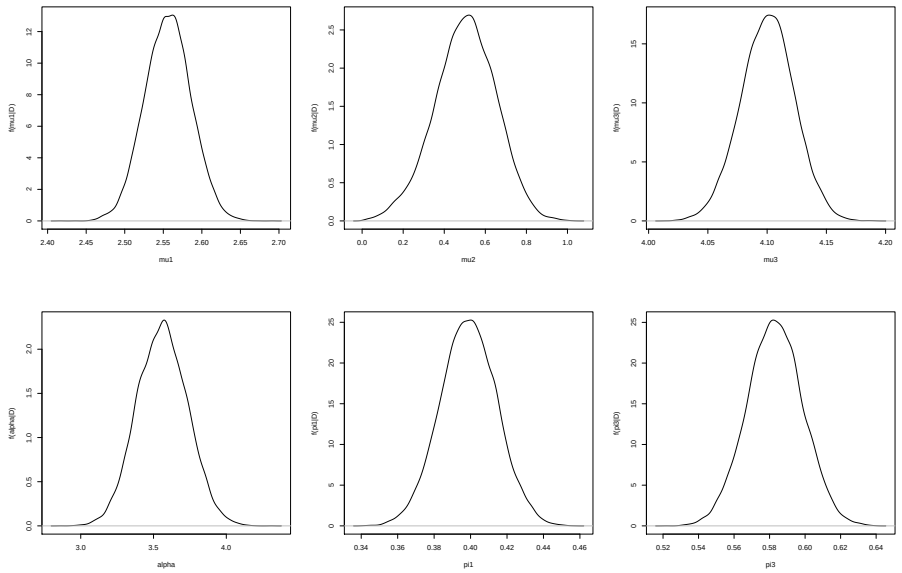

Figure 20: Posterior Parameters Density Plots for the Lognormal Mixture Model with Three Components

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 22 where the shrink factor approaches 1 as the number of iterations increases.

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.

3.5.2 Case 2: Stock Exchange Customers

Another customer profile for which we have investigated abandonment behavior is the stock exchange customers. As shown in Figure 23, stock exchange customers exhibit a different time to abandonment behavior as opposed to the regular customers.

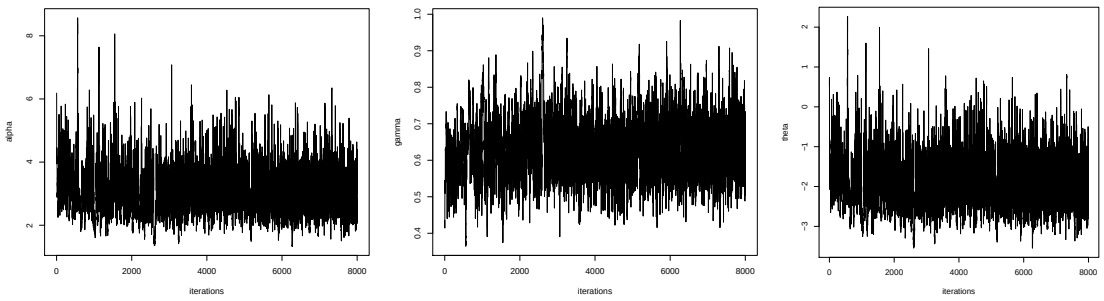

Figure 21: Trace Plots of α (left), γ (middle) and θ (right)

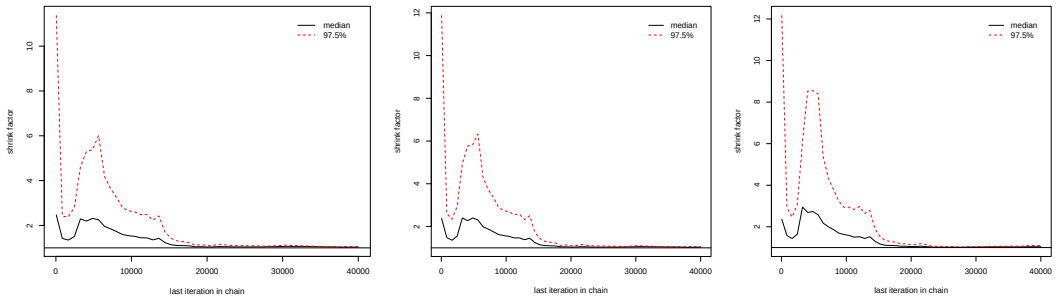

Figure 22: Brooks and Gelman Plots

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 (3.20), were very close to 0 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 (3.43). 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 the stock exchange service is time sensitive and does 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.

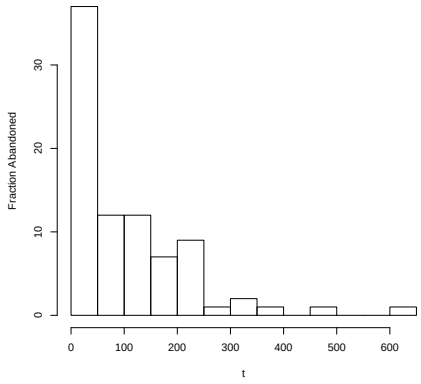

Figure 23: Histogram for Stock Exchange Customers in January

3.6 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 families 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 aban-

donment 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.

4 Essay 3: Bayesian Modeling of Queues in Call Centers

4.1 Summary

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 uncertainy about system primitives probabilistically. We consider models with both patient and impatient customers and discuss their further extensions. We develop details of Bayesian inference for queues with abandonment such as M/M/s+M models (also referred to as the Erlang-A) and present relevant operating characteristics. We illustrate the implementation of the Bayesian models using actual arrival, service, and abandonment data from call centers.

4.2 Introduction

A queuing system can be described as a system where customers arrive to a facility looking for service, wait for service if not served upon arrival, then leave after either by being served or by not being served; the latter is also referred to as abandoning. The first known formal study in queuing is due to the Danish mathematician A.K. Erlang who studied telephone traffic congestion.

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 is mainly defined

by three processes: arrival, service and abandonment. Given that one has a good understanding of these processes, several relevant operating characteristics can be obtained using existing queuing theory (see Gross and Harris (1998) for a detailed coverage of well known queuing results). For instance, 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. One of the relevant operating characteristics, the fraction of abandoned customers in a queuing system, can be considered as a proxy for the quality of service, that is a signal from customer to the call center management. Therefore, being able to make statistical inference on the fraction of abandonment in the long run is of utmost importance to call center managers. There are several operating characteristics such as the 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 random processes with certain statistical properties.

Most of the literature in queuing in call center modeling has been conducted purely from a probabilistic point of view and there is a lack of literature from a statistical inference perspective. Brown et al. (2005) point out the need for such research and refer to it as queuing science rather than queuing theory. We would like to extend on the discussion given in Brown et al. (2005) further by looking at call center operations from a Bayesian perspective. To the best of our knowledge, this would be the first

study in call center queuing science from a Bayesian point of view. In this paper we intend to develop a Bayesian analysis of the the three fundamental processes which in turn will yield the distributions of relevant operating characteristics and their expectations.

The most commonly used queuing model in call center analysis is the M/M/s model. Using Kendall's notation, the first M stands for Markovian arrival and second M for Markovian service process and s represents the number of servers. By Markovian arrival and service processes we mean that the interarrival and the service times are exponential with constant rates λ and μ, respectively. This model assumes that customers do not abandon, in other words 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 times are allowed to form, the customers are assumed to be cleared out 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 further properties of the model introduced in Palm (1957), and defined the models M/M/s+M and M/M/s+G, where in the last argument M stands for exponential distribution for time to abandonment (or patience time) and G for gen-

eral distribution of abandonment. The M/M/s+M model is now 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. Koole and Mandelbaum (2002) present a survey of recent research on classical queuing theory in call centers.

Brown et al. (2005) analyze an anonymous call center from a queuing science perspective by providing statistical analysis for three fundamental processes in call center operations; arrival, service and abandonment. Our aim is to provide additional insights on these fundamental processes as well as on the relevant operating characteristics from a Bayesian perspective. Furthermore, issues that can provide additional managerial insights for call center practitioners such as correlated system primitives, prior assumption for steady state queues, posterior steady state hypothesis, prior and posterior implications of the operating characteristics implied by different models, application of Little's formula in posterior Bayesian analysis and staffing implications will be investigated.

A quick summary of our paper is as follows. In section 3 a brief literature review of call center queuing models with abandonment (or with impatient customers) and closed form operating characteristics for M/M/1, M/M/s, M/M/s+M (also referred to as Erlang-A) queues with extensions will be discussed. In section 4, recent Bayesian queuing models will be introduced. Section 5 will present Bayesian modeling of both

independent and dependent priors for M/M/1, M/M/s and M/M/s+M queues with their extensions. In Section 6, 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 7 under the numerical illustrations.

4.3 Queuing Models in Call Centers

4.3.1 Some General Results for G/G/1 and G/G/s Queues

Prior to the development of more specific models in the upcoming sections, we introduce general results for G/G/1 and G/G/s queues, that is, for single and multiple server queues with general arrival and service processes. Following Gross and Harris (1998), let λ represent the average rate of customer arrivals to a call center and μ the average rate of service of a single server. An important operating characteristic for traffic congestion (also referred to as utilization) is given by

$$\rho = \frac{\lambda}{s\mu}, \tag{4.1}$$

where s is the number of identical servers available in the call center. In order for the system to be in steady state we need that the condition

$$\rho = \frac{\lambda}{s\mu} < 1, \tag{4.2}$$

to be satisfied. Most of the classical literature in queuing theory is based on the assumption in (4.2) which is also referred to as the ergodicity condition. Queuing models which assume that the system is in steady state are called steady state queues, whereas other models where (4.2) is not invoked are referred to as transient queues. The focus of this study will be on the former. A detailed discussion of the latter can be found in Gross and Harris (1998).

There are several operating characteristics, or the so called measures of performance, which can be obtained based on the system primitives, λ and μ. The most common operating characteristic is the total number of customers in the system at time t, denoted by $N(t)$. Also, let $N_q(t)$ represent the number of customers in the queue and $N_s(t)$ the number of customers being served at time t. Thus for any t,

$$N(t) = N_q(t) + N_s(t). \tag{4.3}$$

Assuming that the system is in steady state, let P_n represent the probability of having n customers in the system, that is

$$\lim_{t\to\infty} Pr(N(t) = n|\lambda, \mu) = P(N = n|\lambda, \mu) = P_n. \tag{4.4}$$

Furthermore, let L denote the expected number in the system, that is

$$L = E(N) = \sum_{n=0}^{\infty} nP_n, \tag{4.5}$$

and

$$L_q = E(N_q) = \sum_{n=s+1}^{\infty} (n-s)P_n, \tag{4.6}$$

where L_q denotes the expected number in the queue and s represents the number of servers in the system. Another widely known result in queuing theory is called Little's formula. Let the random variable T represent the time a customer spends in the system and T_q the time spent in the queue. Little's formula implies that

$$L = \lambda W, \tag{4.7}$$

and

$$L_q = \lambda W_q, \tag{4.8}$$

where $W = E(T)$ and $W_q = E(T_q)$. We note that all the above results are conditional on the arrival and service rates, λ and μ respectively. A quick note about (4.7) and (4.8) is that, as noted by Armero and Bayarri (1996), Little's formula may not be used in posterior analysis of Bayesian queues since the relationship may or may not hold a posteriori depending on the prior structure of λ and μ. Its inclusion here is for completeness and its usefulness in Bayesian queuing analysis will be investigated in the sequel. Next we introduce operating characteristics for the most commonly used queuing models in call centers, such as M/M/1, M/M/s and M/M/s+M models.

4.3.2 Operating Characteristics for M/M/1 Queues

We start our discussion of queuing models in call centers with the most basic model, the so called M/M/1 queue where the interarrival times are exponential with constant rate λ and service times are exponential with constant rate μ. In what follows we summarize the most common operating characteristics for the M/M/1 queue. For detailed discussions of the results, see a standard text in queuing theory such as Gross and Harris (1998), where under the steady state assumption the queuing system is treated as a birth and death process and relevant results are obtained using stochastic balance equations.

Given that the system is in steady state, P_n as defined in (4.4) can be obtained as follows

$$(P_n|\lambda,\mu) = P(N = n|\lambda,\mu) = (1-\rho)\rho^n, \tag{4.9}$$

where $\rho < 1$. Note that (4.9) is the geometric distribution with parameter ρ. We can also show that the cumulative probability distribution of T_q is given by

$$P(T_q < t|\lambda,\mu) = 1 - \rho e^{-\mu(1-\rho)t}, \tag{4.10}$$

where $t > 0$. Another operating characteristic of interest is the total time a customer spends in the system, denoted by T. We can obtain the density function of T as

$$p(t|\lambda,\mu) = (\mu-\lambda)e^{-(\mu-\lambda)t}, t > 0. \tag{4.11}$$

The above implies that the total time a customer spends in the system is exponential with rate $\mu - \lambda$. We note here that the results presented in (4.9), (4.10) and (4.11) are all given conditional on the system primitives λ and μ. In classical queuing theory these primitives are assumed to be known and calculations are carried out accordingly. However, as will be presented in the sequel, in the Bayesian paradigm λ and μ will be treated as unknown quantities. We will define our uncertainty about λ and μ via probability distributions, and update our uncertainty as we learn about the queuing system. Furthermore, conditional expectations in terms of λ and μ such as $E(N|\lambda, \mu)$ can be calculated using classical results. Given the probabilistic structure on λ and μ, what would be if interest are unconditional measures such as $E(N)$ given data. These issues will be addressed in the Bayesian queuing models section in detail. Next we introduce the operating characteristics of the M/M/s queues.

4.3.3 Operating Characteristics for M/M/s Queues

The most commonly used model in call center operations is the M/M/s queuing model (also referred to as the Erlang-C). Given that the system is in steady state, for an M/M/s queue (P_n) as defined in (4.4) can be obtained as

$$(P_n|\lambda, \mu) = \begin{cases} \frac{r^n}{n!}(P_0|\lambda, \mu) & \text{for} \quad 1 \leq n < s \\ \\ \frac{r^n}{s^{n-s}s!}(P_0|\lambda, \mu) & \text{for} \quad n \geq s, \end{cases} \tag{4.12}$$

where $\rho = \frac{\lambda}{s\mu} < 1$, $r = \frac{\lambda}{\mu}$ (in general r is called the offered load for queues with patient customers and is an important measure used in call center staffing problems). Also $(P_0|\lambda, \mu)$ is given by

$$(P_0|\lambda, \mu) = (\sum_{n=0}^{s-1} \frac{r^n}{n!} + \frac{r^s}{(1-\rho)s!})^{-1}. \tag{4.13}$$

Similarly, the cumulative density function for $(T_q|\lambda, \mu)$, waiting time in the queue can be obtained as

$$P(T_q \leq t|\lambda, \mu) = 1 - \frac{r^s P_0}{c!(1-\rho)} e^{-(s\mu-\lambda)t}. \tag{4.14}$$

The distribution of total waiting time in the system is a mixture of two exponentials and is given by

$$P(T \leq t|\lambda, \mu) = \frac{s(1-\rho) - Pr(T_q = 0)}{s(1-\rho) - 1}(1 - e^{-\mu t}) - \frac{1 - Pr(T_q = 0)}{s(1-\rho) - 1}(1 - e^{-(s\mu-\lambda)t}), \tag{4.15}$$

where $P(T_q = 0|\lambda, \mu)$ is the probability that a customer will be served upon arrival and can be obtained as

$$Pr(T_q = 0|\lambda, \mu) = 1 - \frac{r^s P_0}{c!(1-\rho)}. \tag{4.16}$$

As shown in (4.12) - (4.15), given the system primitives λ and μ all operating characteristics have closed form solutions. If posterior distributions for λ and μ are

available then the posterior distributions for the operating characteristics can be computed as will be shown in the next sections.

4.3.4 Operating Characteristics for M/M/s+M Queues

Consider an M/M/s+M queue system, where interarrival 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. In the call center literature, this system is referred to as the Erlang-A model (where A stands for abandonment) with the abandonment rate θ. To the best of our knowledge, Palm (1957) is the first one who introduced the idea of abandonment by extending the M/M/s model (Erlang-C) with a patience distribution. 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 stability conditions will 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

$$A(x, y) = \frac{xe^y}{y^x}\gamma(x, y), \tag{4.17}$$

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

$$\gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt. \tag{4.18}$$

Given that the system is in steady state, P_n as defined in (4.4), now conditioned on θ as well, is given by

$$(P_n|\lambda, \mu, \theta) = \begin{cases} (P_s|\lambda, \mu, \theta)\frac{s!}{n! r^{s-n}} & \text{for} \quad 0 \leq n \leq s, \\ (P_s|\lambda, \mu, \theta)\frac{(\frac{\lambda}{\theta})^{n-s}}{\prod_{k=1}^{n-s}(\frac{s\mu}{\theta}+k)} & \text{for} \quad n \geq s+1, \end{cases} \tag{4.19}$$

where

$$(P_s|\lambda, \mu, \theta) = \frac{E_{1,s}}{1 + [A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta}) - 1]E_{1,s}}, \tag{4.20}$$

and

$$E_{1,s} = \frac{\frac{r^s}{s!}}{\sum_{j=0}^{s} \frac{r^j}{j!}}. \tag{4.21}$$

The term $E_{1,s}$ is also called the blocking probability in the M/M/s/s queue (Erlang-B). Furthermore, a commonly used operating characteristic in call center operations, the probability that a customer will wait in the queue, $Pr(T_q > 0|\lambda, \mu)$ can be calculated as follows

$$Pr(T_q > 0|\lambda, \mu, \theta) = \sum_{n=s}^{\infty} (P_n|\lambda, \mu, \theta), \tag{4.22}$$

where $(P_n|\lambda, \mu, \theta)$ for $n = s, \cdots, \infty$ can be obtained from (4.19). Garnett et al. (2002) discuss three call center operational regimes for limiting cases of the delay probability

($P(T_q > 0|\lambda, \mu, \theta) \to 0$ is called the quality driven regime and $P(T_q > 0|\lambda, \mu, \theta) \to 1$ the efficieny-driven regime).

In addition, 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

$$Pr_j(Ab|\mu, \theta) = \frac{(j+1)\theta}{s\mu + (j+1)\theta},\ j \geq 0. \tag{4.23}$$

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

$$Pr(Ab|T_q > 0, \lambda, \mu, \theta) = \frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}. \tag{4.24}$$

The operating characteristics, (4.23) and (4.24), are directly related to the perceived quality of the call center operation from a customer perspective. Consequently, the fraction of abandoning can be calculated as the product of (4.24) and (4.22) as

$$Pr(Ab|\lambda, \mu, \theta) = (\frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}) \sum_{n=s}^{\infty} (P_n|\lambda, \mu, \theta). \tag{4.25}$$

Alternatively, the average waiting time given that the customer is not served upon arrival can be obtained as

$$E(T_q|T_q > 0, \lambda, \mu, \theta) = \frac{1}{\theta}[\frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}]. \tag{4.26}$$

Therefore, the expected wait in the queue can be obtained as the product of (4.26)

and (4.22) as follows

$$E(T_q|\lambda,\mu,\theta) = \frac{1}{\theta}[\frac{1}{\rho A(\frac{s\mu}{\theta},\frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}]\sum_{n=s}^{\infty}(P_n|\lambda,\mu,\theta). \tag{4.27}$$

A quick note about (4.26) and (4.27) is that they are results of the Little's formula (4.7) and their applicability in Bayesian analysis will be investigated in the sequel.

Extensions for the Erlang A model are discussed in Bacelli and Hebuterne (1981), Brandt and Brandt (1997) and Zeltyn and Mandelbaum (2005). These are referred to as the M/M/s+G queues where customers abandon the queue according to a general abandonment distribution. A detailed summary of operating characteristics for the M/M/s+G queue is given in Zeltyn and Mandelbaum (2005). We note here that having a general patience distribution function and calculating operating characteristics based on its posterior distribution is a potential area for future work. We will not discuss its application in this study any further.

4.4 Bayesian Queuing Models

As mentioned previously there is a lack of Bayesian research literature in call center queuing models. Some of the recent work in Bayesian queues 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, provide additional insights regarding inference in queues and discuss how Shannon's measure of information can be used to compare different queuing experiments. An interesting issue pointed out by McGrath

et al. (1987a) is that the assumption of a steady state queue violates Cromwell's Rule (see Lindley (2006)) since no random event can have a probability of one or zero a priori. In addition, prior assumptions of independent and dependent arrival rate (λ) and service rate (μ) are discussed. More specifically, Erlang distributed λ and μ are introduced for the independent case and bivariate lognormal for the dependent case. However, the implications of these prior parameters on posterior operating characteristics are not discussed.

Following Gross and Harris (1998), for the M/M/1 queues, let the system size be denoted by $N(t)$, therefore given the arrival rate (λ) and the service rate (μ), as introduced in (4.4), the distribution of $N(t)$ in steady state is given by

$$\lim_{t\to\infty} Pr(N(t) = n|\lambda, \mu) = P(N = n|\lambda, \mu). \tag{4.28}$$

Therefore an M/M/1 system is said to be in steady state if (4.28) exists and (4.28) will exists if and only if $\rho < 1$. This is referred to as the ergodic condition. In classical queuing theory $\rho < 1$ is assumed and computations are carried out accordingly. Armero and Bayarri (1994b) make the argument that the quantity $Pr(\rho < 1)$ is a very important measure and has no parallel in classical queuing theory. They also point out that the Bayesian paradigm can handle restrictions in the parameter space in an easier manner as implied by the case of $Pr(\rho < 1)$. What would be of interest to call center practitioners is to develop a mechanism from which one can calculate $Pr(\rho < 1|D)$, in other words the probability that the system is in steady state given

that we have observed arrival and service times and keep updating it as new data become available. We discuss this issue in more detail in our numerical examples with real call center data.

The first known full treatment of an M/M/1 queue from a Bayesian point of view is due to Armero and Bayarri (1994a). The focus of the article is on the inference of the system primitives and the predictive distributions of main operating characteristics of the M/M/1 queue where the priors for the interarrival rate (λ) and service rate (μ) are gamma distributed

$$\lambda \sim G(\alpha_0, \beta_0), \tag{4.29}$$

and

$$\mu \sim G(a_0, b_0), \tag{4.30}$$

where λ and μ are assumed to be independent a priori. This prior structure along with exponential interarrival and service times, leads to a conjugate gamma posterior family. In our treatment of call center queuing models, the independent gamma setup introduced by Armero and Bayarri (1994a) will also be considered.

An attractive feature of the model proposed by Armero and Bayarri (1994a) is that inference regarding the traffic intensity, ρ can be carried out by calculating the probability that the system is in steady state. As pointed out by the authors there is no analogous measure in classical queuing theory. More specifically, it can be shown

that with proper transformation, ρ is F distributed as follows

$$(\frac{\rho}{R}|D) \sim F(2\alpha, 2a), \tag{4.31}$$

where $R = E(\lambda|D)/E(\mu|D)$, i.e. the ratio of the posterior means of λ and μ. Furthermore, $Pr(\rho < 1|D)$ is obtained in closed form so that a formal hypothesis test of stationarity can be carried out ($H_0 : \rho < 1$ versus $H_1 : \rho \geq 1$). A formal test for stationarity for Bayesian M/M/s queues is introduced in Armero (1994) and results are illustrated for different values of s for a hypothetical queuing system. An important conclusion reached by Armero (1994) is that when $E(\mu)$=2.2 and $E(\lambda)$=4.3, and s=2, the $Pr(\rho < 1|D)$ is calculated as 0.3421, whereas in an analogous scenario the classical theory would have concluded that the system was in steady state with probability 1.

Another important issue discussed in Armero and Bayarri (1994a) is that restrictions on the parameter space such as $\rho < 1$ can easily be implemented into the system in the form of $Pr(\rho|D, \rho < 1)$, that is if the queue is assumed to be in steady state then a posteriori one can condition all densities on $\rho < 1$. The authors obtain the predictive distributions of the operating characteristics conditioned on $\rho < 1$ and make a note about that most predictive distributions do not have moments due to the nature of the priors. Alternatively Armero and Bayarri (1994b) consider a Gauss hypergeometric prior for ρ and a gamma prior for μ where moments for the operating characteristics exist.

Finally, Armero and Bayarri (1996) extend the results for M/M/1 queues to M/M/s queues where λ and μ are again assumed to be gamma distributed as given in (4.29) and (4.30). The only difference is due to the new definition of ρ which is given in (4.1) as $\rho = \frac{\lambda}{s\mu}$ for M/M/s queues. In order for the system to be in steady state, (4.2) needs to hold. As introduced in (4.12), the offered load, r, becomes an important characteristic and its posterior distribution needs to be computed. Therefore, in order for the system to be in steady state, $r < s$ must hold and the relevant steady-state test for the M/M/s queues is now given by

$$Pr(r < s|D), \tag{4.32}$$

where posterior r is shown to have a gamma-gamma distribution with the following form

$$p(r|a, b, c, D) = \frac{c^b \Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{r^{a-1}}{(r+c)^{a+b}}. \tag{4.33}$$

Based on the distribution given in (4.33), similar to M/M/1 queue results, Armero and Bayarri (1996) obtain the form of $Pr(r|D, r < s)$ and obtain the predictive distributions of the operating characteristics conditional on the restriction, $r < s$. The authors show that under certain conditions, moments for some of the predictive distributions of the operating characteristics can be calculated.

Wiper (1998) introduces Bayesian analysis of Er/M/1 and Er/M/s queues where interarrival times are assumed to be Erlang distributed and the service times are

exponential and estimates the operating characteristics using Monte-Carlo methods. Armero and Bayarri (1997) consider a queue where customers are served immediately upon arrival by modeling it as a queue system with infinite number of servers, i.e. M/M/∞ and consider several experiments for data observation to carry out their Bayesian analysis. Most recently, Armero and Conesa (2006) use Bayesian hierarchical models in order to analyze the congestion for several manufacturing bulk service queues and use Markov chain Monte Carlo methods to obtain posterior predictive distributions of most of the relevant operating characteristics.

4.5 Proposed Bayesian Queuing Models in Call Centers

As mentioned previously, mainly three stochastic processes define a call center queuing system; arrival, service and abandonment. The rates of these three processes are referred to as the system primitives, λ, μ and θ, respectively. In classical queuing theory in call centers, these rates are assumed to be fixed but unknown and are estimated from historical data. From a Bayesian point of view, since these rates are unknown quantities we need to define our uncertainty about them via probability distributions. In this section we introduce different modeling strategies for the system primitives in M/M/1, M/M/s and M/M/s+M queues in call centers. To the best of our knowledge this is the first attempt in modeling call center queuing operations from a Bayesian perspective and the first treatment of system primitives as uncertain quantities. We present modeling of both independent and dependent priors for the

system primitives, λ, μ and θ.

4.5.1 Independent Priors

One can specify either dependent or independent priors for system primitives. Studies in Bayesian queues assume independent priors for λ and μ and the most commonly used form is the gamma distribution. For M/M/1 and M/M/s queues, we can use the priors introduced in (4.29) and (4.30) and for M/M/s+M queues we can assume that the abandonment rate, θ follows a gamma prior due to its conjugate nature with the exponential distribution. Armero and Bayarri (1994a) and Armero and Bayarri (1996) discuss prior and posterior implications of using gamma priors for λ and μ in M/M/1 and M/M/s queues. We introduce posterior analysis of M/M/s+M queues (queues with abandonment), with gamma priors in the sequel.

4.5.2 Dependent Priors

Another modeling strategy one can pursue is to assume a dependent prior distribution between two or more of the system primitives. We consider two cases, the first assumes that λ, μ and θ jointly follow a trivariate lognormal density and the second one considers a bivariate gamma prior for λ and μ. We note here that assuming a dependent prior structure on the system primitives does not violate the assumptions for M/M/1, M/M/s and M/M/s+M queues since the arrival, service and abandonment processes will still stay independent conditional on the λ, μ and θ. Next, we discuss the properties of these two priors and their implications on the operating

characteristics.

Trivariate Lognormal Prior

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, creating 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

$$p(\lambda, \mu, \theta) = (2\pi)^{-3/2}|\boldsymbol{\Sigma}|^{-3/2}(\frac{1}{\lambda\mu\theta})exp(-\frac{1}{2}[\boldsymbol{\Phi} - \boldsymbol{\nu}]'\boldsymbol{\Sigma}^{-1}[\boldsymbol{\Phi} - \boldsymbol{\nu}]), \tag{4.34}$$

where $\boldsymbol{\nu} = (\nu_1\ \nu_2\ \nu_3)'$ and the variance covariance matrix for λ, μ and θ is given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_{3,3} \end{bmatrix}. \tag{4.35}$$

The joint density (4.34) 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 (4.35) ($\sigma_{i,j}$ where $i \neq j$) are all equal to zero, then λ, μ and θ are said to be independent a priori. What would be of interest is to study the behaviour of the dependency structure as new data become available so that inference regarding the dependency between the posterior system primitives can be carried out. Also, for M/M/1 and M/M/s queues, θ does not exist and (4.34) is a bivariate log normal distribution of λ and μ. In addition it can be shown that the distribution of λ/μ for the bivariate case is also lognormal, in other words the prior density of ρ (or r in the multiple server queues) is given by

$$p(\rho) = (2\pi)^{-1/2}\beta^{-1}(\frac{1}{\rho})exp\{-\frac{1}{2}[(\frac{log(\rho-\alpha)}{\beta})^2]\}, \tag{4.36}$$

where $\alpha = \nu_1 - \nu_2$ and $\beta = \sigma_{1,1} + \sigma_{2,2} - 2\sigma_{1,2}$. This implies that depending on the prior parameters, there is a non zero probability that the system is not ergodic. Therefore by obtaining the posterior distribution of ρ, a formal ergodicity test for the system can be conducted and steady state behavior can be inferred using call center data.

Bivariate Gamma Prior

As pointed out previously, classical queuing theory assumes that the system is in steady state. If a priori, call center management believes that the system is in steady state then a similar approach can be introduced in M/M/s queues. Let us first consider an M/M/1 queue in steady state.

In order to obtain a prior structure implying that $\lambda < \mu$, that is the steady state

condition for M/M/1 queues, the following can be used. Let

$$p(\lambda, \mu) = \frac{\theta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\lambda^{\alpha_2-1}(\mu-\lambda)^{\alpha_1-1}e^{-\theta\mu}, \tag{4.37}$$

where $\mu > \lambda$ and $\theta, \alpha_1, \alpha_2 > 0$. (4.37) is also known as the McKay's bivariate gamma distribution as discussed in Kotz et al. (2000). As a consequence of (4.37), the following can be obtained

$$\lambda \sim G(\alpha_2, \theta), \tag{4.38}$$

$$\mu \sim G(\alpha_1 + \alpha_2, \theta). \tag{4.39}$$

Also for $\rho = \frac{\lambda}{\mu}$, it is straightforward to show that

$$\rho \sim B(\alpha_1, \alpha_2), \tag{4.40}$$

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).

Furthermore, the conditional distribution of λ given μ can be obtained as a Truncated Beta density

$$\lambda|\mu \sim TrB(\alpha_1, \alpha_2, 0, \mu), \tag{4.41}$$

where $0 < \lambda < \mu$. An interesting property of McKay's bivariate gamma is the

following correlation structure

$$corr(\lambda, \mu) = \sqrt{\frac{\alpha_1}{\alpha_1 + \alpha_2}}. \tag{4.42}$$

The above implies that a priori there is always a positive correlation between λ and μ which is consistent with our steady state assumption. For instance when arrivals to the system become more frequent (a decrease in λ) the service times are adjusted accordingly (a decrease in μ).

For multiple server queues such as the M/M/s model, the above structure can still be used by assuming that

$$s\mu \sim G(\alpha_1 + \alpha_2, \theta), \tag{4.43}$$

where s is the number of servers in the system. Therefore the distribution of $\rho = \lambda/s\mu$ will be the same as (4.40)

$$\rho \sim B(\alpha_1, \alpha_2), \tag{4.44}$$

where $0 < \rho < 1$ which is consistent with the steady state assumption in M/M/s queues.

Further Implications of the Bivariate Gamma Prior

The prior structure implied by McKay's bivariate gamma makes prior predictive distributions of some of the operating characteristics analytically available. The number of customers in the system and waiting time distribution for M/M/1 queues can be

obtained as follows.

Given (4.9) and (4.40), the predictive prior density for the number of customers in the system in steady state (N) can be obtained as

$$\begin{aligned} P(N=n) &= \int_0^1 P(N=n|\rho)p(\rho)d\rho \\ &= \int_0^1 (1-\rho)\rho^n \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \rho^{\alpha_1-1}(1-\rho)^{\alpha_2-1}d\rho. \end{aligned}$$

It follows that

$$P(N=n) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1+n)\Gamma(\alpha_2+1)}{\Gamma(\alpha_1+\alpha_2+n+1)}, \tag{4.45}$$

is a Polya distribution. This can be used for call center design purposes before observing the data, that is, if λ and μ are believed to have the prior structure implied by (4.38) and (4.39) then (4.45) represents the distribution of customers in the system anytime in steady state. As the new arrivals and service times are observed in a call center the posterior predictive distributions of the operating characteristics can be obtained using Markov chain Monte Carlo methods.

Another measure of interest is the waiting time distribution in the system, T, for M/M/1 queues given by (4.11). Therefore using (4.48) and (4.11), the prior predictive

distribution of T can be obtained as follows

$$\begin{aligned} p(t) &= \int_0^\infty \int_0^\mu p(t|\lambda,\mu)p(\lambda,\mu)d\lambda d\mu \\ &= \int_0^\infty \int_\lambda^\infty (\mu-\lambda)e^{-(\mu-\lambda)t}\frac{\theta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\lambda^{\alpha_2-1}e^{-\theta\lambda}(\mu-\lambda)^{\alpha_1-1}e^{-\theta(\mu-\lambda)}d\lambda\, d\mu \\ &= \int_0^\infty \frac{\theta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\lambda^{\alpha_2-1}e^{-\theta\lambda}\int_\lambda^\infty (\mu-\lambda)^{\alpha_1}e^{-(\theta+t)(\mu-\lambda)}d\lambda\, d\mu \\ &= \frac{\alpha_2\theta^{\alpha_2}}{(\theta+t)^{\alpha_2+1}}, \end{aligned}$$

which implies that the waiting time prior predictive distribution in the system, $p(t)$, is a Pareto density with parameters θ and α_2.

4.6 Bayesian Analysis of Queues in Call Centers

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 start our discussion by introducing notation and preliminaries. Let us assume that we have observed n_x interarrival 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 interarrival 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. In the rest of the

paper, we use $D = (x_1, \cdots, x_{n_x}, y_1, \cdots, y_{n_y})$ for the vector of observed quantities in M/M/1 and M/M/s queues or $D = (x_1, \cdots, x_{n_x}, y_1, \cdots, y_{n_y}, z_1, \cdots, z_{n_z}, z'_1, \cdots, z'_{n'_z})$ for the vector of observed quantities in M/M/s+M queues.

Therefore given D, one can obtain the likelihood function of λ and μ for M/M/1 and M/M/s queues as

$$L(\lambda, \mu; D) \propto \lambda^{n_x} e^{-\lambda \sum_{i=1}^{n_x} x_i} \mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i}, \tag{4.46}$$

and the likelihood function of λ, μ and θ for M/M/s+M queues as

$$L(\lambda, \mu, \theta; D) \propto \lambda^{n_x} e^{-\lambda \sum_{i=1}^{n_x} x_i} \mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i} \theta^{n_z} e^{-\theta \sum_{i=1}^{n_z} z_i} e^{-\theta \sum_{i=1}^{n'_z} z'_i}. \tag{4.47}$$

In (4.47) 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 (in other words customers who did not abandon the queue) in inferring the abandonment rate θ. We note here that there are other ways in which call center data can be collected. McGrath et al. (1987b) and Armero and Bayarri (1996) discuss different experimental designs for queue systems and also point out that regardless of the design scheme the likelihood functions given in (4.46) and (4.47) 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.

4.6.1 Posterior Analysis of Independent Gamma Priors

First we introduce a conjugate Bayesian analysis of gamma priors for the system primitives. Given the priors introduced in (4.29) and (4.30), the joint prior density can be written as the product of the gamma marginals for M/M/1 and M/M/s queues,

$$p(\lambda, \mu) = p(\lambda|\alpha_0, \beta_0)p(\mu|a_0, b_0), \tag{4.48}$$

where $\lambda \sim G(\alpha_0, \beta_0)$ and $\mu \sim G(a_0, b_0)$. Using (4.46) and (4.48), the joint posterior of λ and μ can be obtained as follows

$$p(\lambda, \mu|D) = p(\lambda|D)p(\mu|D), \tag{4.49}$$

where $\lambda \sim G(\alpha, \beta)$, $\mu \sim G(a, b)$ and $\alpha = \alpha_0 + n_x$, $\beta = \beta_0 + (\sum_{i=1}^{n_x} x_i)$, $a = a_0 + n_y$, $b = b_0 + (\sum_{i=1}^{n_y} y_i)$.

For M/M/s+M queues we need to specify a prior for the abandonment rate θ as well, for analytical tractability we can assume a gamma prior on θ as follows

$$\theta \sim G(c_0, d_0), \tag{4.50}$$

which implies the following joint prior of λ, μ and θ

$$p(\lambda, \mu, \theta) = p(\lambda|\alpha_0, \beta_0)p(\lambda|a_0, b_0)p(\theta|c_0, d_0), \tag{4.51}$$

where $\lambda \sim G(\alpha_0, \beta_0)$, $\mu \sim G(a_0, b_0)$. Therefore given (4.47) and (4.51), the joint posterior of λ, μ and θ can be obtained as follows

$$p(\lambda, \mu, \theta|D) = p(\lambda|D)p(\mu|D)p(\theta|D), \tag{4.52}$$

where $\lambda \sim G(\alpha, \beta)$, $\mu \sim G(a, b)$, $\theta \sim G(c, d)$ and $\alpha = \alpha_0 + n_x$, $\beta = \beta_0 + (\sum_{i=1}^{n_x} x_i)$, $a = a_0 + n_y$, $b = b_0 + (\sum_{i=1}^{n_y} y_i)$, $c = c_0 + n_z$, $d = d_0 + (\sum_{i=1}^{n_z} z_i) + (\sum_{i=1}^{n'_z} z'_i)$. We note here that a priori λ, μ and θ are assumed to be independent as a result they remain independent a posteriori as well. Another interesting property which is a result of the likelihood structure introduced in (4.47) is due to the posterior θ whose expected value is calculated as

$$E(\theta|D) = \frac{c_0 + n_z}{d_0 + (\sum_{i=1}^{n_z} z_i) + (\sum_{i=1}^{n'_z} z'_i)}, \tag{4.53}$$

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 now be carried out regardless of the ergodicity assumption and steady state conditions can now be tested formally using the distributions introduced in (4.31) and (4.33) for M/M/1 and M/M/s queues in

call centers, respectively. We will illustrate these stationarity tests in the sequel. 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.

4.6.2 Posterior Analysis of Trivariate Lognormal Prior

Unlike the independent gamma prior model, closed form results cannot be obtained for the posterior distributions of the system primitives, λ, μ and θ. Therefore we can use Markov chain Monte Carlo methods to obtain posterior samples. Next we discuss the implementation of the Metropolis-Hastings algorithm for the trivariate lognormal prior model. Recall the likelihood introduced in (4.47) and the trivariate lognormal prior density from (4.34). Therefore the joint posterior of λ, μ and θ can be obtained as

$$p(\lambda, \mu, \theta|D) \propto L(\lambda, \mu, \theta; D)p(\lambda, \mu, \theta). \tag{4.54}$$

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

$$p(\lambda, \mu, \theta) = p(\lambda|\mu, \theta)p(\mu|\theta)p(\theta), \tag{4.55}$$

where

$$p(\lambda|\mu,\theta) \sim LN(\nu_1 + \begin{bmatrix} \sigma_{1,2} & \sigma_{1,3} \end{bmatrix} \begin{bmatrix} \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{3,2} & \sigma_{3,3} \end{bmatrix} \begin{bmatrix} \mu - \nu_2 \\ \theta - \nu_3 \end{bmatrix}, \sigma_{1,1} - \begin{bmatrix} \sigma_{1,2} & \sigma_{1,3} \end{bmatrix} \begin{bmatrix} \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{3,2} & \sigma_{3,3} \end{bmatrix} \begin{bmatrix} \sigma_{1,2} \\ \sigma_{1,3} \end{bmatrix}),$$

$$p(\mu|\theta) \sim LN(\nu_2 + \frac{\sigma_{2,3}}{\sigma_{3,3}}(\theta - \nu 3), \sigma_{2,2} - \frac{\sigma_{2,3}^2}{\sigma_{3,3}}),$$

and

$$p(\theta) \sim LN(\nu_3, \sigma_{3,3}).$$

In order to obtain samples from (4.54) we can use the random walk Metropolis-Hasting algorithm whose proposal density is multivariate normal. 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-Hasting algorithm can be summarized as follows

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

  Repeat for $t > 0$,

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

- 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

$$\alpha(\boldsymbol{\theta^{(t)}}, \boldsymbol{\theta}^*) = min\left\{1, \frac{\pi(\boldsymbol{\theta^*})q(\boldsymbol{\theta^{(t)}}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta^{(t)}})q(\boldsymbol{\theta}^*|\boldsymbol{\theta^{(t)}})}\right\}. \tag{4.56}$$

In (4.56), $q(.|.)$ is the multivariate normal proposal density and $\pi(\boldsymbol{\theta}) = p(log(\lambda), log(\mu), log(\theta)|D)$ as in (4.54). If we repeat the above a large number of times then we obtain samples from $p(\boldsymbol{\theta}|D)$.

4.6.3 Posterior Analysis of Bivariate Gamma Prior

Following the model introduced in section 4.5.2 for an M/M/s system with the likelihood from (4.46) and the joint prior given by (4.37), we can obtain the joint posterior for λ and μ as

$$p(\lambda, \mu|D) \propto L(\lambda, \mu; D)p(\lambda, \mu), \tag{4.57}$$

where $\mu > \lambda$. More explicitly it can be written as

$$p(\lambda, \mu|D) \propto \lambda^{n_x} e^{-\lambda \sum_{i=1}^{n_x} x_i} \mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i} \frac{\theta^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \lambda^{\alpha_2-1}(\mu-\lambda)^{\alpha_1-1} e^{-\theta\mu}, \tag{4.58}$$

$$\propto \lambda^{n_x+\alpha_2-1} e^{-\lambda \sum_{i=1}^{n_x} x_i} \mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i+\theta} (\mu-\lambda)^{\alpha_1-1}, \tag{4.59}$$

where $\mu > \lambda$. Similar to Case 2, (4.59) is not a well known density. Therefore in order to generate samples from (4.59) we can use Markov chain Monte Carlo methods. We can use the rejection sampling within Gibbs algorithm (see Smith and Gelman (1992)) since the modes from the conditional likelihoods are available. First we need to obtain the full conditional densities, $p(\lambda|\mu, D)$ and $p(\mu|\lambda, D)$ as

$$p(\lambda|\mu, D) \propto \{\lambda^{n_x+\alpha_2-1}(\mu-\lambda)^{\alpha_1-1}\}\{e^{-\lambda \sum_{i=1}^{n_x} x_i}\} I(0, \mu), \tag{4.60}$$

and

$$p(\mu|\lambda, D) \propto \{\mu^{n_y} e^{-\mu \sum_{i=1}^{n_y} y_i+\theta}\}\{(\mu-\lambda)^{\alpha_1-1}\} I(\lambda, \infty). \tag{4.61}$$

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

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

$$\hat{\lambda} = \begin{cases} \mu, & \text{if} \quad \mu \leq \frac{n_x}{\sum_{i=1}^{n_x} x_i} \\ \frac{n_x}{\sum_{i=1}^{n_x} x_i}, & otherwise. \end{cases} \tag{4.62}$$

and

$$\hat{\mu} = \begin{cases} \lambda, & \text{if} \quad \lambda \geq \frac{n_y}{\sum_{i=1}^{n_y} y_i} \\ \frac{n_y}{\sum_{i=1}^{n_y} y_i}, & otherwise. \end{cases} \tag{4.63}$$

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

4.6.4 Posterior Predictive Analysis

Next we focus on the estimation of the posterior predictive distributions of the operating characteristics introduced in section 4.3. In order to use the closed form solutions

of the operating characteristics for M/M/1, M/M/s and M/M/s+M queues, first the joint posterior distribution of λ, μ, θ needs to be calculated. The joint posterior for M/M/1 and M/M/s queues can be obtained via

$$p(\lambda, \mu|D) \propto L(\lambda, \mu; D)p(\lambda, \mu), \tag{4.64}$$

and for M/M/s+M queues as,

$$p(\lambda, \mu, \theta|D) \propto L(\lambda, \mu, \theta; D)p(\lambda, \mu, \theta), \tag{4.65}$$

where $L(\lambda, \mu; D)$ and $L(\lambda, \mu, \theta; D)$ are given as in (4.46) and (4.47) respectively. 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 that are introduced in section 4.3 for M/M/1, M/M/s and M/M/s+M queues. First let us introduce a general notation for a given operating characteristic conditional on system primitives. Let $f(X|\lambda, \mu)$ represent any operating characteristic of an M/M/1 or M/M/s queue, and $f(X|\lambda, \mu, \theta)$ of an M/M/s+M queue. Therefore, in general the expectation of $f(X)$

or $f(X|D)$ for M/M/1 or M/M/s queues can be obtained as follows

$$f(X|D) = \int_{\lambda} \int_{\mu} f(X|\lambda, \mu) p(\lambda, \mu|D) d\lambda \, d\mu, \tag{4.66}$$

and for M/M/s+M queues can be obtained as

$$f(X|D) = \int_{\lambda} \int_{\mu} \int_{\theta} f(X|\lambda, \mu, \theta) p(\lambda, \mu, \theta|D) d\lambda \, d\mu \, d\theta. \tag{4.67}$$

Finally, in order to evaluate $f(X|D)$, one can use straight Monte Carlo approximation for a large sample of λ, μ and θ. For instance in order to obtain $E(Pr(Ab|D))$ from (4.25) we can use the following

$$E(Pr(Ab|D)) \approx \frac{1}{M} \sum_{i=1}^{M} Pr(Ab|\lambda^{(i)}, \mu^{(i)}, \theta^{(i)}) P(\lambda^{(i)}, \mu^{(i)}, \theta^{(i)}|D), \tag{4.68}$$

where M represents the number of samples generated, $\lambda^{(i)}$, $\mu^{(i)}$, $\theta^{(i)}$ are the ith generated samples for λ, μ and θ, respectively. Estimation issues for different measures of performance in different call center queuing models will be addressed in the sequel.

4.7 Numerical Illustrations and Findings

In order to illustrate how the proposed models work, 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 times and abandonment times in the mornings of month February (between 7:00AM -10:30AM) for stock exchange customers whose abandonment times exhibit exponential type of behavior. Three cases will be presented in the sequel. The first two are the independent gamma prior model and the trivariate lognormal prior model for M/M/s+M queues. The third is the bivariate gamma prior model for M/M/s queues.

4.7.1 Case 1: Independent Gamma Priors Case

In order to carry out the inference discussed in Section 4.5.1 we have used flat but proper priors for λ, μ and θ. A summary of the most common operating characteristics (using the notation introduced in Section 3) implied by the posterior distributions of the system primitives for different number of servers is shown in Tables 9 and 10. A quick note about the summary statistics is that they are all conditioned on the updated posterior distributions of the uncertain system primitives. For instance, the expected value of P_0 from Table 9 is in fact $E(P_0|D)$.

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	0.2	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	0.2644	0.1685	0.0521	0.014
	Standard Deviation	0.023	0.0016	0.011	0.0066	0.0027

Table 9: Summary of Operating Characteristics-I for Case 1

As shown in Tables 9 and 10, the probability of not getting served upon arrival ($P(T_q > 0|D)$) is decreasing as the number of servers increase. Similarly, the

Summary of Operating Characteristics-II				
Number of Servers	Statistics	$Pr(T_q > 0)$	$Pr(Ab\|T_q > 0)$	$Pr(Ab)$
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	0.00897
s=3	Expected Value	0.07148	0.1304	0.0093
	Standard Deviation	0.0106	0.0187	0.0021

Table 10: Summary of Operating Characteristics-II for Case 1

probability of abandoning ($P(Ab|D)$) is decreasing as there are more servers. Both measures provide managerial insights to the call center management since they can be interpreted as measures of perceived service quality of the call center operation. Furthermore, there is a trade-off between the number of servers and the fraction abandoning ($P(Ab|D)$) and depending on the relative importance of each an optimal level of servers can be identified. This trade-off issue will be discussed in the sequel.

One of the important features of treating the system primitives as uncertain quantities is that all the operating characteristics will have their own distributions along with their respective statistical properties. Instead of solely observing 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 24, 25 and 26, some of the density plots are not always symmetric, therefore using a point estimate of the mean might be misleading. From a practical point of view, the call center management can instead use the mode as a proxy along with a 95 % credibility interval to make operational decisions.

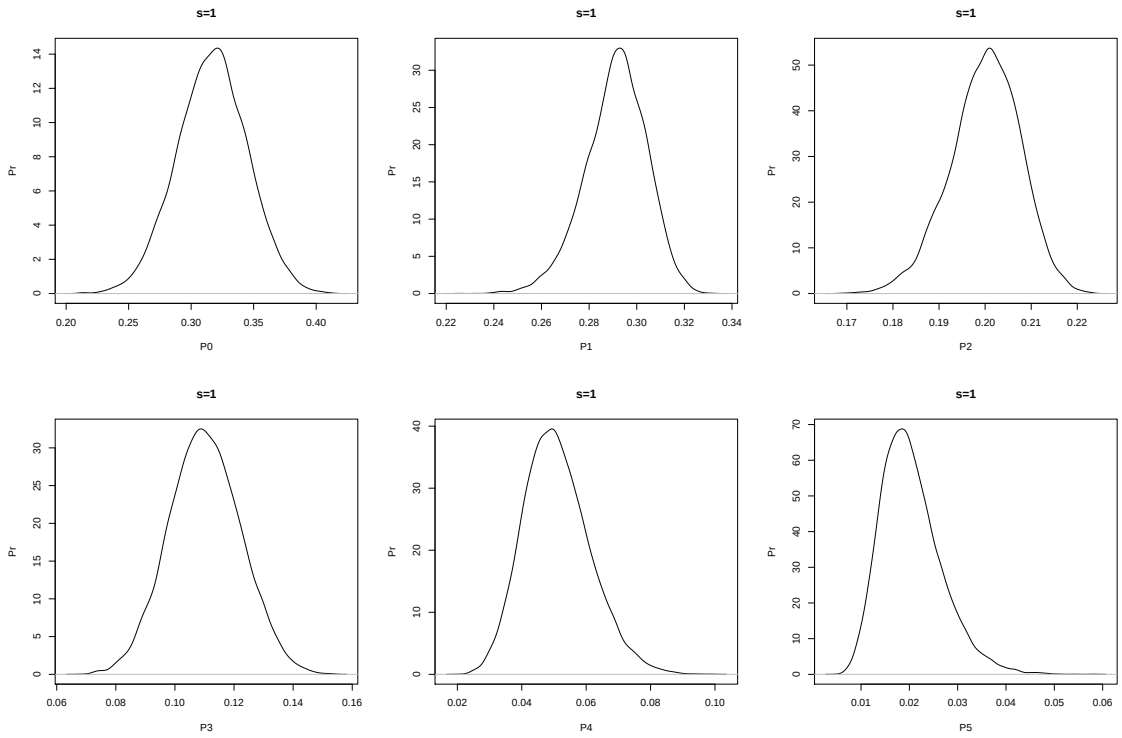

Figure 24: Posterior Distributions of $(P_i|D)$ for $i = 1, \cdots, 5$ when s=1

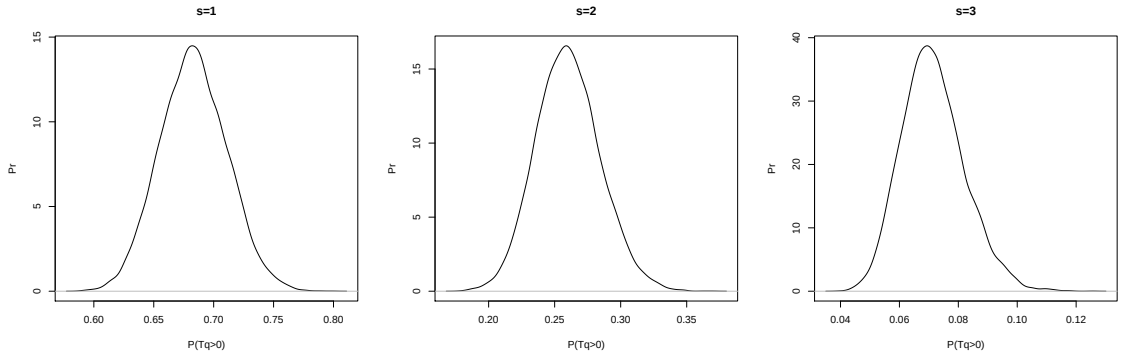

Figure 25: $Pr(T_q > 0|D)$ for s=1,s=2 and s=3

Another implication of the proposed independent gamma prior case is the relationship between λ, μ and the probability of abandoning $(Pr(Ab|D))$. As shown in Figure 27, $Pr(Ab|D)$ is very sensitive to the changes in the values of λ and μ regardless of the number of servers. As expected, as arrivals become more frequent the

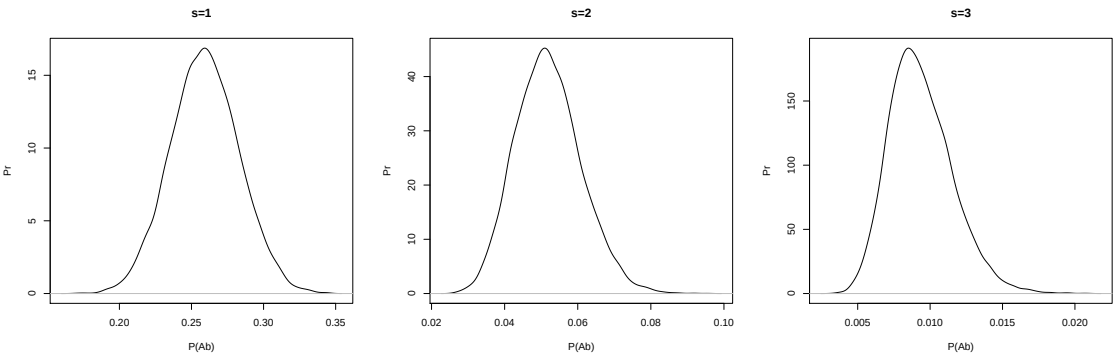

Figure 26: $Pr(Ab|D)$ for s=1,s=2 and s=3

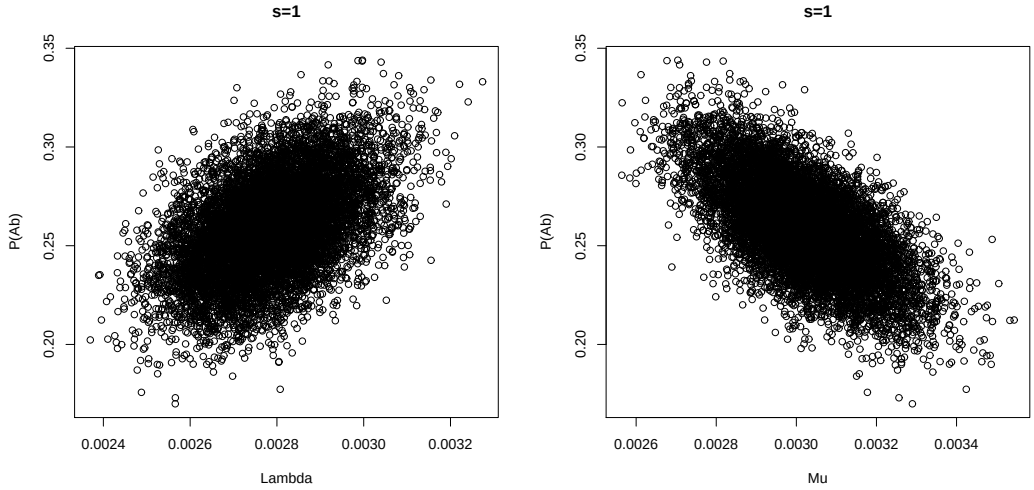

Figure 27: $Pr(Ab|D)$ vs. λ (left) and $Pr(Ab|D)$ vs. μ (rigtht) for s=1

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.

4.7.2 Case 2: Trivariate Lognormal Prior Case

In Case 1, all system primitives were assumed to be independent from each other. Using the structure introduced in Section 4.5.2, a trivariate lognormal prior (Case 2), we will investigate if the system primitives are in fact correlated 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 (4.54), 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 28 and the respective density plots in Table 29. Based on the trace plots we can fairly conclude that convergence has been attained.

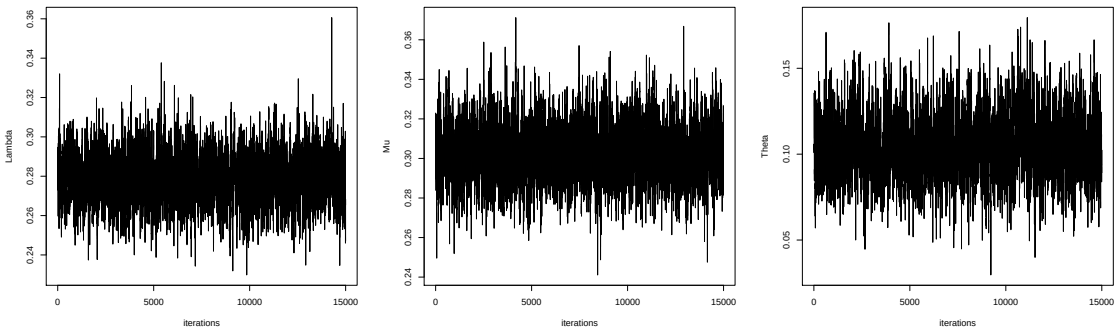

Figure 28: Trace plots for λ (left), μ (middle) and θ (right)

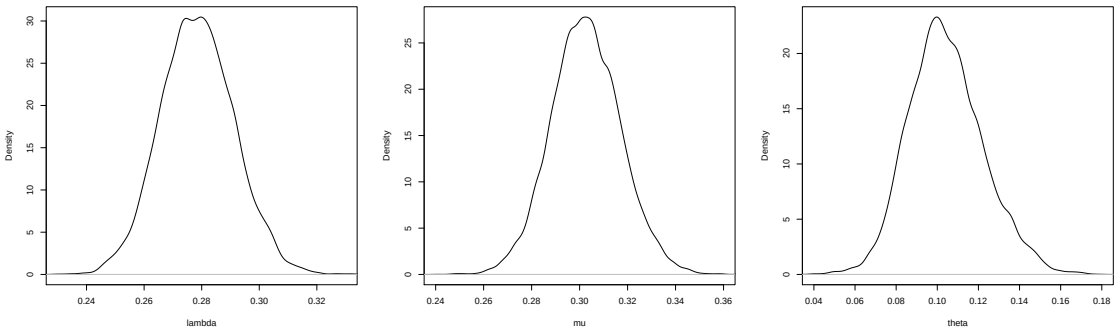

Figure 29: Posterior density plots for λ (left), μ (middle) and θ (right)

Next, similar to Case 1 we obtain the summary of the operating characteristics using the samples generated from (4.54). As shown in Table 11 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 24, 25 and 26. A boxplot of the steady state probabilities for $s = 1$ is shown in Figure 30.

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	0.008	0.0126	0.0108

Table 11: Summary of Operating Characteristics-I for Case 2

Summary of Operating Characteristics-II				
Number of Servers	Statistics	$Pr(T_q > 0)$	$Pr(Ab\|T_q > 0)$	$Pr(Ab)$
s=1	Expected Value	0.68822	0.3798	0.2591
	Standard Deviation	0.02954	0.032	0.025

Table 12: Summary of Operating Characteristics-II for Case 2

The fact that these results are identical provides evidence that the posterior distributions were very similar for Case 1 and Case 2. The only difference between these two is that the latter allows for correlations between system primitives. As shown in Table 13, the posterior sample correlations for λ, μ and θ are all very close to zero. This shows evidence in favor of non-correlated system primitives. For this particular call center data, Case 1 seems to be a better model due to its conjugate and easy to implement nature. However if the system primitives are suspected to be correlated a priori, then one can always test it using Case 2. If evidence is found in favor of

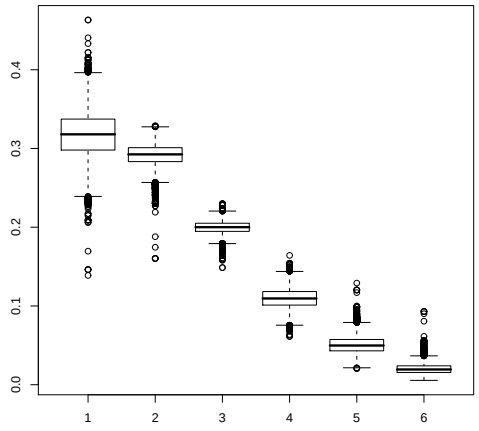

Figure 30: Boxplots for the Steady State Probabilities for Case 2 when s=1

correlation 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 estimating the operating characteristics for the independent gamma model.

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

Table 13: Posterior Sample Correlations for λ, μ and θ

4.7.3 Case 3: Bivariate Gamma Prior Case

Classical queuing theory makes the assumption of steady state for M/M/s type of queues. An analogous scenario can be developed using the Bayesian paradigm. In other words, if a priori call center management believes that the system is in steady state then a bivariate gamma prior structure (Case 3) can be used to model system primitives. In order to show the implementation of Case 3 using the respective sam-

pling algorithm, we have used the same arrival and service dataset as in Case 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 31 shows the trace plots and the density plots for λ and μ. Based on the trace plots we can fairly conclude that convergence has been attained.

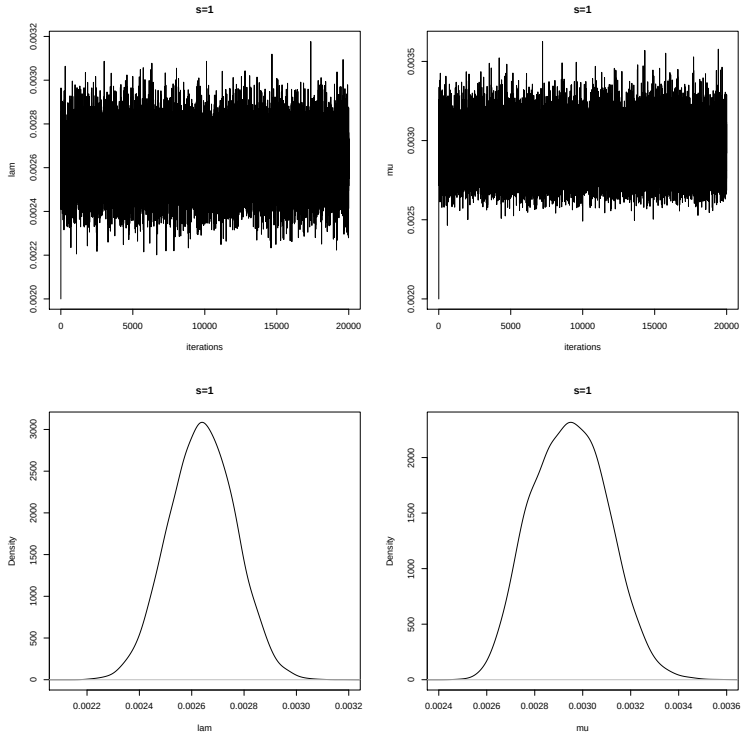

Figure 31: Trace Plots and Posterior Density Plots for λ (left) and μ (right) for Case 3

A summary of the most common operating characteristics (using the notation introduced in Section 3) for M/M/s queues as implied by the posterior samples of λ and μ for different number of servers is shown in Tables 14. An important point about M/M/s queues is that they ignore the abandonment behavior of customers. For instance, as seen in Table 14 (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 9). This under/over estimation in the operating characteristics emphasizes the importance of modeling the abandonment process in call center queuing applications.

Summary of Operating Characteristics						
Servers	Statistics	P_0	P_1	P_2	P_3	$Pr(T_q = 0)$
s=1	Expected Value	0.1011	0.086	0.073	0.062	0.1018
	Standard Deviation	0.07	0.0559	0.045	0.0366	0.069
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	0.049	0.9294
	Standard Deviation	0.0293	0.0024	0.0135	0.0078	0.0130

Table 14: Summary of Operating Characteristics for Case 3

Furthermore, the posterior distributions of the utilization, $p(\rho|D)$, for $s = 1, 2, 4$ are shown in Figure 32. An interesting finding is that for $s = 1$, there are several ρ values that are close to 1 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 fixed in classical queuing theory, being able to obtain the distribution of utilization is another important feature of the Bayesian queuing approach.

4.7.4 Additional Insights from Bayesian Analysis

Posterior steady state hypothesis testing (whether the system is in steady state given the data) in M/M/s queues is an important contribution of the Bayesian queues and

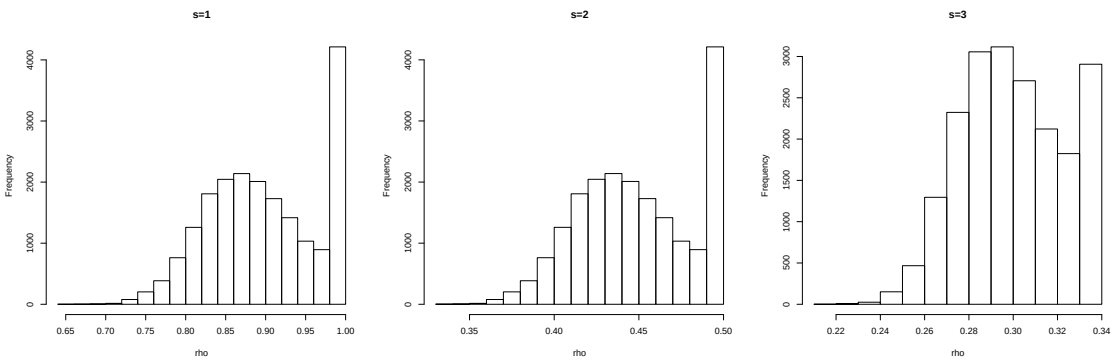

Figure 32: Posterior ρ for $s = 1, 2, 3$ in Case 3

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. an M/M/s type of queue), we can use the posterior distributions of λ and μ obtained for Case 1. Formally, the hypothesis testing can be setup as

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

where $Pr(H_0 : \rho > 1|D)$ when $s = 1$ is 0.09786, as calculated using (4.31). 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 practical for this particular example since there is abandonment in the system which already implies steady state behavior

for M/M/s+M as discussed previously.

Another point of interest is the investigation of the Little's formula introduced in (4.7) in Bayesian analysis. We have investigated this issue using the output obtained for the M/M/s model, i.e. the output for Case 3. Our findings indicate that given the prior structure at hand Little's formula does not hold a posteriori. In order to check if Little's formula hold, we need to compute the following for each sample i.

$$L^{(i)} = \lambda^{(i)} W^{(i)}, \tag{4.69}$$

Alternatively one can also compute $Pr(L^{(i)} = \lambda^{(i)} W^{(i)})$ in order to infer if Little's formula hold. For our numerical example this probability was obtained to be zero, that is given the prior structure for Case 3, Little's formula does not hold. However, as pointed out by Armero and Bayarri (1996), one can obtain results using different priors (for instance non-informative) which might make Little's formula hold. This seems to be an interesting area for future work for Bayesian queuing and will not be pursued any further in this paper.

In addition, the 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). Therefore, following Mandelbaum and Zeltyn (2007), let the

average operational cost (per unit of time) be defined as

$$C(s, \lambda, \mu, \theta) = cs + a\lambda Pr(Ab|\lambda, \mu, \theta) \tag{4.70}$$

where s is the number of servers, c is the staffing cost, a is the abandonment cost and $Pr(Ab|\lambda, \mu, \theta)$ is the fraction of abandoning (as shown in (4.25)). Therefore, given λ, μ and θ one can calculate the value of s which will minimize the cost function given in (4.70). Since in our proposed models λ, μ and θ are unknown quantities and will have their own posterior distributions we can obtain the distribution of the average operational cost as well as its the expected value, i.e. $E_{\lambda,\mu,\theta}(C(s, \lambda, \mu, \theta))$. Thus, for a given ratio of c/a, we can obtain the value of s, say s^* which will minimize $E_{\lambda,\mu,\theta}(C(s, \lambda, \mu, \theta))$. 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 33 and 34. 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 in the boxplot Figures 33 and 34.

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

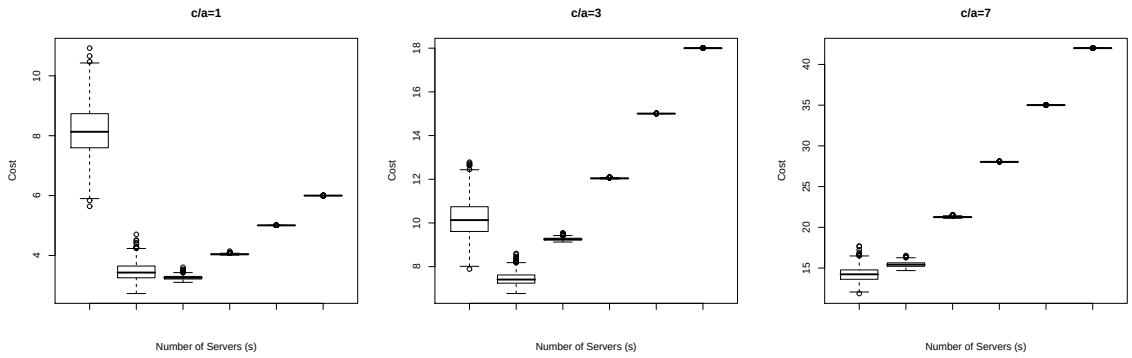

Figure 33: Cost Function vs. Number of Servers for fixed a/c

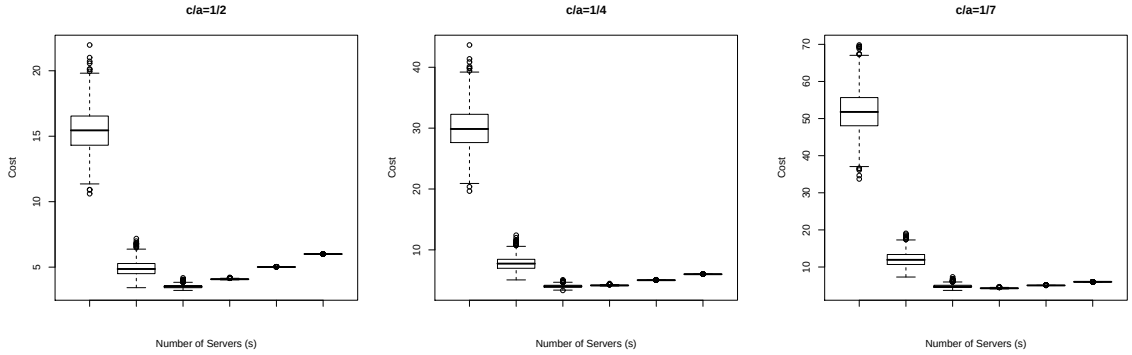

Figure 34: Cost Function vs. Number of Servers for fixed a/c

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

$$Pr(C \leq c|\lambda, \mu, \theta, D) \geq \alpha_{\text{tolerance}}, \tag{4.71}$$

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,

$$s^* = r + \beta\sqrt{r}, \tag{4.72}$$

where s^* (an integer) represents the appropriate staffing level and β is a positive constant which is a function of level of service. (4.72) is said to be practically 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 ($P(W > 0)$). The term $\beta\sqrt{r}$ is usually referred to as the excess capacity in the call center literature and the fact that the excess capacity grows as a square root function of the offered load is referred to as the statistical economies of scale. Garnett et al. (2002) discuss the staffing rule for M/M/s+M queues, the so called Erlang-A model. The staffing rule (4.72) for the Erlang-A model stays the same. However, the coefficient β is now calculated as a function of both $P(W > 0)$ (probability of waiting before service) and θ (the abandonment rate). Also β can now take negative values, namely the optimal staffing level is now allowed to be less than the offered wait due to abandonment. For limiting values of the performance characteristics ($P(W > 0)$ and $P(Abandon)$), 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 (4.72) 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 (4.70) on the implied beta index which can be obtained using (4.72). 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 in the range (-1,2) (see Garnett et al. (2002)).

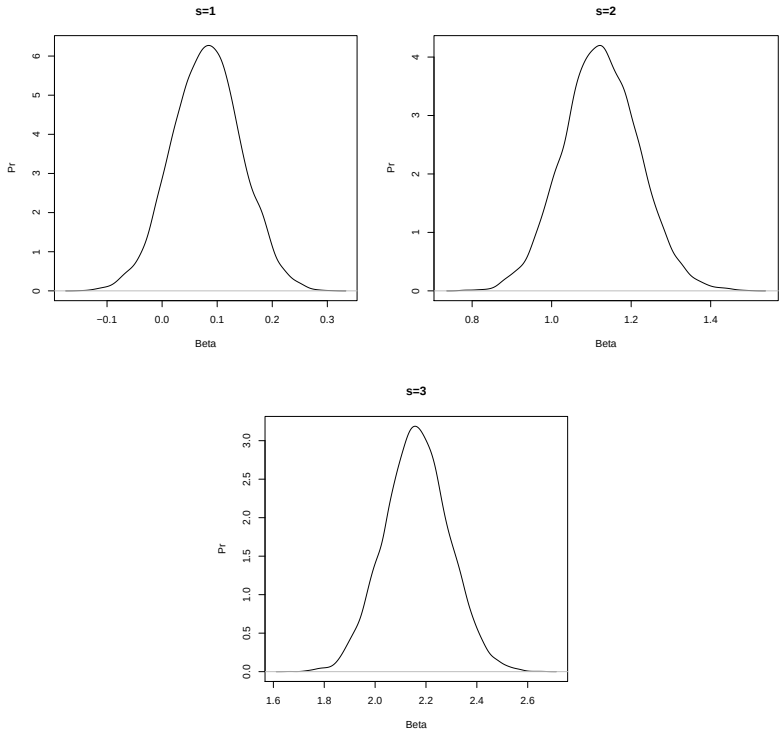

Figure 35: β vs. Optimal Number of Servers for Case 1 and 2

For the M/M/s+M queues from Case 1 and 2, density plots for the implied level of service β for different optimal number of servers are shown in Figure 35. The

posterior means of β $(E(\beta|D, s = s^*))$ 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.

4.8 Concluding Remarks

In this paper we have introduced Bayesian queuing models for call center operations. Three cases were considered, the independent gamma prior and trivariate gamma prior models for M/M/s+M queues and the bivariate gamma prior for M/M/s queues.

To the best of our knowledge this is the first study in the call center literature from a Bayesian perspective where the system primitives are uncertain quantities and the first Bayesian treatment of queues with abandonment. 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.

We have addressed issues such as correlated system primitives and steady state hypothesis tests which are not available in classical queuing models. In the queuing literature, λ, μ and θ are assumed to be un-correlated with each other. We have investigated whether the un-correlation assumption was adequate using real call center data. In doing so, we have developed the trivariate lognormal prior model to handle cases where correlated 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 un-correlated a posteriori as well. 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 identified 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 which in turn validates the M/M/s+M model, further extensions where the abandonment process as a general distribution are possible and are currently being investigated by the authors. Furthermore, 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.

References

Aksin, Z., Armony, M., and Mehrotra, V. (2007). Managing uncertainty in call centers using Poisson mixtures. *Production and Operations Management*, 16(6):665–688.

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

Armero, C. (1994). Bayesian inference in Markovian queues. *Queueing Systems*, 15:419426.

Armero, C. and Bayarri, M. (1994a). Bayesian prediction in M/M/1 queues. *Queueing Systems*, 15:401417.

Armero, C. and Bayarri, M. (1994b). Prior asessments for prediction in queues. *The Statistician*, 43(1):139153.

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

Armero, C. and Bayarri, M. (1997). A Bayesian analysis of a queueing system with unlimited service. *Journal of Statistical Planning and Inference*, 58:241261.

Armero, C. and Conesa, D. (2006). Bayesian hierarchical models in manufacturing bulk service queues. *Journal of Statistical Planning and Inference*, 136:335354.

Avramidis, A. N., Deslauriers, A., and L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908.

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

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

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.

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters*, 49:155–161.

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.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management*, 5(2):79–141.

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.

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

Hassin, R. and Haviv, M. (1995). Equilibrium strategies for queues with impatient customers. *Operations Research Letters*, 17:41–45.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univeriate Distributions: Volume 1.* John Wiley & Sons.

Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318.

Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, (113):4159.

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

Lindley, D. V. (2006). *Understanding Uncertainty.* Wiley-Interscience.

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.

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. *Queueing Systems*, 1:317333.

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. *Queueing Systems*, 1:335353.

Morali, N. and Soyer, R. (2003). Optimal stopping in software reliability. *Naval Research Logistics*, 50.

Mudholker, G. S. and Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactitons on Reliability*, 42(2).

Nassar, M. and Eissa, F. H. (2003). On the exponentiated Weibull distribution. *Communications in Statistics*, 32(7):1317–1336.

Nassar, M. and Eissa, F. H. (2004). Bayesian estimation for the exponentiated Weibull model. *Communications in Statistics*, 33(10):2343–2362.

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

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

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.

Shen, H. and Huang, J. Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management*, 10(3):391–410.

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

Smith, A. F. M. and Gelman, A. E. (1992). Bayesian statistics without tears: A Sampling perspective. *The American Statistician*, 46(2):8488.

Smith, R. and Miller, J. (1986). A non-Gaussian state space model and application to prediction of records. *Journal of the Royal Statistical Society. Series B*, 48(1):79–88.

Soyer, R. and Tarimcilar, M. (2008). Modeling and analysis of call center arrival data: A Bayesian approach. *Management Science*, 54(2):226–278.

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

Weinberg, J., D.Brown, L., and Stroud, J. R. (2007). Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. *Journal of the American Statistical Association*, 102(102).

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

Whitt, W. (1999a). Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212.

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

Whitt, W. (1999c). Predicting queueing 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*, 69:6579.

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.