The Institute for Integrating Statistics in Decision Sciences

Technical Report TR-2016-1

Bayesian Analysis of Markov Modulated Queues with Abandonment

Joshua Landon
Department of Statistics
The George Washington University, USA

Suleyman Ozekici
Department of Industrial Engineering
Koc University, Turkey

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

Bayesian Analysis of Markov Modulated Queues with Abandonment

Joshua Landon*, Süleyman Özekici† and Refik Soyer‡

October 13, 2016

Abstract

We consider a Markovian queueing model with abandonment where customer arrival, service and abandonment processes are all modulated by an external environmental process. The environmental process depicts all factors that affect the exponential arrival, service, and abandonment rates. Moreover, the environmental process is a hidden Markov process whose true state is not observable. Instead, our observations consist of only of customer arrival, service and departure times during some period of time. The main objective is to conduct Bayesian analysis in order to infer the parameters of the stochastic system. This also includes the unknown dimension of the environmental process. We illustrate the implementation of our model and the Bayesian approach by using actual data on call centers.

Keywords. Queueing, Erlang-A model, Hidden Markov model, Call centers, Bayesian inference, Gibbs sampler

1 Introduction

Research on queueing models provide perhaps the oldest and largest pool of literature within operations research, management science, and related fields. The number of articles published on theory and applications as well as the variety of the models considered in these articles are truly amazing. This is undoubtedly due to

*The George Washington University, Department of Statistics, Washington, DC 20052, USA

†Koç University, Department of Industrial Engineering, 34450 ˙Istanbul, Turkey

‡Corresponding Author, The George Washington University, Department of Decision Sciences, Washington, DC 20052, USA (soyer@gwu.edu)

the fact that many real life situations can be represented by queueing models. In recent years, there seems to be an additional interest that activates this field that has been somewhat stagnant for some time. The emergence of new service systems, like call centers and health care, is the main motivation behind this surge. In this paper, we revisit a typical Markovian queue in order to perform a tractable statistical analysis with emphasis on Bayesian learning based on observed data on the system.

Bayesian methods in queues have been considered in literature starting in late 80s where initial work focused on Markovian queues; see for example, [15], [16], and [3]. Extensions to Er/M/1 and Er/M/s queues have been introduced by Wiper [29] who used Monte Carlo methods to obtain operating characteristics of the systems. Rios-Insua et al. [26] developed Bayesian analysis for M/Er/1 and M/H_k/1 queues. Queues with bulk arrivals were analyzed in [4] and [5] using a Bayesian approach. More recently, Ramirez-Cobo et al. [25] introduced Bayesian methods to analyze queues where arrival and/or service distributions with heavy tails and Bayesian analysis of queuing networks was considered by [28]. However, to the best of our knowledge Bayesian analysis of Markov modulated queues with abandonment has not been previously considered.

An important feature of our models is that it involves modulation by a hidden Markov process that represents the possibly unobserved environmental conditions that affect the arrival, abandonment, and service processes. The concept of modulation is now widely accepted and used in a variety of areas in order to make the models more realistic. This introduces additional uncertainty, yet the models remain computationally tractable in many cases. The main idea behind modulation is that the model parameters change randomly with respect to some environmental process that affects these parameters. In the context of queueing models with impatient customers who abandon the queue after the waiting time exceeds a random patience treshold, there is randomness in the customer arrival, service and abandonment processes. For a Markovian queue, interarrival times, service durations, and patience tresholds all have exponential distributions. But, their rates are not necessarily constant over time and, in many cases, they change randomly depending on various environmental factors. We will focus on such a model where the environmental factors form a hidden Markov process the states of which are not necessarily observed. An important outcome of modulation in queueing applications is that the arrival, service and abandonment processes are dependent due to the common environmental factors that they are both subjected to. This is indeed an important contribution by itself to a literature which almost unanimously supposes independence of these processes. In queueing networks, modulation makes it possible to model dependent arrival processes to the different nodes. It is only natural to expect such dependence, since an environmental variation that increases arrivals to a node will very likely have the same (or reverse) effect on another entry node of the

network.

Markovian queues with abandonment have gained considerable attention in call centers where abandonment rate is considered a proxy for quality of service; see for example, Aktekin and Soyer [1]. Implications of abandonments in call center design is discussed by Garnett, Mandelbaum, and Reiman [20]. A queue with Markovian arrivals, service and abandonments, is known as a M/M/c+M queue. It is also referred to as an *Erlang-A model* in the literature. Mandelbaum and Zeltyn [14] present results on the Erlang-A models. A Bayesian analysis of the Erlang-A models is considered in Aktekin and Soyer [1].

There exist considerable literature on Markov modulated queues (MMQ). Prabhu and Zhu [23] discusses such systems and provides a survey of earlier papers including Eisen and Tainiter [10], Neuts [18], and Purdue [24]. Zhu [30] discusses MMQ networks and shows that the steady-state distribution of the queue length has a product form solution. Although we will not be considering queues in discrete time, we should mention in passing that there has been more interest in modulation of discrete-time queues in recent years. The concept of modulation by a random environment does not only apply to queueing models. There are other stochastic models where modulation is used. Arifoğlu and Özekici [2] analyze an inventory model operating in a partially observable random environment where the demand process is modulated by a process that represents the stochastic variations in an economy. First consideration of modulation in software reliability applications is due to Özekici and Soyer [21] who assume that the failures of the software depend on its operational profile, which is now the environmental process that represents the sequence of operations that the software performs. In a recent article, Landon et al. [13] present a tractable Bayesian approach Markov modulated Poisson model for software reliability. Applications also include hardware reliability where a device performs a stochastic mission and its failure rate depends on the stage of the mission. Çekyay and Özekici [8] discuss issues related to mean time to failure and availability when the mission or environmental process is semi-Markovian. Finally, modulation also occurs in portfolio optimization problems when the random asset returns are modulated in a so-called "regime-switching" market, as in Çanakoğlu and Özekici [7].

Our primary objective is to conduct Bayesian analysis of Markovian queues modulated by a hidden Markov process based on observed data of customer arrivals, services, and abandonments. The inference will include not only the arrival, service, and abandonment rates of the customers, but the holding rates and the transition probabilities of the hidden Markov process. Our analysis will also focus on the unknown number of states of the environmental process. The details of our model will be presented in Section 2 where the stochastic structures of the modulating and queueing processes are described. In Section 3 we will assume that the number of

states of the hidden process is known, and show how we can estimate the arrival, service and abandonment rates as well as the transition rates of the Markov process. Then, in Section 4 we will consider the case where we do not know the number of states of the hidden Markov process, and will present an approach to obtain the marginal likelihood based on Chib [9] that will enable us to infer the unknown number of states. Finally, our results will be demonstrated using actual arrival, service and abandonment data from a call center in Section 5.

2 Markov Modulated Queueing Model

Let $Z = \{Z_t; t \geq 0\}$ be a modulated queueing process such that Z_t depicts the total number of customers in the system at time t. Moreover, $N = \{N_t; t \geq 0\}$ is the customer arrival process where N_t denotes the total number of arrivals until time t, and $\Gamma = \{\Gamma_n; n = 1, 2, \cdots\}$ is the service process where Γ_n is the duration of the service for the nth customer. Moreover, $X = \{X_n; n = 1, 2, \cdots\}$ is the patience process where X_n is the patience for the nth customer such that the customer abandons the queue if the waiting time in the queue exceeds this random treshold. The model is Markovian in the sense that interarrivals, service durations, and patience tresholds have exponential distributions with randomly changing rates that depend on the state of an environmental process that modulates the queueing system. The environmental process represents the prevailing conditions or factors that affect the arrival, service, and abandonment rates. We suppose that there are c servers that work in parallel with the "first-come first-served" service discipline. Therefore, our model can be classified as a Markov modulated $M/M/c$ queue with abandonment.

The environmental process is $Y = \{Y_t; t \geq 0\}$ where Y_t represents the state of the environmental at time t, and Y is a latent or hidden process. We assume that $Y = \{Y_t; t \geq 0\}$ is a continuous-time Markov process with a finite state space $E = \{1, 2, \cdots, K\}$ where K is the number of states. The modulation is such that when the state of the environment is $i \in E$, customers arrive according to an ordinary Poisson process with rate λ_i, the service rate of each working server is μ_i, and the abandonment rate of each customer waiting in the queue is θ_i. Therefore, the random rate of arrivals at time t is λ_{Y_t} while the random service rate is μ_{Y_t} and the random abandonment rate is θ_{Y_t}. Since Y is a Markov process, it is well-known that the sequence of states visited by Y form Markov chain with state space E and some transition matrix P such that $P_{ii} = 0$ for all i. Moreover, the amount of time spent in any state i is exponentially distributed with some holding rate ρ_i. Therefore, the transition rate matrix or generator of the Markov process Y is

$$G_{ij} = \begin{cases} -\rho_i & \text{if } j = i \\ \rho_i P_{ij} & \text{if } j \neq i \end{cases} \tag{1}$$

or $G_{ij} = \rho_i(P_{ij} - I_{ij})$ where I_{ij} is (i, j) element of the identity matrix I.

It is clear that the arrival process is a Markov modulated Poisson process (or a doubly stochastic Poisson process) and we refer the reader to Özekici and Soyer [22] for a detailed discussion on these processes. In particular, we can write

$$P[N_t = k|Y] = \frac{e^{-A_t} A_t^k}{k!} \tag{2}$$

where

$$A_t = \int_0^t \lambda(Y_s)ds \tag{3}$$

for all $k = 0, 1, \cdots$ and $t \geq 0$. It follows that, given Y, the customer arrival process N is a nonstationary Poisson process with mean value function $E[N_t|Y] = A_t$. Letting $T = \{T_n; n = 0, 1, 2, \cdots\}$ denote the arrival time process so that T_n is the time of the nth arrival, we have the conditional distribution

$$P[T_{n+1} - T_n > t|Y, T_n] = e^{-(A_{T_n+t} - A_{T_n})}. \tag{4}$$

The modulated process reduces to the ordinary Poisson process with rate λ if the arrival rate vector is $\lambda_i = \lambda$ independent of the state of Y. In this case, $A_t = \lambda t$ deterministically.

It is clear that the bivariate process $(Y, Z) = \{(Y_t, Z_t); t \geq 0\}$ is a Markov process with state space $E \times \mathbb{N}$ where $\mathbb{N} = \{0, 1, 2, \cdots\}$ is the set of all nonnegative integers. Moreover, the generator Q of (Y, Z) is

$$Q_{(i,n),(j,m)} = \begin{cases} \rho_i P_{ij} & j \neq i, m = n \\ \lambda_i & j = i, m = n + 1 \\ (n \curlywedge c)\mu_i + (n - c)^+\theta_i & j = i, m = n - 1 \\ -(\rho_i + \lambda_i + (n \curlywedge c)\mu_i + (n - c)^+\theta_i) & j = i, m = n \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

for all $i, j \in E$ and $n, m \in \mathbb{N}$ with $n \geq 1$. Here and later, we use the notation $a \curlywedge b = \min\{a, b\}$ and $(a - b)^+ = \max\{a - b, 0\}$ for any real numbers a and b. If $n = 0$ and the system is empty, Q still satisfies (5), but state $m = n - 1 < 0$ is not possible. One can also easily obtain the generator for some special cases and variations. For example, when $c = +\infty$, we get $n \curlywedge c = n, (n - c)^+ = 0$ and we have the modulated $M/M/+\infty$ queue where there is no abandonment. Taking $c = 1$ yields the modulated $M/M/1$ queue with abandonment. Note that the size of the matrix Q may be infinite since the cardinality of $\mathbb{N}$ is infinite. It will become computationally tractable if it has a finite dimension. This will indeed be the case if there is some finite capacity C such that customers can not enter the system if there are C customers already present. In this case, we have the Markov modulated $M/M/+\infty/C$ queue and it suffices to define the generator by (5) for all $i, j \in E$ and

$n, m \in \{0, 1, 2, \cdots, C\}$ with $1 \leq n < C$. One must however note that $m = n - 1 < 0$ is not possible when $n = 0$, and $m = n + 1 = C + 1$ is not possible when $n = 0$. The remedy for the latter case is to set $\lambda_C = 0$ in (5).

Since (Y, Z) is a Markov process with generator matrix Q, its transition function

$$P_{(i,n),(j,m)}(t) = P[Y_t = j, Z_t = m | Y_0 = i, Z_0 = n] \tag{6}$$

is given by the exponential matrix defined via the Taylor expansion

$$P(t) = \exp(Qt) = \sum_{n=0}^{+\infty} \frac{t^n}{n!} Q^n. \tag{7}$$

The exponential matrix form of the transition function is very useful in computations on the queueing system since there are many tractable procedures that can be used. Moler and van Loan [17] presents many methods along this direction. Moreover, these matrices are very useful in the analysis of stochastic models with Markovian modulation. We refer the reader to Neuts [19] for details and various results on the exponential matrix that we will be using in our analysis. Asmussen [6] provides a survey on Markovian point processes and discusses how they are used in applied probability calculations.

It is now a direct consequence of (6) that the transient distribution of the number of customers in the system at any time t is

$$P[Z_t = m | Y_0 = i, Z_0 = n] = \sum_{j \in E} P_{(i,n),(j,m)}(t) \tag{8}$$

so that the expected number of customers becomes

$$E[Z_t | Y_0 = i, Z_0 = n] = \sum_{m \in \mathbb{N}} \sum_{j \in E} m P_{(i,n),(j,m)}(t). \tag{9}$$

One can also obtain similar formulas for unconditional probabilities and expectations. For example if $\alpha_i = P[Y_0 = i]$ is the initial distribution of Y and the system is empty at time 0, then

$$P[Z_t = m] = \sum_{i,j \in E} \alpha_i P_{(i,0),(j,m)}(t). \tag{10}$$

If the Markov process (Y, Z) is ergodic, then the steady-state distribution

$$\pi_{(j,m)} = \lim_{t \to +\infty} P[Y_t = j, Z_t = m | Y_0 = i, Z_0 = n] \tag{11}$$

can be computed by solving the system of linear equations $\pi Q = 0$ with $\sum_{j \in E, m \in \mathbb{N}} \pi_{(j,m)} = 1$. Then, it follows that the steady-state distribution of the number of customers is

$$\lim_{t \to +\infty} P[Z_t = m] = \sum_{j \in E} \pi_{(j,m)}. \tag{12}$$

with expectation

$$L = \lim_{t \to +\infty} E[Z_t] = \sum_{m \in \mathbb{N}} \sum_{j \in E} m \pi_{(j,m)}. \tag{13}$$

Moreover, using Little's formula, the average waiting time in the system becomes $W = L/\bar{\lambda}$ where the effective arrival rate is

$$\bar{\lambda} = \sum_{m \in \mathbb{N}} \sum_{j \in E} \lambda_j \pi_{(j,m)}. \tag{14}$$

In summary, using the matrix exponential form (7) one can obtain many important performance measures associated with the transient and ergodic behaviour of the MMQ system. Our primary objective is to develop statistical inference for the MMQ system using a Bayesian framework and obtain the performance measures for the system. It is important to note that in our model the Y process is latent and, therefore, in addition to the unknown parameters we also need to make inference about the latent states.

3 Bayesian Analysis of the MMQ

In this section we will illustrate how we can estimate all the parameters as well as the latent states in the MMQ model. The approach is based on an extension of the Markov Chain Monte Carlo (MCMC) method given in Fearnhead and Sherlock [11] for pure birth processes. This method is based on a Gibbs sampler and requires a three-stage process. We first introduce some notation that will be used in describing the three stages. We denote the customer arrival rates as $\lambda = \{\lambda_i; i \in E\}$, service rates as $\mu = \{\mu_i; i \in E\}$, abandonment rates as $\theta = \{\theta_i; i \in E\}$, holding rates of the Y process as $\rho = \{\rho_i; i \in E\}$, and transition probabilities as $\mathbf{P} = \{P_{ij}; i, j \in E\}$. Since there are K states in E and $P_{ii} = 0$ the total number of parameters to be estimated is $K(K-2) + 4K = K(K+2)$.

We suppose without loss of generality that $Z_0 = 0$ and the system is empty at time 0 for simplicity. We assume that the system is observed until some time t_{obs} to obtain the dataset $\mathcal{D} = \{z_t; 0 \leq t \leq t_{obs}\}$ where z_t is the number of customers observed to be in the system at time t. During the observation period, we suppose that n customers arrived while $m \leq n$ customers have departed either due to service completion or abandonment. Note that our observations contain all the information on arrival times, service durations, patience tresholds as well as customer departure times and the number of customers in the system and service during $[0, t_{obs}]$. We let $t^{(1)}, t^{(2)}, \cdots, t^{(n+m)}$ denote the times at which there has been a change in z_t during $[0, t_{obs}]$. Clearly, these are exactly those times at which there has been an arrival or a departure. To simplify the notation, we will set $z^{(k)} = z_{t^{(k)}}$ denote the number of

customers in the system after the kth change of state. For any time $0 \leq t^{(k)} \leq t_{obs}$, we further define binary indicators a_k, b_k, and c_k as 1 or 0 depending on whether the kth change is due to an arrival, departure after service completion or departure due to abandonment respectively. It is clear that $\{t^{(k)}\}, \{z^{(k)}\}$, $\{a_k\}, \{b_k\}$ and $\{c_k\}$ are all contained in the history or data set $\mathcal{D}$.

In Stage 1, we will simulate the state of the hidden Markov process at each of the event times $t^{(1)}, t^{(2)}, \cdots, t^{(n+m)}$ given in our data set $\mathcal{D}$. Thus, all the results are conditional on the parameters $\lambda, \mu, \rho, \theta$, and $\mathbf{P}$. In Stage 2, we will simulate the entire hidden Markov process, and in Stage 3 we will simulate a new set of parameter values using the Gibbs sampler. In the rest of this section, we will outline what each of the three stage involves.

Stage 1: Simulation of the hidden Markov process at event times.

At any event time $t^{(k)}$, we also observe whether that event is an arrival ($a_k = 1$) or departure after service completion ($b_k = 1$) or departure due to abandonment ($c_k = 1$). It is clear that $a_k + b_k + c_k = 1$ for all k. Moreover, as long as the state of the Markov process is i, arrivals occurs exponentially with rate λ_i, departures occur exponentially with rate $(z \curlywedge c)\,\mu_i$, and abandonments occur exponentially with rate $(z - c)^+ \theta_i$ when z customers are present. We now define diagonal matrices

$$\Lambda_{ij} = \begin{cases} \lambda_i, & \text{if } j = i \\ 0, & \text{if } j \neq i \end{cases}, \Pi_{ij} = \begin{cases} \mu_i, & \text{if } j = i \\ 0, & \text{if } j \neq i \end{cases}, \Theta_{ij} = \begin{cases} \theta_i, & \text{if } j = i \\ 0, & \text{if } j \neq i. \end{cases} \tag{15}$$

We also let the matrix exponential

$$\mathbf{T}^{(k)} = \exp\left[\left(\mathbf{G} - \left(\mathbf{\Lambda} + \left(z^{(k-1)} \curlywedge c\right)\mathbf{\Pi} + \left(z^{(k-1)} - c\right)^+ \mathbf{\Theta}\right)\right)\left(t^{(k)} - t^{(k-1)}\right)\right] \tag{16}$$

represent the transition probabilities of the states of the hidden Markov process Y over the interval $(t^{(k-1)}, t^{(k)})$.

Since Y is a Markov process, the states $\{S_k = Y_{t^{(k)}}\}$ at times $0 = t^{(0)} \leq t^{(1)} \leq t^{(2)} \leq \cdots \leq t^{(n+m)} \leq t^{(n+m+1)} = t_{\text{obs}}$ satisfy the transition probabilities

$$\mathbf{T}^{(k)}_{s_{k-1}, s_k} = P[S_k = s_k,\ Z_t = Z_u \text{ for all } u, t \in (t^{(k-1)}, t^{(k)}) | S_{k-1} = s_{k-1}]. \tag{17}$$

The definition (17) implies that $\mathbf{T}^{(k)}_{i,j}$ is the probability that the state is j at time $t^{(k)}$ and no events (arrivals, departures, or abandonments) occurred during the interval $(t^{(k-1)}, t^{(k)})$ given that the state is i at time $t^{(k-1)}$. Then, (16) follows by noting that arrivals occur exponentially with rate λ_i, departures due to service completions occur exponentially with rate $\left(z^{(k-1)} \curlywedge c\right)\mu_i$, and abandonments occur exponentially with rate $\left(z^{(k-1)} - c\right)^+ \theta_i$ during $(t^{(k-1)}, t^{(k)})$. Since $t^{(n+m+1)} = t_{\text{obs}}$ is not an event time $\mathbf{T}^{(n+m+1)}$ gives the transition probabilities with no events during $(t^{(n+m)}, t^{(n+m+1)} = t_{\text{obs}})$.

We recursively define the matrices

$$\mathbf{A}^{(n+m+1)} = \mathbf{T}^{(n+m+1)} \tag{18}$$
$$\mathbf{A}^{(k)} = \mathbf{T}^{(k)} \left(a_k \mathbf{\Lambda} + b_k \left(z^{(k-1)} \curlywedge c \right) \mathbf{\Pi} + c_k \left(z^{(k-1)} - c \right)^+ \mathbf{\Theta} \right) \mathbf{A}^{(k+1)}$$

for $k = n+m, n+m-1, \cdots, 1$. Note that these matrices denote the likelihoods

$$\mathbf{A}^{(k)}_{s_{k-1}, s_{n+m+1}} = P\left[\left\{ \left((a_l, b_l, c_l), t^{(l)} \right); l = k, k+1, \cdots, n+m \right\}, S_{n+m+1} = s_{n+m+1} | S_{k-1} = s_{k-1} \right].$$

So we first of all calculate $\{\mathbf{T}^{(k)}\}$ using (16), and then we calculate $\{\mathbf{A}^{(k)}\}$ using (18), starting with $\mathbf{A}^{(n+m+1)}$, and going backwards until we have

$$\mathbf{A}^{(1)}_{s_0, s_{n+m+1}} = P[\mathcal{D}, S_{n+m+1} = s_{n+m+1} | S_0 = s_0].$$

We will assume that we know $S_0 = s_0$ and $S_{n+m+1} = s_{n+m+1}$, the states of the Markov process at times 0 and t_{obs}, respectively. If they are unknown then we can adjust this algorithm slightly by putting a prior distribution on the state of the process at these times, but in our example we will assume that these states are known. Then, the state S_k of the Markov chain at time $t^{(k)}$ can be simulated using the conditional distribution $P[S_k = s | \mathcal{D}, S_{k-1} = s_{k-1}, S_{n+m+1} = s_{n+m+1}]$ which is given by

$$\frac{T^{(k)}_{s_{k-1,s}} \left(a_k \lambda_s + b_k \left(z^{(k-1)} \curlywedge c \right) \mu_s + c_k \left(z^{(k-1)} - c \right)^+ \theta_s \right) A^{(k+1)}_{s, s_{n+m+1}}}{A^{(k)}_{s_{k-1}, s_{n+m+1}}}. \tag{19}$$

This is done recursively by proceeding forwards through the event times $t^{(1)}, \cdots, t^{(n+m)}$.

Stage 2: Complete simulation of the hidden Markov process.

After completing Stage 1 we will have our simulated states of the hidden Markov process $\{S_k\}$ at each of our observation times $\{t^{(k)}\}$. We will now use these to simulate the entire hidden Markov process Y. To do this we first of all simulate it over the interval $(t^{(0)}, t^{(1)})$, then $(t^{(1)}, t^{(2)})$ and so on until $(t^{(n+m)}, t^{(n+m+1)})$. The simulation over each interval is done using the uniformization of the Markov process Y supposing that $\rho = \max_{i \in E} \rho_i$ is finite. It is well-known (see, for example, Ross [27]) that the Markov process Y can be represented as a Markov chain $\hat{X}$ subordinated to a Poisson process $\hat{N}$ with arrival rate ρ so that $Y_t = \hat{X}_{\hat{N}_t}$ and

$$\begin{aligned} P[Y_t = s_t | Y_0 = s_0] &= P[\hat{X}_{\hat{N}_t} = s_t | \hat{X}_{\hat{N}_0} = s_0] \\ &= \sum_{n=0}^{+\infty} P[\hat{N}_t = n] P[\hat{X}_n = s_t | \hat{X}_0 = s_0] \\ &= \sum_{n=0}^{+\infty} \frac{e^{-\rho t} (\rho t)^n}{n!} \mathbf{M}^n_{s_0, s_t} \end{aligned}$$

where

$$\mathbf{M} = \frac{1}{\rho}\mathbf{G} + \mathbf{I}. \tag{20}$$

is the transition matrix corresponding to the Markov chain $\hat{X}$. Over any interval $(t^{(k-1)}, t^{(k)})$, we already obtained the simulated states $Y_{t^{(k-1)}} = s_{k-1}$ and $Y_{t^{(k)}} = s_k$ in Stage 1. Therefore, the conditional distribution of the number arrivals of $\hat{N}$ during $(t^{(k-1)}, t^{(k)})$ is

$$\begin{aligned} P[\hat{N}_{t^{(k)}} - \hat{N}_{t^{(k-1)}} = n \quad | Y_{t^{(k-1)}} = \quad & s_{k-1}, Y_{t^{(k)}} = s_k] \\ = \quad & \left(\frac{e^{-\rho(t^{(k)} - t^{(k-1)})} \left(\rho(t^{(k)} - t^{(k-1)})\right)^n}{n!} \right) \\ & \times \frac{\mathbf{M}^n_{s_{k-1}, s_k}}{\exp\left[\mathbf{G}(t^{(k)} - t^{(k-1)})\right]_{s_{k-1}, s_k}} \end{aligned} \tag{21}$$

since

$$P[Y_{t^{(k)}} = s_k | Y_{t^{(k-1)}} = s_{k-1}] = \exp\left[\mathbf{G}(t^{(k)} - t^{(k-1)})\right]_{s_{k-1}, s_k}.$$

Therefore, the number of arrivals $\hat{N}_{t^{(k)}} - \hat{N}_{t^{(k-1)}}$ can be simulated using the distribution (21). If simulation yields $\hat{N}_{t^{(k)}} - \hat{N}_{t^{(k-1)}} = r$, then the r arrival times $\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_r$ of $\hat{N}$ over the interval $(t^{(k-1)}, t^{(k)})$ are simulated by generating r uniform variates over $(t^{(k-1)}, t^{(k)})$ and ordering them. Now, we know that $Y_{t^{(k-1)}} = s_{k-1}$ and $Y_{t^{(k)}} = s_k$ and the states of hidden Markov process at $\hat{t}_1 \leq \hat{t}_2 \leq \cdots \leq \hat{t}_r$ are simulated recursively by using the conditional distributions

$$P[Y_{\hat{t}_j} = s | Y_{\hat{t}_{j-1}} = \hat{s}_{j-1}, Y_{t^{(k)}} = s_k] = \frac{\mathbf{M}_{\hat{s}_{j-1}, s} \mathbf{M}^{r-j}_{s, s_k}}{\mathbf{M}^{r-j+1}_{\hat{s}_{j-1}, s_k}} \tag{22}$$

for $j = 1, 2, \cdots, r$. For $j = 1$, one should set $\hat{t}_{j-1} = \hat{t}_0 = t^{(k-1)}$ and $\hat{s}_{j-1} = \hat{s}_0 = s_{k-1}$. It also follows from the conditional distribution (22) that $Y_{\hat{t}_r} = Y_{t^{(k)}} = s_k$ at the last time point when $j = r$ since $\mathbf{M}^0$ is the identity matrix.

Stage 3: Generation of new parameters using Gibbs sampling.

Having completed Stages 1 and 2, we should now have the entire simulated hidden Markov process, as well as our data $\mathcal{D}$ on the observed MMQ. Let $\mathcal{F} = \{Y_t; 0 \leq t \leq t_{n+1}\}$ denote the environmental process generated using the procedure in Stage 2. Thus, we can write out our conditional likelihood function of the parameters and then obtain the full conditionals to generate a new set of values for our parameters at each step of the Gibbs sampler.

Let τ_i be the total time that the hidden Markov process spends in state i, n_i be the total number of customer arrivals during τ_i, τ_i^{μ} be the total amount of time that

the c servers actually spent serving customers during state i, and n_i^μ be the total number of services completed during state i. We let τ_i^θ denote the total amount of time that the customers, who abandoned as well as those who received service, had spent waiting during state i. It is important to note that the abandonment time for those who received service is censored. In other words, we only know that they have not abandoned during their waiting time for service. We denote the total number of abandonments during state i by n_i^θ and the number of times the environmental process makes a transition from state i to state j by r_{ij}.

It is clear that $\tau_i, n_i, \tau_i^\mu, n_i^\mu, \tau_i^\theta, n_i^\theta$ and r_{ij} are in $\mathcal{F} \cup \mathcal{D}$ for all $i, j \in E$. Given data $\mathcal{D}$ and the entire history $\mathcal{F}$ of the Markov process simulated in Stage 2, the conditional likelihood function, $\mathcal{L}(\lambda, \mu, \rho, \mathbf{P}; \mathcal{F}, \mathcal{D})$, of the parameters λ_i's, μ_i's, θ_i's, ρ_i's, and P_{ij}'s is given by

$$\prod_{i \in E} \left[\left(\rho_i^{\sum_{j \in E} r_{ij}} \exp(-\rho_i \tau_i) \right) (\lambda_i^{n_i} \exp(-\lambda_i \tau_i)) \left(\mu_i^{n_i^\mu} \exp(-\mu_i \tau_i^\mu) \right) \left(\theta_i^{n_i^\theta} \exp(-\theta_i \tau_i^\theta) \right) \prod_{j \in E} P_{ij}^{r_{ij}} \right].$$

Assuming conjugate independent priors for the unknown parameters the full conditional distributions can be easily obtained. More specifically, for a given state $i = 1, \ldots, K$, we assume independent gamma priors for λ_i's, μ_i's, θ_i's and ρ_i's, denoted as $\lambda_i \sim \mathcal{G}(a_i^\lambda, b_i^\lambda)$, $\mu_i \sim \mathcal{G}(a_i^\mu, b_i^\mu)$, $\theta_i \sim \mathcal{G}(a_i^\theta, b_i^\theta)$,and $\rho_i \sim \mathcal{G}(a_i^\rho, b_i^\rho)$ respectively. For the ith row of the transition matrix P, we assume a Dirichlet prior, independent of the other rows, as $\mathbf{P}_i \sim \mathcal{D}ir(\alpha_{i1}, \ldots, \alpha_{iK})$ where $\mathbf{P}_i = (P_{i1}, \ldots, P_{iK})$. Note that in $\mathbf{P}_i$ we have $P_{ii} = 0$ and the corresponding parameter $\alpha_{ii} = 0$. Using standard Bayesian results we can show that given the full history of the hidden Markov process, the full conditional distributions of the parameters can be obtained as

$$\lambda_i | \lambda_i^{-i} \sim \mathcal{G}(a_i^\lambda + n_i, b_i^\lambda + \tau_i), \;\; \mu_i | \mu_i^{-i} \sim \mathcal{G}(a_i^\mu + n_i^\mu, b_i^\mu + \tau_i^\mu),$$

$$\theta_i | \theta_i^{-i} \sim \mathcal{G}(a_i^\theta + n_i^\theta, b_i^\theta + \tau_i^\theta), \;\; \rho_i | \rho_i^{-i} \sim \mathcal{G}(a_i^\rho + \sum_{j \in E} r_{ij}, b_i^\rho + \tau_i)$$

and

$$\mathbf{P}_i | \mathbf{P}_i^{-i} \sim \mathcal{D}ir(\alpha_{i1} + r_{i1}, \ldots, \alpha_{iK} + r_{iK})$$

where $\alpha_{ii} = r_{ii} = 0$.

We then generate new values for these parameters from their posterior distribution and then repeat the whole process again, starting with Stage 1.

4 Assessment of the Number of Environmental States

Our analysis in Section 3 assumed that the number of states K in the hidden Markov process was known. However, in general, the actual number of states may be unknown to us, so it is important to be able to determine how many states there are. The problem of determining K can be considered as a model selection problem in the Bayesian approach where the model choice is made using Bayes factors; see Kass and Raftery [12] for a review. The computation of the Bayes factors requires the evaluation of marginal likelihood for a given model, that is, for given value of K in our case. More specifically, if we let $\mathcal{D}$ denote our observed data, we want to obtain the marginal likelihood $p(\mathcal{D}|K)$. The model with the highest value of $p(\mathcal{D}|K)$ is the one most supported by the data and this can be used as the criterion for determining the value of K. Alternatively, assuming a support for K and specifying prior probabilities $P[K = k]$ for different models such that $\sum_k P[K = k] = 1$ we can obtain posterior model probabilities $P[K = k|\mathcal{D}]$ using the marginal likelihood.

Evaluation of the marginal likelihood $p(\mathcal{D}|K)$ analytically is not possible in many problems since it requires integrating out the unknown parameters. Since draws from prior distributions of the parameters result in unstable estimation, the use of Monte Carlo methods emphasize use of posterior Monte Carlo samples to evaluate $p(\mathcal{D}|K)$. Although this is not straightforward in many cases, when the full posterior conditional distributions are known forms, the marginal likelihood terms can be approximated using the approach proposed by Chib [9]. Since the Bayesian analysis of the MMQ in Section 3 is based on known full conditionals, we can adopt Chib's procedure to our problem as will be discussed in the sequel.

In our case, the marginal likelihood for a specific model with dimension K is given by

$$p(\mathcal{D}) = \frac{p(\mathcal{D}|\lambda, \mu, \theta, \rho, \mathbf{P}, \mathcal{F})\ p(\lambda, \mu, \theta, \rho, \mathbf{P}, \mathcal{F})}{p(\lambda, \mu, \theta, \rho, \mathbf{P}, \mathcal{F}|\mathcal{D})} \tag{23}$$

where λ, μ, θ and ρ are K-dimensional vectors of λ_i's, μ_i's, θ_i's and ρ_i's and $\mathbf{P}$ is the transition probability matrix of dimension K with zeros on the diagonal. We can rewrite (23) as

$$p(\mathcal{D}) = \frac{p(\mathcal{D}|\lambda, \mu, \theta, \rho, \mathbf{P}, \mathcal{F})\ p(\mathcal{F}|\rho, \mathbf{P}) p(\lambda, \mu, \theta, \rho) p(\mathbf{P})}{p(\lambda, \mu, \theta, \rho, \mathbf{P}|\mathcal{F}, \mathcal{D}) p(\mathcal{F}|\mathcal{D})}. \tag{24}$$

Equation (24) holds for any values of $(\lambda, \mu, \theta, \rho, \mathbf{P}, \mathcal{F})$ such as $(\lambda^*, \mu^*, \theta^*, \rho^*, \mathbf{P}^*, \mathcal{F}^*)$ which is typically chosen as the mean or mode values of the posterior distributions. We note that all the terms in the numerator are available to us analytically and therefore can be evaluated at $(\lambda^*, \mu^*, \theta^*, \rho^*, \mathbf{P}^*, \mathcal{F}^*)$.The tricky part to evaluate is

the second term in the denominator

$$p(\mathcal{F}^*|\mathcal{D}) = \int p(\mathcal{F}^*|\mathcal{D}, \lambda, \mu, \theta, \rho, \mathbf{P}) p(\lambda, \mu, \theta, \rho, \mathbf{P}|\mathcal{D}) d(\lambda, \mu, \theta, \rho, \mathbf{P})$$

which can be evaluated using G samples from the posterior distribution via

$$p(\mathcal{F}^*|\mathcal{D}) = \frac{1}{G} \sum_{g=1}^{G} p(\mathcal{F}^*|\lambda^{(g)}, \mu^{(g)}, \theta^{(g)}, \rho^{(g)}, \mathbf{P}^{(g)}, \mathcal{D}). \tag{25}$$

The first term $p(\lambda^*, \mu^*, \theta^*, \rho^*, \mathbf{P}^*|\mathcal{F}^*, \mathcal{D})$ in the denominator of (24) can easily be written down as product of gamma and Dirichlet densities. Thus, for each value of K, we can approximate (24) and determine the model with the highest support of the data. As previously mentioned, using the marginal likelihood we can also compute posterior model probabilities $P[K = k|\mathcal{D}]$ to infer the value of K.

5 Numerical Illustration

In this section, the implementation of the Bayesian approach for the Markov modulated Erlang-A model will be illustrated using actual call center data from an anonymous bank considered in Aktekin and Soyer [1]. The data includes all arrival, service and abandonment information for stock exchange customers whose abandonment times seem to exhibit an exponential type of behavior. For illustrative purposes, we have used data from a single day in our analysis and used an environmental process with $K = 2$ states. Thus, the transition matrix for the embedded Markov chain in our case is given by

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

For the two-states case the inference involves parameters $(\mu_1, \mu_2, \lambda_1, \lambda_2, \theta_1, \theta_2, \rho_1, \rho_2)$ as well as the the latent states of the environmental process. In our Bayesian analysis we use proper but highly diffused priors for the unknown parameters. Specifically, we choose $a_i^{\mu} = b_i^{\mu} = 0.01$, $a_i^{\lambda} = b_i^{\lambda} = 0.01$, $a_i^{\theta} = b_i^{\theta} = 0.01$,and $a_i^{\rho} = b_i^{\rho} = 0.01$ in the respective gamma distributions for $i = 1, 2$.

In the Gibbs sampler, after an initial burn-in run, 10,000 simulations were performed. No convergence problems were experienced in running the Gibbs sampler. The trace plots for $(\mu_1, \lambda_1, \theta_1, \rho_1)$ are shown in Figure 1. Similar behavior was also observed for the environment 2 parameters.

The resulting density plots for the posterior distributions of parameters under both environments are shown in Figure 2. We note that the density plots for $1/\rho_1$

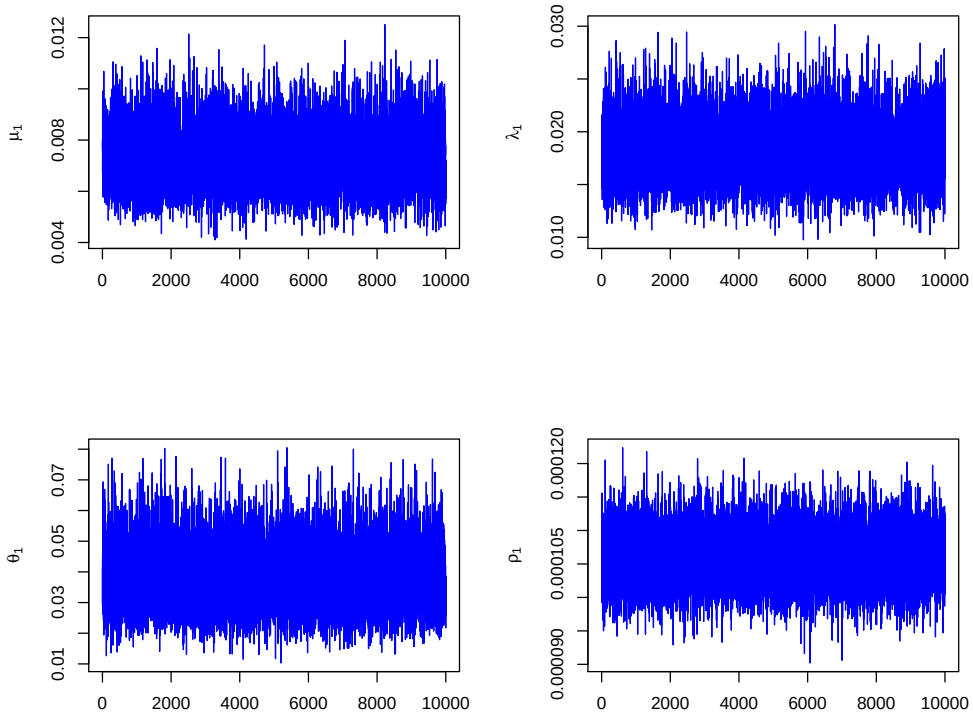

Figure 1: Trace plots of Gibbs sampling for μ_1 λ_1, θ_1 and ρ_1.

and $1/\rho_2$ are shown in the figure as they represent the expected holding times for the states of the environmental process. We note that expected holding time in state 2 is lot larger than that in state 1. In fact posterior probability that $\rho_1 > \rho_2 \approx 1$. We can see from the figure that service rate under environment 1 is higher than that under environment 2. The posterior probability that $\mu_1 > \mu_2 \approx 0.913$. Similarly, the posterior probability that $\lambda_2 > \lambda_1 \approx 0.884$. Finally, we can see from the figure that the abandonment rate under environment 1 is lot higher than under environment 2 with the posterior probability that $\theta_1 > \theta_2 \approx 0.983$. We can also see from Figure 2 that uncertainty for μ and θ is higher under environment 1 than under environment 2. This can be explained by the fact that the environmental process spends most of the time in state 2.

It is also possible to obtain the posterior correlations of μ, λ and θ under environments 1 and 2 using the joint posterior samples. We have observed that for both environments the posterior correlations for the three rates were less than 0.02 in absolute value suggesting their posterior independence. The same conclusion was reached by Aktekin and Soyer Aktekin and Soyer [1].

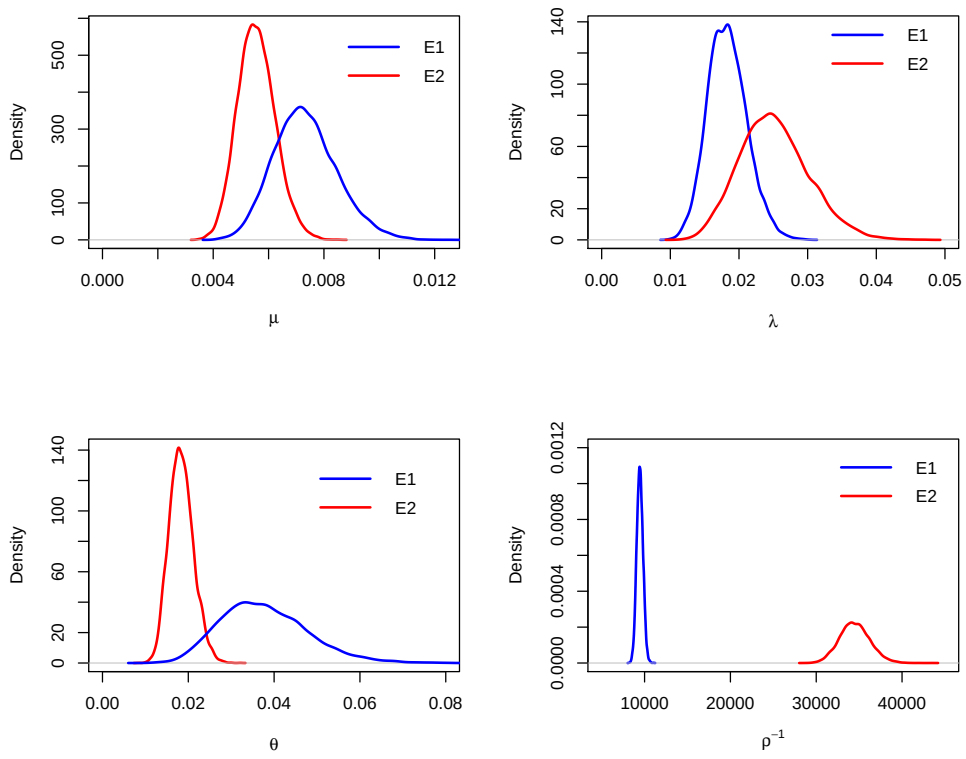

Figure 2: Posterior distributions of μ, λ, θ and $1/\rho$ for environments 1 (E1) and 2 (E2).

Using the posterior samples of the parameters we can evaluate the distribution of the number customers in the system given by (10) or the posterior steady-state distribution given by (12) as a Monte Carlo average. Evaluation of (10) requires the matrix exponential form in (7) for the generator matrix Q for each realization of the parameters. But this can be easily computed using the methods described in Moler and van Loan [17]. Figure 3 gives the posterior steady-state distribution of the number of people in the system. We can estimate the expected value of the steady-state distribution of the number of customers given by (13) as $L = 4.09$.

We can compare the case of two environmental states with the single environment case using marginal likelihoods as discussed in Section 4. In order to do this we can update the posterior distributions of single arrival service and abandonment rates independently using the data. The posterior distributions in each case can be obtained analytically as gamma densities. The comparison of the marginal likelihoods for $K = 1$ and $K = 2$ favors the two-state case.

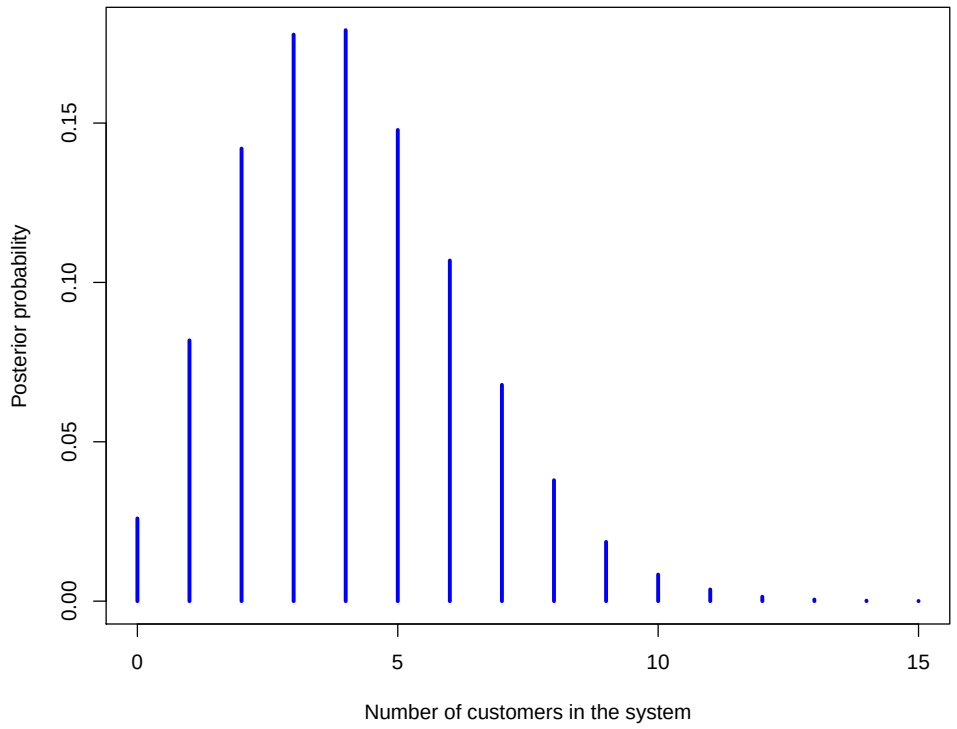

Figure 3: Posterior distribution of number of people in the system in the long-run.

6 Conclusion

In this paper we considered a Bayesian analysis of Markov modulated queueing systems with abandonment. We presented a Bayesian analysis of the queueing system using a Gibbs sampler and illustrated how Bayesian inferences can be made about the system measures such as the number of customer in the system at any time. We discussed how to assess the dimension of the hidden environmental process by computing the marginal likelihood of the data. We illustrated the implementation of our approach using real data from a call center with impatient customers.

References

[1] T. Aktekin and R. Soyer, *Bayesian analysis of queues with impatient customers: Applications to call centers*, Naval Research Logistics **59** (2012), 441–456.

[2] K. Arifoğlu and S. Özekici, *Optimal policies for inventory systems with finite capacity and partially observed Markov-modulated demand and supply processes*, European Journal of Operational Research **204** (2010), 421–483.

[3] C. Armero, *Bayesian inference in Markovian queues*, Queuing Systems **15** (1994), 419–426.

[4] C. Armero and D. Conesa, *Prediction in Markovian bulk arrival queues*, Queueing Systems **34** (2000), 327–350.

[5] ———, *Bayesian hierarchical models in manufacturing bulk service queues*, Journal of Statistical Planning and Inference **136** (2006), 335–354.

[6] S. Asmussen, *Matrix-analytic models and their analysis*, Scandinavian Journal of Statistics **27** (2000), 193–226.

[7] E. Çanakoğlu and S. Özekici, *Portfolio selection in stochastic markets with HARA utility functions*, European Journal of Operational Research **201** (2010), 520–536.

[8] B. Çekyay and S. Özekici, *Mean time to failure and availability of semi-Markov missions with maximal repair*, European Journal of Operational Research **207** (2010), 1442–1454.

[9] S. Chib, *Marginal likelihood from the Gibbs output*, Journal of the American Statistical Association **90** (1995), 1313–1321.

[10] M. Eisen and M. Tainiter, *Stochastic variations in queuing processes*, Operations Research **11** (1963), 922–927.

[11] P. Fearnhead and C. Sherlock, *An exact Gibbs sampler for the Markov-modulated Poisson process*, Journal of the Royal Statistical Society: Series B **68** (2006), 767–784.

[12] R.E. Kass and A.E. Raftery, *Bayes factors*, Journal of the American Statistical Association **90** (1995), 773–795.

[13] J. Landon, S. Özekici, and R. Soyer, *A Markov modulated Poisson model for software reliability*, European Journal of Operational Research **229** (2013), 404–410.

[14] A. Mandelbaum and S. Zeltyn, *Advances in service innovations*, ch. Service Engi- neering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers, pp. 17–48, Springer, 2007.

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

[16] _____, *A subjective Bayesian approach to the theory of queues ii - inference and information in M/M/1 queues*, Queuing Systems **1** (1987), 335–353.

[17] C. Moler and C. van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Review **45** (2003), 3–49.

[18] M.F. Neuts, *A queue subject to extraneous phase changes*, Advances in Applied Probability **3** (1974), 78–119.

[19] M.F. Neuts, *Matrix-geometric solutions in stochastic models, an algorithic approach*, Dover Publications, Inc., New York, 1994.

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

[21] S. Özekici and R. Soyer, *Reliability of software with an operational profile*, European Journal of Operational Research **149** (2003), 459–474.

[22] _____, *Semi-Markov modulated Poisson process: Probabilistic and statistical analysis*, Mathematical Methods of Operations Research **64** (2006), 125–144.

[23] N.U. Prabhu and Y. Zhu, *Markov-modulated queueing systems*, Queueing Systems **5** (1989), 215–246.

[24] P. Purdue, *The M/M/1 queue in a Markovian environment*, Operations Research **22** (1974), 562–569.

[25] P. Ramirez-Cobo, R. E. Lillo, S. Wilson, and M. P. Wiper, *Bayesian inference for double pareto lognormal queues*, Annals of Applied Statistics **4** (2010), 1533–1557.

[26] D. Rios-Insua, M. P. Wiper, and F. Ruggeri, *Bayesian analysis of M/Er/1 and $M/h_k/1$ queues*, Queuing Systems **30** (1998), 289–308.

[27] S. Ross, *Stochastic processes*, 2. ed., Wiley, New York, 1996.

[28] C. Sutton and M. I. Jordan, *Bayesian inference for queuing networks and modeling of internet services*, Annals of Applied Statistics **5** (2011), no. 1, 254–282.

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

[30] Y. Zhu, *Markovian queueing networks in a random environment*, Operations Research Letters **15** (1994), 11–17.