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

Technical Report TR-2012-7
April 2, 2012

Bayesian Analysis of the Discrete Time Single and Multiple Server Queues

Toros Caglar
Department of Decision Sciences
The George Washington University

Refik Soyer
Department of Decision Sciences
The George Washington University

Bayesian Analysis of the Discrete Time Single and Multiple Server Queues

Toros Caglar
Department of Decision Sciences
The George Washington University

Refik Soyer
Department of Decision Sciences
The George Washington University

April 2, 2012

Abstract

In this work, we develop a Bayesian framework for the analysis of a geometric discrete time queue with single and multiple servers. We bring inference methods for the system parameters as well as performance measures of the discrete time queues. We also release the assumption of stationarity in a system. In doing so, we construct a dependent structure between the arrival and service processes that creates ergodicity a priori. We also provide a framework to handle the inference and analysis of systems with batch arrivals in order to account for longer time slots. We finalize the paper by providing a numerical illustration of our methodologies.

1 Introduction

Applications of discrete time queueing models are especially abundant in computer and communication systems, where the time horizon is divided into equal-length slots. Meisling (1958) was the first to consider the simplest discrete time queues. The paper by Kobayashi and Konheim (1977) presents a review of discrete time queueing systems and their networks. A more recent discussion of these queues may be found in Daduna (2001).

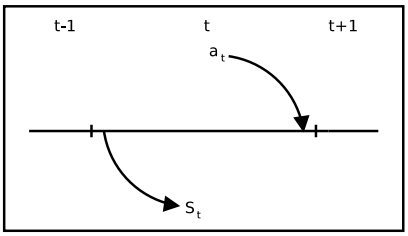

Figure 1.1: Late arrival queue scheme

Arrival and departure processes are assumed to obey one of the two predefined structures. The late arrival model in Figure 1.1, assumes that arrivals occur immediately before the end of a time slot. Therefore, even if the system is empty upon an arrival, the service does not start until the following time slot. The complementing model is the early arrival model, where an arrival occurs early in a time slot. This allows start of service in the same time slot if an arrival occurs to an empty system (Takagi, 1993). Both the early and the late arrival models allow us to treat the inter-arrival times as integer multiples of the length of the time slots. Similarly, the service times can only take values that are integral multiples of the time slots. In this work, we present our results based on the late arrival model with a first come first serve queue policy where the customers are served in the order they enter the queue. One can easily extend the methods presented here to the early arrival model if needed, but such extension will not be discussed here.

In most of the literature on discrete time queues, the systems considered can be viewed as being inherently discrete, due to dependence on cyclical processing devices such as CPUs. However, the application of discrete time queues are not limited to such systems. The advantages of utilizing discrete time queueing models in the analysis of continuous systems have been discussed dating as far back as Dafermos and Neuts (1971) and Neuts (1973). Even though time is a continuous construct, our interpretation of time can be discrete in many applications. For example, staff

schedules in emergency rooms and hospitals usually adhere to shifts made up of discrete blocks of time. We also discretize time to blocks (minutes, hours, days, weeks etc.) in communicating and recording it. Discrete time models will naturally present discrete results, which can be used without conversion in decision making problems concentrating on discrete time blocks. The computational requirements of a discrete time queueing model can be higher than its continuous counterpart in simple systems which have known closed form solutions. However, this disadvantage diminishes as one concentrates on models of higher complexity. It is also easier to incorporate supplementary variables into discrete queues, which may allow the representation of more realistic systems. Finally, with the increasing computational power that computers possess, analysis of discrete time queues are becoming much more accessible.

The foundations and the motivation behind Bayesian analysis of queues are covered very thoroughly in McGrath et al. (1987) and McGrath and Singpurwalla (1987). The main aspect that sets the Bayesian approach apart from the classical analysis of queues is the handling of the unknown parameters describing the system. As opposed to considering the parameters of the system describing the arrival and the service processes as fixed but unknown, the Bayesian approach describes them via probability distributions. This allows the usage of personal probabilities to describe the uncertainties about the parameters and coherently updating these uncertainties as more information is received. The probabilistic inference provided by the Bayesian framework also allows one to attack decision making problems through applications of Bayesian utility theory. For example, system design problems deciding on number of servers and investments on increasing service rates, one can construct the appropriate utility measures and provide solutions maximizing expected utility. The Bayesian approach also allows the predictions of unobserved quantities to be made easily. These

quantities are not limited to the future inter-arrival or service times, but also include performance measures of the queueing system such as the number of customers and waiting times. Armero and Bayyari (1994) derive closed form expressions for most performance measures of the M/M/1 queue. The authors then provide an analysis of the M/M/c queue in Armero and Bayyari (1994). Although the preceding works focus on continuous time queues, the main ideas regarding the inferential statistics and the Bayesian methodologies transfer perfectly to the discrete case.

The Bayesian analysis of discrete time queues is very scarce in the literature. In Conti (1999, 2004), the author concentrates on non-parametric Bayesian analysis of $Geo/G/1$ queues, motivated by ATM systems. Probability generating functions for the delay distributions are calculated and approximations used are shown to converge to true results under the large sample assumption. Such approximations are necessitated by the large buffer sizes the communication systems posses, which make exact computations difficult.

We will start this paper by first providing a simple framework for modeling single and multiple server discrete time queues. We will then introduce batch arrivals to handle longer time slots necessary in modeling continuous systems. We will also provide the performance measure calculations present in the literature for these queues. Then we will present the Bayesian inferential methods for the arrival and service rates of the queues discussed. In doing so, we will utilize conjugate priors for closed form results in the independent prior case. We will introduce a dependent prior case to relax the stationarity assumption prevalent in the literature. Finally, we will illustrate how to obtain the posterior predictive distributions of the queue performance measures and provide a numerical example. Through this paper, we aim to contribute to the discrete time queueing literature by providing inferential methods and a Bayesian framework for performance measurement that allows for probabilistic results and de-

pendent prior analysis. We also show that continuous time systems can be analyzed through the use of discrete time queues.

2 Modeling the discrete time queues

In discrete time queues, the time horizon is divided into slots of equal length Δt. We will consider two approaches to modeling the discrete time single server queues. The first approach will assume an experimental design where geometric inter-arrival and service times provide the information about the system. The second approach will collect arrival and service information through separate binary sequences, resulting in independent Bernoulli models describing the arrival and service processes of the system. Queues modeled using the first approach will be referred to as geometric queues and the second approach will be called Bernoulli queues in the rest of this paper.

2.1 The single and multiple server geometric queues

Meisling (1958) approached the discrete single server queueing system with the following assumptions.

1. No more than one customer may arrive at a given time-slot.
2. The arrival of a customer at a given time-slot is an event which is statistically independent of the arrival of customers at any previous slots. The probability of arrival of a customer at a time slot is a fixed value λ.
3. Customers are served in the order of their arrivals.
4. The service times for different customers are independent and identically distributed random variables, and independent from the arrival process.

By assumptions 1 and 2, the arrival process to the system can be described by a binomial probability distribution. That is,

$$P(m \text{ arrivals in } k \text{ time-slots}) = \begin{cases} \binom{k}{m}\lambda^m(1-\lambda)^{k-m}, & (0 \leq m \leq k) \\ 0, & (m > k) \end{cases}$$

We can also see that the time between arrivals follow a geometric distribution with parameter λ in this setup, and we have

$$P(\text{time between two arrivals is } k\Delta t) = (1-\lambda)^{k-1}\lambda.$$

Meisling's treatment of the arrival process conforms with the arrival process of a continuous time M/M/1 queue. The service process however, does not follow a similar geometric distribution. Instead, Meisling proposes a discrete probability distribution to the service times, $P(\text{service time is } k\Delta t) = s_k$, where $\sum_k s_k = 1$.

For a geometric service process, we will turn to the review paper by Kobayashi and Konheim (1977). Single server systems with geometric arrival and service processes are often referred to as Geo/Geo/1 in the literature. The geometric service process, like the exponential service process in the continuous M/M/1 enjoys the *memoryless property* and can be characterized as follows:

$$P(\text{service time is } k\Delta t) = (1-\mu)^{k-1}\mu$$

where μ is the probability of a service completion in a time slot. This implies that if the state of the system is described by the random process $\mathbf{N} = \{N_j : 0 \leq j \leq \infty\}$ where N_j is the number of customers in the system (waiting or in service) at time-slot j, then, $\mathbf{Q}$ given in (2.1) is a transition matrix representing the birth-death chain $\mathbf{N}$.

Births to the system occur if an arrival occurs, and no service is completed during a time slot, which has a probability of $u = \lambda(1-\mu)$ for all states but $N_j = 0$. The only way of leaving the empty state is in the case of an arrival in a time slot, with probability λ. The system stays empty if no arrivals occur, with probability $(1-\lambda)$. Deaths occur similarly, as a result of a service completion and non-arrival in a given time slot, with a probability of $d = \mu(1-\lambda)$. If, in a time slot, there are no arrivals and service completions, or both an arrival and a service completion, the number of customers in the system does not change, with probability $r = (1-\lambda)(1-\mu) + \lambda\mu$.

$$\mathbf{Q} = \begin{array}{c} \mathbf{N} \\ 0 \\ 1 \\ 2 \\ \dots \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & \dots \end{array} \\ \begin{pmatrix} 1-\lambda & \lambda & 0 & 0 & \dots \\ (1-\lambda)\mu & (1-\lambda)(1-\mu)+\lambda\mu & \lambda(1-\mu) & 0 & \dots \\ 0 & (1-\lambda)\mu & (1-\lambda)(1-\mu)+\lambda\mu & \lambda(1-\mu) & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix} \end{array} \tag{2.1}$$

In the case of the multiple server discrete time queue, the arrival process behaves exactly like the single server geometric queue. However in this model, there are c identical, geometric servers. Consequently, the service times of the customers are identically and independently distributed geometric random variables with parameter μ. As a result of these assumptions, we need to allow for multiple service completions in a single time slot. Due to this modification, the Markov chain describing the state of the system is more complicated.

Let $d_{i,j}$ denote the probability of the number of customers in the system going down by j, if the system has i servers occupied in the immediately preceding time slot, for $i \geq j$, $i > 0$, and $i \leq c$. This jump can occur in two ways. We either need to have no arrival to the system (with probability $1-\lambda$) and j service completions out of i occupied servers (binomial probability of j out of i), or, we have an arrival (with

probability λ) alongside $j+1$ service completions out of i servers. Also let r_i and u_i be the probabilities of a system with i servers occupied having the same number of customers and having one additional customer, respectively, in the immediately following time slot, for $i \geq 0$. Then, $r_0 = 1 - \lambda$ since the only way an empty system will stay empty is if no arrivals occurs in a time slot, and $u_0 = \lambda$ since the empty system can jump to a system with a single customer with an arrival only. For all other i, j pairs, we have:

$$
\begin{aligned}
d_{i,j} &= \binom{i}{j}\mu^j(1-\mu)^{i-j}(1-\lambda) + \binom{i}{j+1}\mu^{j+1}(1-\mu)^{i-j-1}\lambda \qquad (2.2)\\
r_i &= \mu(1-\mu)^{i-1}\lambda + (1-\mu)^i(1-\lambda)\\
u_i &= (1-\mu)^i\lambda
\end{aligned}
$$

We can then obtain the probability transition matrix modeling the Markov chain for the system size as given in (2.3):

$$
\mathbf{Q} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \dots \\ c \\ c+1 \\ \dots \end{array}
\begin{array}{c}
\begin{array}{cccccccccccc} \mathbf{N} & 0 & 1 & 2 & 3 & 4 & \dots & c-1 & c & c+1 & c+2 & c+3 & \dots \end{array}\\
\begin{pmatrix}
r_0 & u_0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots \\
d_{1,1} & r_1 & u_1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots \\
d_{2,2} & d_{2,1} & r_2 & u_2 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
d_{c,c} & d_{c,c-1} & \dots & \dots & \dots & \dots & d_{c,1} & r_c & u_c & 0 & \dots & \dots \\
0 & d_{c,c} & d_{c,c-1} & \dots & \dots & \dots & \dots & d_{c,1} & r_c & u_c & 0 & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots
\end{pmatrix}
\end{array}
\qquad (2.3)
$$

If the load factor $\rho = u/d < 1$, then the process $\mathbf{N}$ given in 2.1 is recurrent and has a stationary distribution which can be obtained by solving the system of equations $\mathbf{\Pi} = \mathbf{\Pi Q}$, where $\mathbf{\Pi}$ is the limiting distribution vector of the number of customers in the system. By letting π_i be the probability that there are i customers in a system that has reached its steady state, we can write the system of equations as shown in

(2.4).

$$\begin{aligned}
\pi_0 &= \pi_0(1-\lambda) + \pi_1\mu(1-\lambda) \\
\pi_1 &= \pi_0\lambda + \pi_1(\lambda\mu + (1-\lambda)(1-\mu)) + \pi_2\mu(1-\lambda) \\
\pi_2 &= \pi_1\lambda(1-\mu) + \pi_2(\lambda\mu + (1-\lambda)(1-\mu)) + \pi_3\mu(1-\lambda) \\
&\dots \\
\pi_i &= \pi_{i-1}\lambda(1-\mu) + \pi_i(\lambda\mu + (1-\lambda)(1-\mu)) + \pi_{i+1}\mu(1-\lambda) \\
&\dots
\end{aligned} \tag{2.4}$$

The probability distribution of the number of customers in the system can also be calculated by the fundamental birth-death process characteristics and given by (Daduna, 2001).

$$\pi_i = P(N=i) = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda(1-\mu)}{\mu(1-\lambda)}\right)^i \left(\frac{1}{1-\mu}\right)^{I'(0,i)},\ i \in \{0,1,2,...\} \tag{2.5}$$

where $I'(a,b)$ is the complementary indicator function defined as:

$$I'(a,b) = \begin{cases} 1\ if\ a \neq b \\ 0\ if\ a = b \end{cases} \tag{2.6}$$

The stationarity condition $u < d$ is equivalent to $\lambda < \mu$.

The waiting time distribution in a single server geometric queue can be obtained by conditioning on the number of customers an arrival sees when it enters the system at time t. Since we are assuming a late arrival system, in order to calculate the number of customers in front of an arrival, we need to consider the number of customers in the previous time slot, $t-1$, and subtract the number of customers served in the current time slot t. If an arrival occurs and the system was empty in the previous time slot, or there was a single customer in the system whose service just completed,

the customer's waiting time in the queue is zero since the service process will start in the immediately following time slot. If a customer enters the system at time t, and there were $n > 1$ customers in the system at time $t-1$, then one of two things will occur: Either the service of the customer in service will end in this time slot, and there will be $n-1$ other customers in front of the arriving customer, or no service will be completed in this time slot, meaning that the arriving customer will wait for n customers. The wait time associated with such a customer is the remaining service time of the customer in service, in addition to the service times of the customers in the queue. We know that the service time associated with the customers in the queue are distributed by independent geometric distributions with rate μ. Additionally, by the memoryless property of the geometric distribution, the remaining service time of the customer in service is also geometric with the same rate. Therefore, a customer entering a system at time t, of size n in the previous time slot, will wait an amount of time determined by the sum of n independent geometrically distributed random variables with probability $1-\mu$, or $n-1$ independent geometrically distributed random variables with probability μ which are negative binomial distributions with parameters (n, μ) and $(n-1, \mu)$, respectively.

$$P(W = w|N = n) = \binom{w+n-1}{n-1}(1-\mu)^n\mu^w, w \in \{0, 1, 2, \ldots\}, for\ n \in \{1, 2, \ldots\} \tag{2.7}$$

By the law of total probability, we can then obtain the marginal distribution of the waiting time of a customer as shown in (2.8). In the case of systems with inherently limited capacity, the summation argument in (2.8) becomes a finite sum which can be calculated exactly after a slight modification to $P(N = n)$ to take into account the limited capacity.

$$
\begin{aligned}
P(W=w) &= \sum_{n=1}^{\infty} P(W=w|N=n)P(N=n) \qquad (2.8)\\
&= \sum_{n=1}^{\infty} \binom{w+n-1}{n-1}(1-\mu)^n\mu^w\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda(1-\mu)}{\mu(1-\lambda)}\right)^n\left(\frac{1}{1-\mu}\right)^{I'(0,n)},\\
& w \in \{0,1,2,\ldots\}
\end{aligned}
$$

For the multiple server queue, the number of servers is denoted as c and the familiar equilibrium condition is given by $\lambda < c\mu$. The equilibrium distribution of the number of customers in the system, N, is not obtainable in closed form. The probability generating function of N presented in Gao et al. (2004) is given by

$$
\mathcal{N}(z) = \frac{\mathcal{A}(z)z^c\mathcal{H}(z)^c\sum_{i=0}^{c-1}[\mathcal{H}(z)^{i-c} - z^{i-c}]n(i)}{z^c - \mathcal{H}(z)^c\mathcal{A}(z)} \qquad (2.9)
$$

where $\mathcal{H}(z) = \mu + (1-\mu)z$ and constants $n(i)$ are probabilities defined as $n(i) \equiv P(N=i), i = 0,1,\ldots,c-1$. Under Bernoulli arrivals, $\mathcal{A}(z) = 1+\lambda(z-1)$.

The customer waiting time W and the delay time D distributions are also presented in Gao et al. (2004) and described by their respective probability generating functions as

$$
\mathcal{W}(z) = (z-1)\sum_{p=0}^{c-1}\frac{x_p(z)^{c-1}}{z\mathcal{T}_c'(x_p(z))(1-x_p(z))}\mathcal{Q}\left(\frac{1}{x_p(z)}\right) \qquad (2.10)
$$

and,

$$
\mathcal{D}(z) = \frac{\mu(z-1)}{1-(1-\mu)z}\sum_{p=0}^{c-1}\frac{x_p(z)^{c-1}}{z\mathcal{T}_c'(x_p(z))(1-x_p(z))}\mathcal{Q}\left(\frac{1}{x_p(z)}\right) \qquad (2.11)
$$

where $\mathcal{T}_i(z) = (1-\mu+\mu z)^i$, $x_p(z)$'s are the solutions of the equation $1 - z\mathcal{T}_c(x) = 0$, and $\mathcal{Q}(z) = \frac{(\mathcal{A}(z)-1)\mathcal{N}(z)}{\lambda(z-1)\mathcal{A}(z)}$. We can obtain the distributions of the system size, customer waiting time and the delay time by obtaining numerical inversions of the probability generating functions using the techniques presented in Abate and Whitt (1992).

2.2 The single and batch arrival Bernoulli queues

In the Bernoulli queue, the arrival data is observed at each time-slot, recording whether a customer has arrived to the system or not, with a 1 or a 0, respectively. The service process will output 1 if a service completion occurs at a time slot, and 0 otherwise. However, it has to be handled slightly differently since the lack of a service completion at a time-slot may either result from an empty system or a customer's unfinished service. To resolve this ambiguity, we need to condition the Bernoulli random variable S_t representing the service completion at time slot t on the number of customers in the system at time $t-1$ as shown in (2.12). We will refer to the single server Bernoulli queue as Bern/Bern/1 throughout this work.

$$P(S_t|N_{t-1}) = \mu^{S_t}(1-\mu)^{1-S_t} I'(N_{t-1}, 0) \tag{2.12}$$

where the $I'(*)$ function is defined in (2.6). Given past data, the number of customers in the system at time t can be calculated as:

$$N_t = \sum_{k=1}^{t-1} (A_k - S_k) \tag{2.13}$$

A_t is a pure Bernoulli process with the rate λ, independent of S_t, and is given as $P(A_t) = \lambda^{A_t}(1-\lambda)^{1-A_t}$, where $A_t = 1$ if an arrival occurs at time slot t and 0 otherwise.

Similar to the geometric queue, an arrival during a time slot occurs with constant probability λ, and a customer's service is completed during a time slot with constant probability μ.

The geometric and the Bernoulli queues are fundamentally equivalent in the single server case. The arrival and service rates are interpreted similarly, and the results obtained will be identical. The data for the geometric queue can easily be converted

to be used in the Bernoulli queue, and vice versa. The size of the data to be used in the geometric queue is also smaller than that of the Bernoulli queue. The Bayesian inference for the Bernoulli queue parameters also slightly differs from that of the geometric queue. The Bernoulli queue does allow modeling systems where parameters can modulate based on calendar time, which cannot be done with the geometric queue, and therefore will be considered in our analysis. We will discuss Markov modulation of these queues in our second paper. We should also note that the Bernoulli queues as we present here can not be utilized in queues with multiple servers without considering batch services.

So far, we have assumed that the maximum number of arrivals and service completions possible in a single time slot is one. Allowing for batch arrivals and service completions relaxes this assumption and allows the use of longer time slots in the analysis of the system. The $Geo^x/Geo/c$ model presented in Rubin and Zhang (1991) provides a very workable foundation to incorporate Bayesian analysis and present a methodology for modeling batch arrival multiple geometric server queues and obtaining the distributions of the number of customers in the system and their waiting times.

In Rubin and Zhang's work, the process describing the number of customers arriving in a time slot, A_t, is assumed to be a general discrete distribution, similar to the discrete time queueing model in Meisling (1958)'s analysis. Consequently, the system sees no arrival at time slot t with probability $P(A_t = 0)$, and arrivals occur with probability $P(A_t > 0) = 1 - P(A_t = 0)$. The resulting process is named the *batch geometric process* and denoted Geo^x. The service process is also geometric, i.e., the service times of each customer is distributed according to a geometric probability distribution with a constant rate μ. In our analysis, we assume a Poisson distribution to describe the number of customers arriving in a time slot.

The pgf of the number of customers in the system is presented in Rubin and Zhang (1991) and can be given as

$$\mathcal{N}(z) = \frac{\mathcal{E}(z)z^c[(1-\mu)z+\mu]^c \sum_{k=0}^{c-1} \left([(1-\mu)z+\mu]^{k-c} - z^{k-c}\right) P(N_k)}{z^c - \mathcal{E}(z)[(1-\mu)z+\mu]^c}, |z| < 1 \tag{2.14}$$

where $\mathcal{E}(z)$ is the pgf of the number of customers arriving in a time slot. For our Poisson arrival model, this function is given as $\mathcal{E}(z) = e^{\lambda(z-1)}$. In order to fully define $\mathcal{N}(z)$, $P(N_k)$, the probability that there are k customers in the system, needs to be obtained. Rubin and Zhang present a way of obtaining $P(N_k)$ by solving the system of equations formed by setting numerator of (2.14) equal to zero at points where the denominator is zero, in addition to the normalizing condition $\mathcal{N}(1) = 1$.

The waiting time distribution is also represented by its pgf given by:

$$\mathcal{W}(z) = \frac{\mathcal{N}(\mathcal{B}(z))}{\mathcal{E}(\mathcal{B}(z))} \frac{1-\mathcal{E}(\mathcal{B}(z))}{\lambda(1-\mathcal{B}(z))} \tag{2.15}$$

where $\mathcal{B}(z) = \frac{\mu z}{1-(1-\mu)z}$.

3 Bayesian inference on the parameters

In order to make inferences on the parameters of a discrete time queue, we have to decide on an experimental design. For the geometric queue, we will assume the design used in Armero and Bayyari (1994) where n_a inter-arrival and n_s service times are observed. For the Bernoulli queue, two separate binary sequences, $a^{(T)}$ and $s^{(T)}$, will represent the arrival and the service processes, respectively. We will also present the analysis for batch arrival queues where the arrival process will be represented by a count process for the number of customers coming into the system at each of the n time slots, and the service process will be observed by n_s individual service times of

the customers.

3.1 Analyzing the geometric queue

Let us denote X_i as the inter-arrival time between customer i and $i-1$, where $i = 1, 2, \ldots, n_a$. Additionally, let Y_j denote the service time of the jth customer, where $j = 1, 2, \ldots, n_s$. By the standard assumptions of the Geo/Geo/1 queue, the arrival process is independent of the service process, and $X_1, \ldots, X_{n_a}$ and $Y_1, \ldots, Y_{n_s}$ are independent random variables. We have already discussed that the inter-arrival times behave according to a Geometric process with parameter λ, resulting in the model

$$P(X_i = x|\lambda) = (1-\lambda)^x \lambda \tag{3.1}$$

where we have x time-slots without arrivals followed by one time-slot with an arrival. Similarly, modeling the service times via a Geometric process with parameter μ gives us the model

$$P(Y_j = y|\mu) = (1-\mu)^{y-1} \mu \tag{3.2}$$

Then the joint likelihood function associated with the observed data is given by

$$L(\lambda, \mu | x^{(n_a)}, y^{(n_s)}) = \lambda^{n_a} (1-\lambda)^{\sum_{i=1}^{n_a} x_i} \mu_s^n (1-\mu)^{\sum_{j=1}^{n_s} y_j - n_s} \tag{3.3}$$

where $(x^{(n_a)}, y^{(n_s)}) = \{x_1, \ldots, x_{n_a}, y_1, \ldots, y_{n_s}\}$ is the set of inter-arrival and service time observations.

We now need to specify prior distributions on the parameters λ and μ in order to carry out a Bayesian analysis. Using independent conjugate Beta priors for our parameters, where $\lambda \sim Beta(\alpha_0, \beta_0)$ and $\mu \sim Beta(\gamma_0, \delta_0)$, the resulting joint prior

distribution is

$$P(\lambda, \mu) = Beta(\alpha_0, \beta_0) \times Beta(\gamma_0, \delta_0) \tag{3.4}$$

where $Beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$.

The prior specification in (3.4) along with (3.3) allow us to compute the joint posterior distribution of λ and μ as

$$P(\lambda, \mu|x^{(n_a)}, y^{(n_s)}) = Beta\,(\alpha^*, \beta^*) \times Beta\,(\gamma^*, \delta^*) \tag{3.5}$$

where $\alpha^* = \alpha_0 + n_a$, $\beta^* = \beta_0 + \sum_{i=1}^{n_a} x_i$, $\gamma^* = \gamma_0 + n_s$, and $\delta^* = \delta_0 + \sum_{j=1}^{n_s} y_j - n_s$.

By the help of the Bayesian paradigm, the posterior predictive distributions of quantities of interest can easily be obtained. All one has to do to reach $P(g|Data)$, the posterior predictive distribution of variable g, is to calculate the integral

$$P(g|Data) = \int P(g|\theta)P(\theta|Data) \tag{3.6}$$

where $P(\theta|Data)$ is the posterior distribution of the parameter set θ. In some cases, the predictive distributions can be obtained in closed form, whereas in others, results can be reached through simulation.

We can obtain the posterior predictive distributions of the next inter-arrival time and next service time by

$$\begin{aligned} P(x_{n_a+1}|x^{(n_a)}) =& \int_0^1 (1-\lambda)^{x_{n_a+1}} \lambda \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \lambda^{\alpha^*-1}(1-\lambda)^{\beta^*-1} d\lambda \\ =& \frac{\Gamma(\alpha^*+1)\Gamma(\beta^*+x_{n_a+1})}{\Gamma(\alpha^*+\beta^*+x_{n_a+1}+1)} \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \\ P(y_{n_s+1}|y^{(n_s)}) =& \frac{\Gamma(\gamma^*+1)\Gamma(\delta^*+y_{n_s+1}-1)}{\Gamma(\gamma^*+\delta^*+y_{n_s+1})} \frac{\Gamma(\gamma^*+\delta^*)}{\Gamma(\gamma^*)\Gamma(\delta^*)} \end{aligned} \tag{3.7}$$

where α^*, β^*, γ^*, and δ^* are defined as in (3.5).

The argument in (3.6) can be used to obtain the posterior predictive distributions of system sizes in (2.5) and waiting time distribution given in (2.8), as well as the steady state distribution of system size obtained from (2.3). We can also obtain posterior predictive distributions using the pgfs in (2.9), (2.10), (2.11), (2.14), and (2.15). However, closed form solutions are not efficiently obtainable for these distributions, and we must resort to simulation methods. A simple Monte Carlo simulation is sufficient to perform these calculations numerically.

The purpose of the Monte Carlo simulation is to numerically calculate the integral in (3.6). The right hand side of (3.6) can be interpreted as the posterior expectation of the function $P(g|\theta)$. Instead of calculating this expectation by solving the integral, we can first obtain Θ^*, a random sample of size J, from the posterior distribution of the parameter set $P(\theta|Data)$. Each sample realization can then be plugged into $P(g|\theta)$ and averaged out as

$$\frac{1}{J}\sum_{\theta\in\Theta^*} P(g|\theta) \tag{3.8}$$

which converges to $P(g|Data)$ as J increases for any value of g.

The inference for the multiple server discrete time queue is identical to the single server queue presented here since we still work with inter-arrival and service time observations.

3.2 Analyzing the Bernoulli queue

For the Bernoulli queue, suppose we have two binary sequences, $a^{(T)} = \{a_1, \ldots, a_T\}$ and $s^{(T)} = \{s_1, \ldots, s_T\}$ created after observing the system for n time slots. Given the model in section 2.2, we have the joint likelihood function associated with the data

formed by two independent, updated Bernoulli processes in (3.9).

$$L(\lambda, \mu | a^{(T)}, s^{(T)}) = \lambda^{\sum_{t=1}^{n} a_t} (1-\lambda)^{n - \sum_{t=1}^{n} a_t} \mu^{\sum_{t=1}^{n} s_t} (1-\mu)^{\sum_{t=1}^{n} (s_t - I'(N_{t-1}, 0))} \tag{3.9}$$

We can once again employ independent Beta distributions to the arrival ($Beta(\alpha_0, \beta_0)$) and service ($Beta(\gamma_0, \delta_0)$) processes to serve as conjugate priors. Doing so results in the posterior distribution given in (3.10)

$$P(\lambda, \mu | a^{(T)}, s^{(T)}) = Beta\,(\alpha^*, \beta^*) \times Beta\,(\gamma^*, \delta^*) \tag{3.10}$$

where $\alpha^* = \alpha_0 + \sum_{t=1}^{n} a_t$, $\beta^* = \beta_0 + n - \sum_{t=1}^{n} a_t$, $\gamma^* = \gamma_0 + \sum_{t=1}^{n} s_t$, and $\delta^* = \delta_0 + \sum_{t=1}^{n} (s_t - I'(N_{t-1}, 0))$. We can easily see that the joint posterior distribution obtained in (3.5) is equivalent to the joint posterior distribution in (3.10).

Similar to (3.7), the posterior predictive distribution of the arrival and service probability in the next time slot is given by

$$\begin{aligned} P(a_{n+1} | a^{(T)}) =& \frac{\Gamma(\alpha^* + a_{n+1})\Gamma(\beta^* - a_{n+1} + 1)}{\Gamma(\alpha^* + \beta^* + 1)} \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \\ P(s_{n+1} | s^{(T)}, N_n > 0) =& \frac{\Gamma(\gamma^* + s_{n+1})\Gamma(\delta^* - s_{n+1} + 1)}{\Gamma(\gamma^* + \delta^* + 1)} \frac{\Gamma(\gamma^* + \delta^*)}{\Gamma(\gamma^*)\Gamma(\delta^*)} \end{aligned} \tag{3.11}$$

We can also obtain the distribution of the number of arrivals in the next T time slots as

$$P(m | a^{(T)}) = \binom{T}{m} \frac{\Gamma(\alpha^* + m)\Gamma(\beta^* + T - m)}{\Gamma(\alpha^* + \beta^* + T)} \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \tag{3.12}$$

which turns out to be a beta-binomial distribution.

The Bernoulli model in this setup does not lend itself to analyze multiple server systems without considering batch service completions, which is discussed in Sec-

tion 3.4.

3.3 Dependent prior analysis

Let us go back to considering the single server geometric queue, $Geo/Geo/1$, with $a^{(n_a)}$ inter-arrival observations and $s^{(n_s)}$ service time observations for this section. Computation of the performance measures requires the system to be in steady state. We can deal with this restriction in one of two ways. The inference can be carried out as if the arrival and the service processes are independent as we have so far done. This requires the computation of the performance measures to be carried out under the stationarity condition $\lambda < \mu$. Alternatively, we can assume a dependence structure between the arrival and service processes in a single server queue a priori by following the ideas discussed in Mazzuchi and Soyer (1992) and Erkanli et al. (1998).

One option for such a dependence structure is to use an ordered Dirichlet distribution for the arrival and service rates, which is given by

$$P(\lambda, \mu) = \frac{\Gamma(\beta)}{\Gamma(\beta\alpha_\lambda)\Gamma(\beta\alpha_\mu)\Gamma(\beta\alpha_3)} \lambda^{\beta\alpha_\lambda - 1}(\mu - \lambda)^{\beta\alpha_\mu - 1}(1 - \mu)^{\beta\alpha_3 - 1} \tag{3.13}$$

where $\alpha_\lambda + \alpha_\mu + \alpha_3 = 1$, and $\lambda < \mu < 1$.

The resulting marginal prior distributions of the arrival and service rates turn out to be following Beta distributions:

$$\lambda \sim Beta\left(\beta\alpha_\lambda, \beta(1 - \alpha_\lambda)\right) \tag{3.14}$$

$$\mu \sim Beta\left(\beta(\alpha_\lambda + \alpha_\mu), \beta(1 - (\alpha_\lambda + \alpha_\mu))\right) \tag{3.15}$$

with expected values $E(\lambda) = \alpha_\lambda$ and $E(\mu) = \alpha_\lambda + \alpha_\mu$.

The distribution of the difference between the two rates, $\mu - \lambda$ can be obtained as

a $Beta(\beta(\alpha_\mu), \beta(1-\alpha_\mu))$ distribution with expected value $E(\mu - \lambda) = \alpha_\mu$. The ratio of the two rates λ/μ which is a measure for the stationarity of the queue can also be shown to have a $Beta(\beta(\alpha_\lambda), \beta(\alpha_\mu))$ distribution, a priori.

The posterior distribution of the arrival and service rates can be obtained as

$$P(\lambda, \mu | a^{(n_a)}, s^{(n_s)}) \propto (1-\lambda)^{\sum_{i=1}^{n_a} a_i} \lambda^{\beta\alpha_\lambda + n_a - 1} (1-\mu)^{\beta\alpha_3 + \sum_{i=1}^{n_s} s_i - n_s - 1} \mu^{n_s} (\mu - \lambda)^{\beta\alpha_\mu - 1} \tag{3.16}$$

which is not analytically available. Therefore, we need to resort to numerical methods such as the one discussed in Erkanli et al. (1998). Samples from the joint distribution of λ and μ can be obtained using the Gibbs sampler. The Gibbs method that the authors describe iteratively samples from the full conditional distributions of the parameters. After achieving convergence and discarding the values obtained during the warm-up period, the remaining sample behaves according to the unconditional joint posterior distribution of the parameters. The simulation starts by a set of initial values for the parameters, λ^0 and μ^0. The Gibbs steps are given below:

Set counter $[ctr] = 1$.

Step 1: Draw $\lambda^{[ctr]}$ from $P(\lambda | \mu^{[ctr-1]}, a^{(n_a)}, s^{(n_s)})$

Step 2: Draw $\mu^{[ctr]}$ from $P(\mu | \lambda^{[ctr]}, a^{(n_a)}, s^{(n_s)})$

Step 3: Increment $[ctr]$ by 1 and go to *Step 1*

The conditional distributions in *Step 1* and *Step 2* are not available in closed form unless we choose a uniform joint prior for (λ, μ) by selecting $\beta = 3$ and $\alpha_\lambda = \alpha_\mu = \alpha_3 = 1/3$. In this case, the distributions $P(\lambda | \mu^{[ctr-1]}, a^{(T)}, s^{(T)})$ and $P(\mu | \lambda^{[ctr]}, a^{(n_a)}, s^{(n_s)})$ become truncated Beta distributions from which direct sampling is possible. Otherwise, we can obtain the samples in *Step 1* and *Step 2* above by a rejection sampling method similar to the one presented in Erkanli et al. (1998). The steps of the rejection sampling method is give below:

Step 1: Draw $\lambda^{[ctr]}$ from the prior conditional distribution $P(\lambda|\mu^{[ctr-1]})$, which is a truncated $Beta(\beta\alpha_\lambda, \beta\alpha_\mu)$ distribution in the interval $(0, \mu^{[ctr-1]})$.

Step 2: Draw an independent $uniform(0,1)$ random variate u.

Step 3: Calculate the maximum likelihood estimator $\hat{\lambda} = min(\mu^{[ctr-1]}, \frac{n_a}{n_a + \sum_{i=1}^{n_a} a_i})$

Step 4: If $u > \frac{(\lambda^{[ctr]})^{n_a}(1-\lambda^{[ctr]})^{\sum_{i=1}^{n_a} a_i}}{\hat{\lambda}^{n_a}(1-\hat{\lambda})^{\sum_{i=1}^{n_a} a_i}}$ then go to *Step 1*

Step 5: Draw $\mu^{[ctr]}$ from the prior conditional distribution $P(\mu|\lambda^{[ctr]})$, which is a truncated $Beta(\beta\alpha_\mu, \beta\alpha_3)$ distribution in the interval $(\lambda^{[ctr]}, 1)$.

Step 6: Draw an independent $uniform(0,1)$ random variate u.

Step 7: Calculate the maximum likelihood estimator $\hat{\mu} = max(\lambda^{[ctr]}, \frac{n_s}{n_s + \sum_{i=1}^{n_s} s_i})$

Step 8: If $u > \frac{(\mu^{[ctr]})^{\sum_{i=1}^{n_s} s_i - n_s}(1-\mu^{[ctr]})^{n_s}}{\hat{\mu}^{n_s}(1-\hat{\mu})^{\sum_{i=1}^{n_s} s_i - n_s}}$ then go to *Step 5*

The extension to the c-server case is straightforward. All one has to do is to modify the ordered Dirichlet prior to force the ordering $\frac{\lambda}{c} < \mu$, which will modify the posterior results slightly. Additionally, the simulation methods discussed here will apply to the multiple server case as well.

3.4 Discrete queues with batch arrivals and multiple servers

The Bayesian inference for this model is quite straightforward. Let us denote the data counting the number of arrivals each time slot $t = 1 \ldots n$ given by $a^{(T)} = \{a_1, \ldots, a_T\}$. The joint likelihood function for the arrival process is

$$L(\lambda|a^{(T)}) = \frac{\lambda^{\sum_{t=1}^{n} a_t} e^{-\lambda}}{\prod_{t=1}^{n} a_t!}. \tag{3.17}$$

Coupled with a conjugate $Gamma(\alpha, \beta)$ prior, we can obtain the posterior distribution for the arrival rate λ as $Gamma(\alpha + \sum_{t=1}^{n} a_t, \beta + n)$. The service process is observed via the individual service times of each customer, and therefore can be analyzed identically to the single server queue in Section 3.1.

As discussed earlier, posterior predictive distributions of queue occupancy and

customer waiting times can be obtained by utilizing the Monte Carlo simulation method presented in Section 3.1. We can also utilize the Bayesian framework we bring to this analysis to obtain posterior predictive distribution of the next arrival batch size and discuss the stationarity of the queue probabilistically.

4 Numerical demonstration

For our numerical analysis, we will use part of a dataset collected at the emergency department of York Hospital in York, PA consisting of 565 inter-arrival and service times from 11/11/1999 to 11/18/1999. The continuously collected data is discretized to 1-minute long time slots for the non-batch model. For the batch arrival and service processes, we have set the time slot lengths to 30 minutes, resulting in 336 consecutive slots.

We employed non-informative $Beta(1, 1)$ priors for both the arrival and the service rates of the geometric and Bernoulli queues analyzing the 1-minute interval dataset. The resulting independent beta distributions given in (3.5) provide the probability distributions illustrated in Figure 4.1.

As expected, since the system on hand is a multiple server queue, the arrival rate is much higher than the service rate of a single server. Since we are using data from a system with multiple servers, we need to use the multiple server model results to obtain distributions of performance measures.

For a more general demonstration, we have chosen to use the $Geo^x/Geo/c$ model to obtain the inference and performance characteristics of the system. The posterior distributions are plotted in Figure 4.2. The batch arrival rate has a mean of 1.6895, and the service rate of a single server has a mean of 0.1898. Using samples from these distributions, we can numerically calculate the distribution of the number of people

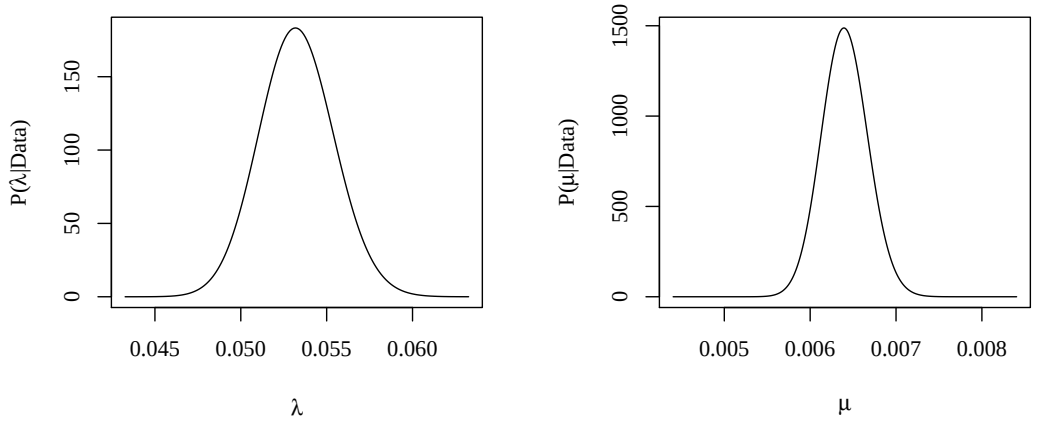

Figure 4.1: Posterior arrival and service rate distributions

in the system as well as the waiting time. The resulting distributions can be seen in Figure 4.3.

We can also obtain the probability of the queue being in steady state by investigating the distribution of $\rho = \frac{\lambda}{c\mu}$ where $c = 10$ in our case. The resulting distribution calculated by a straight Monte Carlo simulation using the posterior samples of λ and μ, and is given in Figure 4.4.

5 Concluding remarks

In this essay, we provide a framework for conducting Bayesian analysis of the single and multiple server discrete time queues with geometric arrival and service processes. Bayesian analysis of multiple server queues with batch arrivals are also considered. Through the use of a Bayesian approach, we are able to obtain posterior distributions of the system parameters. We also obtain the distribution of the load factor of the system which determines whether or not the system is stationary. While in

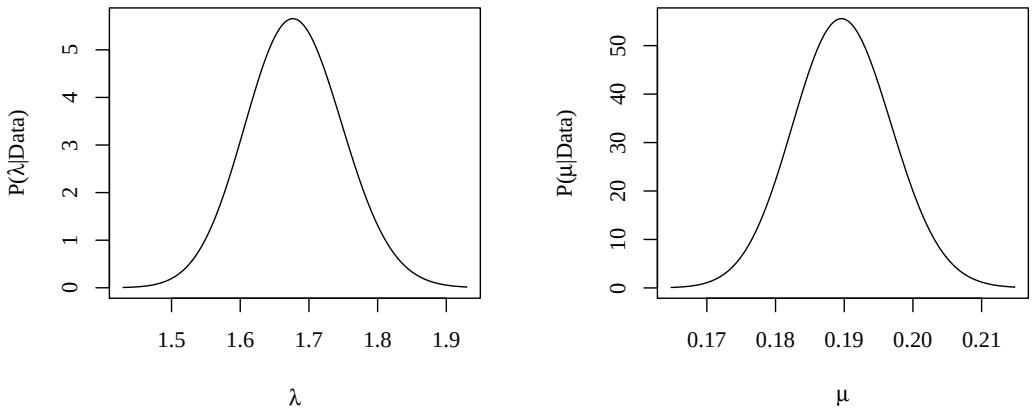

Figure 4.2: Posterior batch arrival and service rate distributions

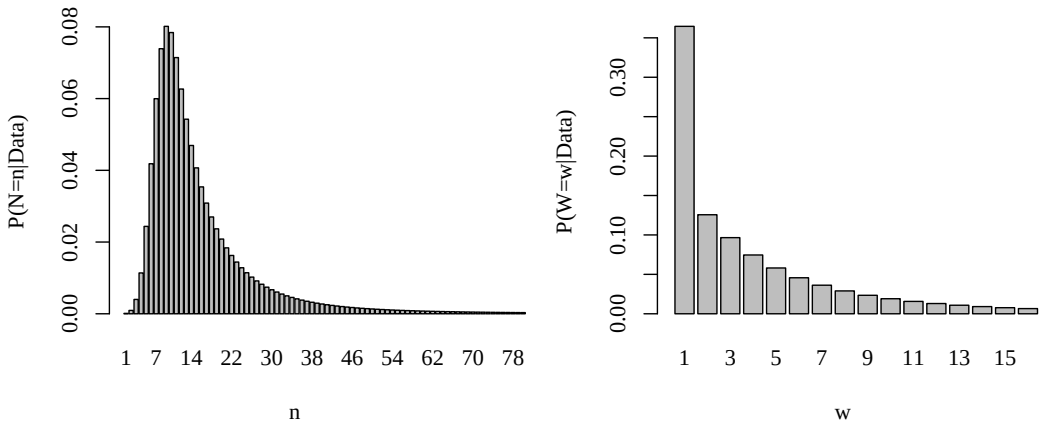

Figure 4.3: System occupancy and waiting time distributions

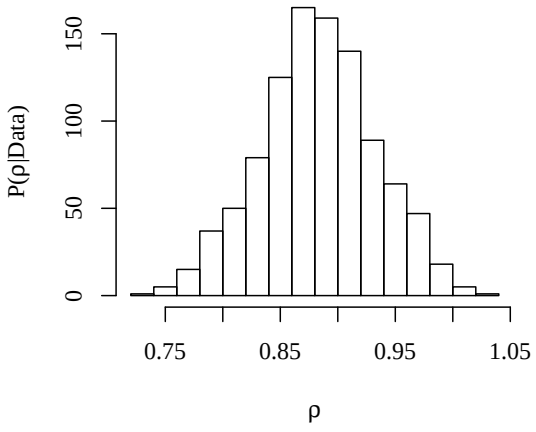

Figure 4.4: Distribution of $\rho = \frac{\lambda}{c\mu}$

the classical (non-Bayesian) analysis of discrete time queues stationarity is assumed, our approach allows us to describe the stationarity of the system probabilistically. Additionally, numerical calculation methods of system occupancy and delay time distributions are presented.

One advantage of these distributional results is the ease in which they lend themselves to decision making problems. A key question in system design in queues is what the system capacity should be. The decision is the product of a trade-off between customer service, which can be measured as a function of the delay time, and service cost. Upon constructing an appropriate penalty function combining the customer waiting time and service costs based on the distribution of delay time obtained by our methods, one can easily obtain the distribution of this penalty function. As a result, decisions on the system capacity can be made by satisfying certain criteria (e.g. minimizing the expected penalty).

The current state of research in discrete time queues is mostly from a non-Bayesian

perspective and is concentrated on performance measure calculations based on known system parameters. While this assumption may be valid for most computer and communication systems, when dealing with systems requiring human interaction and performance, these parameters may not be known or fixed. Our approach complements the current state of the discrete time queue analysis by providing an inferential tool as well as performance evaluation methods based on the Bayesian framework. We also contribute to the Bayesian analysis of discrete time queues by introducing batch arrival and multiple server extensions to the single server queue. Additionally, we allow for a dependence structure between the arrival and service rates of a system that provides a queue that is always stationary, instead of assuming these conditions aposteriori.

References

Abate, J. and Whitt, W. (1992). Numerical inversion of probability generating functions. *Operations Research Letters, Vol. 12, :4, 245-251*, 12, 4:245–251.

Armero, C. and Bayyari, M. J. (1994). Bayesian prediction in m/m/1 queues. *Queueing Systems*, 15:401–417.

Conti, P. L. (1999). Large sample bayesian analysis for geo/g/1 discrete-time queueing models. *The Annals of Statistics*, 27(6):1785–1807.

Conti, P. L. (2004). Bootstrap approximations for bayesian analysis of geo/g/1 discrete-time queueing models. *Journal of Statistical Planning and Inference*, 120(1-2):65 – 84.

Daduna, H. (2001). *Queueing Networks with Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Networks*, volume 2046 of *Lecture notes in computer science*. Springer.

Dafermos, S. C. and Neuts, M. F. (1971). A single server queue in discrete time. *Cahiers du Centre d'Etude de Recherche Operationnelle*, 13:23–40.

Erkanli, A., Mazzuchi, T. A., and Soyer, R. (1998). Bayesian computations for a class of reliability growth models. *Technometrics*, 40:14–23.

Gao, P., Wittevrongel, S., and Bruneel, H. (2004). Discrete-time multiserver queues with geometric service times. *Computers and Operations Research*, 31:81–99.

Kobayashi, H. and Konheim, A. G. (1977). Queueing models for computer communications system analsyis. *IEEE Transactions on Communications*, 25(1):2–29.

Mazzuchi, T. and Soyer, R. (1992). Reliability assessment and prediction during product development. In *1992 Proceedings of the Annual Reliability and Maintainability Symposium*, pages 468–474, New York, USA. Institute of Electrical and Electronics Engineers.

McGrath, M. F., Gross, D., and Singpurwalla, N. D. (1987). A subjective bayesian approach to the theory of queues i – modeling. *Queueing Systems*, 1(4):317–333.

McGrath, M. F. and Singpurwalla, N. D. (1987). A subjective bayesian approach to the theory of queues ii – inference and information in m/m/1 queues. *Queueing Systems*, 1(4):335–353.

Meisling, T. (1958). Discrete-time queuing theory. *Operations Research*, 6:96–105.

Neuts, M. F. (1973). The single server queue in discrete time-numerical analysis i. *Naval Research Logistics Quarterly*, 20(2):297–304.

Rubin, I. and Zhang, Z. (1991). Message delay and queue-size analysis for circuit-switched tdma systems. *IEEE Transactions on Communications*, 39(6):905–914.

Takagi, H. (1993). *Queueing Analysis: A Foundation of Performance Evaluation, Volume 3: Discrete-Time Systems*, volume 3. North-Holland.