Working Paper 01-41
Statistics and Econometrics Series 26
October 2001

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

BAYESIAN INFERENCE AND PREDICTION FOR THE GI/M/1 QUEUEING SYSTEM.

M.C. Ausín, R.E. Lillo and M.P. Wiper*

Abstract

This article undertake Bayesian inference and prediction for GI/M/1 queueing systems. A semiparametric model based on mixtures of Erlang distributions is considered to model the general interarrival time distribution. Given arrival and service data, a Bayesian procedure based on birth-death Markov Chain Monte Carlo methods is proposed. An estimation of the system parameters and predictive distributions of measures such as the stationary system size and waiting time is given.

Keywords: Queueing systems; bayesian inference; Erlang mixtures; birth-and-death MCMC.

*Ausin, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, e-mail:causin@est-econ.uc3m.es, Tfno: 91-6249972; Lillo, Dpto. de Estadística y Econometría, Univ. Carlos III de Madrid, e-mail: lillo@est-econ.uc3m.es; Wiper, Dpto. de Estadística y Econometría, Univ. Carlos III de Madrid, e-mail: mwiper@est-econ.uc3m.es.The work of all the authors was partially supported by the project BEC2000-0167 of the Spanish Ministry of Science and Technology.

Bayesian inference and prediction for the $GI/M/1$ queueing system

M. C. Ausín*, (causin@est-econ.uc3m.es) R. E. Lillo
(lillo@est-econ.uc3m.es) and M. P. Wiper
(mwiper@est-econ.uc3m.es)
*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid,
Calle Madrid 126, 28903 Getafe, Madrid, Spain.*

Abstract. This article undertake Bayesian inference and prediction for *GI/M/1* queueing systems. A semiparametric model based on mixtures of Erlang distributions is considered to model the general interarrival time distribution. Given arrival and service data, a Bayesian procedure based on birth-death Markov Chain Monte Carlo methods is proposed. An estimation of the system parameters and predictive distributions of measures such as the stationary system size and waiting time is given.

Keywords: queueing systems, Bayesian inference, Erlang mixtures, birth-and-death MCMC.

AMS Classification: 62F15, 60K25

1. Introduction

Bayesian analysis of queueing systems is a relatively recent research area. As far as we know, the first papers on Bayesian estimation in queueing models are Bagchi and Cunningham (1972), Muddapur (1972) and Reynolds (1973). In the 1980's and 1990's, there has been much work on Bayesian inference for the simple Markovian *M/M/c* queue, see Armero (1985), McGrath et al. (1987), McGrath and Singpurwalla (1987) and Armero and Bayarri (1994, 1995, 1997). More recently, the development of modern numerical integration methods have allowed the development of inference and prediction in more general queueing systems, see Armero and Conesa (1998), Wiper (1998), Rios Insua et al. (1998) and Ausin et al. (2001).

Up to now, most analyses have considered queueing systems where the customers arrive according to a Poisson process. To the best of our knowledge, the only exception is Wiper (1998) where inference for the *Er/M/1* (and *Er/M/c*) model, with Erlang interarrival times, was considered. Although the Erlang distribution has often been used in

[*] The work of all the authors was partially supported by the project BEC2000-0167 of the Spanish Ministry of Science and Technology

the queueing literature to fit interarrival (or service) time data with coefficient of variation greater than one, see Allen (1990), it will not be an appropriate model if the data have low coefficient of variation or are multimodal. Our objective in this paper is to consider inference and prediction for these, more general, *G/M/1* queues. To do this, we consider a semiparametric approximation to the general interarrival time distribution based on a mixture of Erlang distributions. Note that this family includes the Erlang, hyperexponential and exponential distributions, which are commonly used in the queueing literature, as special cases. It is also dense over the set of distributions on the positive reals, see Asmussen (1987).

The use of mixture distributions to model data is very common and the Bayesian approach provides an important tool for semiparametric density estimation, see, for example, Diebolt and Robert (1994). Markov chain Monte Carlo methods (MCMC), see Robert (1996), have been developed for Bayesian analyses for mixture models. Recently, MCMC methods for exploring mixture models of unknown dimension have been proposed. Richardson and Green (1997) introduced the reversible jump technique to analyze normal mixtures. This type of algorithm was used by Rios et al. (1998) for exponential mixtures and Wiper et al. (2001) for mixtures of gamma distributions. More recently, an alternative approach to reversible jump based on a birth-death process has been proposed by Stephens (2000). In this paper we make use of the latter methodology to study mixtures of Erlang distributions.

Throughout the paper we consider a queueing system with a single server and FIFO discipline. Service times S are distributed exponentially with mean $1/\mu$. Interarrival and service times are considered to be independent. Customers are assumed to arrive singly with interarrival times identically and independently distributed. Defining T to be a typical interarrival time then we assume that T is distributed as random variable distributed as a mixture of k Erlang distributions with parameters $\mathbf{w} = (w_1, ..., w_k)$,$\lambda = (\lambda_1, ..., \lambda_k)$ and $\nu = (\nu_1, ..., \nu_k)$. The corresponding density function is given by,

$$f(t \mid k, \mathbf{w}, \lambda, \nu) = \sum_{i=1}^{k} w_i Er(t \mid \nu_i, \lambda_i),$$

where

$$Er(t \mid \nu_i, \lambda_i) = \frac{(\nu_i/\lambda_i)^{\nu_i}}{\Gamma(\nu_i)} t^{\nu_i - 1} \exp(-\frac{\nu_i}{\lambda_i} t)$$

represents an Erlang density parameterized to have mean λ_i.

Given interarrival and service time data, we wish to make inference on the system parameters and to predict the stationary distributions.

For the $G/M/1$ system, a stationary distribution exists if the traffic intensity ρ is less than one, see, for example, Kleinrock (1975). In our model, the traffic intensity is given by

$$\rho = \left(\mu \sum_{i=1}^{k} w_i \lambda_i\right)^{-1}$$

In Section 2, we describe a simple experiment and develop a method to make inference for the system parameters. We define prior distributions and propose a birth and death algorithm to obtain a sample from the joint posterior distribution of the system parameters. In Section 3, we describe how to estimate the predictive distributions of the system size and the waiting time using the data generated from the MCMC algorithm described in Section 2.

In Section 4, we illustrate this methodology with various simulated examples and a real data set. Conclusions and a discussion of possible extensions are included in Section 5.

2. Data observation and Bayesian inference

Assume that given a single sample of m_a interarrival times, $\mathbf{t} = \{t_1, ..., t_{m_a}\}$, and m_s service times, $\mathbf{s} = \{s_1, ..., s_{m_s}\}$, we wish to make inference for the system parameters, $k, \mathbf{w}, \lambda, \nu$ and μ. Given this simple experiment, the likelihood function is,

$$L\left(k, \mathbf{w}, \lambda, \nu, \mu \mid \mathbf{t}, \mathbf{s}\right) \propto \left[\prod_{j=1}^{m_a} \left(\sum_{i=1}^{k} w_i Er(t_j \mid \nu_i, \lambda_i)\right)\right] \mu^{m_s} \exp\left(-\mu \sum_{j=1}^{m_s} s_j\right).$$

Thus, the likelihood separates into two parts, one concerning the arrival parameters, $(k, \mathbf{w}, \lambda, \nu)$ and another concerning the service parameter, μ. Hence, assuming independent prior distributions for the arrival and service parameters, the corresponding posterior distributions will also be independent a posteriori.

This experiment has also been used for many inference problems in queueing systems, see for example, Thiruvaiyaru and Basawa (1992), Rios et al. (1998) and Armero and Bayarri (1994). However, we could use other experiments which would result in different forms for the likelihood as proposed by Lehoczky (1990).

In the next subsection, we will introduce a prior model for the system parameters, $k, \mathbf{w}, \lambda, \nu$ and μ and show how to estimate the joint posterior distribution, $f\left(k, \mathbf{w}, \lambda, \nu, \mu \mid \mathbf{t}, \mathbf{s}\right)$. Exact inference is impossible in this case but we can calculate the necessary conditional posterior distributions to implement a Markov chain Monte Carlo (MCMC) algorithm to sample the joint posterior distribution.

2.1. Prior specification and updating

Here, we will assign prior distributions for the system parameters, $k, \mathbf{w}, \lambda, \nu$ and μ. As indicated earlier, we will consider independent prior distributions for the interarrival and service parameters. For the service rate, μ, we assume a gamma prior distribution, $\mu \sim G\left(a, b\right)$. It is straightforward to show that the prior to posterior updating is conjugate so that,

$$\mu \mid \mathbf{s} \sim G\left(a + m_s, b + \sum_{j=1}^{m_s} s_j\right). \tag{1}$$

In order to make inference for the interarrival distribution parameters, following Diebolt and Robert (1994), it is first convenient to introduce a missing data formulation in which each observed interarrival time, t_j, is assumed to arise from a specific component of the mixture. Thus, we define missing data $\mathbf{z} = \{z_1, ..., z_{m_a}\}$ assumed to be realizations from i.i.d. missing variables, $Z_1, ..., Z_{m_a}$, such that

$$P(Z_j = i \mid k, \mathbf{w}) = w_i, \qquad \text{for } i = 1, ..., k.$$

Conditional on these variables, we have

$$f\left(t_j \mid Z_j = i, k, \mathbf{w}, \lambda, \nu\right) = Er(t_j \mid \nu_i, \lambda_i), \qquad \text{for } j = 1, ..., m_a.$$

Following Stephens (2000), we define a hierarchical model for the joint prior distribution on the remaining mixture parameters, $(k, \mathbf{w}, \lambda, \nu)$ of the form

$$f\left(k, \mathbf{w}, \lambda, \nu, \mathbf{z}\right) \propto f\left(k\right) f\left(\mathbf{z} \mid k, \mathbf{w}\right) f\left(\mathbf{w} \mid k\right) f\left(\lambda \mid k\right) f\left(\nu \mid k\right). \tag{2}$$

We now define a truncated Poisson prior distribution for the mixture size, k, taking values from 1 to $k_{\max}$,

$$f\left(k\right) \propto \frac{\gamma^k}{k!} \tag{3}$$

For the examples of Section 4, we have used the values $\gamma = 2$ and $k_{\max} = 10$ but, in principle, any values could be considered. Other prior structures such as a discrete uniform defined on $[1, k_{\max}]$ could also be used.

Using the hierarchical structure (2), we can define the following prior distributions for the remaining parameters conditional on k,

$$\mathbf{w} \mid k \sim D(\phi, ..., \phi),$$

$$\lambda_i \mid k \sim IG(\alpha, \beta), \qquad \text{for } i = 1, ..., k,$$

$$\nu_i \mid k \sim GE(\vartheta), \qquad \text{for } i = 1, ..., k,$$

where $D(\phi, ..., \phi)$ denotes a symmetric Dirichlet distribution of dimension k; $IG(\alpha, \beta)$ denotes an inverted gamma and $GE(\vartheta)$ is a geometric distribution with mean $1/\vartheta$. Typically, we might set $\phi = 1$, which implies a uniform prior for $\mathbf{w} \mid k$ and $\alpha = 1$ and $\beta = 1$ and $\vartheta = 0.01$ giving fairly diffuse priors for λ_i and ν_i.

Given k, the required conditional posterior distributions which are used within the MCMC algorithm can be shown to be;

$$P(Z_j = i \mid \mathbf{t}, k, \mathbf{w}, \lambda, \nu) \propto w_i \frac{(\nu_i/\lambda_i)^{\nu_i}}{\Gamma(\nu_i)} t_j^{\nu_i - 1} \exp(-\frac{\nu_i}{\lambda_i} t_j), \qquad \text{for } i = 1, ..., k,$$

$$\mathbf{w} \mid \mathbf{t}, \mathbf{z}, k \sim D(\phi + m_1, ..., \phi + m_k),$$

$$\lambda_i \mid \mathbf{t}, \mathbf{z}, k \sim IG(\alpha + m_i \nu_i, \beta + T_i \nu_i),$$

where $m_i = \#\{Z_j = i\}$ and $T_i = \sum_{j:Z_j=i} t_j$, for $i = 1, ..., k$, and,

$$f(\nu_i \mid \mathbf{t}, \mathbf{z}, k, \mathbf{w}, \lambda) \propto \frac{\nu_i^{m_i \nu_i}}{\Gamma(\nu_i)^{m_i}} \exp\left\{-\nu_i \left(-\log(1-\vartheta) + \frac{T_i}{\lambda_i} + m_i \log \lambda_i - \log P_i\right)\right\} \tag{4}$$

where $P_i = \prod_{j:Z_j=i} t_j$.

In the following subsection, we construct an MCMC algorithm in order to estimate the joint posterior distribution of the interarrival parameters, $(k, \mathbf{w}, \lambda, \nu)$.

2.2. BDMCMC algorithm

Here, we obtain a sample from the joint posterior distribution of the interarrival parameters, $k, \mathbf{w}, \lambda$ and ν. To analyze the Erlang mixture model, we propose a birth-death MCMC (BDMCMC) algorithm. This method is based on a birth-death process and was introduced by Stephens (2000) in the context of normal mixtures. With this approach, the model parameters are interpreted as observations from a marked point process and the mixture size, k, changes so that births and deaths of the mixture components occur in continuous time. The rates at which this happens determine the stationary distribution of the process.

To create a Markov chain with stationary distribution $f\left(k,\mathbf{w},\lambda,\nu \mid \mathbf{t}\right)$, a birth-death process (BD) is combined with a standard MCMC method where the mixture size, k, is kept fixed.

In the BD process, births of the mixture components occur at a constant rate which we might set equal to the parameter γ, of the prior distribution of k in (3). A birth increases the number of components by one and the parameters of the new component are generated from the prior distribution.

The death rate of every mixture component is a likelihood ratio of the model with and without this component. Thus, death rates are very low if the corresponding component explains a lot of data and high if it does not. The total death rate of the process at any time is the sum of the individual death rates. A death decreases the number of mixture components by one.

Then, we define an algorithm, based on Stephens (2000), as follows:

```
1. Set initial values $k^{(0)}, \mathbf{w}^{(0)}, \lambda^{(0)}, \nu^{(0)}$.

Birth Death process.

2. Run the birth-death process for a fixed time $t_0$.
    2.1. Start from $k^{(n)}, \mathbf{w}^{(n)}, \lambda^{(n)}, \nu^{(n)}$.
    2.2. Compute the death rates.
    2.3. Simulate the exponential time to next jump.
    2.4. Simulate the type of jump (birth or death).
    2.5. Modify the mixture components and
    2.6. If the run time is less than $t_0$ go to 2.2.

MCMC algorithm.

3. Update the allocation by sampling from $\mathbf{z}^{(n+1)} \sim \mathbf{z}|\ \mathbf{t},k^{(n)}, \mathbf{w}^{(n)}, \mu^{(n)}, \nu^{(n)}$.
4. Update the weights by sampling from $\mathbf{w}^{(n+1)} \sim \mathbf{w}|\ \mathbf{t}, \mathbf{z}^{(n+1)}, k^{(n)}$.
5. For $i = 1, ..., k$,
    5.1. Update the means by sampling from $\mu_i^{(n+1)} \sim \mu_i|\ \mathbf{t},\mathbf{z}^{(n+1)}, k^{(n)}$.
    5.2. Update $\nu_i$ using a Metropolis step.
6. $n = n + 1$. Go to 2.
```

Step 2 of the algorithm is the BD process described above. Following Stephens (2000), we have fixed in our examples $t_0 = 1$. We have also chosen a birth rate equal to the parameter, γ. As should be expected, we have found in practice that larger values of the birth rate produce better mixing but require more time in the computation of the algorithm.

Steps 3 to 5 are standard Gibbs sampling, see, for example, Gelfand and Smith (1990) whereby the model parameters are updated conditional on the mixture size, k. The only slightly complicated step is 5.2. where we introduce a Metropolis Hasting method, see Hastings (1970), to sample from the posterior distribution of ν. To do this, we generate candidate values for ν from a negative binomial proposal distribution. We have chosen this proposal distribution because, for large values of ν, the conditional distribution in (4) has a similar form to a negative binomial distribution.

Given the MCMC output of size N, we can estimate the predictive density of the interarrival time distribution using

$$f(t \mid \mathbf{s}, \mathbf{t}) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{k^{(n)}} w_i^{(n)} Er(s \mid \nu_i^{(n)}, \lambda_i^{(n)}).$$

For further details of this type of algorithm in the context of Bayesian inference for a normal mixture model, see Stephens (2000) or Hurn et al (2001).

3. Application to queues

Suppose now that we have obtained Monte Carlo samples of size N from the posterior distribution of the arrival parameters, via the BDMCMC algorithm, and the service parameter μ via direct sampling of the gamma density $f(\mu \mid \mathbf{s})$ Then we can make inference about the system. In particular, we can estimate the posterior distribution of the traffic intensity ρ. For example, the probability that the system is stable can be estimated with,

$$P\left(\rho < 1 \mid \mathbf{s}, \mathbf{t}\right) \approx \frac{1}{N} \# \left\{\rho^{(n)} < 1\right\},$$

where

$$\rho^{(n)} = \left(\mu^{(n)} \sum_{i=1}^{k^{(n)}} w_i^{(n)} \lambda_i^{(n)}\right)^{-1},$$

$\left\{\left(k^{(1)}, \mathbf{w}^{(1)}, \lambda^{(1)}, \nu^{(1)}\right), ..., \left(k^{(N)}, \mathbf{w}^{(N)}, \lambda^{(N)}, \nu^{(N)}\right)\right\}$ is the sample obtained from the BDMCMC algorithm and $\left\{\mu^{(1)}, ..., \mu^{(N)}\right\}$ is the sample generated from the posterior distribution of μ. If this probability is large enough, it may be reasonable to consider inference assuming that the system is stable.

It is well known, see Kleinrock (1975), that in queueing systems with non-Markovian interarrival process, the probability that an arriving customer finds j customers in the system, $\pi\left(j\right)$, differs from the probability that a random arrival finds j customers in the system, $p\left(j\right)$.

Specifically, for the $G/M/1$ system, the stationary probability describing the number of customers in the system at the arrival instants is a geometric distribution given by,

$$\pi\left(j \mid \sigma\right) = \left(1-\sigma\right)\sigma^{j}, \qquad j = 0, 1, 2, ...$$

where σ is the unique root in the interval $(0,1)$ of the equation,

$$\sigma = A^{*}\left(\mu - \mu\sigma\right), \tag{5}$$

and A^{*} is the Laplace transform of the interarrival time distribution. For our model, the arrival distribution is an Erlang mixture, therefore,

$$A^{*}\left(s\right) = \sum_{i=1}^{k} w_i \left(\frac{\nu_i/\lambda_i}{s+\nu_i/\lambda_i}\right)^{\nu_i}.$$

Given the model parameters, $\mu, k, \mathbf{w}, \lambda$ and ν, it is easy to approximate σ using the Newton-Raphson method or a similar procedure.

On the other hand, the steady-state distribution of the number of customers found by a random arrival is,

$$p\left(j \mid \rho, \sigma\right) = \begin{cases} 1-\rho, & \text{si } j=0 \\ \rho\pi\left(j-1 \mid \sigma\right), & \text{si } j \geq 1 \end{cases}$$

Given the MCMC output, we can approximate the predictive stationary distribution $\pi\left(j\right)$ using,

$$\pi\left(j\right) \approx \frac{1}{R}\sum_{n:\rho^{(n)}<1}\left(1-\sigma^{(n)}\right)\left(\sigma^{(n)}\right)^{j}$$

where $R = \#\left\{\rho^{(n)} < 1\right\}$ and $\sigma^{(n)}$ can be computed by solving (5) for every $k^{(n)}, \mathbf{w}^{(n)}, \lambda^{(n)}, \nu^{(n)}$ and $\mu^{(n)}$. Analogously, we can estimate the predictive distribution function of $p\left(j\right)$.

The probability of having an empty system is $\pi\left(0 \mid \sigma\right) = 1-\sigma$. Thus, the probability that a customer has to wait is σ. Using this fact and the conditional distribution of the waiting time in the queue, given that the queue exists, the unconditional waiting time distribution can be obtained. This is exponential with a jump of height $1-\sigma$ at $t=0$. The distribution function is,

$$F_W\left(t \mid \sigma, \mu\right) = 1 - \sigma e^{-\mu(1-\sigma)t}, \qquad t \geq 0.$$

As above, we can use the following approximation,

$$F_W\left(t\right) \approx \frac{1}{R}\sum_{n:\rho^{(n)}<1} F_W\left(t \mid k^{(n)}, \mathbf{w}^{(n)}, \lambda^{(n)}, \nu^{(n)}, \mu^{(n)}\right)$$

Wiper (1998) shows that, for any given $G/M/1$ system, given independent, continuous priors on the arrival and service rates with positive density in $\rho = 1$, the moments of the predictive distributions of waiting time and queue size do not exist. This is thus the case here.

4. Examples

In this section, we illustrate our method with both simulated and real data from several $G/M/1$ systems. In each case, we estimate the general interarrival time distribution and perform inference for the mixture size, k. We also obtain approximations of the predictive system size and waiting time distributions.

4.1. Simulated data

We consider three examples of $G/M/1$ queueing systems with the same traffic intensity $\rho = 1/3$ and the following interarrival times,

1. Exponential distribution with mean 3. (M)
2. Mixture of Erlang distributions with $\mathbf{w} = (0.2, 0.1, 0.7)$; $\lambda = (1.5, 2.5, 3.5)$ and $\nu = (200, 400, 600)$. (HEr)
3. Continuous uniform distribution defined on $(0, 6)$. (U)

We have included two examples of Erlang mixtures, viz cases 1 and 2. The third case, the uniform distribution, is unrelated to the Erlang mixture family.

For the exponential service time distribution, we here assume a fixed, known service rate, $\mu = 1$.

For each data set, we generated samples of 200 interarrival times and carried out the Bayesian analysis described in section 2. We ran 200000 iterations of the BDMCMC algorithm with 100000 for burn-in. The starting point used for the BDMCMC algorithm was, in each case, a long way from the true values of the mixture parameters, including the initial mixture size. For each case, we have set the same birth rate, $\gamma = 2$.

Figure 1 illustrates histograms of the generated arrival data superimposed (in dotted line) with the estimation of the predictive interarrival time densities for each example. The true densities are given in solid lines. In the first two cases, we observe that estimated and generating distributions are very similar. In the uniform example, the method predicts a relatively large number of components in order to fit the data and, as we might expect, the fit is a little bit worse.

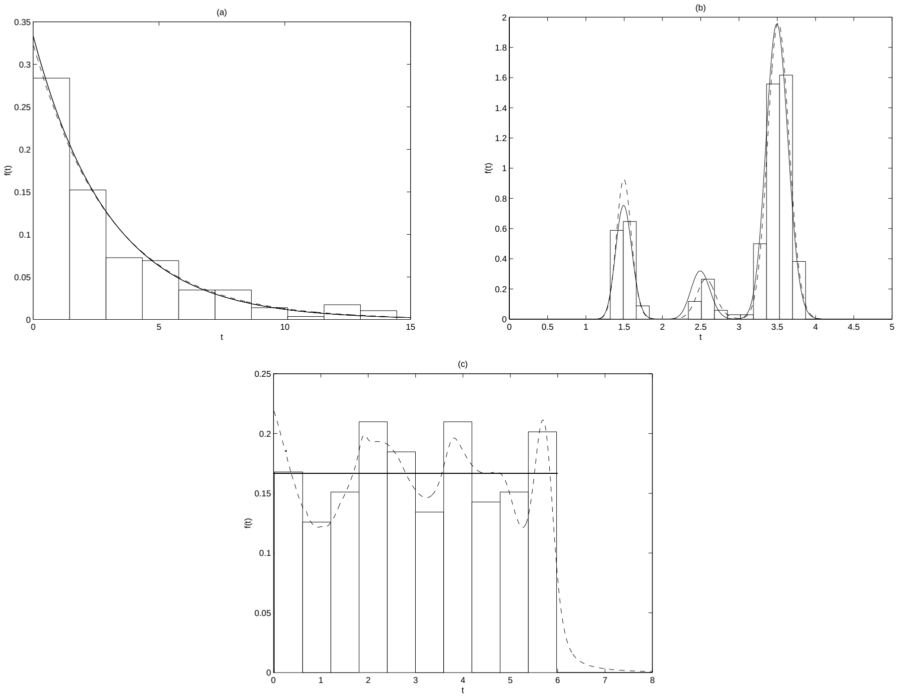

Figure 1. Predictive (dotted line) and true (solid line) interarrival time densities for (a) exponential (b) Erlang mixture and (c) uniform data sets.

In Table I, we tabulate the first few posterior probabilities of the mixture size, k. Note that $P\,(k = 1 \mid \mathbf{t}) \simeq 0.93$ for the exponential data set. Also, given $k = 1$, the posterior probability that the inter arrival time distribution is exponential is estimated to be $P\,(\nu = 1 \mid k = 1, \mathbf{t}) \approx 0.9999$. Thus, it is clear that the correct *M/M/1* model has been well identified in this case. In the Erlang mixture case, the method identifies also the correct number of component although with some uncertainty. In the uniform case, there is much more uncertainty.

The chain appears to be mixing quite well, visiting many states, in the three cases.

For all three systems, the estimated posterior probability that the traffic intensity is less than one is larger than 0.99. Thus, we can assume the system is stable and compute the stationary distributions of the queue. Table II gives the estimated posterior probabilities of the numbers of customers in each system at the arrival instants. The true stationary probabilities given the system parameters are also shown. Observe that these are close to the estimated ones, even in the uniform case.

Table I. Posterior probabilities for different mixture sizes.

$P(k \mid \mathbf{t})$	$M/M/1$	$HEr/M/1$	$U/M/1$
$k = 1$	0.928	0.000	0.000
$k = 2$	0.066	0.000	0.001
$k = 3$	0.004	0.859	0.083
$k = 4$	0.001	0.127	0.474
$k = 5$	0.000	0.012	0.319
$k = 6$	0.000	0.001	0.096

Table II. Estimated posterior probabilities $\pi(j \mid data)$ (upper) and true probabilities (lower) of the number of customers j in the system at the arrival instants.

j	$M/M/1$	$HEr/M/1$	$U/M/1$
π_0	.676 .666	.904 .909	.797 .791
π_1	.218 .222	.086 .083	.161 .165
π_2	.079 .074	.008 .007	.033 .034
π_3	.023 .025	.000 .000	.006 .007
π_4	.007 .008	.000 .000	.001 .001

We have also estimated the predictive distribution of the number of customers found by a random observer. For example, the estimated probability of finding an empty system is 0.676, 0.664 and 0.673, for cases 1 to 3 respectively. Recall that the true stationary probability is given by $p(0 \mid \rho, \sigma) = 1 - \rho = 2/3$ in all three cases.

Figure 2 shows in dotted line the estimation for the predictive waiting time distribution in the queue in each example. The true distribution is also illustrated in solid line. Observe that the jump of height $1 - \sigma$ at $t = 0$ is not represented. This value was given in Table II because, as is well known, it is equal to π_0.

4.2. Real data problem

Here, we undertake Bayesian analysis of a real data problem. We consider data on arrivals and service at a cashpoint in a bank in Madrid.

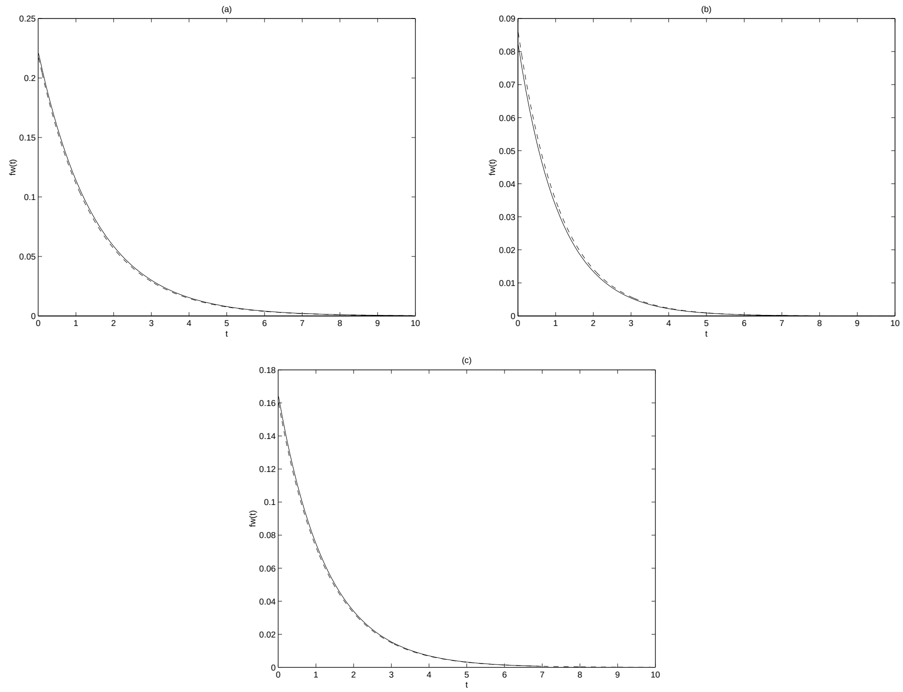

Figure 2. Predictive (dotted line) and true (solid line) waiting time densities for (a) exponential (b) Erlang mixture and (c) uniform interarrival time systems.

Interarrival and service times of 98 customers were recorded from 10:00 a.m. to 11:30 a.m. during three days. The mean service time was approximately 93.6 seconds. Our Bayesian density estimation method predicts a single, exponential distribution for the service time distribution. Thus, we assume this model for the service time. We also use a non-informative prior in (1) by setting a and b equal to zero. Thus, the posterior distribution of the service rate parameter, μ, is $G\left(98, 9172\right)$.

Figure 3 shows the histogram of the 98 interarrival times. The estimated density function using the mixture of Erlang distributions with the BDMCMC algorithm has been superimposed. None of the times is larger than two minutes and the distribution appears to be bimodal. In fact, the posterior probability of having two components is $P\left(k = 2 \mid \mathbf{t}\right) \approx 0.958$.

Given this arrival and service data, the posterior probability of having a stable system is estimated to be $P\left(\rho < 1 \mid \mathbf{s}, \mathbf{t}\right) \approx 0.85$. Assuming equilibrium, the predicted probability that an arriving customer has to wait is estimated to be

$$E\left[\sigma \mid \rho < 1, \mathbf{t}, \mathbf{s}\right] \approx 0.805$$

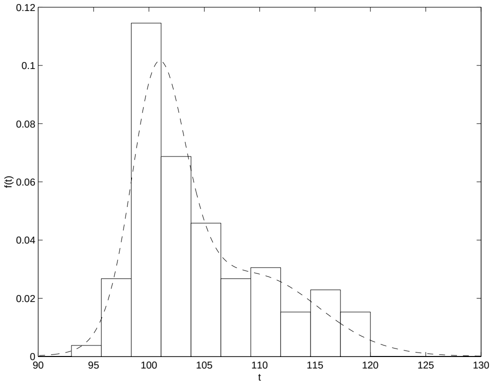

Figure 3. Histogram of interarrival time data and the estimated interarrival time density.

and the estimated expected value of the traffic intensity is

$$E\left[\rho \mid \rho < 1, \mathbf{t}, \mathbf{s}\right] \approx 0.896.$$

Figure 4 illustrates the estimated probabilities describing the number of customers found by an arriving customer and by a random observer.

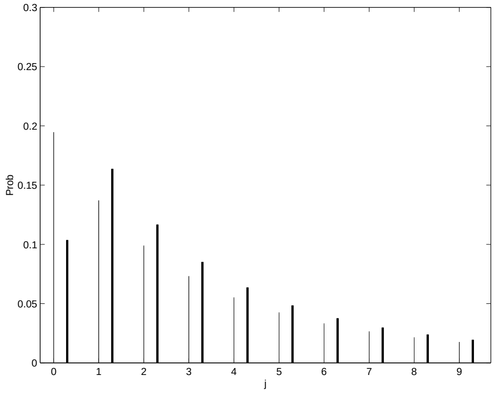

Figure 4. Predictive probabilities for the system size found by an arriving customer (solid line) and by a random observer (bold solid line).

5. Conclusions

In this paper, we have developed a Bayesian approach to make inference and prediction for *G/M/1* systems. We have developed a density estimation method based on mixtures of Erlang distributions in order to approximate the general interarrival time distribution. To make inference on the arrival parameters, we have implemented an MCMC algorithm based on births and deaths of mixture components making use of the BDMCMC technique proposed by Stephens (2000). Some important measures of the system, such as the system size or the waiting time, has been predicted. We have illustrated this methodology with simulated and real data.

It is possible to extend our approach to queueing systems with c servers. Given the system parameters, the stationary distribution of the system size and waiting time can be easily derived, see Allen (1990). Then, the predictive distributions can be estimated via the BDMCMC algorithm as for the *G/M/1* system. Note however that in such cases, the computational cost can increase dramatically as the number of servers c increases. See also Wiper (1998).

An alternative to the BDMCMC methodology is the "reversible jump" introduced by Richardson and Green (1997). This type of this algorithm had been used in a previous work to make inference on the general service time distribution for a *M/G/1* system, see Ausin et al. (2001) . In practice, we have found that there both schemes perform similarly. A disadvantage of the reversible jump scheme is that we often find that mixing of the chain and thus convergence of the algorithm is very slow because of problems due to the discrete support of ν. In the BDMCMC algorithm, as we have indicated, larger values of the birth rate produce better mixing, but also increasing the computational cost. We have also found some problems of non-convergence of the algorithm if the birth rate elected is very high. Thus, it would be useful to explore methods for election of this parameter in order to optimize the algorithm.

In this article, we have approximated the general interarrival time using a mixture of Erlang distributions as an extension of the analysis of the *Er/M/1* system in Wiper (1998). The assumption that the elements of the parameter ν are integers can be removed considering mixtures of gamma distributions which is a more flexible model. Wiper et al. (2001) consider this family and describe a density estimation method using reversible jump techniques although they do not apply their results to $G/M/1$ queues. An advantage of using the Erlang mixture structure as here is that we can directly calculate the probability that a simpler interarrival time model (exponential, Erlang or hyperexponential) is appropriate given the MCMC output. For example, we can estimate the posterior probability that the interarrival distribution is a single Erlang distribution by the proportion of times that the mixture size $k = 1$ in the MCMC sample. See also Ausin et al (2001).

Many other dense families can be used to model a general distribution, for example, the mixed generalized Erlang distributions. An advantage of this class is that the Laplace transforms of the queue length, waiting time and busy period distributions of a system with mixed generalized Erlang distributions for the interarrival and service times are known, see Bertsimas and Nazakato (1992). Theoretically it is possible to combine these results with variable dimensional MCMC methods to make inference and prediction for *G/G/1* systems. Work on this problem is currently in progress.

References

Allen, A: *Probability, Statistics and Queueing Theory with Computer Science Applications.* Academic Press, USA, 1990.

Abramowitz, M., Stegun, I.A: *Handbook of Mathematical Functions.* Dover, USA, 1964.

Armero, C: Bayesian analysis of *M/M/1/*∞/FIFO queues. In J. Bernardo, M. DeG-root, D. Lindley and A. Smith, editors, *Bayesian Statistics 2*, pp. 613-618, North Holland, Amsterdam, 1985.

Armero, C. and M.J. Bayarri: 1994. Bayesian prediction in M/M/1 queues. *Queueing Systems*, **15**, 401-417.

Armero, C. and M.J. Bayarri: Bayesian questions and Bayesian answers in queues. In J. Bernardo, J. Berger, A. Dawid and A. Smith, editors, *Bayesian Statistics 5*, pp. 3-23. Oxford, University Press, 1995.

Armero, C. and M.J. Bayarri: 1997. A Bayesian analysis of a queueing system with unlimited service. *Journal of Statistical Planning and Inference*, **58**, 241-261.

Armero, C. and D. Conesa: 1998. Inference and prediction in bulk arrival queues and queues with service in stages. *Applied Stochastic Models and Data Analysis*, **14**, 35-46.

Asmussen, S: *Applied probability and queues.* New York, Wiley, 1987.

Ausin, M.C., M.P. Wiper and R.E. Lillo: . Bayesian estimation for the M/G/1 queue using a phase type approximation. Working Paper, 01-30, Statistics and Econometrics Series 43, Universidad Carlos III de Madrid, 2001.

Bagchi, T.P. and A.A. Cunningham: 1972. Bayesian approach to the design of queueing systems. *Informs*, **10**, 36-46.

Bertsimas D.J. and D.N. Nazakato: 1992. Transient and busy period of the GI/G/1 queue: The method of stages. *Queueing Systems*, **10**, 153-184.

Diebolt, J. and C.P. Robert: 1994. Estimation of finite mixture distributions. *Journal of the Royal Statistical Society, B*, **56**, 363-375.

Gelfand, A.E. and A.F.M. Smith: 1990. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Green, P: 1995. Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82**, 711-732.

Hastings, W.K: 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Hurn, M., A. Justel and C.P. Robert: 2001. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, in press.

Kleinrock, L: *Queueing Systems.Volume I: Theory.* New York, JohnWiley, 1975.

Lehoczky, J: 1990. Statistical methods. In D. Heyman and M. Sobel, editors, *Stochastic Models*, 225-294, North-Holland. Amsterdam.

McGrath, M.F. and N.D. Singpurwalla: 1987. A subjective Bayesian approach to the theory of queues II - inference and information in M/M/1 queues. *Queueing Systems*, **1**, 335-353.

McGrath, M.F., D. Gross and N.D. Singpurwalla: 1987. A subjective Bayesian approach to the theory of queues I - Modeling. *Queueing Systems*, **1**, 317-333.

Muddapur, M.V: 1972. Bayesian estimates of parameters in some queueing models. *Annals of the Institute of Mathematics*, **24**, 327-331.

Reynolds, J.F: 1973. On estimating the parameters of a birth-death process. *Australian Journal of Statistics*, **15**, 35-43.

Richardson, S. and P. Green: 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, B*, **59**, 731-792.

Rios, D., M.P. Wiper and F. Ruggeri: 1998. Bayesian analysis of M/Er/1 and M/H$_k$/1 queues. *Queueing Systems*, **30**, 289-308.

Robert , C: Mixtures of distributions : inference and estimation. In W.R. Gilks, S.Richardson and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, 441-464. London: Chapman & Hall, 1996.

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

Wiper, M.P., D. Rios Insua and F. Ruggeri: 2001. Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, in press.