Working Paper 01-30
Statistics and Econometrics Series 19
June 2001

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

BAYESIAN ESTIMATION FOR THE M/G/1 QUEUE USING A PHASE TYPE APPROXIMATION

M.C. Ausín, M.P. Wiper, R.E. Lillo*

Abstract

This article deals with Bayesian inference and prediction for M/G/1 queueing systems. The general service time density is approximated with a class of Erlang mixtures which are phase type distributions. Given this phase type approximation, an explicit evaluation of measures such as the stationary queue size, waiting time and busy period distributions can be obtained. Given arrival and service data, a Bayesian procedure based on reversible jump Markov Chain Monte Carlo methods is proposed to estimate system parameters and predictive distributions.

Keywords: Queues; Bayesian mixtures; reversible jump MCMC; phase type distributions, matrix geometric methods.

*Ausín Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), Spain, e-mail: causin@est-econ.uc3m.es; Wiper, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Tfno: 91-6249852, e-mail: mwiper@est-econ.uc3m.es ; Lillo, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Tfno: 91-6249857, e-mail: lillo@est-econ.uc3m.es.

Bayesian Estimation for the $M/G/1$ queue using a phase type approximation

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

Departamento de Estadística y Econometría

Universidad Carlos III de Madrid

Madrid, 126, 28903 Getafe, Madrid, Spain

June 11, 2001

Abstract

This article deals with Bayesian inference and prediction for *M/G/1* queueing systems. The general service time density is approximated with a class of Erlang mixtures which are phase type distributions. Given this phase type approximation, an explicit evaluation of measures such as the stationary queue size, waiting time and busy period distributions can be obtained. Given arrival and service data, a Bayesian procedure based on reversible jump Markov Chain Monte Carlo methods is proposed to estimate system parameters and predictive distributions.

Keywords: queues, Bayesian mixtures, reversible jump MCMC, phase type distributions, matrix geometric methods.

AMS Classification: 62F15, 60K25

1 Introduction

Bayesian analysis of queueing systems seems to have first been considered in the early 1970's; see Bagchi and Cunningham (1972), Muddapur (1972) and Reynolds (1973). In the mid 1980's there was a revival of interest in the subject, see Armero (1985), McGrath et al. (1987) and McGrath and Singpurwalla (1987) and in recent years, there have been an increasing number of papers using Bayesian techniques. Some useful references are Armero and Bayarri (1994, 1995, 1997), Armero and Conesa (1998), Wiper (1998) and Rios Insua et al. (1998).

Up to now, most papers have considered the simple Markovian queue $M/M/c$ with exponential interarrival and service times and various numbers of servers. Exceptions are, for example, Wiper (1998) where the $Er/M/c$ queue (with Erlang interarrival times) was analyzed and Ríos Insua et al. (1998) where the $M/Er/1$ and $M/H_k/1$ (hyperexponential service time) systems were considered. One reason for these choices is that in practice, given service data, the service distribution is often modelled by an exponential, Erlang and hyperexponential approximation according to whether the sample coefficient of variation is approximately equal, less than or more than 1, respectively, see e.g. Allen (1990) or Nelson (1995). It is not always clear that such simple approximations will be adequate, see e.g. Rios et al. (1998). Thus, the objective for this paper is to

consider Bayesian inference for queues with more general service distributions, i.e. the *M/G/1* family.

In this paper, we will consider a semiparametric model for the service distribution based on a mixture of Erlang densities. One advantage of this model is that the Erlang mixture family is dense over the set of distributions on the positive reals, see Asmussen (1987). Another advantage in the queueing context is that this family includes the Erlang, hyperexponential and exponential densities as special cases. However, the main reason for choosing a hyperErlang distribution is that this is a continuous phase-type (PH) distribution. Thus, it is possible to apply some useful results obtained by Neuts (1981) for queueing systems with phase-type service time distribution and exponential interarrival time.

There has been much previous work on Bayesian density estimation using mixture models. See, for example Diebolt and Robert (1994) and Richardson and Green (1997) for normal mixtures, Gruet et al. (1998) and Rios et al. (1998) for exponential mixtures and Wiper et al. (2001) for mixtures of gamma distributions. Note that the advantage of using a mixture of Erlang distributions, as opposed to a gamma mixture, to model the service distribution is that in the first case, assuming that the queue is stable, queue size, waiting time and busy period distributions of the queue can all be evaluated whereas using a gamma mixture, it is only possible to evaluate the queue size distribution. See Wiper et al. (2001).

An outline of our paper is as follows. Throughout we will assume the

following FIFO queueing system. Let T be the interarrival time, then T has an exponential distribution conditional on some (unknown) parameter λ, i.e.

$$f\left(t \mid \lambda\right) = \lambda \exp\left(-\lambda t\right), \qquad 0 < t < \infty.$$

For service times, S, we assume a mixture of k Erlang distributions with parameters $\mathbf{w}, \boldsymbol{\mu}$ and $\boldsymbol{\nu}$,

$$f(s \mid k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \sum_{i=1}^{k} w_i Er(s \mid \nu_i, \mu_i),$$

where

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

and the mixture size, k, and all other parameters are unknown.

In Section 2, we describe a simple experiment for observing arrival and service data. Prior distributions are defined for the unknown model parameters and a reversible jump MCMC algorithm (see Green 1995, Richardson and Green 1997) for sampling the posterior service distribution is introduced. We also comment on the selection of simpler models for the service distribution.

In Section 3, we consider the problem of estimating the equilibrium characteristics of the queue. Firstly, we briefly review the definitions and basic properties of phase-type distributions. Secondly, we illustrate how to obtain the stationary distribution of the queue size, the waiting time and the length of a busy period given the system parameters. Finally, we compute the predictive distributions of these measures 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 A simple experiment and inference for the system parameters

We wish to make inference for the system parameters $\lambda, k, \mathbf{w}, \boldsymbol{\mu}$ and $\boldsymbol{\nu}$. A simple experiment providing complete information consists in observing m_a interarrival times $\mathbf{t} = \{t_1, ..., t_{m_a}\}$ and m_s service times $\mathbf{s} = \{s_1, ..., s_{m_s}\}$. This experiment has also been considered in, for example, Thiruvaiyaru and Basawa (1992), Rios et al. (1998) and Armero and Bayarri (1994). Given this $\mathbf{t}$ and $\mathbf{s}$, the likelihood is of form

$$L(\lambda, \boldsymbol{\theta} \mid \mathbf{t}, \mathbf{s}) \propto L(\lambda \mid \mathbf{t})\, L(\boldsymbol{\theta} \mid \mathbf{s})$$

where $\boldsymbol{\theta} = (k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu})$ are the service parameters,

$$L(\lambda \mid \mathbf{t}) = \lambda^{m_a} \exp\left(-\lambda \sum_{j=1}^{m_a} t_j\right)$$

and

$$L(\boldsymbol{\theta} \mid \mathbf{s}) = \prod_{j=1}^{m_s} \left(\sum_{i=1}^{k} w_i Er(s_j \mid \nu_i, \mu_i)\right).$$

Accordingly, given independent prior distributions for λ and $\boldsymbol{\theta}$, then the arrival and service parameters are independent a posteriori as well.

Suppose that we define a gamma prior distribution for λ; say $\lambda \sim G(a, b)$. Then, it is straightforward to show that the posterior distribution will be

$$\lambda \mid \mathbf{t} \sim G\left(a + m_a, b + \sum_{j=1}^{m_a} t_j\right),$$

see also, Armero and Bayarri (1994) and Rios et al. (1998).

For the service parameter, $\boldsymbol{\theta}$, given a prior distribution, Bayesian inference may be performed using MCMC methods, which involve the construction of a Markov chain with the posterior distribution $f(\boldsymbol{\theta} \mid \mathbf{s})$ as its stationary distribution, see e.g. Tierney (1996). Given suitable regularity conditions, a sample of the full posterior distribution of the mixture parameters, including the mixture size, k, can be obtained.

In the following subsection, we define a suitable prior distribution for the service parameter, $\boldsymbol{\theta}$, and describe an MCMC algorithm that can be used to sample from the posterior distribution. This algorithm is similar to that used in Wiper et al. (2001) for a mixture of gamma distributions but we carry out some modifications due to the discrete support of the parameter $\boldsymbol{\nu}$. We also introduce a birth death move following Richardson and Green (1997).

2.1 Prior distribution for $\boldsymbol{\theta}$ and an MCMC algorithm to sample $f(\boldsymbol{\theta} \mid \mathbf{s})$

Here, we will define a prior distribution for $\boldsymbol{\theta}$. We shall treat the constituent parts, $k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}$, separately. Firstly, we define a discrete prior for the mixture size, k, for example, a discrete uniform taking values from 1 to $k_{\max}$. Condi-

tional on k, we can now define prior distributions for the remaining parameters.

Firstly, as in Diebolt and Robert (1994), we introduce indicator variables, $Z_1, ..., Z_{m_s}$, for each observed service time $s_1, ..., s_{m_s}$, with the following prior distributions,

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

and thus, we have that the conditional distribution of the service time, T_j, is,

$$T_j \mid Z_j = i \sim Er(\nu_i, \mu_i), \qquad \text{for } j = 1, ..., m_s.$$

Now we can define a prior distribution for $\boldsymbol{\theta} \mid \mathbf{z}$,$k$. We will assume that the joint prior distribution can be factorized as,

$$f(\boldsymbol{\theta} \mid \mathbf{z},k) \propto f(\mathbf{z} \mid k, \mathbf{w})\, f(\mathbf{w} \mid k)\, f(\boldsymbol{\mu} \mid k)\, f(\boldsymbol{\nu} \mid k)$$

and we now define proper but diffuse distributions for $\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}$.

Firstly, following Wiper et al. (2001), we assume

$$\mathbf{w} \mid k \sim D(\phi_1, ..., \phi_k), \qquad \text{(a Dirichlet distribution)}$$

Typically, we might set $\phi_i = 1$, for all $i = 1, ..., k$, giving a uniform prior for the weights.

Secondly, we assume that the joint prior for the mean parameters $\boldsymbol{\mu}$ is proportional to a product of inverted gamma distributions:

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

restricted to the set $\mu_1 < ... < \mu_k$ for identifiability purposes. Typically we might set $\alpha = 1.1$ and $\beta = 1$. Note that if $\alpha \leq 1$ then the prior moments of μ do not exist which also implies that the posterior moments will not exist either. Here, in particular, we are concerned with predicting the traffic intensity of the queue which is given by

$$\rho = \lambda \sum_{i=1}^{k} w_i \mu_i.$$

Clearly, if the moments of μ do not exist then neither will the moments of ρ.

Finally, we use a geometric prior distribution with mean $1/\vartheta$ for the integer parameters, ν_i. We set, for example, $\vartheta = 0.01$.

Given k and the prior distributions, then it is straightforward to calculate the conditional posterior distributions. Firstly,

$$P(Z_j = i \mid \mathbf{s}, k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}) \propto w_i \frac{(\nu_i/\mu_i)^{\nu_i}}{\Gamma(\nu_i)} s_j^{\nu_i - 1} \exp(-\frac{\nu_i}{\mu_i} s_j), \qquad \text{for } i = 1, ..., k,$$

and also we have,

$$\mathbf{w} \mid \mathbf{s}, \mathbf{z}, k \sim D(\phi_1 + m_1, ..., \phi_k + m_k)$$

$$\mu_i \mid \mathbf{s}, \mathbf{z}, k \sim IG(\alpha + m_i \nu_i, \beta + S_i \nu_i)$$

where $m_i = \#\{Z_j = i\}$ and $S_i = \sum_{j: Z_j = i} s_j$, for $i = 1, ..., k$. Finally,

$$f(\nu_i \mid \mathbf{s}, \mathbf{z}, k, \mathbf{w}, \boldsymbol{\mu}) \propto \frac{\nu_i^{m_i \nu_i}}{\Gamma(\nu_i)^{m_i}} \exp\left\{-\nu_i \left(-\log(1-\vartheta) + \frac{S_i}{\mu_i} + m_i \log \mu_i - \log P_i\right)\right\} \tag{2}$$

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

Now, we can define an MCMC algorithm, based on Richardson and Green (1997), consisting of modified Gibbs sampling as follows:

1. Set initial values for $k^{(0)}, \mathbf{w}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\nu}^{(0)}$.
2. Update the allocation by sampling from $\mathbf{z}^{(n+1)} \sim \mathbf{z}|\mathbf{s}, k^{(n)}, \mathbf{w}^{(n)}, \boldsymbol{\mu}^{(n)}, \boldsymbol{\nu}^{(n)}$.
3. Update the weights by sampling from $\mathbf{w}^{(n+1)} \sim \mathbf{w}|\mathbf{s}, \mathbf{z}^{(n+1)}, k^{(n)}$.
4. For $i = 1, ..., k$,
   a. Update the means by sampling from $\mu_i^{(n+1)} \sim \mu_i|\mathbf{s}, \mathbf{z}^{(n+1)}, k^{(n)}$.
   b. Update ν_i using a Metropolis step.
5. Order $\boldsymbol{\mu}^{(n+1)}$ and sort $\mathbf{w}^{(n+1)}$ and $\boldsymbol{\nu}^{(n+1)}$.
6. Split one mixture component into two, or combine two into one.
7. Birth or death of an empty component.
8. $n = n + 1$. Go to 2.

In step 4b, we generate candidate values for $\tilde{\nu}_i$ from a Negative Binomial proposal distribution,

$$f_{NB}(\nu) = \begin{pmatrix} r+\nu-2 \\ \\ \nu-1 \end{pmatrix} p^r (1-p)^{\nu-1}, \qquad \nu = 1, 2, ...$$

It is easy to see that, for large values of ν_i, the conditional posterior distribution in (2) is of a similar form to this distribution. We propose choosing the parameters (r, p) such that the mode of the Negative Binomial is equal to the previous value of ν_i and the variance produces a favourable acceptance rate.

For example, we have found in practical tests that setting $r = \nu_i^{(n)} + 1$ and $p = (r-1)/\nu_i^{(n)} + r - 1.5$, (where $\nu_i^{(n)}$ represents the value of the previous iteration) works well. The candidate, $\tilde{\nu}_i$, is accepted with probability $\alpha = \min\{1, A\}$, where

$$A = \frac{f(\nu_i \mid \mathbf{s}, \mathbf{z}^{(n+1)}, k^{(n)}, \mathbf{w}^{(n+1)}, \mu^{(n+1)}) p(\tilde{\nu}_i, \nu_i^{(n)})}{f(\nu_i^{(n+1)} \mid \mathbf{s}, \mathbf{z}^{(n+1)}, k^{(n)}, \mathbf{w}^{(n+1)}, \mu^{(n+1)}) p(\nu_i^{(n)}, \tilde{\nu}_i)},$$

where $p(\nu_i^{(n)}, \tilde{\nu}_i)$ is the probability of generating $\tilde{\nu}_i$ given the previous value $\nu_i^{(n)}$.

In steps 6 and 7, we introduce a reversible jump to let the chain move through the posterior distribution of the mixture size, k; see Green (1995) and Richardson and Green (1997). Firstly, in step 6, a mixture size candidate, $\tilde{k}$, is generated making a random choice between splitting a mixture component into two or combining two components.

If the combine move is selected, we choose at random two adjacent components (i_1, i_2) to be merged, reducing k by one. Then, we modify the other parameters as follows,

1. $\tilde{w} = w_{i_1} + w_{i_2}$.
2. $\tilde{w}\tilde{\mu} = w_{i_1}\mu_{i_1} + w_{i_2}\mu_{i_2}$.
3. $\tilde{\nu} = \nu_{i_1}$.

Using these transformations, we preserve the moments of order 0 and 1 of the service time distribution.

For the split move, a component, i, is elected at random to be split into two. To specify the new parameters we generate two values, u_1 and u_2 from an

$U(0,1)$ and a third one, u_3, from a Negative Binomial distribution with mode ν_i. Then, we define

1. $\tilde{w}_{i_1} = u_1 w_i, \qquad \tilde{w}_{i_2} = (1-u_1)w_i.$

2. $\tilde{\mu}_{i_1} = \mu_{i-1} u_1 w_1, \qquad \tilde{\mu}_{i_2} = \frac{1}{1-u_1}(1-u_1u_2)\mu_i - u_1(1-u_2)\mu_{i-1}.$

3. $\tilde{\nu}_{i_1} = \nu_i, \qquad \tilde{\nu}_{i_2} = u_3.$

Finally, we accept the move with probability

$$\min\left\{1, \frac{f(\tilde{\boldsymbol{\theta}} \mid \mathbf{s})p(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta})}{f(\tilde{\boldsymbol{\theta}} \mid \mathbf{s})p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})}\right\}, \tag{3}$$

where $f(\boldsymbol{\theta} \mid \mathbf{s})$ is the posterior service time parameters distribution and $p(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is the probability of moving from $\boldsymbol{\theta} = (k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu})$ to $\tilde{\boldsymbol{\theta}} = (\tilde{k}, \tilde{\mathbf{w}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}})$.

It would have been natural to think about a transformation preserving the moment of second order, but it can be shown that this is not possible because of the discrete distribution of ν.

In step 7, we again generate a mixture size candidate, $\tilde{k}$, making a random choice between a birth or a death of a mixture component. If a birth is elected, we generate a weight from a beta distribution (be) and the proposed new component parameters from the prior distribution,

$$w_j^* \sim be\,(1,k)\,, \qquad \nu_j^* \sim geo\,(\vartheta)\,, \qquad \mu_j^* \sim IG\,(\alpha,\beta)\,.$$

Otherwise, for a death, an empty component is randomly chosen from the existing ones and it is deleted. In both cases, the weights are rescaled to sum 1. The acceptance probability is the corresponding expression for (3) .

2.2 Estimating parameters from the MCMC output and model selection

Given a sample realization of the Markov chain we have just defined and a sample of equal size from $f(\lambda \mid \mathbf{t})$, we can estimate various quantities of interest. For example, given the sample data, we will often wish to assess whether or not the model is stable. The queue is stable if and only if the traffic intensity, ρ is less than one. Thus we can estimate the probability of having a stable queue with,

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

where

$$\rho^{(n)} = \lambda^{(n)} \sum_{i=1}^{k^{(n)}} w_i^{(n)} \mu_i^{(n)}$$

and $\left\{ \left(k^{(1)}, \mathbf{w}^{(1)}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\nu}^{(1)}\right), ..., \left(k^{(N)}, \mathbf{w}^{(N)}, \boldsymbol{\mu}^{(N)}, \boldsymbol{\nu}^{(N)}\right) \right\}$ is a sample of size N obtained from the MCMC algorithm and $\left\{ \lambda^{(1)}, ..., \lambda^{(N)} \right\}$ is a sample of size N generated from the posterior distribution of λ. A consistent estimator of the traffic intensity, ρ, is

$$E[\rho \mid \mathbf{t}, \mathbf{s}] \approx E[\lambda \mid \mathbf{t}] \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{k^{(n)}} w_i^{(n)} \mu_i^{(n)},$$

where $E[\lambda \mid \mathbf{t}] = (a + m_a) \left(b + \sum_{j=1}^{m_a} t_j \right)^{-1}$.

We can also perform inference for the mixture size, k, estimating the marginal posterior distribution by

$$P(k \mid \mathbf{t}, \mathbf{s}) \approx \frac{1}{N} \# \left\{ n : k^{(n)} = k \right\}.$$

This provides a tool for service time model selection. For example, the posterior probability of having a single Erlang distribution for the service time distribution would be $P(k=1 \mid \mathbf{s})$. If this probability is big enough, we can model the system as an *M/Er/1* queue and use a simpler algorithm to make inference about the system parameters, see Rios et al. (1998). Analogously, we could estimate the posterior probability that the service time distribution is distributed as a hyperexponential distribution with

$$P(\boldsymbol{\nu}=1 \mid \mathbf{s}) \approx \frac{1}{N}\#\left\{n : \nu_i^{(n)}=1 \text{ for } i=1,..,k^{(n)}\right\}.$$

That is the probability a posteriori of having an $M/H_k/1$ system. Again, simpler schemes for this model are available in Rios et al. (1998). The posterior probability that the service time distribution is exponential can be approximated using

$$P(\boldsymbol{\nu}=1, k=1 \mid \mathbf{s}) \approx \frac{1}{N}\#\left\{n : k^{(n)}=1 \text{ and } \nu^{(n)}=1\right\}.$$

This case corresponds to an *M/M/1* system. Exact results for this system are given in Armero and Bayarri (1994).

3 Prediction in Equilibrium

Henceforth, we will assume that the model is stable and then, a stationary distribution exists. We are interested in observable quantities such as the number of customers in the queue, the waiting time in the queue and the length of busy periods.

In this section, we will first assume that the system parameters $(\lambda, \boldsymbol{\theta})$ are known. Firstly, we briefly introduce the notation and more useful properties for the class of phase-type distributions. Then, we use the matrix-geometric approach developed in Neuts (1981) to obtain the stationary distributions related to the $M/PH/1$ model. We conclude by showing how to estimate the predictive stationary distributions given a sample of data from the posterior distribution of $(\lambda, \boldsymbol{\theta})$.

3.1 The phase-type distributions

A continuous phase-type distribution of order m is defined as the distribution on $[0, \infty)$ of the time until absorption in a finite Markov process on the states $\{1, ..., m+1\}$ with infinitesimal generator

$$Q = \begin{bmatrix} T & \mathbf{T}^0 \\ \mathbf{0} & 0 \end{bmatrix},$$

where the $m \times m$ matrix T satisfies $T_{ii} < 0$, for $1 \leq i \leq m$, and $T_{ij} \geq 0$, for $i \neq j$ and also $T\mathbf{e} + \mathbf{T}^0 = \mathbf{0}$. The initial probability vector of Q is $(\boldsymbol{\alpha}, \alpha_{m+1})$, with $\boldsymbol{\alpha}\mathbf{e} + \alpha_{m+1} = 1$, where $\mathbf{e}$ is an $m \times 1$ unit vector. The distribution function is given by,

$$F(x) = 1 - \boldsymbol{\alpha} \exp\{Tx\}\, \mathbf{e}, \qquad \text{for } x \geq 0,$$

where $\exp\{Tx\}$ is the matrix exponential of Tx. The distribution is defined by the pair $(\boldsymbol{\alpha}, T)$.

The simplest phase-type distribution is the exponential distribution. In this case, $m = 1$ and $(\alpha, T) = (1, -\frac{1}{\mu})$, where μ is the mean parameter.

The Erlang distribution $Er(\nu, \mu)$ is phase-type of order ν with representation $(\boldsymbol{\alpha}_{Er}, T_{Er})$, where $\boldsymbol{\alpha}_{Er}$=$(1, 0, ..., 0)_{(1\times\nu)}$ and

$$T_{Er} = -\frac{\nu}{\mu}\begin{bmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & ... & & & \\ & & & 1 & -1 & \\ & & & & 1 & \end{bmatrix}_{(\nu\times\nu)}.$$

An useful property is that a finite mixture of phase-type distributions is a phase-type distribution. Thus a mixture of Erlang distribution (HEr) is phase-type. If $(w_1, ..., w_k)$ are the weights of the mixture and component i has the representation $\left(\boldsymbol{\alpha}_{Er}^i, T_{Er}^i\right)$, $1 \leq i \leq k$, then the Erlang mixture has the representation $\boldsymbol{\alpha}_{HEr}$= $\left[w_1\boldsymbol{\alpha}_{Er}^k, ..., w_k\boldsymbol{\alpha}_{Er}^k\right]$, and

$$T_{HEr} = \begin{bmatrix} T_{Er}^1 & 0 & \cdots & 0 \\ 0 & T_{Er}^2 & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & T_{Er}^k \end{bmatrix}.$$

In the following subsections, assuming that the model parameters, $(\lambda, \boldsymbol{\theta})$, are known, we show how to compute the stationary distribution of measures of performance of the queue such as the number of customers in the system, the queueing time and the length of a busy period using the well-known results for the $M/PH/1$ queue.

3.2 Number of customers in the *M/HEr/1* queue

Suppose that $\pi(j)$ is the equilibrium probability that there are j customers in the system. Then it can be shown that

$$\pi(j) = \pi(0)\,\delta_j + \sum_{i=1}^{j+1} \pi(j)\,\delta_{j-i+1}, \qquad j = 0, 1, ...$$

with $\pi(0) = 1 - \rho$, where δ_j is the probability that j customers arrive during a service interval, see e.g. Nelson (1995). For the Erlang mixture form of the service distribution, we have, see also Wiper et al. (2001),

$$\delta_j = \int_0^\infty \frac{(\lambda s)^j}{j!} e^{-\lambda s} f(s|\boldsymbol{\theta})\, ds = \lambda^j \sum_{i=1}^{k} w_i \begin{pmatrix} j+\nu_i-1 \\ j \end{pmatrix} \left(\frac{1}{1+\lambda\frac{\mu_i}{\nu_i}}\right)^{\nu_i} \left(\frac{\nu_i}{\mu_i}+\lambda\right)^{-j}.$$

3.3 Waiting time in the *M/HEr/1* queue.

The stationary queueing time in the $M/PH/1$ system is phase-type, see Neuts (1981, p. 57). In particular, given the Erlang mixture form of the service distribution, the stationary queueing time distribution is phase-type with representation, (α_W, T_W), where,

$$\boldsymbol{\alpha}_W = \rho\boldsymbol{\psi}, \qquad T_W = T_{HEr} + \rho\mathbf{T}^0_{HEr}\boldsymbol{\psi},$$

where $\boldsymbol{\psi} = (\boldsymbol{\psi}_{\nu_1}, ..., \boldsymbol{\psi}_{\nu_k})$ is the stationary probability vector of $T_{HEr} + \mathbf{T}^0_{HEr}\boldsymbol{\alpha}_{HEr}$. Thus, $\boldsymbol{\psi}$ is the unique solution to the equations

$$\boldsymbol{\psi}(T_{HEr} + \mathbf{T}^0_{HEr}\boldsymbol{\alpha}_{HEr}) = 0, \qquad \boldsymbol{\psi}\mathbf{e} = 1.$$

It is straightforward to show that

$$\boldsymbol{\psi}_{\nu_i} = \left(\sum_{i=1}^{k} w_i \mu_i \right)^{-1} \frac{w_i \mu_i}{\upsilon_i} \mathbf{e}'_{\nu_i},$$

where $\mathbf{e}'_{\nu_i}$ is the $1 \times \nu_i$ unit vector. Therefore,

$$\boldsymbol{\alpha}_W = \rho \boldsymbol{\psi} = \lambda [\frac{w_1 \mu_1}{\upsilon_1} \mathbf{e}'_{\nu_1}, ..., \frac{w_k \mu_k}{\upsilon_k} \mathbf{e}'_{\nu_k}],$$

and

$$T_W = T_{HEr} + \lambda M,$$

where M is a block matrix where each block M_{ν_i,ν_j} is a $\nu_i \times \nu_j$ matrix such that,

$$M_{\nu_i,\nu_j} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ w_j \frac{\upsilon_i \mu_j}{\mu_i \upsilon_j} & \cdots & w_j \frac{\upsilon_i \mu_j}{\mu_i \upsilon_j} \end{bmatrix}.$$

Now, we can calculate the waiting time density function $f_W(t)$ given the parameters,

$$f_W(t) = \begin{cases} 1 - \rho, & \text{if } t = 0 \\ \boldsymbol{\alpha}_W \exp\{T_W t\} \mathbf{T}_W^0, & \text{if } t > 0 \end{cases}$$

where $\exp\{T_W t\}$ is the matrix exponential of $T_W t$.

For large k or any large $\{\nu_i, i = 1, ..., k\}$, the computation of $\exp\{T_W t\}$ becomes impractical due to storage requirements. In such cases, we can obtain $\boldsymbol{\alpha}_W \exp\{T_W t\}$ by solving the linear system of differential equations

$$\chi'(t) = \chi(t) T_W, \qquad \text{with } \chi(0) = \boldsymbol{\alpha}_W$$

using a classical Runge-Kutta method of low order, see e.g. Abramowitz and Stegun (1964).

3.4 Busy Period in the *M/HEr/1* queue.

Neuts (1977) shows how to obtain the distribution of the busy period in a stable *M/PH/1* queue. The busy period distribution can be considered as that of the time till absorption in an infinite state Markov chain. Thus, the computations are reduced to the numerical solution of an infinite system of linear equations which can be truncated by using the distribution of the maximum queue length during a busy period.

For the *M/HEr/1* queue, the busy period is an "infinite" phase-type distribution with representation $(\boldsymbol{\alpha}_B, T_B)$ of infinite order given by

$$\boldsymbol{\alpha}_B = (\boldsymbol{\alpha}_{HEr}, \mathbf{0}, \mathbf{0}, ...), \qquad T_B = \lambda \begin{bmatrix} T_{HEr} - I & I & & \\ \mathbf{T}^0_{HEr}\boldsymbol{\alpha}_{HEr} & T_{HEr} - I & I & \\ & \mathbf{T}^0_{HEr}\boldsymbol{\alpha}_{HEr} & T_{HEr} - I & I \\ & & & \ddots \end{bmatrix},$$

where $\mathbf{0}$ represents a $1 \times \sum_{i=1}^{k} \nu_i$ zero vector and I is the $\sum_{i=1}^{k} \nu_i \times \sum_{i=1}^{k} \nu_i$ identity matrix. The infinite state space is $S = \left\{ (i,j)\,,\ i = 1, 2, ...,\ 1 \leq j \leq \sum_{i=1}^{k} \nu_i \right\}$.

We can truncate the system by taking a phase-type distribution of order $n_{\max} \times \sum_{i=1}^{k} \nu_i$ where $n_{\max}$ is chosen according to a criterion of Neuts (1977) by replacing the unbounded queue by a finite one with maximum queue length $n_{\max}$, in which all customers in excess of $n_{\max}$ are lost. Given this approxima-

tion, we must solve the following system of differential equations

$$\left(\boldsymbol{\chi}_1', \boldsymbol{\chi}_2', ..., \boldsymbol{\chi}_{n_{\max}}'\right) = \left(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, ..., \boldsymbol{\chi}_{n_{\max}}\right) \lambda \begin{bmatrix} T_{HEr} - I & I & & \\ \mathbf{T}_{HEr}^0 \boldsymbol{\alpha}_{HEr} & T_{HEr} - I & I & \\ & & \ddots & \\ & & \mathbf{T}_{HEr}^0 \boldsymbol{\alpha}_{HEr} & T_{HEr} - I \end{bmatrix}$$

with $\boldsymbol{\chi}(0) = (\boldsymbol{\alpha}_{HEr}, \mathbf{0}, \mathbf{0}, ...)$. Again, the solution to these equations can be calculated using Runge Kutta.

Given the solution, the distribution function of the length of a busy period can be computed as,

$$F_B(t) \approx 1 - \sum_{i=1}^{n_{\max}} \boldsymbol{\chi}_i(t)\, \mathbf{e}.$$

There are some practical problems in the computation of the busy period, and to a lesser extent, the waiting time distribution in cases where $\sum_{i=1}^{k} \nu_i$ is large. In such cases, the time consumed in computation can become prohibitive. In practice, we have found that truncating $\max\{\nu_i\} \leq 50$ or 100 is sufficient to give good approximations in most situations.

3.5 Estimation of predictive stationary densities from the MCMC output

Given a sample realization from the posterior distribution of $(\lambda, \boldsymbol{\theta})$ we can estimate the predictive probabilities of the number of customers in the system

using Rao-Blackwellization,

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

where $R = \#\left\{\rho^{(n)} < 1\right\}$.

Similarly, we can estimate the predictive distribution of the waiting time in the queue by

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

and the predictive distribution of the busy period can be approximated analogously.

The estimation of the moments of the queue size, waiting time or busy period distributions is impossible given our prior distribution structure, these do not exist. See for example, Wiper (1998).

4 Examples

4.1 Simulated examples

In this section, we illustrate the performance of the metodology with examples of several *M/G/1* queues. For simplicity, we assume that the interarrival rate, λ, is known and equal to 1. We consider samples of 100 service data from various service time distributions with a common mean 0.6, as follows,

1. 100 data simulated from a single exponential distribution.

2. 100 service times simulated from a mixture of Erlang distributions with, $\mathbf{w} = (1/3, 1/3, 1/3)$, $\boldsymbol{\mu} = (0.2, 0.5, 1.1)$ and $\boldsymbol{\nu} = (10, 20, 30)$.

3. 100 data equal to 0.6 from a degenerate distribution.

4. 100 data from a lognormal distribution $LN\left(\eta, \sigma^2\right)$ with parameters $\eta = -1.01$ and $\sigma^2 = 1$.

In each case, we suppose a discrete uniform prior distribution for k, with $k_{\max} = 10$. Note that the service distributions in cases 1 and 2 are Erlang mixtures, whereas the degenerate distribution in case 3 can be thought of as the limit of an Erlang distribution $Er(\nu, 1/0.6)$ where $\nu \to \infty$. The lognormal service distribution in case 4 is not a phase-type distribution.

The MCMC algorithm introduced in Section 2.1 was run for each data set with 100000 burn-in iterations and 100000 iterations "in equilibrium". Figure 1 illustrates the predictive service time densities (dotted line) for all four data sets. Also shown are the true densities in the non degenerate cases (solid line). We can see that in the non degenerate cases, the density estimates and the true density functions are very similar. In the degenerate case, as we might expect, the estimated density is concentrated around the point $s = 0.6$.

Table 1 gives the estimated posterior probabilities for various mixture sizes of k for the three data sets. Note that $P\left(k = 1 \mid \mathbf{t}\right) \simeq 0.98$ for the first data set. Also, given $k = 1$, the posterior probability that the service distribution is exponential is $P\left(\nu = 1 \mid k = 1, \mathbf{t}\right) = 0.9999$. Thus, it is clear that the correct *M/M/1* model has been well predicted in this case. In the Erlang mixture case, the correct mixture size has also been identified although with some uncertainty. In case 3, the density estimate for the degenerate service distribution has one

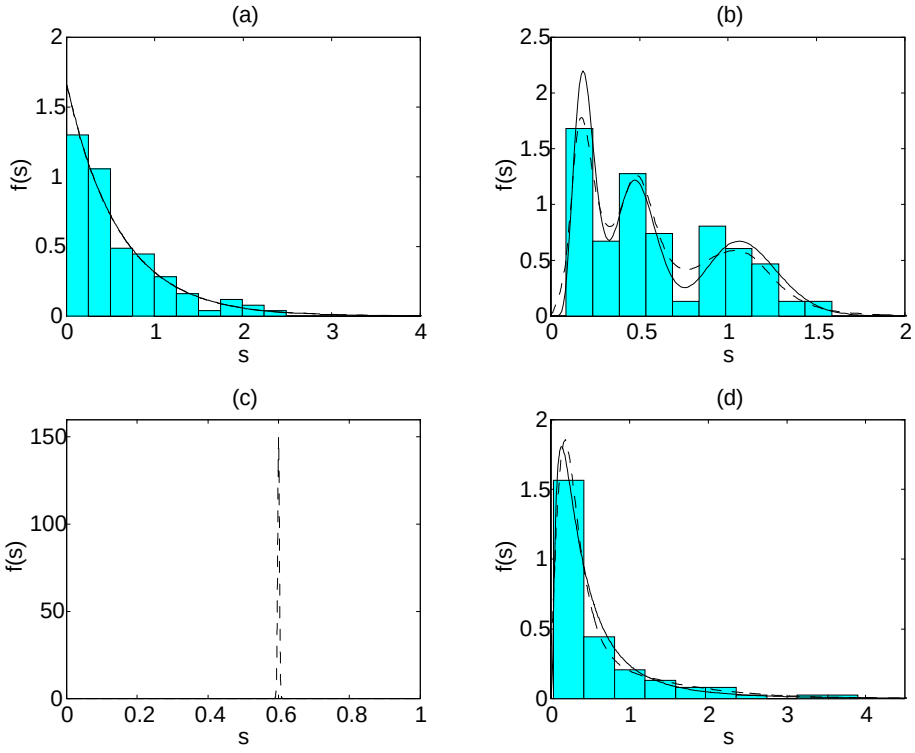

Figure 1: Predictive service time densities (dotted line) and the true densities (solid line) for (a) the exponential, (b) the hypererlang, (c) the degenerate and (d) the lognormal data sets.

$P(k \mid \mathrm{s})$	$M/M/1$	$M/HEr/1$	$M/D/1$	$M/LN/1$
$k=1$	0.979	0.149	1.000	0.097
$k=2$	0.012	0.263	0.000	0.843
$k=3$	0.005	0.507	0.000	0.033
$k=4$	0.001	0.058	0.000	0.015
$k=5$	0.000	0.015	0.000	0.007

Table 1: Posterior probabilities for different mixture sizes.

component with very small variance. The posterior mean service time is estimated to be 0.600 (s. d. .0003) and the posterior mean estimate for ν is around 50000 with large variance. This seems natural given our earlier comment that the degenerate distribution is a limiting case of an Erlang distribution.

	$M/M/1$	$M/HEr/1$	$M/D/1$	$M/LN/1$
$P(\rho < 1 \mid \mathrm{data})$	0.9998	0.9996	1.000	0.9993
$E[\rho \mid \rho < 1, \mathrm{data}]$	0.6039	0.5956	0.600	0.6104

Table 2: Posterior probability that the system is stable and the posterior mean values for the traffic intensity.

For all 4 data sets, the estimated posterior probability that the system is stable is extremely high (> 0.95) and the posterior mean values for ρ are, in all cases close to the true value of 0.6, as shown in Table 2. Thus, it seems reasonable to estimate the equilibrium distributions. In Table 3, the equilibrium probabilities of queue sizes between 0 and 4 are given for each of the 4 cases.

These are compared with the theoretical queue size probabilities.

j	$M/M/1$	$M/HEr/1$	$M/D/1$	M/LN
0	.396	.404	.400	.390
	.400	.400	.400	.400
1	.235	.277	.329	.221
	.240	.278	.329	.226
2	.141	.155	.162	.131
	.144	.158	.162	.128
3	.085	.085	.067	.084
	.086	.083	.067	.078
4	.052	.041	.026	.055
	.051	.042	.026	.051

Table 3: Estimated posterior probabilities $\pi(j|data)$ (upper) and true probabilities (lower) of the number of customers j in the system.

The estimated posterior queue size probabilities compare well with the theoretical probabilities for all cases where these can be evaluated.

Figure 2 shows the estimated predictive waiting time distribution functions (dotted line) for the first three systems. The true waiting time distribution functions are also illustrated (solid line). In the case of the lognormal service distribution this has not been done, as the theoretical waiting time distribution is unknown. The queueing time distribution function is not differentiable for the

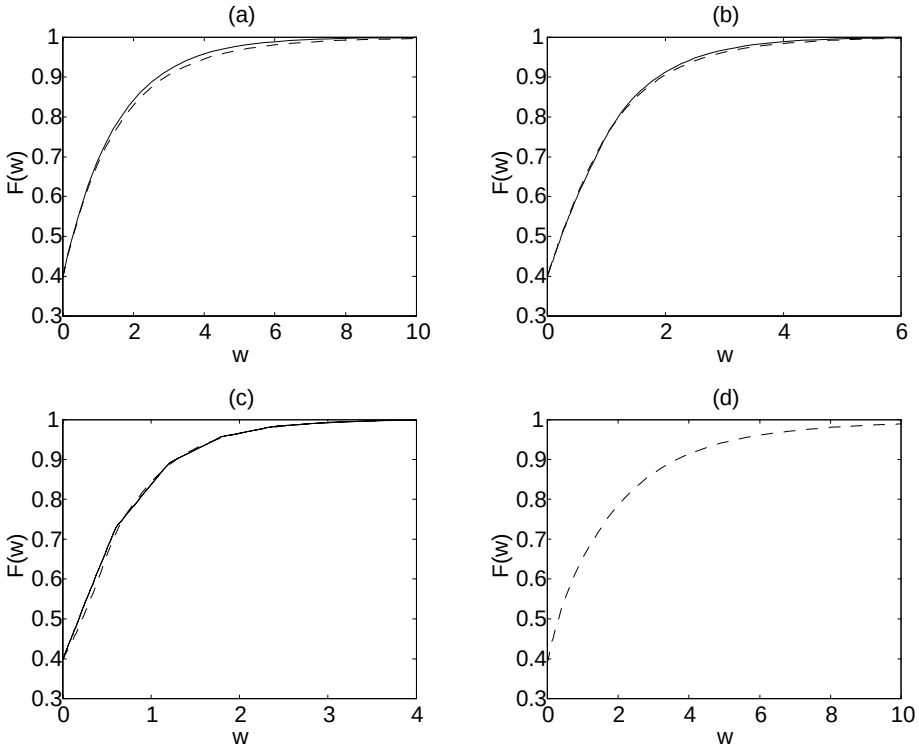

Figure 2: Predictive waiting time distribution functions (dotted line) and true distribution functions (solid line) for (a) the exponential, (b) the hypererlang, (c) the degenerate and (d) the lognormal data sets.

M/D/1 system, but it is fairly well approximated by the predictive distribution. In this case, we have used a smaller MCMC sample (of 100) to approximate the waiting time density truncating in $\nu = 50$ as commented earlier.

Figure 3 illustrates some single distribution functions of the length of the busy period selected at random from the MCMC sample for some of the mixture sizes (dotted line). Also shown is the true distribution (solid line). Numerical computations are more difficult at this point because we have to compute the maximum queue length distribution during a busy period for every MCMC sample in order to truncate the system of linear differential equations. Moreover, the order of the system can become very high. The busy period distribution function is not continuous for the degenerate case, but we can observe that, even in this case, the estimation is very similar. In this case, we compare the true distribution with some samples truncating in $\nu_{\max} = 1000$.

4.2 Real data example

We consider data collected at a cash register supplied by Professor Todd Schultz, Augusta State University's College of Business Administration, in his personal web page,

{http://www.aug.edu/~sbatas/resources/technicalnotes/JoeJava/mon6am/qdatidx.htm}

The number of customers arriving in 15-minute increments are recorded during 9 hours. The total number of arrivals during this time is 204. Using a non-informative prior distribution for the interarrival rate (per minute), λ, the posterior distribution is a gamma, $G(204, 540)$.

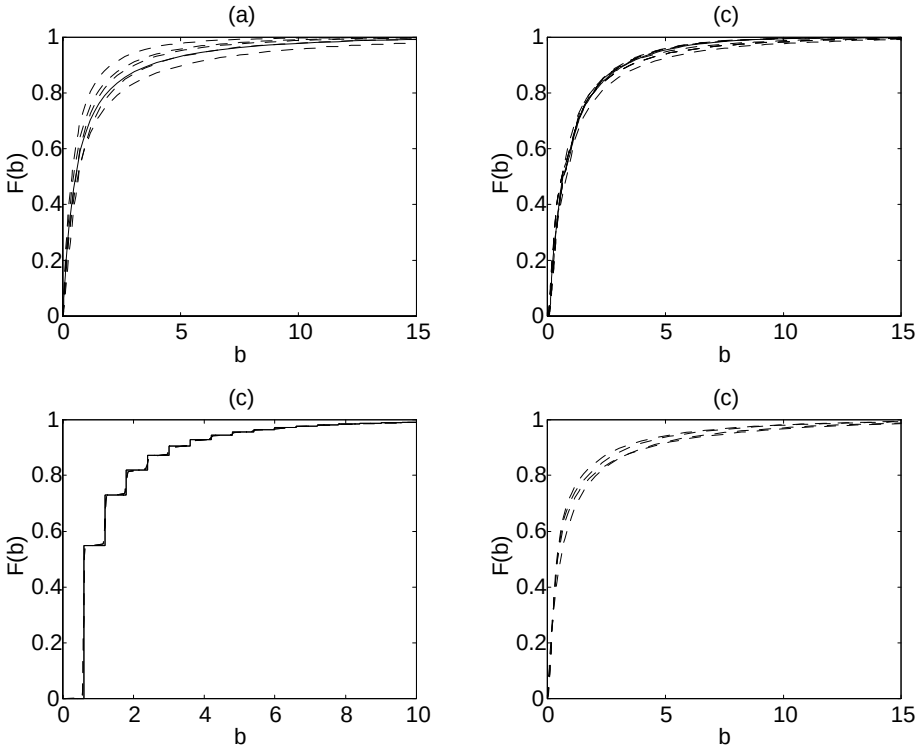

Figure 3: Single distribution functions of the length of the busy period (dotted line) and true distribution (solid line) for (a) the exponential, (b) the hyper-erlang, (c) the degenerate and (d) the lognormal data sets.

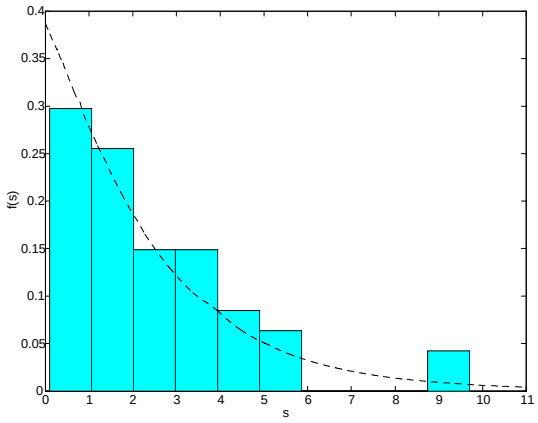

Figure 4: Histogram of service time data and estimated predictive service time density.

50 service times are also available. The majority of services take under 3 minutes and the largest service time is just under 10 minutes. Fitting this data using a mixture of Erlang distributions with the MCMC algorithm, the posterior probability of having a single Erlang distribution for the service time distribution is very high, $P(k = 1 \mid \mathbf{s}) \approx 0.926$. Conditioning on $k = 1$, there is some posterior uncertainty about the value of ν, being 1 or 2 with probabilities approximately equal to 0.918 and 0.082 respectively. Figure 4 shows the histogram of service time data and the estimated density. The Erlang mixture model seems to fit reasonably well. Also, the simple Markovian model appears reasonable.

Given the arrival data, the posterior probability of having a stable queue is estimated to be 0.777. Assuming equilibrium, the estimated expected value of

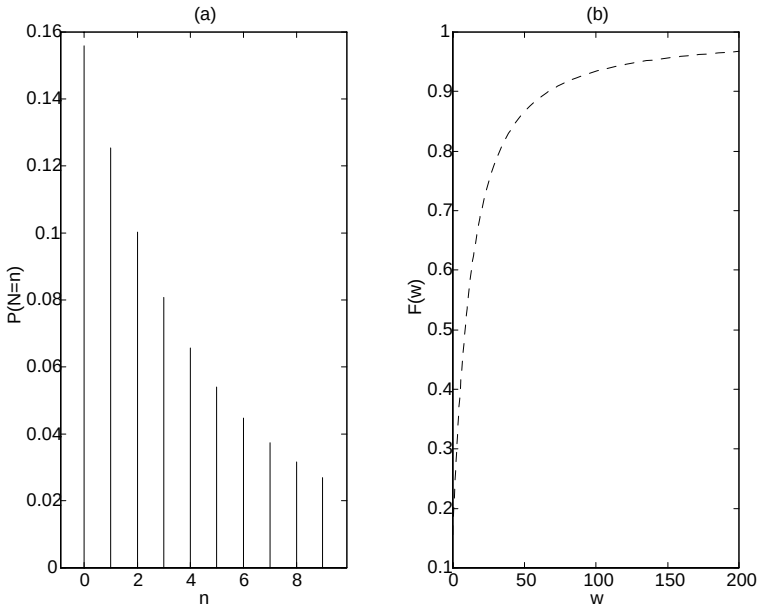

Figure 5: Estimated queue size distribution (a) and estimated waiting time distribution in equilibrium (b).

the traffic intensity is

$$E\left[\rho \mid \rho < 1, \mathbf{t}, \mathbf{s}\right] \approx 0.844.$$

Then, the predictive probability that the queue is empty is estimated to be 0.156. Figure 5 shows the estimated probabilities of the queue size in equilibrium and the estimated waiting time distribution function given the data. Analogously, the predictive density of the length of a busy period can be estimated.

5 Conclusions

We have developed a Bayesian analysis of *M/G/1* systems by modelling the general service time distribution using a mixture of Erlang distributions. We have constructed an MCMC algorithm making use of the "reversible jump" methodology and have combined this with matrix-analytic methods which has allowed us to make inference and predictions of various system quantities. We have illustrated our procedure with both simulated and real data.

The reversible jump algorithm used here to sample the posterior distribution is similar to the algorithm of Richardson and Green (1997) for normal mixtures. One important point here is that in this application, without the use of a birth death move for empty components, we have observed convergence problems due to continued splitting of empty components. This phenomenon was not observed by Wiper et al. (2001) who used a somewhat different algorithm to model a mixture of gamma distributions.

We have found some particular problems due to the discrete support of ν. Preservation of the second moments of the mixture in the reversible jump scheme is not possible. This often produces low acceptance rates (1 or 2%) for proposed changes in the mixture size, especially when ν takes values near 1. One possibility is to consider an alternative to the reversible jump algorithm. A method based on births and deaths of mixture components has been proposed by Stephens (2000). We are currently working on implementing this procedure for the Erlang mixture model.

Although, in this article, we have used a mixture of Erlang distributions to

model services times, we could also consider other classes of phase-type distributions, such as acyclic phase-type distributions or the class of Coxian distributions. In some cases, we can only obtain satisfactory approximations of the service distribution using Erlang mixtures of very high order. This complicates the numerical computations of predictive distributions. The class of acyclic phase-type distributions may lead to good estimations of moderate orders. This occurs in maximum-likelihood based parameter estimation as presented by Bobbio and Telek (1994).

We might compare our results with other maximum-likelihood based methods used to fit phase-type distributions such as that implemented by Asmussen et al. (1996) via EM algorithm. However, as far as we know, nothing seems to have been done to obtain confidence intervals. One possibility is the use of bootstrap techniques, but this is computationally inefficient because it requires many ML estimations.

Our approach can also be extended to other queueing systems. A very similar method can be used for $G/M/c$ systems, where, for example, we could model the general interarrival distribution using the Erlang mixture distribution. In many ways, analysis of this system is much simpler than the $M/G/1$ queue, as the equilibrium distributions of the queue size and waiting time in the $G/M/c$ systems have a relatively simple form. A paper on this problem is currently under preparation. Finally, it is possible to model $G/G/1$ systems using a phase-type approximation for both the service and interarrival time distributions. In this case, the calculation of the queue size and waiting time distributions is

much more complicated. This problem is currently being studied.

References

[1] Allen, A., 1990. Probability, Statistics and Queueing Theory with Computer Science Applications. Acad. Press.

[2] Abramowitz, M., Stegun, I.A. 1964. Handbook of Mathematical Functions. New York: Dover.

[3] Armero, C., 1985. Bayesian analysis of $M/M/1/\infty$/FIFO queues. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 2., pp. 613-618. North Holland, Amsterdam.

[4] Armero, C., Bayarri, M.J., 1994. Bayesian prediction in M/M/1 queues. Queueing Syst., 15, 401-417.

[5] Armero, C., Bayarri, M.J., 1995. Bayesian questions and Bayesian answers in queues. In: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds., Bayesian Statistics 5, 3-23. Oxford: University Press.

[6] Armero, C., Bayarri, M.J., 1997. A Bayesian analysis of a queueing system with unlimited service. J. Statist. Planning and Inference, 58, 241-261.

[7] Armero, C., Conesa, D., 1998. Inference and prediction in bulk arrival queues and queues with service in stages. Appl. Stochast. Models and Data Anal., 14, 35-46.

[8] Asmussen, S., 1987. Applied probability and queues. New York: Wiley.

[9] Asmussen, S., Nerman, O., Olsson, M., 1996. Fitting phase-type distributions via the EM algorithm. Scand. J. Statist., 23, 419-441.

[10] Bagchi, T.P., Cunningham, A.A., 1972. Bayesian approach to the design of queueing systems. Inform., 10, 36-46.

[11] Bobbio, A., Telek., M., 1994. A benchmark for PH fitting algorithms: result for acyclic PH. Comm. Statis. Stochastic Models, 10, 661-667.

[12] Diebolt, J., Robert, C.P., 1994. Estimation of finite mixture distributions. J. Royal Statist. Soc., B, 56, 363-375.

[13] Green, P., 1995. Reversible jump MCMC computation and Bayesian model determination. Biometrika, 82, 711-732.

[14] Gruet, M.A., Philippe, A., Robert, C.P., 1998. MCMC Control Spreadsheets for Exponential mixture Estimation. Tech. Rep., CREST, Paris: INSEE.

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

[16] McGrath, M.F., Gross, D., Singpurwalla, N.D., 1987. A subjective Bayesian approach to the theory of queues I - Modeling. Queueing Syst., 1, 317-333.

[17] Muddapur, M.V., 1972. Bayesian estimates of parameters in some queueing models. Ann. Inst. Math., 24, 327-331.

[18] Neuts, M.F., 1981. Matrix Geometric Solutions in Stochastic Models. Baltimore: John Hopkins University Press.

[19] Neuts, M.F., 1977. Algorithms for the waiting time distributions under various queue disciplines in the M/G/1 queue with service time distributions of phase type. TIMS Studies in the Management Sciences, 7, 177-197

[20] Nelson, R., 1995. Probability, Stochastic Processes and Queueing Theory. New York: Springer-Verlag.

[21] Reynolds, J.F., 1973. On estimating the parameters of a birth-death process. Austral. J. Statist., 15, 35-43.

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

[23] Rios, D., Wiper, M.P., Ruggeri, F., 1998. Bayesian analysis of M/Er/1 and $M/H_k/1$ queues. Queueing Syst., 30, 289-308.

[24] Stephens, M., 2000. Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. Ann. Statist., 28, 40-74.

[25] Thiruvaiyaru, D., Basawa, I.V., 1992. Empirical Bayes estimation for queueing systems and networks, Queueing Syst., 11, 179-202.

[26] Tierney, L., 1996. Introduction to general state-space Markov chain Theory. In Markov Chain Monte Carlo in Practice. In: W.R. Gilks, S.Richardson and D. J. Spiegelhalter, eds., 59-74. London: Chapman & Hall.

[27] Wiper, M.P., Rios Insua, D., Ruggeri, F., 2001. Mixtures of gamma distributions with applications, Journal of Computational and Graphical Statistics,

[28] Wiper, M.P., 1998. Bayesian analysis of Er/M/1 and Er/M/c queues, J. Statist. Planning and Inference, 69, 65-79.