UNIVERSIDAD CARLOS III DE MADRID

working papers

Working Paper 04-69
Statistics and Econometrics Series 17
December 2004

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

BAYESIAN CONTROL OF THE NUMBER OF SERVERS IN A GI/M/C QUEUING SYSTEM

María Concepción Ausín Olivera; Rosa Elvira Lillo; Michael Peter Wiper*

Abstract

In this paper we consider the problem of designing a *GI/M/c* queueing system. Given arrival and service data, our objective is to choose the optimal number of servers so as to minimize an expected cost function which depends on quantities, such as the number of customers in the queue. A semiparametric approach based on Erlang mixture distributions is used to model the general interarrival time distribution. Given the sample data, Bayesian Markov chain Monte Carlo methods are used to estimate the system parameters and the predictive distributions of the usual performance measures. We can then use these estimates to minimize the steady-state expected total cost rate as a function of the control parameter *c*. We provide a numerical example based on real data obtained from a bank in Madrid.

Keywords: Queueing systems, Bayesian design, birth-and-death MCMC, optimal service channels.

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

Bayesian control of the number of servers in a $GI/M/c$ queueing system.

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

Departamento de Estadística.

Universidad Carlos III de Madrid

Madrid, 126, 28903 Getafe, Madrid, Spain

May 21, 2004

Abstract

In this paper we consider the problem of designing a $GI/M/c$ queueing system. Given arrival and service data, our objective is to choose the optimal number of servers so as to minimize an expected cost function which depends on quantities, such as the number of customers in the queue. A semiparametric approach based on Erlang mixture distributions is used to model the general interarrival time distribution. Given the sample data, Bayesian Markov chain Monte Carlo methods are used to estimate the system parameters and the predictive distributions of the usual performance measures. We can then use these estimates to minimize the steady-state expected total cost rate as a function of the control parameter c. We provide a numerical example based on real data obtained from Madrid bank.

1 Introduction

Optimal design and control of queues have been extensively investigated from an operational research point of view, see e.g. Kitaev and Rykov (1995). However, in this framework, the system parameters are typically assumed known. In practice, the system manager is faced with the problem of estimating the system parameters before solving the optimization problem. Furthermore, a common approach consist in selecting a queueing model and estimate the parameters without considering the uncertainty induced from this estimation in the system design. The Bayesian methodology offers a natural way to introduce the uncertainty resulting from the parameter estimation and model selection into a cost function depending on estimated performance measures.

Bayesian analysis of queueing systems is a fairly recent research area. Some recent references are Armero and Conesa (2000), Ausín et al. (2003, 2004). In these works, Bayesian inference and prediction is undertaken for different queueing models ranging from the $M/M/c$ system to more general queues. However, although most Bayesian analyses have considered the estimation of quantities of interest such as queue size, few studies have been devoted to the design and control problem. In one of the first works in Bayesian estimation for queues, Bagchi and Cunningham (1972) develop an optimal design procedure to find the optimum service rate and system capacity in a single server, Markovian queue. Also, Armero and Bayarri (1996) discusses some criteria for deciding the number of servers in a $M/M/c$ queue and Wiper (1998) also for the $Er/M/1$ model, but no systematic procedure for decision making is proposed. These works motivates the formulation of a closed expression based on a cost structure to address the decision problem on the number of servers.

On the other hand, most Bayesian 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 $Er/M/c$ model is considered. However, although the Erlang distribution may be used to fit interarrival (or service) time data with coefficient of variation less than one, it is inappropriate if the data have large coefficient of variation or are multimodal. Our objective in this paper is thus to consider Bayesian control for the general, $GI/M/c$ queueing system.

In $\oint$ 2, we describe the $GI/M/c$ queueing model where 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.

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 Ríos 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 $\oint$ 3, we make use of the latter methodology to make inference for the system parameters. We define prior distributions and propose a birth-and death MCMC algorithm to obtain a sample from the joint posterior distribution of the system parameters and the predictive interarrival time distribution.

In $\oint$ 4 and $\oint$ 5, we describe describe the estimation of various quantities of interest in the system and address the problem of optimizing the number of servers. Firstly, we estimate the traffic intensity and the probability that the equilibrium condition holds. Then, assuming a stable system, we estimate the predictive distributions of the system size and the waiting time in the queue, among other characteristics. Finally, we propose a steady state, average cost function which depends on the number of servers and some performance measures. The predictive cost and the performance measures are all estimated using the data generated from the MCMC algorithm.

In $\oint$ 6, we illustrate the methodology with real data obtained from Madrid bank. Conclusions and a discussion of possible extensions are included in $\oint 7$.

2 Queueing model

Throughout, we will consider a multichannel queueing system with c servers FIFO discipline and independence between interarrival and service times. Furthermore, service times are independent and exponentially distributed with unknown mean $1/\mu$. In order to model the general interarrival time distribution, we use a semiparametric model based on a mixture of Erlang distributions. Thus, customers are assumed to arrive individually with independent interarrival times distributed as a mixture of Erlang distributions. If T is a typical interarrival time, we have,

$$f(t \mid k, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \sum_{r=1}^{k} w_r Er(t \mid \nu_r, \lambda_r), \qquad 0 < t < \infty, \tag{1}$$

where k is the number of mixture components, $\mathbf{w} = (w_1, ..., w_k)$, are weights and $Er(t \mid \nu_r, \lambda_r)$ represents the Erlang density function, which has been parameterized to have mean λ_r, for $r = 1, \ldots, k$, that is,

$$Er(t \mid \nu_r, \lambda_r) = \frac{(\nu_r/\lambda_r)^{\nu_r}}{\Gamma(\nu_r)} t^{\nu_r - 1} \exp(-\frac{\nu_r}{\lambda_r} t). \tag{2}$$

For fixed k, this model includes the usual Erlang, hyperexponential and exponential distributions as special cases and letting $k \to \infty$, essentially any distribution on the positive real line can be modeled as a mixture of Erlang distributions.

We wish to estimate the performance measures and a cost function for the system in equilibrium. The equilibrium condition for a $GI/G/c$ queue is that the traffic intensity, ρ, is less than the number of servers, c, see, for example, Gross and Harris (1985). In the $GI/M/c$ model as outlined above, the traffic intensity is given by,

$$\rho = \left(\mu \sum_{r=1}^{k} w_r \lambda_r \right)^{-1}. \tag{3}$$

3 Bayesian inference.

In this section, we develop Bayesian inference techniques for the unknown arrival parameters, $k, \mathbf{w}, \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k), \boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)$ and for the service parameter μ.

We consider throughout the simple experiment of observing n_s service times, $\mathbf{s} = \{s_1, ..., s_{n_s}\}$, and n_a interarrival times, $\mathbf{t} = \{t_1, ..., t_{n_a}\}$, which has been considered in a number of earlier articles; see e.g. Armero and Bayarri (1996). Given this experiment, the likelihood function separates into two parts, one concerning the arrival parameters, $(k, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ 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.

3.1 Prior specification and updating

Here, we assign prior distributions for the system parameters. For the service rate, μ, we can assume a gamma prior distribution, $\mu \sim G(a, b)$, that is

$$f(\mu \mid a, b) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu} \quad \text{for } \mu > 0.$$

It is straightforward to show that, conditional on the service data, the posterior distribution is also gamma so that,

$$\mu \mid \mathbf{s} \sim G\left(a + n_s, b + \sum_{i=1}^{n_s} s_i\right). \tag{4}$$

In order to make inference for the interarrival distribution parameters, following Diebolt and Robert (1994), it is convenient to introduce a missing data formulation in which we define a set of independent and identically distributed (i.i.d.) latent variables, $Z_1, ..., Z_{n_a}$, associated with the interarrival time variables, $T_1, ..., T_{n_a}$, so that,

$$T_i \mid Z_i = r \sim Er(\nu_r, \mu_r), \qquad P(Z_i = r \mid k, \mathbf{w}) = w_r,$$

for $r = 1, ..., k$. With this approach, every interarrival data set, $\mathbf{t} = \{t_1, ..., t_{n_a}\}$, is associated to a missing data set, $\mathbf{z} = \{z_1, ..., z_{n_a}\}$, indicating the specific components of the mixture from which the observed interarrival times are assumed to arise.

Now, we can define a joint prior distribution on the mixture parameters, $(k, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\nu})$. Firstly, we assume a truncated Poisson prior distribution for the mixture size, k, taking values from 1 to $k_{\max}$,

$$P(k) \propto \frac{\gamma^k}{k!}, \tag{5}$$

In practice we take $\gamma = 2$ and $k_{\max} = 10$ in order to penalize against overfitting the data with mixtures with a large number of components. We also define prior distributions for the remaining parameters conditional on k,

$$\mathbf{w} \mid k \sim D(\phi, ..., \phi), \qquad \nu_r \mid k \sim GE(\vartheta), \quad \lambda_r^{-1} \mid k \sim G(\alpha, \beta),$$

for $r = 1, ..., k$, where $D(\phi, ..., \phi)$ denotes a symmetric Dirichlet distribution,

$$f(\mathbf{w} \mid k) = \frac{\Gamma(k\phi)}{\Gamma(\phi)^k} \prod_{i=1}^{k} w_i^{\phi-1}$$

$GE(\vartheta)$ is a geometric distribution with mean $1/\vartheta$, i.e.

$$P(\nu_i) = (1 - \vartheta)^{\nu_i - 1}\vartheta \quad \text{for } \nu_i = 1, 2, \ldots$$

and $G(\alpha, \beta)$ denotes a gamma distribution. Typically, in practice we set, for all $r = 1, ..., k$; $\phi_r = 1$, which implies a uniform prior for $\mathbf{w}$ and $\alpha = 1.1$, $\beta = 1$ and $\vartheta = 0.01$ giving fairly diffuse priors for λ_r and ν_r with finite means.

Conditional on k, and given the interarrival time data, the required posterior conditional distributions for the MCMC algorithm can be shown to be,

$$P(Z_i = r \mid \mathbf{t}, k, \mathbf{w}, \lambda, \nu) \propto w_r \frac{(\nu_r/\lambda_r)^{\nu_r}}{\Gamma(\nu_r)} t_i^{\nu_r - 1} \exp(-\frac{\nu_r}{\lambda_r} t_i), \qquad \text{for } r = 1, ..., k,$$

$$\mathbf{w} \mid \mathbf{t}, \mathbf{z}, k \sim D(\phi_1 + n_1, ..., \phi_k + n_k),$$

$$\lambda_r \mid \mathbf{t}, \mathbf{z}, k \sim IG(\alpha + n_r\nu_r, \beta + T_r\nu_r),$$

and,

$$f(\nu_r \mid \mathbf{t},\mathbf{z},k,\mathbf{w},\lambda) \propto \frac{\nu_r^{n_r\nu_r}}{\Gamma(\nu_r)^{n_r}} \exp\left\{-\nu_r\left(-\log(1-\vartheta)+\frac{S_r}{\lambda_r}+n_r\log\lambda_r-\log P_r\right)\right\}, \tag{6}$$

where $n_r = \#\{Z_i = r\}$, $S_r = \sum\limits_{i:Z_i=r} t_i$ and $P_r = \prod\limits_{i:Z_i=r} t_i$, for $r = 1,...,k$.

3.2 BDMCMC algorithm

In this subsection, we propose a birth-death MCMC (BDMCMC) algorithm to obtain a sample from the joint posterior distribution of the interarrival parameters, $k, \mathbf{w}, \boldsymbol{\lambda}$ and $\boldsymbol{\nu}$. The BDMCMC approach was introduced by Stephens(2000) for normal mixtures and is based on a birth-death process (BD) where the mixture size, k, changes so that births and deaths of the mixture components occur in continuous time. The stationary distribution of the BD process is the joint posterior of the mixture parameters. In order to improve mixing, the BD process can be combined with a standard MCMC method where k is kept fixed, as will be shown further on.

In the BD process, births of the mixture components occur at a constant rate which we might set equal to the parameter, γ, from the prior distribution of k in (5). A birth increases the number of components by one. The weight of the new component are generated from a beta distribution with parameters $(1, k)$ and the remaining parameters are sampled from the prior distribution. The death rate of every mixture component is a likelihood ratio of the model with and without this component, given by,

$$\delta_{r_0} = \prod_{i=1}^{n_a}\left(\frac{\sum_{\substack{r=1\\ r\neq r_0}}^{k} \frac{w_r}{1-w_{r_0}} Er(t_i \mid \nu_r,\lambda_r)}{\sum_{r=1}^{k} w_r Er(t_i \mid \nu_r,\lambda_r)}\right), \qquad \text{for } r_0 = 1,...,k.$$

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, δ, of the process at any time is the sum of the individual death rates. A death decreases the number of mixture components by one. The birth and death processes are independent Poisson processes, thus, the time to next birth/death event is exponentially distributed with mean $1/\left(\delta+\gamma\right)$ and a birth or death occur with probabilities proportional to γ and δ, respectively.

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 $k^{(j)}, \mathbf{w}^{(j)}, \lambda^{(j)}, \nu^{(j)}$.
    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 conditional on $k$.

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

Step 2 of the algorithm is the BD process described above. The BD process is run for a fixed time, t_0, in each iteration of the algorithm. Following Stephens (2000), we have fixed in our examples $t_0 = 1$ because doubling t_0 is equivalent to doubling γ. As should be expected, we have found in practice that larger values of the birth rate, γ, 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 (6) has a similar form to a negative binomial distribution. This part of the algorithm where the mixture size, k, is kept fixed is very similar to that used in Ausín et al. (2004).

This algorithm can be shown to produce a sample from the joint posterior parameter distribution; see e.g. Stephens (2000). Thus, given the MCMC output of size J, we can estimate the predictive density of the interarrival time distribution using,

$$f(t \mid \mathbf{s}, \mathbf{t}) = \frac{1}{J} \sum_{j=1}^{J} \sum_{r=1}^{k^{(r)}} w_r^{(j)} Er(t \mid \nu_r^{(j)}, \lambda_r^{(j)}). \tag{7}$$

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. (2003).

4 Estimation of performance measures in the system

Suppose now that we have obtained a Monte Carlo sample of size J from the posterior distribution of the arrival parameters, via the BDMCMC algorithm, and the service parameter μ via direct sampling of the gamma density $f(\mu \mid \mathbf{s})$ as in (4). Then we can estimate the probability of having a stationary distribution with,

$$P\left(\rho < c \mid \mathbf{s}, \mathbf{t}\right) \approx \frac{1}{J} \# \left\{ \rho^{(j)} < c \right\}, \tag{8}$$

where,

$$\rho^{(j)} = \left(\mu^{(j)} \sum_{r=1}^{k^{(j)}} w_r^{(j)} \lambda_r^{(j)} \right)^{-1}, \tag{9}$$

and $\{(k^{(1)}, \mathbf{w}^{(1)}, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\nu}^{(1)}), ..., (k^{(J)}, \mathbf{w}^{(J)}, \boldsymbol{\lambda}^{(J)}, \boldsymbol{\nu}^{(J)})\}$ is the sample obtained from the BDMCMC algorithm and $\{\mu^{(1)}, ..., \mu^{(J)}\}$ is the sample generated from the posterior distribution of μ given by (4). If this probability is large, it may be reasonable to assume that the system is stable. Assuming equilibrium, we can estimate the traffic intensity, given in (3), as follows,

$$E\left[\rho \mid \mathbf{t}, \mathbf{s}, \rho < c\right] \approx \frac{1}{J_1} \sum_{j:\rho^{(j)}<c} \rho^{(j)}, \tag{10}$$

where $\rho^{(j)}$ is given in (9) and,

$$J_1 = \#\{\rho^{(j)} < c\}, \tag{11}$$

is the size of the MCMC subsample where the equilibrium condition holds.

It is well known, see e.g. Gross and Harris (1985), that in queuing systems with non-Markovian interarrival process, the stationary distribution of the number of customers, N^*, found in the system by an arriving customer differs from the stationary distribution of the number of customers, N, found in the system at an arbitrary time instant. For our $GI/M/c$ model, given the system parameters, $\boldsymbol{\theta} = \{k, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \mu\}$, we have that (see e.g. Allen, 1990),

$$P\left(N^* = n \mid \boldsymbol{\theta}\right) = \begin{cases} \sum\limits_{m=n}^{c-1} (-1)^{m-n} \binom{m}{n} U_m & \text{for } n = 0, 1, ..., c-2, \\ D\sigma^{n-c} & n \geq c-1, \end{cases} \tag{12}$$

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

$$\sigma = f_A^*\left(c\mu\left(1-\sigma\right)\right), \tag{13}$$

and f_A^* is the Laplace transform of the interarrival time distribution,

$$f_A^*\left(s\right) = \sum_{r=1}^{k} w_r \left(\frac{\nu_r/\lambda_r}{s + \nu_r/\lambda_r}\right)^{\nu_r},$$

and,

$$\begin{aligned} g_p &= f_A^*\left(p\mu\right), \qquad \text{for } p = 1, ..., c, \\ C_p &= \begin{cases} 1 & \text{if } p = 0, \\ \prod\limits_{m=1}^{p} \left(\frac{g_m}{1-g_m}\right) & \text{if } p = 1, 2, ..., c, \end{cases} \\ D &= \left[\frac{1}{1-\sigma} + \sum_{p=1}^{c} \frac{\binom{c}{p}}{C_p(1-g_p)} \frac{c(1-g_p)-p}{c(1-\sigma)-p}\right]^{-1}, \\ U_n &= DC_n \sum_{p=n+1}^{c} \frac{\binom{c}{p}}{C_p(1-g_p)} \frac{c(1-g_p)-p}{c(1-\sigma)-p}, \qquad \text{for } n = 0, 1, ..., c-1. \end{aligned} \tag{14}$$

The distribution of N depends on the distribution of N^* and is given by,

$$P\left(N=n\mid\boldsymbol{\theta}\right)=\begin{cases}1-\frac{\rho}{c}-\rho\sum\limits_{p=1}^{c-1}P\left(N^*=p-1\mid\boldsymbol{\theta}\right)\left(\frac{1}{p}-\frac{1}{c}\right) & \text{for } n=0,\\ \frac{\rho}{n}P\left(N^*=n-1\mid\boldsymbol{\theta}\right) & \text{for } n=1,...,c-1,\\ \frac{\rho}{c}P\left(N^*=n-1\mid\boldsymbol{\theta}\right) & \text{for } n\geq c.\end{cases} \tag{15}$$

Assuming equilibrium, Monte Carlo approximations of the predictive stationary distributions of $N^*$and N, can be obtained. For example, we can approximate the predictive distribution of N by,

$$P\left(N=n\mid\mathbf{s},\mathbf{t},\rho<c\right)\approx\frac{1}{J_1}\sum_{j:\rho^{(j)}<c}P\left(N=n\mid\boldsymbol{\theta}^{(j)}\right) \tag{16}$$

where $\boldsymbol{\theta}^{(j)}=(k^{(j)},\mathbf{w}^{(j)},\boldsymbol{\lambda}^{(j)},\boldsymbol{\nu}^{(j)},\mu^{(j)})$ and J_1 is given in (11). Note that equation (13) has to be solved for every $\boldsymbol{\theta}^{(j)}$, but it is easy to approximate $\sigma^{(j)}$ by using the Newton-Raphson method or a similar procedure. Other quantities such as the stationary distribution of the number of busy servers can also be estimated although again, we must distinguish between the number of busy servers at arrival and arbitrary time instants, N_b^* and N_b. Observe that the number of busy servers is equal to the number of customers in the system if there are less customers than servers and equals c in the contrary case. Thus,

$$P\left(N_b=n\mid\mathbf{s},\mathbf{t},\rho<c\right)=\begin{cases}P\left(N=n\mid\mathbf{s},\mathbf{t},\rho<c\right) & \text{if } n<c,\\ P\left(N\geq c\mid\mathbf{s},\mathbf{t},\rho<c\right) & \text{if } n=c.\end{cases} \tag{17}$$

Other important quantities are the predictive distributions of the number of customers in the queue at arrival and arbitrary time instants, N_q^* and N_q. In the first case, we have,

$$P\left(N_q=n\mid\mathbf{s},\mathbf{t}\right)=\begin{cases}P\left(N\leq c\mid\mathbf{s},\mathbf{t}\right) & \text{if } n=0,\\ P\left(N=c+n\mid\mathbf{s},\mathbf{t}\right) & \text{if } n\geq 1.\end{cases} \tag{18}$$

Another measure which is of interest to arriving customers, is the waiting time in the queue, W. Given the system parameters, $\boldsymbol{\theta}$, this is exponentially distributed with a jump of height $P\left(W=0\right)$ at the origin. The distribution function is given by,

$$F_W\left(x\mid\boldsymbol{\theta}\right)=1-P\left(W>0\right)\exp\left\{-c\mu\left(1-\sigma\right)x\right\},\qquad x\geq 0, \tag{19}$$

see Allen (1990), where,

$$P\left(W>0 \mid \boldsymbol{\theta}\right)=\frac{D}{1-\sigma}, \tag{20}$$

and where σ and D are given in (13) and (14), respectively. As above, we can use the following Monte Carlo approximation,

$$F_W\left(x \mid \mathbf{s},\mathbf{t}\right)\approx\frac{1}{J_1}\sum_{j:\rho^{(j)}<1}F_W\left(x \mid \boldsymbol{\theta}^{(j)}\right) \tag{21}$$

Wiper (1998) shows that, for any given $GI/M/1$ system, where independent, continuous priors on the arrival and service rates with positive density in $\rho=1$ are considered, the moments of the predictive distributions of waiting time and queue size do not exist. It is straightforward to see that the moments for N^*, N, N_q, N_q^* and W do not either exist for the multiserver system, $GI/M/c$, with the same prior conditions. Thus, the distributions given in (16) and in (21) do not have finite moments. It is possible however to evaluate the expectations of these predictive distributions if we assume $\rho<c-\varepsilon$ instead of $\rho<c$, see Lehoczky (1990), but we have found in practical examples that this procedure is very sensible to the election of ε. Observe, on the other hand, that the predictive distribution of number of busy servers N_b given in (17) does have finite moments.

5 Cost functions and optimal control for the model.

In this section, we formulate cost functions in order to address the design problem for the $GI/M/c$ queueing model and determine the optimal number of servers in the system. We consider a classical, linear, cost structure evaluated in the stationary state. Each cost function will depend linearly on the expected values of the performance measures considered in the previous section, or equivalently, on their mean values per unit of time (u.t.). Thus, we are dealing with an infinite horizon problem where the objective function is the expected cost per u.t. evaluated in the stationary state.

Also, our aim is to construct cost functions which balance the designer's and the customers' interests. For that reason, we consider two different classes of costs in the queue: on the one hand, costs incurred from servers activities and, on the other hand, costs incurred from the wait of clients. The first group of costs includes the expenses coming from the number of busy and empty servers and the benefits obtained from the number of served clients which are all of them associated to the designers' interests. The second group of costs represents the customers' interests and are related with the number of clients and the period of time they spend waiting in the queue. We introduce the following notation to define the cost structure:

$r_b =$ cost per u.t. per busy server.

$r_e =$ cost per u.t. per empty server.

$r_s =$ cost for each customer that is served.

$r_q =$ cost per u.t. per customer waiting in the queue.

$r_W =$ cost per unit of waiting time in the queue.

These costs can take positive or negatives values on whether they correspond to profits or losses. As the problem of designing a queue is not generally a work for clients but for people supervising the system, we consider performance measures at arbitrary time instants and not at arrival time instants, both described in the previous section. Under this construction, the total cost per u.t. will be,

$$Cost = r_b N_b + r_e \left\{c - N_b\right\} + r_s N_s + r_q L\left(N_q\right) + r_W L\left(W\right), \tag{22}$$

where N_s is the number of customers served per u.t. and N_b, N_q and W are the number of busy servers, the number of customers and the waiting time in the queue, respectively, defined in the previous section. $L\left(N_q\right)$ represents the loss due to the number of people waiting for service and $L\left(W\right)$ is the loss due to the time they spend waiting in queue. For example, we can consider a loss formulation with the following structure,

$$L_1\left(N_q\right) = \begin{cases} 0 & \text{if } N_q \leq n_0, \\ 1 & \text{if } N_q > n_0, \end{cases} \tag{23}$$

where a cost, r_q, is incurred per u.t. if the queue length exceeds a previously specified threshold,

$n_0 > 0$. A more realistic alternative would be to consider a linear cost proportional to the number of waiting customers,

$$L_2\left(N_q\right) = \begin{cases} N_q & \text{if } N_q \leq n_0, \\ n_0 & \text{if } N_q > n_0, \end{cases} \tag{24}$$

where a cost, r_q, is incurred per u.t. per customer in the queue if the queue length does not exceed a threshold, $n_0 < \infty$. Similar loss functions, $L_1\left(W\right)$ and $L_2\left(W\right)$, can be formulated for the waiting time in the queue, for which a threshold, $w_0 < \infty$, have to be fixed. The values of n_0 and w_0 are finite by assumption because, as pointed out, the predictive distribution of W and N_q have no finite moments, and thus, an infinite value for a threshold will lead to an infinite value of the expected cost.

Given that the system parameters verify the equilibrium condition and considering L_1 loss functions, the expected cost per u.t. each c is given by,

$$g\left(c \mid \boldsymbol{\theta}\right) = E\left[Cost \mid \boldsymbol{\theta}\right] = r_e c + \left(r_b - r_e + r_s \mu\right)\rho + r_q P\left(N_q > n_0 \mid \boldsymbol{\theta}\right) + r_W P\left(W > w_0 \mid \boldsymbol{\theta}\right). \tag{25}$$

To understand this expression, note that on average, each busy server attends μ clients per u.t. so that the number of served clients per u.t. is,

$$E\left[N_s \mid \boldsymbol{\theta}\right] = \mu E\left[N_b \mid \boldsymbol{\theta}\right],$$

and also for any $GI/G/c$ system in equilibrium, the expected number of busy servers is,

$$E\left[N_b \mid \boldsymbol{\theta}\right] = \rho, \tag{26}$$

see e.g. Gross and Harris (1985), which in our queuing model is given by (3). Finally, the required probability for N_q can be obtained from (15) by using that,

$$P\left(N_q > n_0 \mid \boldsymbol{\theta}\right) = 1 - P\left(N \leq n_0 + c \mid \boldsymbol{\theta}\right),$$

and for W from (19). As an alternative to (25), expected costs can be derived with loss functions as given in (24). For these cases, the expected losses will be,

$$E\left[L_2\left(N_q\right) \mid \boldsymbol{\theta}\right] = \sum_{n=0}^{n_0} nP\left(N = n + c \mid \boldsymbol{\theta}\right) + n_0\left[1 - P\left(N \leq n_0 + c \mid \boldsymbol{\theta}\right)\right],$$

where the distribution of N is given in (15), and the expected loss for W can be shown to be,

$$E\left[L_2(W)\mid\boldsymbol{\theta}\right]=\frac{D\left[1-\exp\left(-c\mu\left(1-\sigma\right)w_0\right)\right]}{c\mu\left(1-\sigma\right)^2}.$$

For discrete functions, it is possible to find out how many minima there are considering a monotone optimal procedure, see Lillo and Martín (2000) . This consists in finding a point, c_0, of the objective function, $g(c)$, where $g(c_0+1)-g(c_0)>0$, and such that, $g(c+1)-g(c)>0$ for every $c>c_0$. It can be shown that the expected cost function (25) allows a monotone optimal procedure if $r_e>0$. Observe that the probabilities that N_q and W are larger than n_0 and w_0 approaches to zero as c grows and then, $g(c)$ will be approximately linearly increasing for large c. The same argument can be used for expected cost functions with losses with the structure given in (24).

If the system parameters are not known, but we have a sample of interarrival and service times, $\{\mathbf{t},\mathbf{s}\}$, we can estimate the mean cost per u.t. given the MCMC output in the usual way,

$$g\left(c\mid\mathbf{t},\mathbf{s},\rho<c\right)=E\left[Cost\mid\mathbf{t},\mathbf{s},\rho<c\right]\approx\frac{1}{J_1}\sum_{j:\rho^{(j)}<c}E\left[Cost\mid\boldsymbol{\theta}^{(j)}\right],\qquad(27)$$

where J_1 is given in (11).

6 Bank data problem.

In this section, we consider the design of a multiserver real bank in Madrid. Interarrival and service times of 98 customers are recorded from 10:00 to 11:30 in the morning during three days. The mean service time is approximately 275.16 seconds. Our Bayesian density estimation method predicts an exponential distribution for service time distribution. Thus, we assume this model for the service time. We also use a non-informative prior in (4) by setting a and b equal to zero. Then, the posterior distribution of the service rate parameter, μ, is $G(98,26965.6)$.

Figure 1 shows the histogram of the 98 interarrival times. The estimated density function (7) using the Erlang mixture with the BDMCMCM algorithm has been superimposed. None of times is

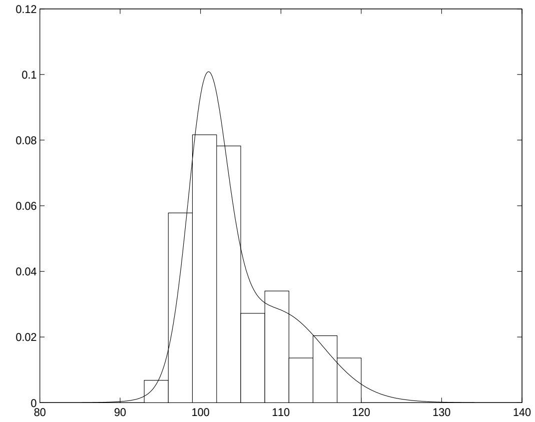

Figure 1: Histogram of interarrival time data and estimated interarrival time density.

larger than two minutes and the distribution seems to be bimodal. In fact, the posterior probability of having two Erlang mixture components is very high, $P(k = 2 \mid \mathbf{t}) \approx 0.958$.

Given these arrival and service data, we estimate the posterior probability of having a stable system, see (8), for different values of c, which are shown in Table 1. Observe that at least, 3 servers are needed to assume that the ergodic condition, $\rho < c$, holds. However, 3 servers may not satisfy the optimal conditions resulting from the balance of costs in the system, as will be shown below.

Table 1 also shows the estimations for the traffic intensity for each c, see (10). Note that using (26), it can be shown that,

$$E\left[N_b \mid \mathbf{s}, \mathbf{t}, \rho < c\right] = E\left[\rho \mid \mathbf{s}, \mathbf{t}, \rho < c\right].$$

Then, when there are only 1 or 2 servers, all of them are almost always busy on average as the system is probably unstable. But, when there are 3 servers or more, the equilibrium condition holds with high probability and there are approximately 2.66 busy servers on average.

Figure 2 illustrates the estimated probabilities describing the number of customers in the system, N, see (16), at arbitrary time instants, for 3, 4 and 5 servers. Note that the probability of having 2 or 3 customers in the system are very similar for each number of servers. We have observed that this feature does not appear in the predictive distribution of N^* where we have identified the mode

c	1	2	3	4	5	6
$P(\rho < c \mid \mathbf{s}, \mathbf{t})$	.00001	.00181	.89194	.99996	1.00	1.00
$E[\rho \mid \mathbf{s}, \mathbf{t}, \rho < c]$	0.999	1.976	2.661	2.660	2.660	2.6580

Table 1: Estimations of the posterior probabilities of having a stable system and the expected values for the traffic intensity for some values of c.

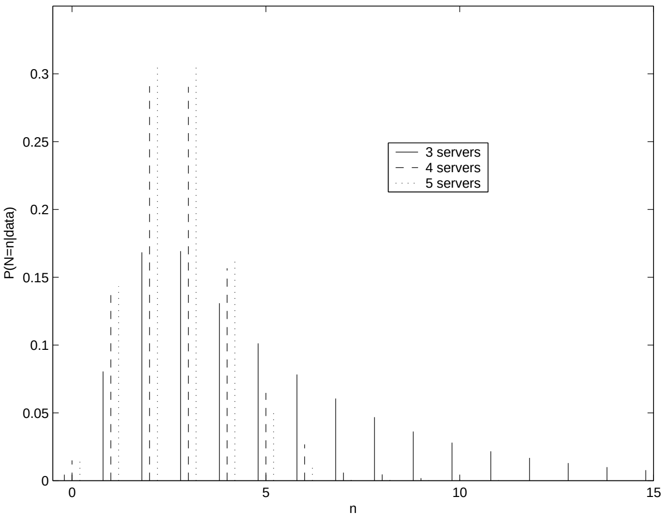

Figure 2: Predictive probabilities for the number of customers in the system at arbitrary time instants for queues with 3, 4 and 5 servers.

in 2 customers for any number of servers. This illustrates the fact that, although the mean number of busy servers at an arbitrary time instant is 2.66, the mean number of busy servers found by an arrival customer is less than 2.66. It can also be seen that the distribution of N in a system with 3 servers has a long tail compared to other systems. Note that just by increasing the number of servers from 3 to 4 the probability of having an empty queue, $P(N_q = 0 \mid \mathbf{s}, \mathbf{t})$, grows from 0.42 to 0.89.

Figure 3 shows the distribution of the waiting time in the queue, W, see (21), for 3, 4 and 5 servers. Observe that, in a system with 3 servers, the probability of having to wait less than 10 minutes (600 seconds) is fairly large, $P(W < 10 \mid \mathbf{s}, \mathbf{t}) \approx 0.85$. However, again, if the value of c is increased from 3

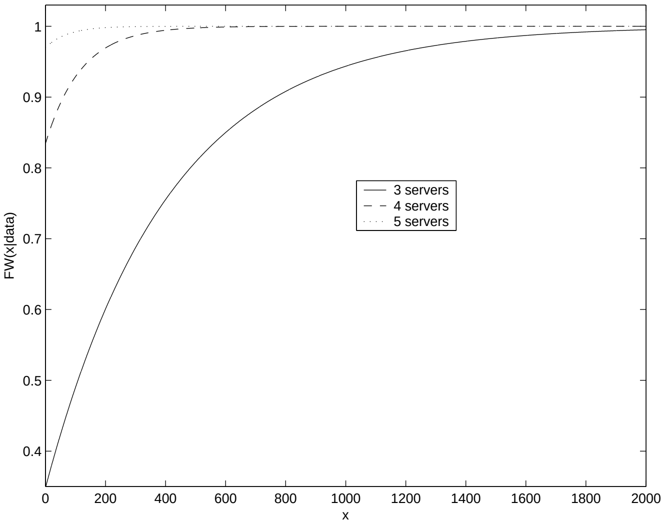

Figure 3: Predictive distribution functions of the waiting time (in seconds) in the queue for systems with 3, 4 and 5 servers.

to 4 the probability of not having to wait, $P\left(W=0 \mid \mathbf{s},\mathbf{t}\right)$, obtained from (21), grows from 0.35 to 0.83.

Now, we can address the optimization problem with the bank data. We formulate different cost functions defined from the minimum number of servers from which we have assumed equilibrium, that is, 3 servers. In practice, it is not easy to assign costs associated with the wait of customers, but, in general, the costs incurred from servers activities are known. Thus, we consider fixed costs per u.t. per busy server, $r_b = 1.5$, per empty server, $r_e = 1$, and per served customer, $r_s = -0.05$, and we consider different values for r_q and r_W. We assume a L_2 loss function for N_q, see (24) and a L_1 loss function for W, see (23). Finally, the thresholds for N_q and W are assumed to be $n_0 = 20$ and $w_0 = 2000$, respectively, as the probability of exceeding these values is very small, see Figures 2 and 3.

Table 2 shows the estimated average cost per u.t. obtained from (27) for different values of r_q and r_W. Each column corresponds to an average cost function depending on c. Optimum values are indicated in bold. Observe that the first two functions are very similar for $c \geq 4$ and same feature is observed for the last two functions. The reason is that, for $c \geq 4$, the increase in r_W does not affect in the cost, because, in this case, the probability of having to wait more than $w_0 = 2000$ seconds is very close to zero. However, the increase in the cost per customer in the queue, r_q, does affect in the average cost function, as the value of $E\left[N_q \mid N_q \leq 20\right]$ is larger in this case. Finally, as expected, Table 2 also shows that for $c \geq 7$ all cost functions are very similar and tend to be the same linear function with slope $r_e = 1$ not influenced by the values of r_q and r_W.

	$E\left[Cost \mid \mathbf{s}, \mathbf{t}, \rho < c\right]$			
c	$r_W = .01$ $r_q = .01$	$r_W = 1$ $r_q = .01$	$r_W = .01$ $r_q = 1$	$r_W = 1$ $r_q = 1$
3	**4.3535**	**4.3584**	6.8607	6.8655
4	5.3300	5.3300	**5.5159**	**5.5159**
5	6.3284	6.3284	6.3493	6.3493
6	7.3282	7.3282	7.3301	7.3301
7	8.3282	8.3282	8.3283	8.3283

Table 2: Estimated average cost per u.t. for different values of r_q and r_W. Optimal values are indicated in bold.

7 Conclusions

In this paper, we have proposed a Bayesian approach for control of the number of servers c in a $GI/M/c$ system. We have developed a BDMCMC method based on mixtures of Erlang distributions to approximate the general interarrival time distribution, performance measures have been predicted and incorporated into average cost functions to determine the optimal number of servers. This methodology have been illustrated with a real data set.

Our Bayesian approach can be extended to the $GI/G/c$ queue considering Erlang mixtures both for the interarrival and the service times. However, in this case, the stationary distributions are not easy to calculate. One possibility is to consider the phase type family of distributions (PH) introduced by Neuts (1981). Some known results of the $GI/PH/c$ model could be used as the Erlang mixture is a PH distribution. Similar ideas are implemented in Ausín et al. (2004) for the $M/PH/1$ queue.

A more general extension consists in the design of the $GI/G/c/K$ model, with $K \leq \infty$, where K is the system capacity. It is possible to extend the cost structure to queues with finite capacity by considering costs based on lost demand. An example for the particular case where the system capacity equals the number of servers can be found in Ausín et al. (2003).

Some modifications of our analysis could also be carried out. 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 Ausín et al. (2004). In practice, we have found that both schemes perform similarly. However, the BDMCMC algorithm is somewhat easier to implement.

We could also have considered approximating the interarrival time with a mixture of gamma distributions which is a more flexible model; see Wiper et al. (2001). However, a disadvantage of this model is that the probability that a simpler model (exponential, Erlang or hyperexponential) cannot be easily calculated; see e.g. Ausín et al. (2004). Another disadvantage is that the gamma mixture is not PH which means that extension to more complex systems with this model is difficult.

Finally, there are some alternatives to the cost structure defined. For example, costs per unit of time in the stationary state could be replaced by costs per busy cycle using the cycle criterion, see Lillo (2000).

Acknowledgements

We acknowledge the financial support partially provided by project DGICYT BEC2000-0167.

References

ALLEN, A.O. 1990. *Probability, Statistics and Queueing Theory with Computer Science Applications.* Academic Press, Boston.

ARMERO, C., AND M.J. BAYARRI. 1996. Bayesian questions and answers in queues. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith, eds. *Bayesian Statistics 5.* Oxford University Press, Oxford, pp. 613-618.

ARMERO, C., AND D. CONESA. 2000. Prediction in Markovian bulk arrival queues. *Queueing Systems.* 34, 327-350.

AUSIN, M.C., R.E. LILLO, F. RUGGERI AND M.P. WIPER. 2003. Bayesian Modeling of Hospital Bed Occupancy Times Using a Mixed Generalized Erlang Distribution. J.M. Bernardo, J.O. Berger, A.P. Dawid, M. West, eds. *Bayesian Statistics 7.* Oxford University Press, Oxford, pp. 443-452.

AUSIN, M.C., M.P. WIPER AND R.E. LILLO. 2004. Bayesian Estimation for the M/G/1 queue using a phase type approximation. *Journal of Statistical Planning and Inference.* 118, 83-101.

BAGCHI, T.P., AND A.A. CUNNINGHAM. 1972. Bayesian approach to the design of queueing systems. *INFOR.* 10, 36-46.

DIEBOLT, J., AND C.P. ROBERT. 1994. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series 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.

GROSS, D., AND C.M. HARRIS. 1985. *Fundamentals of Queueing Theory.* John Wiley & Sons, New York.

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. 2003. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics.* 12, 1-25.

KITAEV, M.Y., AND V.V. RYKOV. 1995. *Controlled queueing systems.* CRC Press, Boca Raton, FL.

LEHOCZKY, J. 1990. Statistical methods. D.P. Heyman and M.J. Sobel, eds. *Stochastic Models.* Elsevier/North-Holland, New York and Amsterdam, pp. 255-293.

LILLO, R.E. 2000. On the optimal control of $M/G/1$ systems under the cycle criterion. *Systems & Control Letters.* 41, 29-39.

LILLO, R.E., AND M. MARTIN. 2000. Characterization and computation of optimal policies for an $M/G/1$ priority queue. *The Belgian Journal of Operations Research, Statistics and Computer Science.* 38, 45-57.

NEUTS, M.F. 1981. *Matrix Geometric Solutions in Stochastic Models.* Johns Hopkins University Press, Baltimore.

RICHARDSON, S., AND P.J. GREEN. 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series 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.P. 1996. Mixtures of distributions : inference and estimation. W.R. Gilks, S. Richardson, D.J. Spiegelhalter, eds. *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London, pp. 441-464.

STEPHENS, M. Bayesian analysis of mixture models with an unknown number of components — an alternative to reversible jump methods. *The Annals of Statistics.* 28, 40-74.

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 AND F. RUGGERI. 2001. Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics.* 10, 440-454.