Bayesian prediction of the transient behaviour and busy period in short and long-tailed GI/G/1 queueing systems.

M. Concepción Ausín,*
Department of Mathematics, Universidade da Coruña,
15071 A Coruña, Spain.
Michael P. Wiper, Rosa E. Lillo
Department of Statistics, Universidad Carlos III de Madrid,
28903 Getafe, Madrid, Spain.

Abstract

Bayesian inference for the transient behavior and duration of a busy period in a single server queueing system with general, unknown distributions for the interarrival and service times is investigated. Both the interarrival and service time distributions are approximated using the dense family of Coxian distributions. A suitable reparameterization allows the definition of a non-informative prior and Bayesian inference is then undertaken using reversible jump Markov chain Monte Carlo methods. An advantage of the proposed procedure is that heavy tailed interarrival and service time distributions such as the Pareto can be well approximated. The proposed procedure for estimating the system measures is based on recent theoretical results for the Coxian/Coxian/1 system. A numerical technique is developed for every MCMC iteration so that the transient queue length and waiting time distributions and the duration of a busy period can be estimated. The approach is illustrated with both simulated and real data.

Keywords: Bayesian inference, Coxian distribution, heavy tails, queueing systems, semiparametric modelling, transient analysis, reversible jump.

[1]Corresponding author. M. C. Ausín, Departamento of Matemáticas, Facultad de Informática, Campus de Elviña, Universidade da Coruña, 15071 A Coruña, Spain. Tel.: +34 981 167 000 (ext. 1301); fax: + 34 981 167 160. E-mail address: mausin@udc.es

1 Introduction

Bayesian estimation of stationary distributions in queueing systems is a well developed research area. Some useful references are Armero and Bayarri (1994, 1997), Rios et al. (1998), Wiper (1998), Armero and Conesa (2000) and Ausín et al. (2003, 2004). However, less progress has been made on the estimation of the transient behaviour although in many practical situations, the analysis of a queueing system in the transient states is of great interest, for example, when the system is regularly stopped and started again, or when convergence to steady-state is very slow. Furthermore, the estimation of the duration of a busy period has also been neglected in the Bayesian literature on queues although the study of the busy period distribution is very important in optimal control problems. Exceptions are Armero and Bayarri (1994, 1997) who consider the busy period for the M/M/1 system and the transient distribution for the M/M/∞ system, respectively, and Ausín et al. (2004) who consider the busy period of the M/G/1 system. Although exact results for the transient behaviour and busy period of Markovian systems are known, see e.g. Gross and Harris (1985), such systems are rarely found in practice. One of the contributions of this article is the estimation of the transient behaviour and busy period distribution of a much more general queueing system.

It is well known that many measures in communication systems such as packet interarrival times, file lengths or intervals between connection requests in Internet traffic, have long-tailed distributions, see e.g. Paxson and Floyd (1995) and Crovella et al. (1997). These measures are usually described with distributions such as the Pareto or Weibull. Unfortunately, queueing systems with long-tailed interarrival or service time distributions are very difficult to analyze due to the non-existence of their moments or the lack of Laplace transforms with explicit expressions, see e.g. Abate et al. (1994). As an alternative, Feldmann and Whitt (1998) and Riska et al. (2004) propose approximating long-tailed distributions using hyperexponential densities and mixtures of Erlang distributions, respectively, and apply these approximations to the analysis of queueing systems using classical statistical methodology. Bayesian inference for hyperexponential distributions and mixtures of Erlang distributions has been considered in Rios et al. (1998) and Ausín et al. (2004), respectively. However, due to the influence of the proper prior distributions, these approaches cannot properly approximate long-tailed interarrival or service time distributions. In contrast, in this article, the model and prior formulation used permits us to well approximate long-tailed distributions.

Throughout, we shall use the Coxian family of distributions as a semiparametric approximation to the interarrival and service time distributions. This family includes the standard distributions usually studied in queueing theory, such as the exponential, Erlang and hyperexponential distribution, as special cases. The Coxian family is dense over the set of distributions on the positive reals, see e.g. Bertsimas (1990), and thus, any continuous and positive distribution can be approximated arbitrarily closely. Furthermore, based on the Cox's method of stages, see Cox (1955), many results for systems with Coxian distributions have been obtained in the queueing literature. In particular, in this paper we will consider results derived by Bertsimas and Nakazato (1992) for the transient behaviour and busy period.

Bayesian inference for the Coxian distribution was considered in Ausín et al. (2003), in the context of a hospital resource allocation problem equivalent to a M/G/c/K, finite capacity, Poisson arrivals, queueing system. In that article, a different parameterization and approach to that used here was applied. There, a proper prior was assumed and semi-conjugate inference was developed using latent variables. Unfortunately, this setup leads to poor approximations of long-tailed distributions. Furthermore, the Coxian model considered in Ausín et al. (2003) is not identifiable which often affects the convergence of the MCMC algorithm and the interpretation of the estimated parameters. Finally, transient and busy period problems were not analyzed. Here, we propose a new parameterization which provides an identifiable Coxian model where an improper prior distribution can be applied which leads to good performance of the MCMC algorithm and which allows us to well approximate both short and long-tailed distributions.

The rest of this paper is organized as follows. In Section 2, we introduce the Coxian distribution model and then describe how to carry out Bayesian inference for this model given an improper prior and using a reversible jump MCMC algorithm, see Green (1995) and Richardson and Green (1997). In Section 3, we describe some results derived by Bertsimas and Nakazato (1992) which allow us to estimate the transient and busy period distributions of the Coxian/Coxian/1 system conditional on the system parameters. Then,

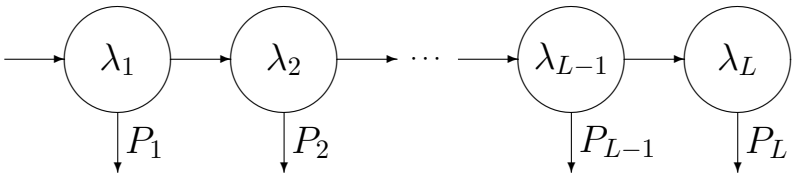

Figure 1: Phase diagram of the Coxian distribution which, with probability P_r, is a sum of r exponential random variables with rates $\lambda_1, \ldots, \lambda_r$, for $r = 1, \ldots, L$.

we explain a numerical procedure which allows us to incorporate these results within the reversible jump algorithm so that we can estimate the predictive transient and busy period distributions given a sample of interarrival and service data. At the end of both Section 2 and Section 3, our approach is illustrated with simulated and real interarrival and service time data. The paper finishes with some brief conclusions and extensions.

2 Modelling and inference for the interarrival and service times

Firstly we will define the Coxian distribution that we shall use throughout to model the interarrival and service time data. The Coxian or Mixed Generalized Erlang distribution (MGE), is defined as follows. A positive random variable, X, has a Coxian distribution with parameters L, $\mathbf{P} = (P_1, ..., P_L)$ and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_L)$, if,

$$X = \begin{cases} Y_1, & \text{with probability} = P_1, \\ Y_1 + Y_2, & \text{with probability} = P_2, \\ \vdots & \vdots \\ Y_1 + ... + Y_L, & \text{with probability} = P_L, \end{cases} \tag{1}$$

where Y_r has an exponential distribution with parameter λ_r, $P_r, \lambda_r > 0$ and $\sum_{r=1}^{L} P_r = 1$, see e.g. Bertsimas and Nakazato (1992). This model has a nice visual representation, see Figure 1, where it can be seen that the interarrival (or service) time of each customer can be represented by a sequence of a random number of exponential stages.

The Coxian distribution model (1) can also be expressed in mixture form,

$$f\left(x \mid L, \mathbf{P}, \boldsymbol{\lambda}\right) = \sum_{r=1}^{L} P_r f_r\left(x \mid \lambda_1, ..., \lambda_r\right), \qquad x \geq 0, \tag{2}$$

where $f_r\left(x \mid \lambda_1, ..., \lambda_r\right)$ is the density of a generalized Erlang distribution, i.e. the density of a sum of r exponential random variables with rates $\lambda_1, \lambda_2, ..., \lambda_r$ respectively. If all rates are unequal, this density is given by,

$$f_r\left(x \mid \lambda_1, ..., \lambda_r\right) = \sum_{j=1}^{r} \left(\prod_{\substack{i=1 \\ i \neq j}}^{r} \frac{\lambda_i}{\lambda_i - \lambda_j} \right) \lambda_j \exp\left(-\lambda_j x\right), \qquad x \geq 0, \tag{3}$$

see e.g. Johnson and Kotz (1970). In some cases where there are two or more equal rates, alternative expressions can be derived, e.g. if all rates are equal, the Coxian density reduces to an Erlang, but a general formula for all possible cases is not known. However, the Laplace transform of the Coxian distribution

defined in (2) is available and is given by,

$$f_r^*(s \mid \lambda_1, ..., \lambda_r) = \prod_{i=1}^{r} \left(\frac{\lambda_i}{\lambda_i + s} \right). \tag{4}$$

Thus, if we wish to approximate the Coxian density when there are two or more equal rates, or when the difference between two rates is small, this can be done by numerical inversion of the Laplace transform using, for example the algorithm by Hosono (1981). In the examples later in this article, we have used Laplace transform inversion rather than direct evaluation of the Coxian density whenever we encountered at least a pair of rates whose difference was less than 10^{-4}.

Cumani (1982) shows that the Coxian distribution model (1) is identifiable up to permutation of the rates and therefore, we can assume without loss of generality that,

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_L. \tag{5}$$

In the following subsection, we will reparameterize the Coxian distribution using this identifiability constraint so that given a sample of data, we can develop a Bayesian inference scheme using a non-informative prior.

2.1 Bayesian inference

Assume now that we undertake the simple experiment of observing independent samples of interarrival times and service times. This experiment has been considered in the majority of Bayesian articles on queueing; see e.g. Armero and Bayarri (1996). We assume independence between the arrival and service processes and will model both as Coxian distributions. Then, given that we use independent prior distributions for the arrival and service parameters, the corresponding posterior distributions will also be independent. Therefore, for the rest of this section, we assume that we have observed a sample of n interarrival (or service) times, $\mathbf{x} = (x_1, ..., x_n)$, following a Coxian distribution and we want to make inference over the Coxian parameters.

Firstly, it is convenient to reparameterize the rates $\boldsymbol{\lambda}$ of the Coxian model as follows,

$$\lambda_r = \lambda_1 \upsilon_2 ... \upsilon_r, \qquad \text{where } 0 < \upsilon_j \leq 1, \quad \text{for } r, j = 2, ..., L. \tag{6}$$

This reparameterization automatically incorporates the ordering restraint given in (5). This kind of reparameterization has also been considered in Robert and Mengersen (1999) for normal mixtures, and in Gruet et al. (1999) for exponential mixtures. There are two main reasons for using this reparameterization. Firstly, the usual requirement of imposing a joint proper prior distribution for the mixture parameters is relaxed. Note that using the standard parameterization $(\lambda_1, \ldots, \lambda_L)$, as in Ausín et al. (2003), it is not possible to use an improper prior for each rate, λ_r, as it leads to an improper posterior distribution, see e.g. Diebolt and Robert (1994). In contrast, using the reparameterization (6), we will be able to assume a non-informative improper distribution for $(\lambda_1, \upsilon_2, \ldots, \upsilon_L)$ which will allow us to approximate heavy-tailed distributions. Secondly, the performance of the MCMC algorithm can be improved avoiding trapping states with empty, or almost empty, components since most of observations will contribute to the estimation of various mixture rates. In particular, every observation will give information about the first rate λ_1.

We now define prior distributions for the model parameters $(L, \mathbf{P}, \lambda_1, \boldsymbol{\upsilon})$ as follows. Supposing first that L is known, we assume the following improper prior distribution,

$$\mathbf{P} \sim \textit{Dirichlet}\,(\phi_1, ..., \phi_L)\,, \tag{7}$$

$$f(\lambda_1) \propto \frac{1}{\lambda_1} \tag{8}$$

$$\upsilon_r \sim \textit{Beta}(a, b), \qquad \text{for } r = 1, ..., L. \tag{9}$$

Although the joint prior distribution is improper, it can be shown that it leads to a proper posterior, see the Appendix. Typically, we might set $\phi_r = 1$, for $r = 1, ..., L$, to give a uniform prior over the weights. Also,

we might set $a = b = 1$ to give a uniform prior over each υ_r, for $r = 1, ..., L$. However, it is necessary to set $a > 1$ in order that the posterior predictive mean of the Coxian variable, X, exists, as is shown in the Appendix. Then, we have set $a = 1.1$ and $b = 1$ in the examples.

This type of parameterization and prior choice allows us to approximate long-tailed distributions as will be shown in the examples. This is because we do not make a strong assumption about the rate of the first component, λ_1, and consequently, the means of the various mixture components can be as small or as large as required.

Our task is now to construct an MCMC algorithm to sample from the joint posterior distribution which is obtained from the prior distributions, (7), (8) and (9), and the likelihood function,

$$l(\boldsymbol{\theta} \mid \mathbf{x}) = \prod_{i=1}^{n} \left(\sum_{r=1}^{L} P_r f_r(x_i \mid \lambda_1, \upsilon_2, \ldots, \upsilon_r) \right), \tag{10}$$

where $\boldsymbol{\theta} = (L, \mathbf{P}, \lambda_1, \boldsymbol{\upsilon})$ and where from now on we use $f_r(x \mid \lambda_1, \upsilon_2, ..., \upsilon_r)$ to denote the reparameterized Coxian density $f_r(x \mid \lambda_1, \lambda_2, .., \lambda_r)$ given in (2). As is usually done in mixture models, see e.g. Diebolt and Robert (1994), we use a data augmentation algorithm, introducing for each datum, x_i, component indicator variables, Z_i, such that,

$$P(Z_i = r \mid L, \mathbf{P}) = P_r, \qquad f(x_i \mid Z_i = r) = f_r(x_i \mid \lambda_1, \upsilon_2, \ldots, \upsilon_r)$$

for $i = 1, ..., n$. Then, we obtain that,

$$P(Z_i = r \mid x_i, \mathbf{P}, \lambda_1, \boldsymbol{\upsilon}) \propto P_r f_r(x_i \mid \lambda_1, \upsilon_2, ..., \upsilon_r), \quad \text{for } r = 1, ..., L, \tag{11}$$

for $i = 1, ..., n$, where f_r is evaluated by using (3) or inverting (4) as commented above. With this approach, every observed data set, $\mathbf{x} = \{x_1, \ldots, x_n\}$ is associated to a missing data set, $\mathbf{z} = \{z_1, \ldots, z_n\}$, indicating the specific components of the mixture from which the observations are assumed to arise. Given the missing data, the likelihood (10) is simplified to,

$$l(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{z}) \propto \prod_{i=1}^{n} P_{z_i} f_{z_i}(x_i \mid \lambda_1, \upsilon_2, \ldots, \upsilon_{z_i}). \tag{12}$$

From (7) and (12), the conditional posterior distribution of the weights is given by,

$$\mathbf{P} \mid \mathbf{x}, \mathbf{z} \sim \mathit{Dirichlet}(1 + n_1, ..., 1 + n_L), \tag{13}$$

where n_r is the number of data assigned to the r'th mixture component for $r = 1, ..., L$. From (8), (9) and (12), we obtain the conditional posterior distributions of λ_1 and $\boldsymbol{\upsilon}$, which do not have explicit expressions but their density functions can be evaluated up to the integration constants as,

$$f(\lambda_1 \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\upsilon}) \propto \prod_{i=1}^{n} f_{z_i}(x_i \mid \lambda_1, \upsilon_2, \ldots, \upsilon_{z_i}) f(\lambda_1), \tag{14}$$

$$f(\upsilon_r \mid \mathbf{x}, \mathbf{z}, \lambda_1, \boldsymbol{\upsilon}_{-r}) \propto \prod_{\substack{i=1 \\ z_i \geq r}}^{n} f_{z_i}(x_i \mid \lambda_1, \upsilon_2, \ldots, \upsilon_{z_i}) f(\upsilon_r), \quad \text{for } r = 2, \ldots, L. \tag{15}$$

Thus, we can now use Gibbs sampling scheme to generate a sample from the posterior parameter distribution. The only slightly complicated steps are sampling the densities of λ_1 and $\boldsymbol{\upsilon}$ in (14) and (15). Firstly, in order to sample from the distribution in (14) we use a Metropolis-Hastings step with a Gamma candidate. This choice is motivated by the fact that λ_1 follows a Gamma posterior distribution when a larger set of latent variables is introduced, see Ausín et al. (2003). To sample υ_r, we use a Metropolis-Hastings step with a Beta mixture candidate distribution,

$$g\left(\tilde{\upsilon}_r \mid \upsilon_r^{(j-1)}\right) = \frac{1}{2} \mathit{Beta}\left(\frac{1}{1 - \upsilon_r^{(j-1)}}, 2\right) + \frac{1}{2} \mathit{Beta}\left(2, \frac{1}{\upsilon_r^{(j-1)}}\right), \quad \text{for } r = 2, ..., L.$$

This mixture has been chosen to avoid getting stuck at indefinitely at values of υ_r near to zero or one and to simultaneously preserve the mode of the value of υ_r in the previous iteration.

We can extend the previous Gibbs sampling algorithm to the case where L is unknown. Firstly, assume a discrete uniform prior defined on $[1, L_{\max}]$. In the examples, we have set $L_{\max} = 20$. In order to let the chain move through the posterior distribution of L, we use the reversible jump algorithm, see Green (1995) and Richardson and Green (1997). Specific moves are designed to allow us to try to change the number of components at each step from L to $L \pm 1$. We consider the so called split and combine moves where one mixture component, r, is split into two adjacent components (r_1, r_2) or two adjacent components are combined into one, respectively. In the combine move the parameters are modified such that,

$$\tilde{P}_r = P_{r_1} + P_{r_2}, \qquad \tilde{\upsilon}_r = \upsilon_{r_1}\upsilon_{r_2},$$

which implies that $\tilde{\lambda}_r = \lambda_{r_2}$. For the case that $r = 1$ we consider $\tilde{\lambda}_1 = \lambda_1\upsilon_2$. For the split move,

$$\begin{array}{ll} \tilde{P}_{r_1} = u_1 P_r, & \tilde{\upsilon}_{r_1} = u_2 + \upsilon_r (1 - u_2), \\ \tilde{P}_{r_2} = (1 - u_1)P_r, & \tilde{\upsilon}_{r_2} = \frac{\upsilon_r}{u_2 + \upsilon_r(1-u_2)}, \end{array}$$

where u_1 and u_2 are uniform $U(0, 1)$. This implies that $\tilde{\lambda}_{r_1} = \lambda_{r-1}u_2 + \lambda_r (1 - u_2)$ and $\tilde{\lambda}_{r_2} = \lambda_r$. For the case that $r = 1$, we generate $\tilde{\lambda}_1 = \lambda_1/u_2$ and $\tilde{\upsilon}_2 = u_2$ where $u_2 \sim U(0, 1)$. Also, every observation such that $z_i = r$ is assigned to each of the two components, r_1 or r_2, with probability,

$$P\left(\tilde{Z}_i = r_j\right) \propto \tilde{P}_{r_1} f_{r_1}\left(x_i \mid \lambda_1, \upsilon_2, ..., \tilde{\upsilon}_{r_j}\right), \quad \text{for } j = 1, 2.$$

Note that the defined split-combine moves are chosen such that the parameters in the remaining mixture components are not modified, as in Gruet et al. (1999), and do not necessarily preserve the moments of the distribution of X. The acceptance probability of a split move is $\min\{1, A\}$ where,

$$A = \frac{\tilde{P}_{r_1}^{\tilde{n}_{r_1}} \tilde{P}_{r_2}^{\tilde{n}_{r_2}}}{P_r^{n_r}} \frac{\prod\limits_{i:\tilde{z}_i \geq r_1} f_{\tilde{z}_i}\left(x_i \mid \lambda_1, \upsilon_2, ..., \tilde{\upsilon}_{\tilde{z}_i}\right)}{\prod\limits_{i:z_i \geq r_1} f_{z_i}\left(x_i \mid \lambda_1, \upsilon_2, ..., \upsilon_{z_i}\right)} \times \frac{d_{L+1}}{b_L \prod\limits_{i:z_i = r} P\left(\tilde{Z}_i = \tilde{z}_i\right)} \times \frac{P_r (1 - \upsilon_r)}{u_2 + \upsilon_r (1 - u_2)}$$

when $r > 1$ and where $\tilde{n}_{r_j}$ is the number of observations assigned to component r_j, for $j = 1, 2$ and d_L and b_L are respectively the probabilities of a combine or a split move. The last factor is the determinant of the Jacobian of the transformation $(P_r, \upsilon_r, u_1, u_2) \rightharpoonup (\tilde{P}_{r_1}, \tilde{P}_{r_2}, \tilde{\upsilon}_{r_1}, \tilde{\upsilon}_{r_2})$. When $r = 1$, the Jacobian determinant of $(P_1, \lambda_1, u_1, u_2) \rightharpoonup (\tilde{P}_1, \tilde{P}_2, \tilde{\lambda}_1, \tilde{\upsilon}_2)$ is given by P_1/u_2. The acceptance probability for the reverse, combine move can be obtained analogously. Note that the acceptance probabilities do not need to incorporate factorial terms due to the natural order of the rates derived from the new parameterization, see Gruet et al. (1999). As usual, given a MCMC sample of size J, the predictive distribution of the interarrival (or service) time can be approximated by,

$$f(x \mid \mathbf{x}) \approx \frac{1}{J} \sum_{j=1}^{J} \sum_{r=1}^{L^{(j)}} P_r^{(j)} f_r\left(x \mid \lambda_1^{(j)}, \upsilon_2^{(j)}, ..., \upsilon_r^{(j)}\right). \tag{16}$$

2.2 Results for simulated and real data sets

In this subsection, we illustrate the performance of our proposed Bayesian density estimation method using different simulated and real data samples.

Simulated data. Firstly, we consider 300 simulated data for each of the following distributions:

1. A single exponential distribution with $\lambda = 1$.

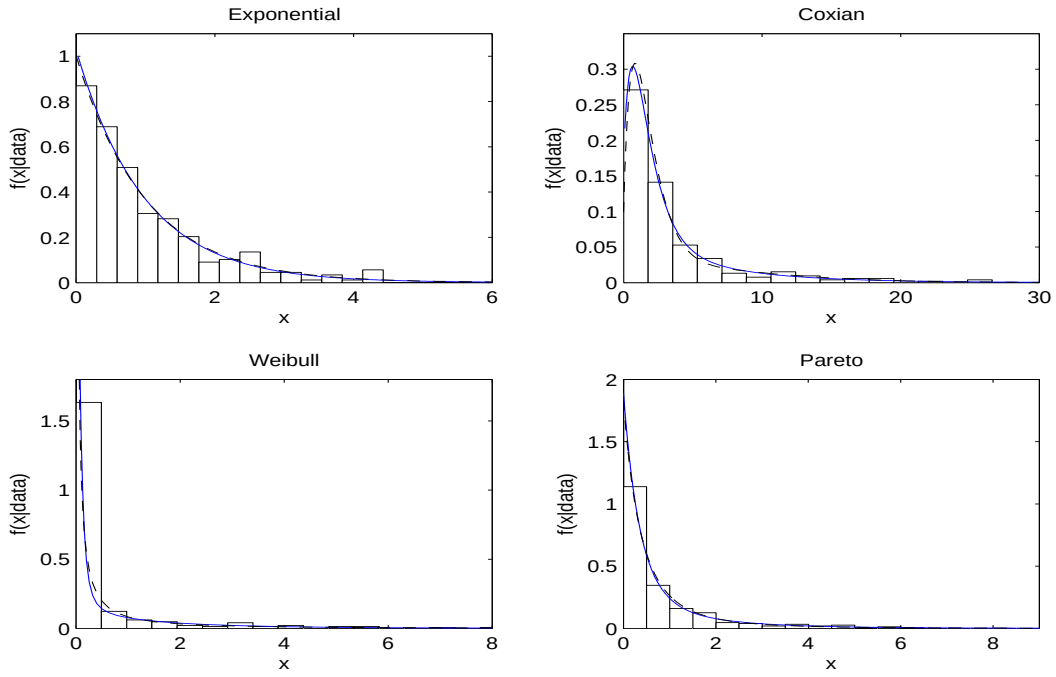

Figure 2: Histograms, predictive densities (solid line) and theoretical densities (dotted line) for the four simulated data sets. The fits are so close to the true densities that it is sometimes complicated to differentiate both lines.

2. A Coxian distribution with $\mathbf{P} = (0.09, 0.7, 0.01, 0.2)$ and $\boldsymbol{\lambda} = (1.1, 1, 0.251, 0.25)$.
3. A Weibull density , $f(x) = ca^c x^{c-1} \exp\{-(ax)^c\}$, with $(c, a) = (0.3, 9.26053)$.
4. A Pareto distribution, $f(x) = ab(1+bx)^{-(a+1)}$, with $(a, b) = (2.2, 0.8333)$.

The exponential distribution is the simplest case of a Coxian distribution. In case 2, we have chosen a Coxian distribution with two pairs of very similar rates to illustrate that this does not affect the stability of the algorithm as commented earlier. These two examples are short-tailed distributions. Cases 3 and 4 are two of the examples of long tailed distributions considered in Feldman and Whitt (1998). The distributions in cases 1, 3 and 4 all have means equal to unity.

Figure 2 illustrates the empirical and predictive cumulative distribution function estimated after running the MCMC algorithm for 10000 burn-in iterations followed by an additional 100000 iterations. We can observe that the fits are quite satisfactory even for the heavy-tailed cases. The proportions of moves accepted are 27.2%, 13.13%, 12.88% and 16.85% for the exponential, Coxian, Weibull and Pareto cases, respectively, which are reasonable values in reversible jump setups, see Richardson and Green (1997). Figure 3 shows the trace plots of the posterior samples of the mixture size, L, for the four simulated data sets, indicating a good mixing performance. The MCMC algorithm was programmed in FORTRAN and the computational cost depend mainly on the number of phases required to approximate the distribution, varying from a few minutes for the exponential distribution to about one hour for the Weibull data.

Figure 4 shows the posterior probabilities of the number of components, L. Observe that in cases 1 and 2, the algorithm identifies the correct mixture size and the posterior mode of L is equal to its true value. The estimated density in case 3 requires a large number of phases to fit the Weibull distribution. However, we need fewer than the 20 mixture components used in Feldman and Whitt (1998) to fit a Weibull distribution using a hyperexponential approximation. This benefit of using a Coxian rather than a hyperexponential

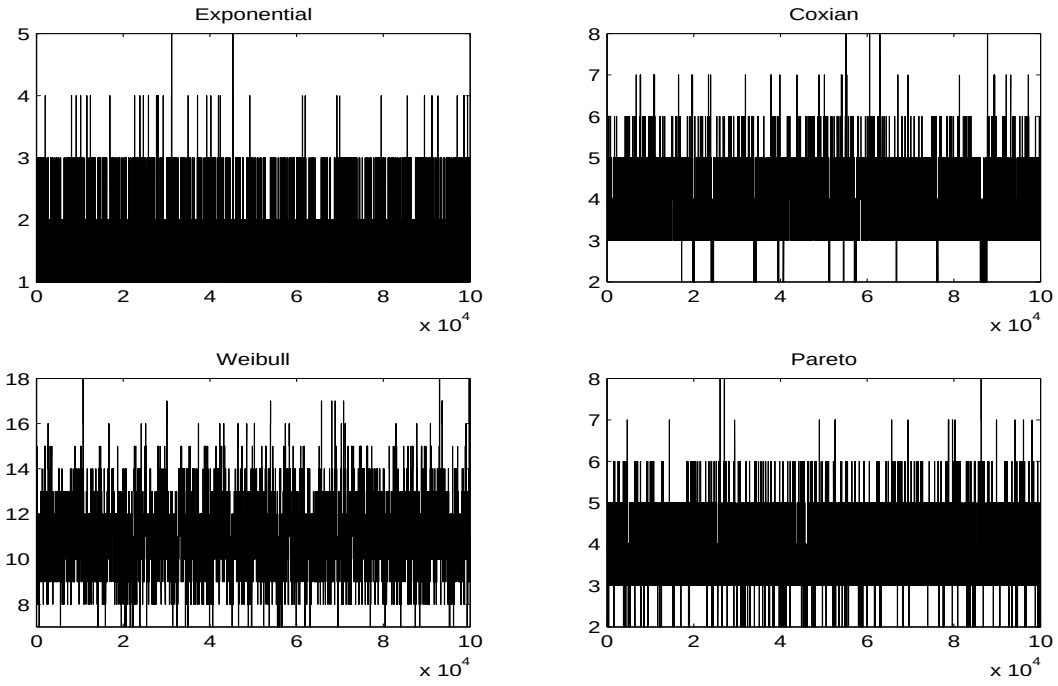

Figure 3: Trace plots of the MCMC samples from the posterior distribution of the mixture size, L, for the four simulated data sets. The plots indicate a good mixing performance visiting many different states.

model is further illustrated in case 4 where the posterior mode of L is 3 in contrast with the 14 exponential components used in Feldman and Whitt (1998).

Figure 5 illustrates the running means for the posterior distribution of the mixture size, L, for the Coxian data set using different starting values in the MCMC algorithm. We can observe that the choice of 10000 burn-in iterations seems adequate and that different initial values lead to similar posterior distributions for L.

Table 1 shows the posterior means and posterior standard deviations of the parameters conditioning on the most probable mixture size. For the exponential data set, the most probable model is a single mixture component (see Figure 4) and, conditioning on $L = 1$, the posterior mean of λ_1 is 1.0366, which is very close to the true rate, 1. For the Coxian data set, the estimated parameters conditioning on the most probable mixture size, $L = 4$, are in general close to the true values. However, note that although the Coxian distribution (1) with the order restriction (5) is an identifiable model, very different values of the parameters may lead to similar density functions. Then, the posterior distributions of various Coxian parameters are not unimodal and the parameters are not significant. Nevertheless, the main interest in practice is the estimation of the density function (see Figure 2) and not the estimation of the parameters. For the Weibull data set, the most probable mixture size is $L = 11$. Observe that the estimated rates, λ_r, vary from extremely large to very small values in order to approximate this heavy-tailed distribution. This is possible because we have assumed an improper prior distribution for the rates, see (8) and (9). Note that using a proper prior for each λ_r as considered in Ausín et al. (2003), the posterior mean of the rates cannot be as small or as large as required. Finally, Table 1 also shows the estimated parameters for the Pareto data set conditioning on $L = 3$.

Real data. We now illustrate the method with real interarrival and service times taken from from a face-to-face bank data base available from `http://iew3.technion.ac.il/serveng`. We consider data of two different kind of services: foreign currency exchange and business banking transactions. These types of

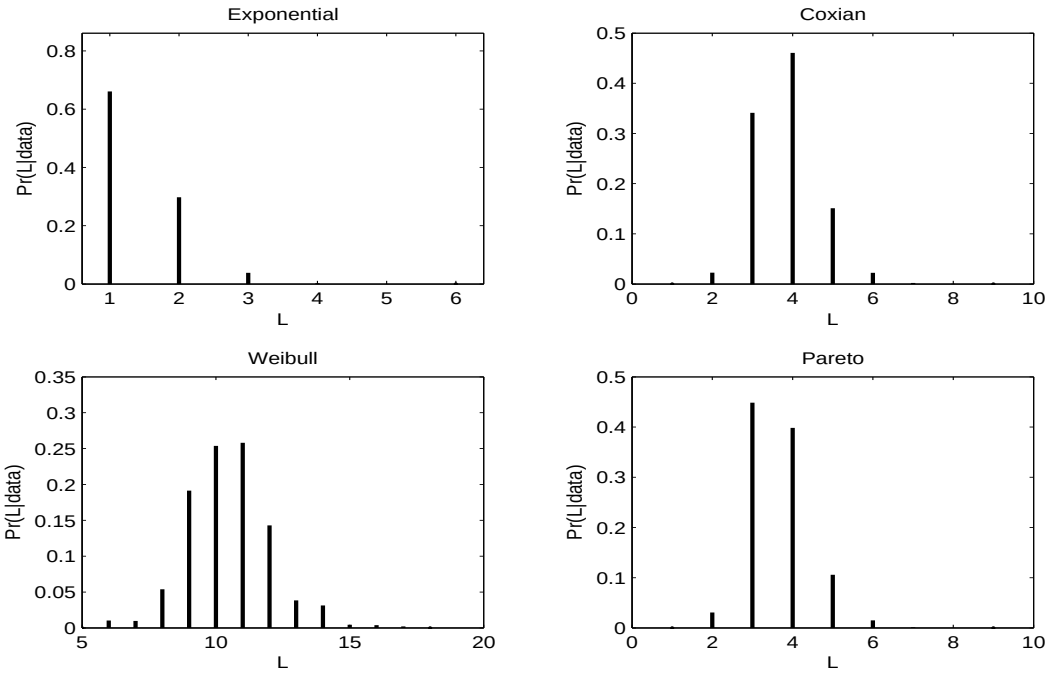

Figure 4: Posterior probabilities of the mixture size, L. The algorithm identifies the correct mixture size for the exponential and Coxian cases. Although the Weibull and Pareto distributions are not Coxian distributions, they can be well approximated using a moderate number of components.

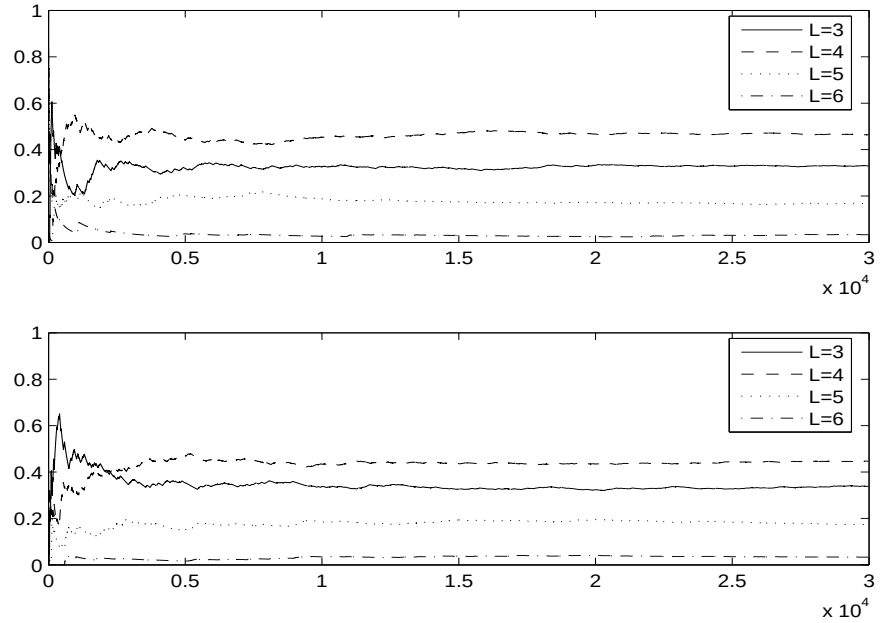

Figure 5: Running means for the posterior distribution of L for the Coxian data using two different initial values: $L^{(0)} = 4$ (top) and $L^{(0)} = 1$ (bottom). Observe that runs from different starting values converge.

	Exponential		Coxian		Weibull		Pareto	
r	P_r	λ_r	P_r	λ_r	P_r	λ_r	P_r	λ_r
1	1.0 [1] 0.0	1.0367 [1] 0.0596	0.1173 [0.09] 0.0734	1.8007 [1.1] 3.3192	0.0477 0.0197	4701652.31 29233277.33	0.6737 0.0882	2.8611 0.6117
2			0.5489 [0.70] 0.1472	0.9981 [1.0] 0.3846	0.0427 0.0224	355710.93 2939331.32	0.2940 0.0895	0.7528 0.2751
3			0.1721 [0.01] 0.1337	0.4320 [.251] 0.2643	0.0567 0.0332	49539.99 307257.03	0.0322 0.0379	0.1723 0.1128
4			0.1615 [0.20] 0.0803	0.1790 [0.25] 0.0418	0.0872 0.0551	6201.52 17800.47		
5					0.1164 0.0659	987.82 3910.61		
6					0.1405 0.0854	174.22 401.97		
7					0.1643 0.0922	41.032 91.229		
8					0.1398 0.0802	9.7372 22.638		
9					0.1164 0.0665	1.7024 2.3922		
10					0.0696 0.0619	0.3095 0.2441		
11					0.0180 0.0113	0.0181 0.0096		

Table 1: Posterior means and posterior standard deviations of the Coxian parameters conditioning on the most probable mixture size. True parameter values are shown in brackets for the Coxian distributions. The estimated rates for the Weibull data vary from extremely large to very small values in order to approximate this heavy-tailed distribution.

service require a single server who works three days a week from 8:30 to 12:00 and two days a week from 8:30 to 12:30 and from 16:00 to 18:00. The data are recorded during 14 days and consist of 249 interarrival and 270 service times for foreign currency exchange and 822 interarrival and 843 service times of business banking transactions.

Figure 6 shows histograms of the observed data and the predictive densities obtained after a run of 100000 iterations in equilibrium. Observe that the Coxian distribution can also fit densities with non zero mode such as the business banking service time density. Interarrival times between customers asking for currency exchange are larger on average than for business transactions. Their predictive means are 12.3 and 4.70 minutes, respectively. Also the required time for exchange currency services is greater on average than for business transaction services. Their predictive means are 7.22 and 3.86 minutes, respectively. However, in this case the business banking service distribution seems quite heterogeneous, with a maximum value of 42.283 minutes, while the currency exchange service seems fairly homogeneous.

Table 2 gives the posterior distribution of the number of components, L, for the four data sets. Observe that our Bayesian density estimation method predicts, with some uncertainty, an exponential distribution for the foreign currency service time and for the business banking interarrival time with estimated rates approximately equal to 0.14 and 0.22, respectively. For the exchange currency interarrival distribution, the algorithm suggests a two component mixture . The first component is an exponential and the second an Erlang distribution both with rates close to 0.1. Finally, fitting the business banking service time distribution needs a fairly large number of components.

3 Bayesian prediction for the $GI/G/1$ queueing model

In this section we are mainly interested in the prediction of the transient characteristics and the busy period of the $GI/G/1$ queueing system where the interarrival and the service times are i.i.d. random variables, denoted respectively by X_A and X_S. It is known that if we approximate a given short or long-tailed interarrival or service time distribution by another distribution which is sufficiently close to the original, then performance measures such as the transient waiting time will also be approximately what they would be with the original

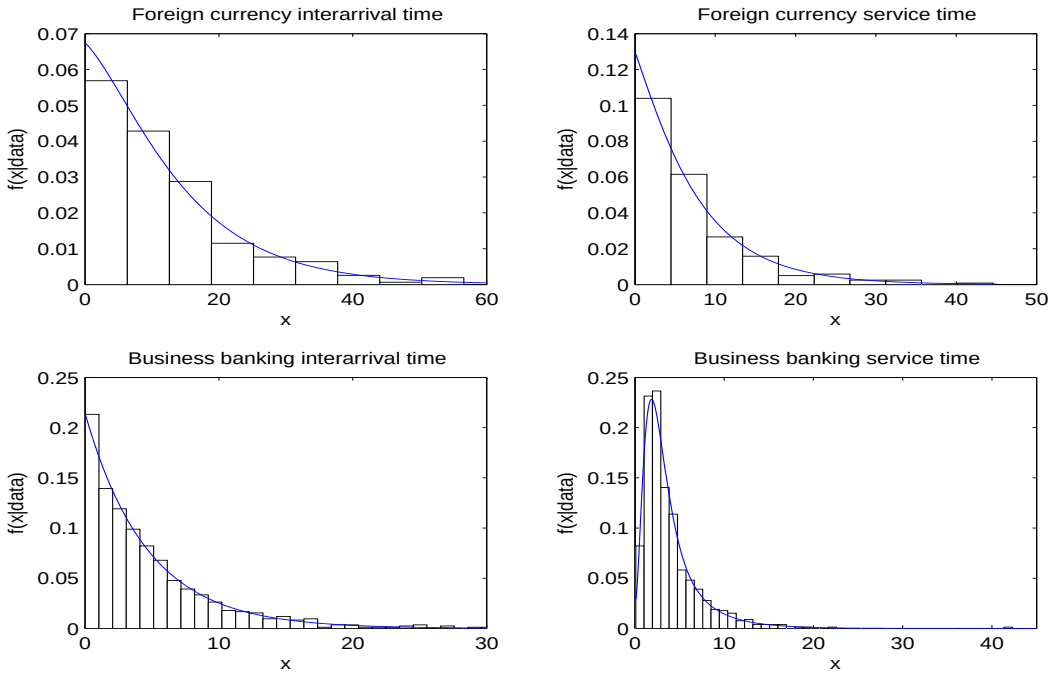

Figure 6: Histograms and predictive interarrival and service time densities for the four real data sets. The fits are quite satisfactory and the Coxian distribution can also approximate densities with non zero mode such as the business banking service time density.

interarrival or service time distribution, see e.g. Feldmann and Whitt (1998). Using these results, we will assume that X_A and X_S are distributed as Coxian distributions with parameters $\{L, \mathbf{P}, \boldsymbol{\lambda}\}$ and $\{M, \mathbf{Q}, \boldsymbol{\mu}\}$, respectively. In the arrival process, each customer must go through up to L exponential stages with rates $\lambda_1, ..., \lambda_L$ and exit probabilities $P_1, ..., P_L$ respectively, before accessing the waiting line or, eventually, the service. Also, each service time is the sum of up to M exponential stages with rates $\mu_1, ..., \mu_M$ and exit probabilities $Q_1, ..., Q_M$, respectively.

Assume now that we observe samples of n_a interarrival and n_s service times, $\mathbf{x}_a = \{x_1^a, ..., x_{n_a}^a\}$ and $\mathbf{x}_s = \{x_1^s, ..., x_{n_s}^s\}$, respectively. Then, assuming that we have generated MCMC samples of size J from the posterior distribution of both interarrival and service parameters, we can use this output to estimate various measures of interest for the queueing system.

An important measure of the average occupancy is the traffic intensity, ρ, that for this queueing model,

	$P(L \mid data)$	1	2	3	4	5	6	7	8
F. C.	Arrival	.2233	.6451	.1181	.0115	.0017	.0003	.0001	.0000
	Service	.5019	.3701	.1048	.0212	.0019	.0001	.0000	.0000
B. B.	Arrival	.6999	.2669	.0320	.0011	.0001	.0000	.0000	.0000
	Service	.0000	.0000	.0000	.3369	.4281	.1896	.0431	.0024

Table 2: Posterior probabilities for the mixture size, L, in foreign currency exchanges (F.C.) and business banking transactions (B.B.). The algorithm predicts an exponential distribution for the F.C. service time and for the B.B. interarrival time. A fairly large number of components is required for the B.B. service time.

conditional on the system parameters is given by,

$$\rho = \frac{E\left[X_S \mid M, \mathbf{Q}, \boldsymbol{\mu}\right]}{E\left[X_A \mid L, \mathbf{P}, \boldsymbol{\lambda}\right]} = \frac{\sum_{r=1}^{M} Q_r \sum_{k=1}^{r} \frac{1}{\mu_k}}{\sum_{r=1}^{L} P_r \sum_{k=1}^{r} \frac{1}{\lambda_k}}. \tag{17}$$

It is well known that the system is stable if $\rho < 1$, see e.g. Gross and Harris (1985). The posterior probability of having a stable queue can be estimated with,

$$P\left(\rho < 1 \mid \mathbf{x}_a, \mathbf{x}_s\right) \approx \frac{R}{J}, \tag{18}$$

where R is the number of times that the value of the traffic intensity calculated from (17) for each element of the MCMC sample $(L^{(j)}, \mathbf{P}^{(j)}, \lambda^{(j)}, M^{(j)}, \mathbf{Q}^{(j)}, \mu^{(j)})$ for $j = 1, ..., J$ is less than unity. Note that the posterior mean of ρ is given by,

$$E\left[\rho \mid \mathbf{x}_a, \mathbf{x}_s\right] = E\left[\sum_{r=1}^{M} Q_r \sum_{k=1}^{r} \frac{1}{\mu_k} \mid \mathbf{x}_s\right] \times E\left[\left(\sum_{r=1}^{L} P_r \sum_{k=1}^{r} \frac{1}{\lambda_k}\right)^{-1} \mid \mathbf{x}_a\right]. \tag{19}$$

Given the prior structure used, this is finite as shown in the Appendix. Thus, the expectation can be estimated by the sample mean of the traffic intensities for each element of the MCMC sample. Analogously, we can estimate the posterior mean of ρ assuming stability, $E\left[\rho \mid \rho < 1, \mathbf{x}_a, \mathbf{x}_s\right]$.

3.1 Estimation of the transient behaviour and busy period distributions

Bertsimas and Nakazato (1992) obtain closed expressions for the Laplace transform of the busy period, the transient waiting time and the transient queue length when the system parameters, $\{L, \mathbf{P}, \boldsymbol{\lambda}\}$ and $\{M, \mathbf{Q}, \boldsymbol{\mu}\}$, are known. The estimation procedure described in this section is based on these results.

Bertsimas and Nakazato (1992) show that, conditional on $\rho < 1$, the Laplace transform of the cumulative distribution function of the length of a busy period, B, is given by,

$$\int_0^{\infty} e^{-st} \Pr\left(B \leq t\right) dt = \frac{1}{s} - \frac{1 - \sum_{r=1}^{M} Q_r \prod_{k=1}^{r} \left(\frac{\mu_k}{\mu_k + s}\right)}{s} \prod_{r=1}^{M} \frac{s + \mu_r}{s - y_r\left(s\right)} \tag{20}$$

where $y_r\left(s\right)$, for $r = 1, ..., M$, are the M roots of the following equation,

$$\left[\sum_{r=1}^{L} P_r \prod_{k=1}^{r} \left(\frac{\lambda_k}{\lambda_k + s - y_r\left(s\right)}\right)\right] \times \left[\sum_{t=1}^{M} Q_t \prod_{k=1}^{r} \left(\frac{\mu_k}{\mu_k + y_r\left(s\right)}\right)\right] = 1, \tag{21}$$

with negative real part, for any complex s with positive real part. Bertsimas and Nakazato (1992) demonstrate that given s, equation (21) has M roots of interest, $y_r\left(s\right)$, for $r = 1, ..., M$, having negative real part and a further L roots with non negative real part.

The roots of the equation (21) cannot be found analytically and Bertsimas and Nakazato (1992) suggest using Mathematica to compute them. This approach is not feasible when using reversible jump methods as the equation is distinct at every MCMC iteration and does not always even have the same number of roots. Therefore, we propose to extract the roots numerically as follows. Note that, given s, the roots $y_r(s)$ of equation (21) are also roots of the following polynomial,

$$\mathcal{P}\left(y_r\left(s\right)\right) = \left[\sum_{r=1}^{L}\sum_{t=1}^{M} P_r Q_t \left(\prod_{k=1}^{r} \lambda_k \prod_{k=1}^{t} \mu_k\right) \mathcal{Q}_{r,t}\left(y_r\left(s\right)\right)\right] - \mathcal{Q}_{0,0}\left(y_r\left(s\right)\right), \tag{22}$$

where $\mathcal{Q}_{r,t}\left(y_r\left(s\right)\right)$ is another polynomial given by,

$$\mathcal{Q}_{r,t}\left(y_r\left(s\right)\right) = \prod_{k=r+1}^{L} \left(\lambda_k + s - y_r\left(s\right)\right) \prod_{k=t+1}^{M} \left(\mu_k + y_r\left(s\right)\right),$$

whose roots are given by $\{(\lambda_{r+1}+s),...,(\lambda_L+s),-\mu_{t+1},...,-\mu_M\}$ and the coefficient of order $(L+M-r-t)$ is $(-1)^{L-r}$. Then, we can apply for example the Laguerre algorithm, see e.g. Ralston and Rabinowitz (1978), which is designed specifically to find the roots of a complex polynomial.

This numerical procedure can be combined with the MCMC algorithm in order to estimate the busy period distribution function as follows. Given a sample realization from the posterior distribution of the parameters, $(L^{(j)},\mathbf{P}^{(j)},\lambda^{(j)},M^{(j)},\mathbf{Q}^{(j)},\mu^{(j)})$ for $j=1,...,J$, the natural way of estimating the predictive distributions is using a Monte Carlo approximation,

$$\Pr\left(B\leq t\mid \mathbf{x}_a,\mathbf{x}_s,\rho<1\right)\approx\frac{1}{R}\sum_{j:\rho^{(j)}<1}\Pr\left(B\leq t\mid\boldsymbol{\theta}^{(j)}\right), \tag{23}$$

where $\Pr(B\leq t\mid\boldsymbol{\theta}^{(j)})$ is obtained by numerically inverting the Laplace transform given in (20) for each element of the MCMC sample, $\boldsymbol{\theta}^{(j)}$. This can be done using a fast numerical, Laplace transform inversion method such as that described in Hosono (1981) and extracting numerically the $M^{(j)}$ roots of equation (21) as describe above. The error of numerical inversion of the Hosono algorithm is less than $10^{-p+1}\left|f(t)\right|$ for any inverted function $f(t)$, given a previously chosen value p, see also Bertsimas and Nakazato (1992). Then, considering that $\Pr(B\leq t\mid\boldsymbol{\theta}^{(j)})$ is bounded by one, the error of numerical inversion for this probability is less than 10^{-p+1} for each t. In our examples, we have chosen $p=6$.

Analogously, we can estimate the distribution function of the transient waiting time and the transient queue length distributions functions using the two following results. Assuming that the system is initially empty and some other mild initial conditions, Bertsimas and Nakazato (1992) show that if $\rho<1$, the Laplace transform of the waiting time, $W(\tau)$, of a customer arriving at time τ is given by,

$$\int_0^\infty e^{-s\tau}\Pr\left(W(\tau)\leq w\right)d\tau=\frac{1}{s}+\sum_{r=1}^{M}\frac{(-1)^M e^{y_r(s)w}}{s}\prod_{k=1}^{M}\frac{\mu_k+y_r(s)}{\mu_k}\prod_{\substack{k=1\\k\neq r}}^{M}\frac{y_k(s)}{y_r(s)-y_k(s)}, \tag{24}$$

where $y_r(s)$, for $r=1,...,M$, are the M roots of equation (21) with negative real part.

And finally, assuming the same initial conditions, Bertsimas and Nakazato (1992) give a expression for the Laplace transform of the number of customers, $N(\tau)$, in the system at time τ,

$$\int_0^\infty e^{-s\tau}\Pr\left(N(\tau)=n\right)d\tau=\begin{cases}\sum_{i=1}^{L}\pi^*_{0,i}(s), & \text{if } n=0,\\ \sum_{i=1}^{L}\sum_{j=1}^{M}\pi^*_{n,i,j}(s), & \text{if } n\geq 1,\end{cases} \tag{25}$$

where $\pi^*_{0,i}(s)$ is the Laplace transform of the probability that at time τ the number of customers in the system is 0 and the arrival stage currently occupied by the arriving customer is i, and $\pi^*_{n,i,j}(s)$ is the Laplace transform of the probability that at time τ the number of customers in the system is n, the arrival stage currently occupied by the arriving customer is i, and the service who is being served is j. The expressions for $\pi^*_{0,i}(s)$ and $\pi^*_{n,i,j}(s)$ are more complicated and are not reported here to save space, although they can be numerically evaluated without difficulties. These depend again on the roots of equation (21).

It can be shown that the predictive moments of the system size and waiting time in equilibrium, i.e. when $\tau\to\infty$, and the moments of the busy period distribution do not exist given any prior formulation that places positive mass on $\rho=1$, see e.g. Wiper (1998). Thus, the moments of the transient distributions will go to infinity as τ goes to infinity. However, as alternatives, we can calculate the median and quantiles of these distributions.

3.2 Results for simulated and real queues

In this section, we illustrate the behaviour of the proposed method with several simulated and real queues based on the data sets analysed in Section 2. Firstly we consider two simulated queueing systems and then real systems in the Israeli bank.

Simulated data. We use interarrival and service data simulated from the following two queueing systems with approximately the same traffic intensity $\rho = 0.289$.

- An $M/M/1$ system where the service times are the exponential data simulated in Section 2 and 300 arrival times were simulated from an exponential distribution with mean 3.456.
- A Coxian/Pareto/1 system where both interarrival and service time data are as in Section 2.

Table 3 shows the probabilities of equilibrium in the system and the posterior means of ρ for both systems. Also shown are the MLE estimates of ρ. As we might expect, the estimates are very close to the true value.

	$P(\rho < 1 \mid data)$	$E[\rho \mid data]$	$E[\rho \mid \rho < 1, data]$	$\hat{\rho}_{MLE}$
$M/M/1$	.99998	.29009	.29008	.28935
$Coxian/Pareto/1$	.99993	.30740	.30736	.28116

Table 3: Posterior probabilities that the system is stable and posterior mean values and MLE estimates for the traffic intensity for the two simulated systems. The equilibrium probability is extremely large in both cases and the estimations of ρ are very close to the true value.

Figure 7 illustrates in dotted lines the estimated transient distributions of queue length and waiting time as a function of time, τ, for both systems. Note that the predictive transient distributions clearly converge as τ goes to infinity. Observe that in the case of the Coxian/Pareto/1 system, the convergence of the transient waiting time distributions to the stationary distribution is much slower than for the $M/M/1$ case. On the contrary the speed of convergence of the transient queue length distribution is similar in both examples. These differences are mostly caused by the long-tailed behaviour of the Pareto service time distribution. It is well known that waiting time distributions are very influenced by the shape of the service time densities, while queue length distributions are mainly dependent only on their first moment, see e.g. Lillo (2000). Note that as τ goes to infinity, the estimated probability that the queue is empty approaches one minus the posterior mean traffic intensity which is coherent with the known result, $P(N = 0) = 1 - \rho$, where N denotes the equilibrium queue length in a $GI/G/1$ model, see e.g. Gross and Harris (1985). Also, the waiting time probability, $P(W(\tau)\,0)$, approaches the estimated traffic intensity in the $M/M/1$ system but not in the Coxian/Pareto/1 system. Note that the result $P(W = 0) = 1 - \rho$ is known to be true for systems with Poisson arrivals.

Figure 8 shows the estimated distribution functions of the busy period. For the $M/M/1$ system, the true system distribution is also given. This coincides very closely with the predictive distribution. Note that the tail of the busy period distribution is shorter for the Markovian queue.

In order to compare our analysis with a simpler approach, we also implemented a naive, classical analysis assuming an $M/M/1$ system for the data generated from the Coxian/Pareto/1 system. Thus, we estimated both the transient queue length and waiting time distributions and the busy period distribution using the MLE estimates of the interarrival and service time distributions. The results are quite different from those derived previously using our approach. This is illustrated in Figure 9 where the estimated waiting time distributions using the two procedures are compared. These differences are more marked for the waiting time and busy period distributions than for the queue length distribution. This should be expected because, as commented earlier, the latter is strongly dependent only on the first moments of the interarrival and service time distributions.

Real data. We now analyze the foreign currency exchange and business banking data discussed in Section 2. As the bank has a single teller for each kind of service, we have two single server, FIFO, queueing systems. For both systems, the estimated posterior probability that the system is stable is extremely high and the posterior mean values for ρ are close to the MLE estimators, as given in Table 4. The traffic intensity is higher for the business banking system implying that the server in this system has more busy time than the server in the currency exchange system.

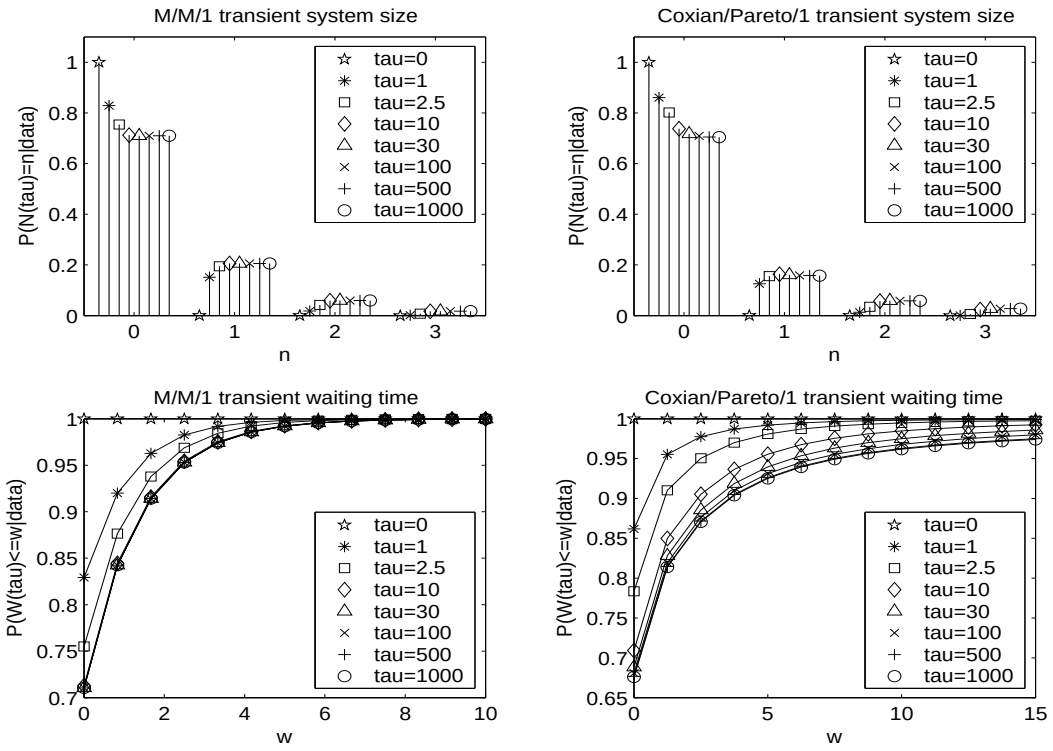

Figure 7: Transient distributions for queue length (top) and waiting time (bottom) for the $M/M/1$ and Coxian/Pareto/1 systems. These approximate the stationary distribution as τ increases. The speed of convergence of the transient waiting time is slower for the Coxian/Pareto/1 than for the $M/M/1$ system.

	$P(\rho < 1 \mid data)$	$E[\rho \mid data]$	$E[\rho \mid \rho < 1, data]$	$\hat{\rho}_{MLE}$
Foreign Currency	.99957	.59328	.59132	.58691
Business Banking	.99954	.82011	.81968	.82149

Table 4: Posterior probabilitities that the system is stable and posterior mean values for ρ in the two real bank systems. The equilibrium probability is extremely large in both cases. The foreign currency server is busy about 60% of the time while the business banking server is busy about 82% of the time.

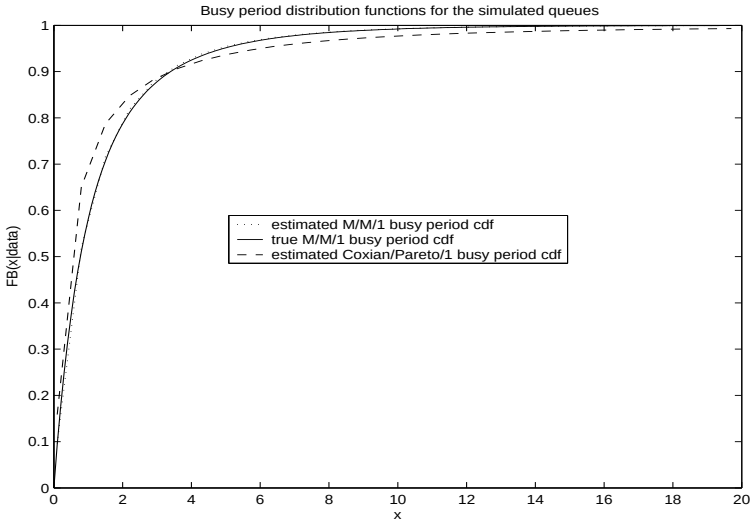

Figure 8: Estimated busy period distribution functions for the $M/M/1$ and Coxian/Pareto/1 systems. The true distribution for the $M/M/1$ system is also plotted and is very close to the estimated distribution. The tail of the busy period distribution is shorter for the $M/M/1$ queue.

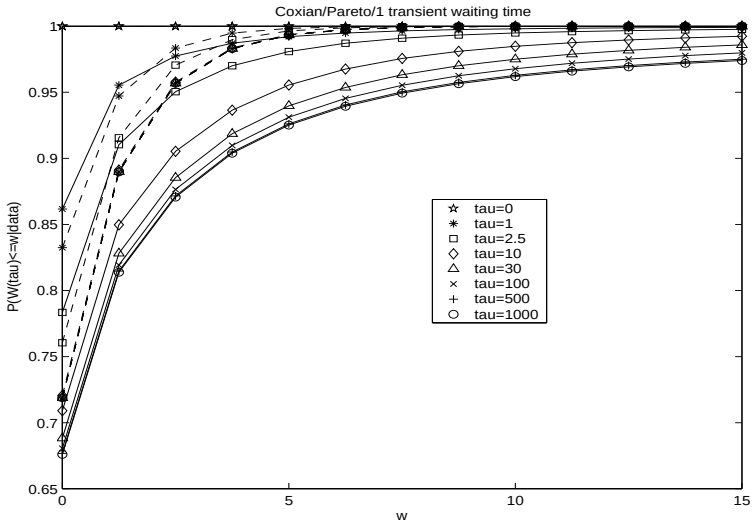

Figure 9: Estimated transient waiting time distribution using the proposed Bayesian $GI/G/1$ model (solid lines) and a simple MLE estimation based on a $M/M/1$ system (dotted lines) for the data generated from the Coxian/Pareto/1 system. The results are rather different using both approaches.

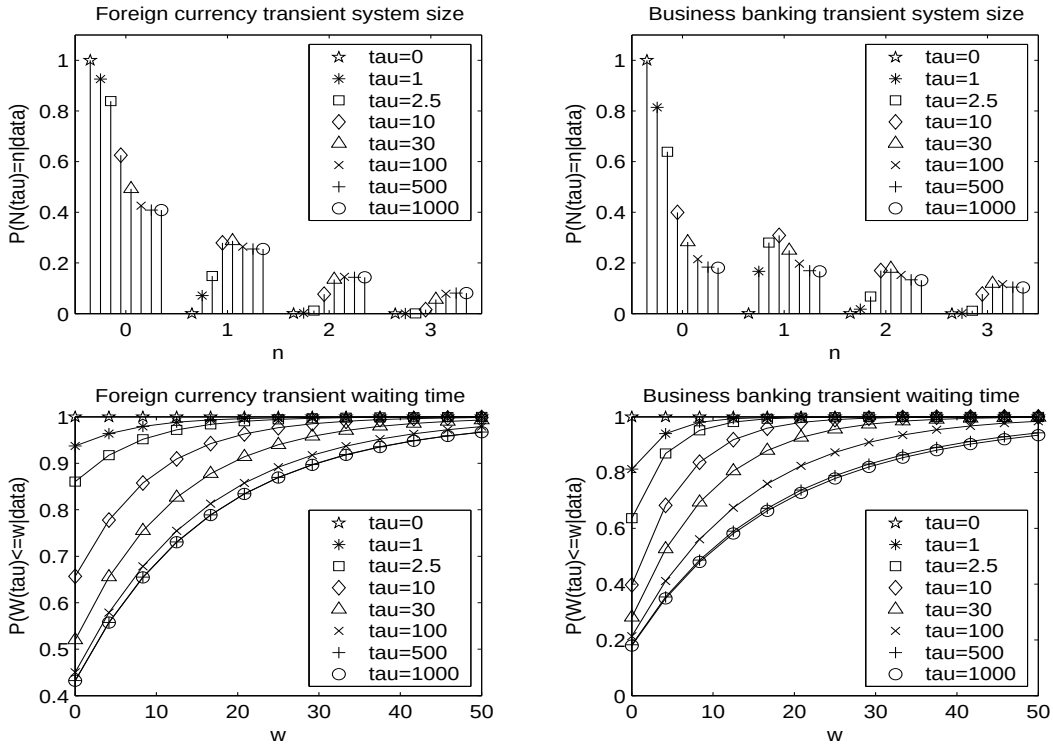

Figure 10: Predictive transient queue length (top) and waiting time (down) distributions for the bank systems. These approximate the stationary distributions as τ increases but the speed of convergence is slow. For the business banking system, there are still differences between the waiting time distributions after 500 and 1000 minutes.

Figure 10 illustrates the estimated distributions of the transient queue length and waiting time for both systems as a function of the time, τ. The assumption that the systems are empty at time $\tau = 0$ is true in this case. We observe that the speed of convergence is slow in both cases. After $\tau = 100$ minutes, the transient distributions have still not converged to the steady-state. In particular, for the business banking system, we can still appreciate differences between the waiting time distributions after 500 and 1000 minutes. Observe that using our approach it is possible to estimate the desired waiting time distribution for a client arriving at any given instant of time, τ. For example, the estimated probability that a customer who arrives at 9am asking for business banking services has to wait more than 8 minutes is approximately 0.3.

Table 5 shows some quantiles of the estimated distributions of the length of the busy period. Observe that the tail of the distribution is heavier for the business bank transactions case.

	0.25	0.50	0.65	0.80	0.90	0.95	0.97
Foreign Currency	2.342	6.323	10.902	21.271	41.108	69.610	96.298
Business Banking	2.113	4.596	8.411	19.452	46.671	95.431	148.290

Table 5: Quantiles of the busy period distribution for the two bank systems. The tail of the distribution is heavier for the business bank transactions case.

The algorithm used to estimate the transient and busy period distributions was programmed in FOR-

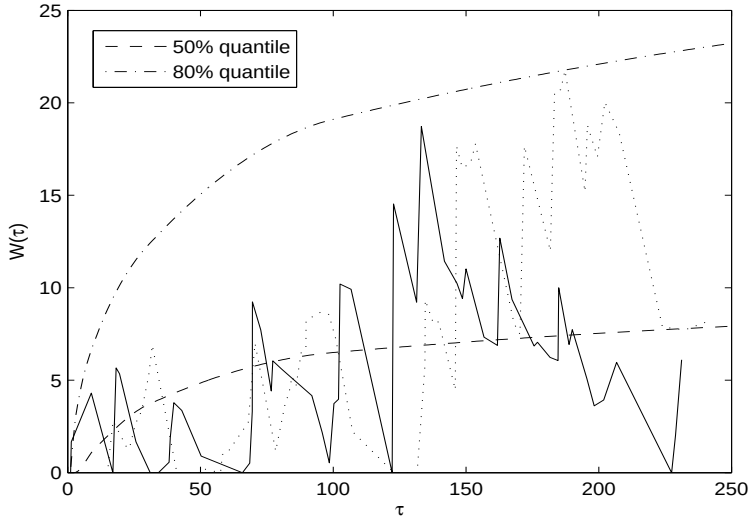

Figure 11: Trajectories of the observed waiting times, $W(\tau)$, during two test mornings and the 50% and 80% quantiles of the predictive, transient waiting time distributions obtained using training data from the first 10 days.

TRAN and the computational cost depends mainly on the number of time instants, τ, for which the transient distributions are estimated, on the number of points where the distributions are evaluated and on the number of phases of the interarrival and service time distributions in the MCMC sample. For the examples, the computational time required to estimate all the quantities associated with each queueing system was less than one and a half hours.

Finally, in order to analyze the performance of the procedure, we develop the following cross-validation mechanism. We consider the interarrival and service data corresponding to the first 10 days as a training data set and estimate the transient waiting time distributions using the training data. Then, the trajectories of the true waiting times of the customers arriving during the last four days can be compared with these predictive distributions. Figure 11 shows two trajectories of the waiting times, $W(\tau)$, observed during the mornings of two of the last four days as a function of the arrival instants, τ. These are compared with the 50% and 80% quantiles of the predictive transient waiting time distributions obtained using the training data. The observed waiting time trajectories are consistent with the quantiles of the predictive distributions. The trajectories for the other periods during the last four days show a similar pattern, which suggests that the predictions from this model are consistent with the data.

4 Conclusions and Extensions

In this article, we have demonstrated that we can carry out Bayesian inference for a $GI/G/1$ queueing system. Firstly, we have shown that the Coxian distribution can be used as a semiparametric model for density approximation. One particular advantage of our approach is that, due to the use of an improper prior, we are able to approximate both long and short tailed interarrival and service time distributions.

We have also shown that it is possible to carry out Bayesian prediction for the transient and busy period behaviour of a general queueing system. In particular, we have illustrated how a root finding algorithm and Laplace transform inversion can be combined within a MCMC algorithm in order to make inference about these characteristics. This is important in real life situations such as banking where the bank is only open for a finite time and thus, equilibrium may not be reached in practice.

A number of extensions are possible. Firstly, we should note that to capture multimodal distributions, a Coxian distribution with a large number of phases and hence a large number of parameters is needed. This can make the semiparametric approximation somewhat inefficient. In such cases alternative approximations which better capture multimodality could be considered. A difficulty is that in the queueing context, results concerning the transient and busy period behaviour are hard to obtain with such systems.

As pointed out by a referee, in order to attempt to simplify the inferential procedure, we might use model selection criteria, such as the Akaike Information Criterion, Bayesian Information Criterion or Bayes factors, for choosing a single queueing model from the Coxian class. An advantage of this approach is that once we have chosen the most probable model for the interarrival and service times, the computation of the performance measures of the queueing system is simplified.

In this article we have only considered a single server queueing system. From the point of view of inference for the interarrival and service processes, considering a multi server system presents no extra problems. However there are more difficulties in estimating the characteristics of the queueing system such as waiting time as the queueing theory for the Coxian/Coxian/c system is less well developed. Some results are available however, see e.g. Bertsimas (1990). Also, we should note that the Coxian distribution is phase type, see e.g. Neuts (1981) and, for example results for the waiting time in steady state of a system with phase type arrivals and services are known, see e.g. Asmussen et al. (2001).

Another important objective in the queueing context is optimal control of the system, e.g. when to open and close the system or optimizing the number of servers. The methods of Bayesian decision theory can be combined with Bayesian inference to do this. For one example of the use of Bayesian methods to control the capacity of a queueing system, see e.g. Ausín et al. (2003).

Finally, a further extension is to apply the Coxian model and prior structure in areas such as insurance risk where long tailed samples are common. In particular, in the context of estimating ruin probabilities, many results from queueing theory can be applied, see e.g. Prabhu (1998). Work is currently underway in this area.

Acknowledgment

We would like to thank three anonymous referees and editor for their valuable comments and suggestions. We also wish to thank Professor Avi Mandelbaum for allowing us to use the real data from his web page. We acknowledge the financial support provided by the projects SEJ2004-03303 and 06/HSE/0181/2004. The first author also acknowledges the financial support provided by Xunta de Galicia under the Isidro Parga Pondal Program.

Appendix: Proof that the posterior of ρ is proper with finite mean

Here, we show that the posterior is indeed proper. Assume that we observe a sample $x_1, ..., x_n$ from a Coxian random variable X with parameters $\boldsymbol{\theta} = (L, \mathbf{P}, \lambda_1, \boldsymbol{\upsilon})$ and that we use the prior distribution defined in Section 2. Then, we need to prove that the following integral is finite,

$$\int f(\boldsymbol{\theta}) \prod_{i=1}^{n} \left(\sum_{r=1}^{L} P_r f_r(x_i \mid \lambda_1, \boldsymbol{\upsilon}) \right) d\boldsymbol{\theta}. \tag{26}$$

Let us denote, for $r = 1, ..., L$,

$$C_{j,r} = \prod_{i \neq j}^{r} \frac{\lambda_i}{\lambda_i - \lambda_j} = \prod_{i \neq j}^{r} \frac{\prod_{k=2}^{i} \upsilon_k}{\prod_{k=2}^{i} \upsilon_k - \prod_{k=2}^{j} \upsilon_k}, \qquad \text{for } j = 1, ..., r.$$

Observe that $\sum_{j=1}^{r} C_{j,r} = 1$, see e.g. McGill et al. (1965). Then, for $n = 1$, the integral (26) is given by,

$$\int \frac{f(L) f(\mathbf{P}|L) f(\boldsymbol{\upsilon}|L)}{\lambda_1} \sum_{r=1}^{L} P_r \sum_{j=1}^{r} C_{j,r} \lambda_1 \prod_{k=2}^{j} \upsilon_k \exp\left(-\lambda_1 x_1 \prod_{k=2}^{j} \upsilon_k\right) d\boldsymbol{\theta}$$
$$= \frac{1}{x} \int f(L) f(\mathbf{P}|L) f(\boldsymbol{\upsilon}|L) \sum_{r=1}^{L} P_r \sum_{j=1}^{r} C_{j,r} d\boldsymbol{\theta}_{-\lambda_1} = \frac{1}{x} < \infty,$$

where we have denoted $\boldsymbol{\theta}_{-\lambda_1} = (L, \mathbf{P}, \boldsymbol{\upsilon})$. Finally, note that it is sufficient to have proved that the integral (26) is finite for $n = 1$, because now, we can define $f(\boldsymbol{\theta} \mid x_1)$ as a new proper *prior* and consider the likelihood based on $\{x_2, ..., x_n\}$, which is regular and proper, in which case the posterior is known to be proper. Then, the integral (26) is finite for $n \geq 1$.

Now, we will prove that the posterior mean of ρ given in (19) is finite. Firstly, observe that the generalized Erlang distribution, whose density is given in (27), is bounded by the first rate,

$$f_r(x \mid \lambda_1, \boldsymbol{\upsilon}) \leq \lambda_1, \tag{27}$$

This can be proved by induction. For $r = 1$, we have that $f_1(x \mid \lambda_1) = \lambda_1 e^{-\lambda_1 x} \leq \lambda_1$. Now, for $i = 2, ..., r-1$ assume that,

$$f_i(x \mid \lambda_1, \boldsymbol{\upsilon}) = f_i(x \mid \lambda_1, ..., \lambda_i) \leq \lambda_1,$$

where $\lambda_i = \lambda_1 \upsilon_2 ..., \upsilon_r$. Then, as f_r is the density of the sum of r exponentials, it can be expressed as the convolution of the r'th exponential density and f_{r-1}, that is,

$$f_r(x \mid \lambda_1, ..., \lambda_r) = \int_0^x \lambda_r \exp(-\lambda_r x) f_{r-1}(x - u \mid \lambda_1, ..., \lambda_{r-1}) du$$
$$\leq \lambda_1 [1 - \exp\{-\lambda_r x\}] \leq \lambda_1.$$

Using this property, we now prove that the first factor of (19), which is the predictive mean of the Coxian distribution, is finite. It is clear that if we do not observe at least two data the predictive mean does not exit. Then, assuming that $\mathbf{x}_s = \{x_1^s, x_2^s\}$ and using (27),

$$E\left[\sum_{r=1}^{M} Q_r \sum_{j=1}^{r} \frac{1}{\mu_j} | \mathbf{x}_s\right] \propto \int \sum_{r=1}^{M} Q_r \sum_{j=1}^{r} \frac{f(\boldsymbol{\theta})}{\mu_1 \prod_{k=2}^{j} \upsilon_k} \prod_{i=1}^{2} \left(\sum_{r=1}^{M} Q_r f_r(x_i \mid \mu_1, \boldsymbol{\upsilon})\right) d\boldsymbol{\theta}$$
$$\leq \int \left(\sum_{r=1}^{M} Q_r \sum_{j=1}^{r} \frac{f(\boldsymbol{\theta})}{\prod_{k=2}^{j} \upsilon_k}\right) \left(\sum_{r=1}^{M} Q_r f_r(x_1 \mid \mu_1, \boldsymbol{\upsilon})\right) d\boldsymbol{\theta}$$
$$= \int \left(\sum_{r=1}^{M} Q_r \sum_{j=1}^{r} \frac{f(M) f(\mathbf{Q}|M) f(\boldsymbol{\upsilon}|M)}{\prod_{k=2}^{j} \upsilon_k}\right) \left(\sum_{r=1}^{M} Q_r \sum_{j=1}^{r} C_{j,r} \prod_{k=2}^{j} \upsilon_k \exp\left(-\mu_1 x_1 \prod_{k=2}^{j} \upsilon_k\right)\right) d\boldsymbol{\theta}$$
$$= \frac{1}{x_1} \int \sum_{r=1}^{M} Q_r \sum_{j=1}^{r} \frac{f(M) f(\mathbf{Q}|M) f(\boldsymbol{\upsilon}|M)}{\prod_{k=2}^{j} \upsilon_k} d\boldsymbol{\theta}_{-\mu_1}$$
$$= \frac{1}{x_1} \int \sum_{r=1}^{M} Q_r \sum_{j=1}^{r} f(M) f(\mathbf{Q}|M) d\mathbf{Q} \int \frac{f(\boldsymbol{\upsilon}|M)}{\prod_{k=2}^{j} \upsilon_k} d\boldsymbol{\upsilon}$$
$$\leq \infty \quad \text{if } a > 1.$$

Finally, as this integral is finite for $n = 2$, we can define a new proper *prior* for $\boldsymbol{\theta}$,

$$g(\boldsymbol{\theta}) \propto \sum_{r=1}^{M} Q_r \sum_{j=1}^{r} \frac{f(\boldsymbol{\theta})}{\mu_1 \prod_{k=2}^{r} \upsilon_k} \prod_{i=1}^{2} \left(\sum_{r=1}^{M} Q_r f_r(x_i \mid \mu_1, \boldsymbol{\upsilon})\right),$$

which is a proper density. With this prior and the likelihood based on $\{x_3, ..., x_n\}$, the posterior is proper and then, the predictive mean is finite for $n \geq 2$.

Finally, let us prove that the second factor of the posterior mean of ρ given in (19) is finite. Firstly, we show that ,

$$\sum_{k=1}^{r} \frac{1}{\prod_{k=2}^{r} \upsilon_k} = \sum_{j=1}^{r} \frac{C_{j,r}}{\prod_{k=2}^{r} \upsilon_k}.$$

This can be proved observing that the expectation of a generalized Erlang, f_r, which is a sum of r exponential random variables with rates $\lambda_1, ..., \lambda_r$, is given by the sum of the inverse of these rates,

$$\sum_{j=1}^{r} \frac{1}{\lambda_j} = \sum_{k=1}^{r} \frac{1}{\lambda_1 \prod_{k=2}^{r} \upsilon_k} = \int_0^{\infty} x \sum_{j=1}^{r} C_{j,r} \lambda_1 \prod_{k=2}^{r} \upsilon_k \exp\left(-\lambda_1 x_1 \prod_{k=2}^{r} \upsilon_k\right) dx = \sum_{j=1}^{r} \frac{C_{j,r}}{\lambda_1 \prod_{k=2}^{r} \upsilon_k}.$$

We now assume that $n_a = 1$,

$$\begin{aligned}
E\left[\left(\sum_{r=1}^{L} P_r \sum_{j=1}^{r} \frac{1}{\lambda_j}\right)^{-1} |\mathbf{x}_a\right] &\propto \int \frac{f(\boldsymbol{\theta})}{\sum_{r=1}^{L} P_r \sum_{j=1}^{r} \left(\lambda_1 \prod_{k=2}^{j} \upsilon_k\right)^{-1}} \left(\sum_{r=1}^{L} P_r f_r(x_1 \mid \lambda_1, \boldsymbol{\upsilon})\right) d\boldsymbol{\theta} \\
&= \int \frac{f(L) f(\mathbf{P}|L) f(\boldsymbol{\upsilon}|L)}{\sum_{r=1}^{L} P_r \sum_{j=1}^{r} \left(\prod_{k=2}^{j} \upsilon_k\right)^{-1}} \left(\sum_{r=1}^{L} P_r f_r(x_1 \mid \lambda_1, \boldsymbol{\upsilon})\right) d\boldsymbol{\theta} \\
&= \int \frac{f(L) f(\mathbf{P}|L) f(\boldsymbol{\upsilon}|L)}{\sum_{r=1}^{L} P_r \sum_{j=1}^{r} \left(\prod_{k=2}^{j} \upsilon_k\right)^{-1}} \sum_{r=1}^{L} P_r \sum_{j=1}^{r} C_{j,r} \lambda_1 \prod_{k=2}^{j} \upsilon_k \exp\left(-\lambda_1 x_1 \prod_{k=2}^{j} \upsilon_k\right) d\boldsymbol{\theta} \\
&= \frac{1}{x_1^2} \int \frac{f(L) f(\mathbf{P}|L) f(\boldsymbol{\upsilon}|L)}{\sum_{r=1}^{L} P_r \sum_{j=1}^{r} \left(\prod_{k=2}^{j} \upsilon_k\right)^{-1}} \sum_{r=1}^{L} P_r \sum_{j=1}^{r} \frac{C_{j,r}}{\prod_{k=2}^{j} \upsilon_k} d\boldsymbol{\theta} \\
&= \frac{1}{x_1^2} \int \frac{f(L) f(\mathbf{P}|L) f(\boldsymbol{\upsilon}|L)}{\sum_{r=1}^{L} P_r \sum_{j=1}^{r} \left(\prod_{k=2}^{j} \upsilon_k\right)^{-1}} \sum_{r=1}^{L} P_r \sum_{j=1}^{r} \left(\prod_{k=2}^{j} \upsilon_k\right)^{-1} d\boldsymbol{\theta}_{-\lambda_1} \\
&= \frac{1}{x_1^2} \int f(L) f(\mathbf{P}|L) f(\boldsymbol{\upsilon}|L) d\boldsymbol{\theta}_{-\lambda_1} = \frac{1}{x_1^2}
\end{aligned}$$

Now, using the same arguments as above, this integral is also finite for $n_a \geq 1$.

References

Abate, J., Choudhury, G. L., and Whitt, W. (1994). Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems.* **16**, 311-338.

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

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

Armero, C., and Conesa, D. 2000. Prediction in Markovian bulk arrival queues. *Queueing Systems.* **34**, 327-350.

Asmussen, S., and Moller, J.R., 2001. Calculation of the steady-state waiting time distribution in $GI/PH/c$ and $MAP/PH/c$ queues. *Queueing Systems.* **37**, 9-29.

Ausín, M.C., Lillo, R.E., Ruggeri, F., and Wiper, M.P. 2003. Bayesian modelling of hospital bed occupancy times using a mixed generalized Erlang distribution. In: J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West (Ed.),*Bayesian Statistics 7*, Oxford University Press, Oxford, 443-452.

Ausín, M.C., Wiper, M.P., and Lillo, R.E., 2004. Bayesian estimation for the M/G/1 queue using a phase type approximation, *Journal of Statistical Planning and Inference.* **118**, 83-101.

Bertsimas, D., 1990. An analytic approach to a general class of $G/G/c$ queueing systems. *Operations Research.* **38**, 139-155.

Bertsimas, D., and Nakazato, D. 1992. Transient and busy period analysis of the $G/G/1$ queue: The method of stages. *Queueing Systems.* **10**, 153-184.

Crovella, M. E., Taqqu, M.S., and Bestavros, A. 1998. Heavy-Tailed Probability Distributions in the World Wide Web. In: R.J. Adler, R.E. Feldman and M.S. Taqqu (Ed.), *A Practical Guide To Heavy Tails*, Chapman and Hall, New York, 3-26.

Cox, D. R. 1955. A use of complex probabilities in the theory of stochastic processes. *Proc. Cam. Phil. Soc.* **51**, 313319.

Cumani, A., 1982. On the canonical representation of homogenous Markov processes modelling failure-time distributions. *Microelectronics and reliability.* **22**, 583-602.

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

Feldmann, A., and Whitt, W. 1998. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation.* **31**, 245-279.

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

Gross, D., and Harris, C.M., 1985. *Fundamentals of queueing theory.* John Wiley & Sons, New York.

Gruet, M.A., Philippe, A., and Robert, C.P., 1999. MCMC control spreadsheets for exponential mixture estimation. *Journal of Computational and Graphical Statistics.* **8**, 298-317.

Hosono, T., 1981. Numerical inversion of Laplace transform and some applications to wave optics. *Radio Science.* **16**, 1015-1019.

Johnson, N.L., and Kotz, S., 1970. *Distributions in statistics. Continuous univariate distributions.* John Wiley & Sons, New York.

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

McGill, W. J., and Gibbon, J., 1965. The general-gamma distribution and reaction time. *Journal of Mathematical Psychology.* **2**, 1-18.

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

Paxson, V. and Floyd, S., 1995. Wide-area traffic: The failure of Poisson modeling. *IEEE Transactions in Networking.* **3**, 226-244.

Prabhu, N.U., 1998. *Stochastic Storage Processes*, Springer Verlag, New York.

Ralston, A., and Rabinowitz, P., 1978. *A first course in numerical analysis.* McGraw-Hill, New York.

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.

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

Riska, A. Diev, V., and Smirni, E., 2004. An EM-based technique for approximating long-tailed data sets with PH distributions. *Performance Evaluation,* **55**, 147-164.

Robert, C.P., and Mengersen, K.L., 1999. Reparameterisation Issues in Mixture Modelling and their bearing on MCMC algorithms. *Computational Statistics & Data Analysis,* **29**, 325-343.

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