Bayesian prediction in $M/M/1$ queues

C. Armero and M.J. Bayarri

Department of Statistics and Operations Research, University of Valencia, Doctor Moliner 50, 46100 Burjassot, Valencia, Spain

Received 7 August 1992; revised 19 January 1993

Simple queues with Poisson input and exponential service times are considered to illustrate how well-suited Bayesian methods are used to handle the common inferential aims that appear when dealing with queue problems. The emphasis will mainly be placed on prediction; in particular, we study the predictive distribution of usual measures of effectiveness in an $M/M/1$ queue system, such as the number of customers in the queue and in the system, the waiting time in the queue and in the system, the length of an idle period and the length of a busy period.

Keywords: Bayesian inference; conjugate families; measures of performance; restricted parameter space; steady-state distributions.

1. Introduction

Although the stochastic modeling of waiting lines has a long history and has been the subject of a considerable body of research, the statistical analysis of queueing systems has received comparatively little attention. A good review of the literature on the subject to date can be found in Bhat and Rao [14] (see also Basawa and Prakasa Rao [7]). Hence, it is not all that surprising that Bayesian methods are almost absent in the statistical literature concerning queues. The most relevant works are those of Muddapur [26] (the first one to our knowledge), Reynolds [27], Armero [3], McGrath et al. [21], McGrath and Singpurwalla [22], Armero [4,5] and Armero and Bayarri [6].

Many reasons have been given in favour of (and also against) Bayesian methodology and we shall not attempt to review them here. (A good introductory account of the Bayesian paradigm applied to queues can be found in the paper by McGrath et al. [21].) We shall not insist on the need to quantify and incorporate prior information into a statistical analysis, nor shall we make it necessary to be coherent in our main point for using Bayesian methods. We shall instead take a pragmatic point of view and show how Bayesian methodology can handle in a natural way some statistical issues of special relevance in the analysis of queueing systems,

namely those of prediction of observable quantities, and the incorporation of restrictions in the parameter space.

In this paper we deal with a very simple queueing system, namely an $M/M/1$ queue. Accordingly, we assume that there is a single server, that the customers arrive according to a Poisson process with mean λ and that service times are independent of the arrivals and follow an exponential distribution with mean $1/\mu$. We shall also assume that the queue is in steady-state or equilibrium. This is a strong requirement. There are, of course, many systems for which such an assumption is very natural (maybe because the system has been functioning for a long period of time, basically in equilibrium, or because the interest lies in the average of some aspects of the system over a long period of time), but even if this assumption is not entirely appropriate studying the steady-state performance of the queue is always worthwhile, if only as an exploratory tool.

For an $M/M/1$ system to be in equilibrium, the parameter $\rho = \lambda/\mu$ (the traffic intensity) has to be strictly less than one. Hence this restriction has to be explicitly incorporated into the analysis. Also, for a steady-state $M/M/1$ queue, the quantities of practical interest are usually not the parameters governing the queue (λ and μ), but the so-called *measures of performance* of the queue. There are many of these measures, sometimes also called measures of congestion, but the three main ones are the number of customers in the system (and in the queue), the waiting time in the queue (and in the system), and the length of busy periods (and of idle periods). Notice that all these quantities are observable quantities; inferences concerning observable quantities are usually referred to as problems of *prediction*. Thus, when facing inferences on a steady-state $M/M/1$ queue, restrictions in the parameter space as well as problems of prediction, have to be addressed.

Contrary to what happens with classical methods of inference, Bayesian methods do handle in a very natural way the restrictions in the parameter space. Also, they are specially well suited to prediction problems. (Recall that there is not a single, generally agreed upon, method of prediction in the classical approach to inference.)

The parametric Bayesian paradigm can be described in a succinct, informal way, as follows: let Z be the random variable (or random vector) to be observed, whose distribution is not completely known but depends on the unknown value θ of some parameter (or parameter vector) with parameter space Θ, so that for each value of θ, Z is distributed according to $p(z|\theta)$. Usually, but not always, Z is assumed to be a random sample, $X_1, \ldots, X_n$ so that $p(z|\theta) = \prod p(x_i|\theta)$. To simplify the notation, here and in the rest of the paper, p will denote a generic density (with respect to Lebesgue or counting measure) with no implications that it is the same density. (We shall not make any attempts to distinguish among the different densities that appear by using different symbols unless it becomes necessary.) From a Bayesian point of view θ is also considered a random variable whose distribution gets updated, via Bayes' theorem, as information is obtained. Before Z is observed,

the a priori information about θ is quantified by the so-called *prior distribution* $p(\theta)$; after $Z = z$ is obtained, the *posterior distribution* of θ is given by

$$p(\theta|z) = \frac{p(z|\theta)p(\theta)}{p(z)} \quad \text{for } \theta \in \Theta, \tag{1.1}$$

where for the observed z, $p(z) = \int p(z|\theta)p(\theta)\,d\theta$ is a constant, and $p(z|\theta)$ is the likelihood function $L(\theta)$ of θ. Hence $p(\theta|z)$ is most usually computed as

$$p(\theta|z) \propto L(\theta)p(\theta) \quad \text{for } \theta \in \Theta, \tag{1.2}$$

where the proportionality constant is the one that makes $p(\theta|z)$ integrate to one.

Restrictions in the possible values of θ are incorporated in a natural way as part of the prior information. Thus, if θ is restricted to lie in $\Theta_0 \subset \Theta$, say, then the prior information would be such that $\Pr(\Theta_0^c) = 0$, and the prior density $p(\theta)$ would integrate to one over Θ_0. This is just the usual way to handle any prior information. What makes restrictions on the values of θ a specially easy information to incorporate is the fact that the analysis with the prior incorporating the restriction is equivalent to the usual, unrestricted analysis in which the restriction is incorporated at the end by making $p(\theta|z)$ integrate to 1 over Θ_0. To see this, let $p(\theta)$ denote an unrestricted prior density over Θ. If we were to incorporate the restriction $\theta \in \Theta_0$ directly in the prior, then θ would be distributed over Θ_0 according to the density

$$p^*(\theta) = \frac{p(\theta)}{\Pr(\theta \in \Theta_0)} \quad \text{for } \theta \in \Theta_0, \tag{1.3}$$

and $p^*(\theta) = 0$ otherwise. Hence, the posterior density would be computed as

$$p^*(\theta|z) \propto L(\theta)p^*(\theta) \quad \text{for } \theta \in \Theta_0. \tag{1.4}$$

But since $\Pr(\theta \in \Theta_0)$ in (1.3) is a constant, (1.4) can be equivalently expressed as

$$p^*(\theta|z) \propto L(\theta)p(\theta) \quad \text{for } \theta \in \Theta_0, \tag{1.5}$$

and $p^*(\theta|z) = 0$ otherwise. It can be seen from (1.5) that this is equivalent to carrying out the analysis in the unrestricted problem, compute $p(\theta|z) \propto L(\theta)p(\theta)$ on $\theta \in \Theta$, and then restrict it to take positive values only on Θ_0, computing the normalizing constant accordingly.

Problems of prediction can also be dealt with in a trivial way. Suppose that we wish to make predictions (point predictors, performance of the predictor, predictive regions, etc.) about an observable Y with distribution given by $p(y|\theta)$. Usually Y is some simple function of the future observations, as the first observation, or the average, but it can also depend on the observations in more complicated ways, as when we wish to predict waiting times. All the predictive aims can be attained from the *posterior predictive distribution* of Y, as given by

$$p(y|z) = \int p(y|\theta)p(\theta|z)\,d\theta. \tag{1.6}$$

This formulation for prediction assumes, as it is most often the case, that Y is independent of Z given θ, but it can trivially be modified to accommodate the lack of independence. Notice that $p(y|z)$ does explicitly incorporate the uncertainty about θ; contrast this to the usual approach in which $p(y|\hat{\theta})$ is used instead, $\hat{\theta}$ being some estimate of θ.

The paper is organized in 6 sections of which this introduction constitutes the first one. Section 2 is devoted to the computation of the posterior distributions needed to derive the desired predictive distributions in the following sections. Section 3 derives the predictive distributions for the steady-state number of customers in the system and for the steady-state number of customers in the queue; probabilities of interest (such as the steady-state probability that the system is empty and/or the steady-state probability that the system is busy) can be directly computed from the former one. Section 4 deals with prediction about the steady-state waiting time of a customer in the queue and the steady-state waiting time of a customer in the system; the former allows prediction about the steady-state probability of not queueing at all. Section 5 derives the predictive distributions of the length of busy periods and the length of idle periods of the queue at steady-state. Section 6 is devoted to illustrate some of the results in a numerical example.

2. Inferences on the parameters of an $M/M/1$ queue

Suppose that we want to make inferences about the arrival rate λ, and service rate μ, of an $M/M/1$ queue system. If there are no restrictions in the observability of the system, we can use a number of different experimental designs. (For simplicity, we shall assume that we do not observe the initial system size.) Among the designs providing complete information about the queue, the most usual ones consist in observing arrival and service times over a continuous period of time $(0, T]$, where T can be either a fixed value determined in advance (Benes [10]; Cox [16]) or determined by a suitable stopping rule. For instance, the system can be observed until the busy time reaches some preassigned fixed value (as in Clarke [15]), until a fixed number of customers have departed from the system (as in Basawa and Prabhu [8,9]), until a fixed number of transitions (arrivals and departures) have been recorded (as in Moran [24,25]), and so on. For various types of designs providing only incomplete information see, for instance, Cox [16], Basawa and Prakasa Rao [7], and Keiding [20].

A very simple and easy experiment that provides complete information about the system consists in observing n_a interarrival times and n_s service completions (the observation of the arrival and service processes do not need to be simultaneous), for fixed n_a and n_s. (This experiment has also been used in Armero [3–5], Armero and Bayarri [6] and Thiruvaiyaru and Basawa [28].) Let X_i denote the service time of the ith customer, $i = 1, 2, \ldots, n_s$, and let Y_j denote the time elapsed between the arrivals of customers j and $j-1, j = 1, 2, \ldots, n_a$ (as a notational

device, assume the customer 0 is the first one entering the queue during the observation period). Then, according to the hypothesis of an $M/M/1$ queue, $Y_1, \ldots, Y_{n_a}$ are i.i.d. random variables having an exponential distribution with parameter λ, and $X_1, \ldots, X_{n_s}$ are i.i.d. random variables, independent of the Ys, having an exponential distribution with parameter μ. Here, and in the rest of the paper, we shall let $z = (y_1, \ldots, y_{n_a}, x_1, \ldots, x_{n_s})^t$ the vector of all the observations in the queue. Hence, when z is observed the likelihood function of λ, μ is given by

$$L(\lambda, \mu) = \lambda^{n_a} e^{-\lambda t_a} \mu^{n_s} e^{-\mu t_s}, \tag{2.1}$$

where $t_a = \sum y_j$ and $t_s = \sum x_i$ are the observed values of the sufficient statistics.

It should be noted that many different experimental designs for observing the queue system result in likelihoods that are proportional to (2.1) (see Armero [5], McGrath and Singpurwalla [22]). Hence, since the Bayesian paradigm is compatible with the *likelihood principle* (for a through investigation of this principle, see Berger and Wolpert [13]), it follows that all of the analyses, and posterior and predictive distributions given in this paper do apply also to these other experimental designs.

Bayesian analysis requires the specification of a prior distribution which quantifies the prior information about the unknowns λ and μ. Here, we shall use a member of the *natural conjugate family of prior distributions* (see, i.e. DeGroot [17]). Accordingly, we assume that λ and μ are a priori independent having distributions $\mathrm{Ga}(\alpha_0, \beta_0)$ and $\mathrm{Ga}(a_0, b_0)$, respectively. Hence, their joint prior density is given by

$$p(\lambda, \mu) = \mathrm{Ga}(\lambda|\alpha_0, \beta_0) \cdot \mathrm{Ga}(\mu|a_0, b_0), \tag{2.2}$$

where $\mathrm{Ga}(x|\alpha, \beta)$ denotes the density at x of a Gamma distribution with parameters α, β as given by

$$\mathrm{Ga}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0. \tag{2.3}$$

From (2.1) and (2.2) the joint posterior distribution for λ and μ is easily calculated to be

$$p(\lambda, \mu|z) \propto p(\lambda, \mu) L(\lambda, \mu) \propto \mathrm{Ga}(\lambda|\alpha, \beta) \cdot \mathrm{Ga}(\mu|a, b), \tag{2.4}$$

where $\alpha = \alpha_0 + n_a, \beta = \beta_0 + t_a, a = a_0 + n_s, b = b_0 + t_s$. Hence, λ and μ are also independent a posteriori with marginal distributions $\mathrm{Ga}(\lambda|\alpha, \beta)$ and $\mathrm{Ga}(\mu|a, b)$, respectively.

Sometimes investigators wish to avoid the quantification of a prior distribution such as (2.2). Most often this is due to the fact that prior information is little and it is thought not to be worth the time and effort spent in its quantification, but other reasons are also adduced (as the need for an "objective" inference, or a multi-user prior, etc.). In these cases, an approximate Bayesian analysis is still possible by using what is called *non-informative prior distributions*. Although the term is somehow misleading (since they are usually improper, and hence not really probability

distributions, as well as they cannot be non-informative about every unknown feature of the model at hand), they are functions of the parameters developed to play the role of the prior distributions in the derivation of the posteriors, and that have been obtained with the aim of representing approximately a very vague prior information about the parameters. An excellent discussion of non-informative priors can be found in Berger [11] and Berger and Bernardo [12].

In our problem, the (improper) non-informative

$$\pi(\lambda, \mu) \propto \lambda^{-1}\mu^{-1} \tag{2.5}$$

results in the posterior distribution

$$\pi(\lambda, \mu|z) = \mathrm{Ga}(\lambda|n_a, t_a) \cdot \mathrm{Ga}(\mu|n_s, t_s)\,. \tag{2.6}$$

It can immediately be seen from (2.6) and (2.4) that the posterior (2.6) is a limiting case of the conjugate posterior (2.4) when the hyperparameters of the prior distribution go to zero. Hence, statistical analyses with the non-informative prior can be deduced from the corresponding analyses with the conjugate posterior by simply taking α_0, β_0, a_0 and b_0 to be zero. It should also be noted that the usual Bayes estimators, the posterior means of λ and μ are, in this case, $E(\lambda|z) = n_a/t_a$, $E(\mu|z) = n_s/t_s$, the MLE's of λ and μ, respectively.

A parameter of special importance for $M/M/1$ queues is the traffic intensity $\rho = \lambda/\mu$. Marginal inferences about ρ will be based on the posterior $p(\rho|z)$, which can be easily deduced from (2.4). Indeed, if $\chi^2(\nu)$ denotes a chi-square distribution with ν degrees of freedom, then it follows from (2.4) that $2\beta\lambda \sim \chi^2(2\alpha)$ and $2b\mu \sim \chi^2(2a)$. Thus, given z, the posterior distribution of the ratio ρ/R,

$$(\rho/R) \sim F(2\alpha, 2a), \quad \rho > 0\,, \tag{2.7}$$

is an F distribution with 2α and $2a$ degrees of freedom, where

$$R = \frac{\alpha b}{\beta a} = \frac{E[\lambda|z]}{E[\mu|z]}\,. \tag{2.8}$$

Hence, an estimator of ρ would be (for $a > 1$)

$$E[\rho|z] = \frac{a}{(a-1)}R\,. \tag{2.9}$$

A non-informative analysis results in $\rho/R \sim F(2n_a, 2n_s)$ so that $E(\rho|z) = Rn_s/(n_s - 1)$ (assuming $n_s > 1$), where here $R = \hat{\lambda}/\hat{\mu}$ is the ratio of the MLE's $\hat{\lambda}, \hat{\mu}$.

In the same way as we are using the means of the posterior distributions as estimators, natural measures of the accuracy of the estimates are provided by the variance of the posterior distributions. (As a matter of fact, these elections can be justified on decision theory grounds.) Thus, when desired, they can easily be computed from (2.4) and (2.7).

The posterior distribution for ρ, (2.7), allows inferences about the stationarity

of the queue. Thus, for instance, a test for stationarity $H_0 : \rho<1$ versus $H_1 : \rho\geqslant 1$ can be carried out (Armero [5]), or the probability that the queue is stationary can be computed as

$$\Pr(\rho<1|z) = \int_0^1 p(\rho|z)\,\mathrm{d}\rho = \frac{\Gamma(a+\alpha)B^\alpha}{\Gamma(\alpha)\Gamma(a)}\int_0^1 \frac{\rho^{\alpha-1}}{(1+B\rho)^{a+\alpha}}\,\mathrm{d}\rho$$
$$= \frac{\Gamma(a+\alpha)B^\alpha}{\Gamma(\alpha)\Gamma(a)}F(a+\alpha,\alpha;\alpha+1;-B)\,, \tag{2.10}$$

where $B=\beta/b$, and $F(a,b;c;z)$ is the hypergeometric function with integral representation

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\,\mathrm{d}t,$$
$$c>b>0\,. \tag{2.11}$$

$F(a,b;c;z)$ is also called Gauss' hypergeometric function, and sometimes it is denoted by ${}_2F_1(a,b;c;z)$ (see Abramowitz and Stegun [1, chap. 15]).

In this paper we assume that the queue is in equilibrium (so that $\rho<1$). As mentioned in the previous section, this restriction on the values of ρ can easily be incorporated into the analysis. Indeed, in a stationary $M/M/1$ queue, the posterior distribution of the traffic intensity ρ is computed from (2.7) and (2.10) as

$$p(\rho|z,\rho<1) = \frac{p(\rho|z)}{\Pr(\rho<1|z)} = C\frac{\rho^{\alpha-1}}{(1+B\rho)^{a+\alpha}} \quad \text{for } \rho<1\,, \tag{2.12}$$

and $p(\rho|z,\rho<1)=0$ otherwise, where the proportionality constant C can be expressed as

$$C = \frac{\alpha}{F(a+\alpha,\alpha;1+\alpha;-B)}\,. \tag{2.13}$$

The posterior expected value of ρ when $\rho<1$ can also be expressed in terms of the hypergeometric function as follows:

$$E(\rho|z,\rho<1) = C\int_0^1 \frac{\rho^\alpha}{(1+B\rho)^{a+\alpha}}\,\mathrm{d}\rho = \frac{\alpha}{\alpha+1}\frac{F(a+\alpha,1+\alpha;2+\alpha;-B)}{F(a+\alpha,\alpha;1+\alpha;-B)}\,. \tag{2.14}$$

To compute the predictive distributions in the next sections, we shall find it convenient to reparameterize in terms of (ρ,μ) instead of working with (λ,μ). The joint posterior density of (ρ,μ) in a stationary queue is given, for any $\rho<1, \mu>0$, by

$$p(\rho,\mu|z,\rho<1) = p(\mu|z,\rho)p(\rho|z,\rho<1)\,, \tag{2.15}$$

where $p(\rho|z,\rho<1)$ is given by (2.12). The conditional $p(\mu|z,\rho)$ is $p(\mu,\rho|z)/p(\rho|z)$, where $p(\mu,\rho|z)$ can easily be computed from (2.4) and $p(\rho|z)$ is given in (2.7), resulting in

$$p(\mu|z,\rho) = \mathrm{Ga}(\mu|a+\alpha, b+\beta\rho)\,. \tag{2.16}$$

From now on, we shall always assume that the $M/M/1$ queue is in steady-state and hence that $\rho<1$. Thus, even though we shall not explicitly display $\rho<1$, it should always be understood that $p(\rho|z)$ refers to $p(\rho|z,\rho<1)$ as given by (2.12) and that $p(\mu|z,\rho)$ is only defined for values of $\rho<1$.

3. Number of customers in the system and in the queue

Let N denote the steady-state number of customers in the system. Then, for an $M/M/1$ queue in steady-state, the distribution of N is geometric with parameter $(1-\rho)$, that is

$$p(N=n|\rho) = (1-\rho)\rho^n, \quad n=0,1,2,\ldots \tag{3.1}$$

for any $\rho<1$ (See, for instance, Medhi [23, pp. 72–74]).

The predictive distribution of N is

$$\begin{aligned} p(N=n|z) &= \int_0^1 p(N=n|\rho)p(\rho|z)\,\mathrm{d}\rho \\ &= C\int_0^1 \frac{\rho^{n+\alpha-1}(1-\rho)}{(1+B\rho)^{a+\alpha}}\,\mathrm{d}\rho, \quad n=0,1,2,\ldots, \end{aligned} \tag{3.2}$$

where $B=\beta/b$ as in the previous section and C is given in (2.13). An expression in terms of the hypergeometric function can also be given. Indeed, since the integral appearing in (3.2) can be expressed as $F(a+\alpha,n+\alpha;2+n+\alpha;-B)/[(n+\alpha+1)(n+\alpha)]$, it follows that an alternative expression for (3.2) is

$$p(N=n|z) = \frac{\alpha}{(n+\alpha+1)(n+\alpha)}\,\frac{F(a+\alpha,n+\alpha;2+n+\alpha;-B)}{F(a+\alpha,\alpha;1+\alpha;-B)}$$
$$\text{for } n=0,1,2,\ldots. \tag{3.3}$$

A non-informative analysis results in

$$p(N=n|z) = \frac{n_a}{(n+n_a+1)(n+n_a)}\,\frac{F(n_s+n_a,n+n_a;n+2+n_a;-t_a/t_s)}{F(n_s+n_a,n_a;1+n_a;-t_a/t_s)}$$
$$\text{for } n=0,1,2,\ldots. \tag{3.4}$$

Probabilities of interest can be computed from (3.3) and (3.4). Thus, for instance, the steady-state probability that the system is empty is given by

$$p(N=0|z) = \frac{1}{(\alpha+1)}\,\frac{F(a+\alpha,\alpha;2+\alpha;-B)}{F(a+\alpha,\alpha;1+\alpha;-B)}\,. \tag{3.5}$$

Alternatively, notice that this probability can also be directly computed as

$$p(N=0|z) = E^{\rho|z}[p(N=0|\rho)] = 1-E(\rho|z)\,, \tag{3.6}$$

where $E(\rho|z)$ is given by (2.14). Thus, another probability of interest, $\Pr(N \geqslant 1|z)$, the steady-state probability that the system is busy, is nothing but $E(\rho|z)$ the posterior expected value of the traffic intensity ρ.

A striking property of this predictive distributions is that it has no moments. To see this, notice that

$$E[N|z] = E^{\rho|z}[E(N|\rho, z)] = E^{\rho|z}[\rho/(1-\rho)]\,. \tag{3.7}$$

But the integral

$$\int_0^1 \frac{\rho^{\alpha}(1-\rho)^k}{(1+B\rho)^{a+\alpha}}\,\mathrm{d}\rho \tag{3.8}$$

converges only if $k > -1$, so that $E(N|z)$ as given by (3.7) does not exist, and hence higher order moments do not exist either. Notice that this will be true no matter how many observations n_a, n_s we take and no matter how large the ratio t_a/t_s is. (For $n_a = n_s$, the ratio t_a/t_s is the ratio of the mean interarrival time to the mean service time.)

The posterior predictive distribution of the steady-state number of customers in the queue, N_q, can easily be expressed in terms of $p(N|z)$. Indeed, for any given value of $\rho(\rho < 1)$, the distribution of N_q is given by

$$\begin{aligned} &p(N_q = 0|\rho) = p(N \leqslant 1|\rho) = 1 - \rho^2\,, \\ &p(N_q = n|\rho) = p(N = n+1|\rho) = (1-\rho)\rho^{n+1}, \quad n = 1, 2, \ldots\,. \end{aligned} \tag{3.9}$$

Since the posterior predictive distribution $p(N_q = n|\rho)$ is computed as $E^{\rho|z}[p(N_q = n|\rho)]$, it follows that

$$p(N_q = n|z) = \begin{cases} p(N \leqslant 1|z) & \text{for } n = 0\,, \\ p(N = n+1|z) & \text{for } n = 1, 2, \ldots\,, \end{cases} \tag{3.10}$$

where $p(N = n|z)$ is given in (3.3) (or in 3.4) if a non-informative analysis is desired). It follows from (3.10) that this predictive distribution does not have any moments either. The lack of moments of the predictive distributions in this section (as well as most predictive distributions that will appear in the following sections), is due to the special form of the prior distribution for ρ (which is the same as the form of the posterior distribution, as given by (2.12)), whose right tail does not go to 0 as ρ goes to 1. This (maybe undesirable) property does not hold when priors with different tail behaviors are used (see Armero and Bayarri [6]).

4. Waiting times

Other measures of performance of the queue in steady-state that are often of interest are the waiting time in the system and the waiting time in the queue. We now proceed to derive their posterior predictive distributions.

First, let T denote the waiting time in the system. Then, for any given values of ρ ($\rho<1$) and μ ($\mu>0$) the distribution of T is exponential with parameter $\mu(1-\rho)$ (see for instance Medhi [23, pp. 74–78]). Hence its density, usually denoted by $w(t)$, is given by

$$w(t|\mu,\rho) = \mu(1-\rho)e^{-\mu(1-\rho)t}, \quad t>0. \tag{4.1}$$

The posterior predictive density is given by

$$w(t|z) = E^{\rho|z}E^{\mu|\rho,z}[w(t|\mu,\rho)]. \tag{4.2}$$

Let's compute first the conditional $w(t|\rho,z)$. Recall from (2.16) that $\mu|z,\rho \sim \text{Ga}(\mu|a+\alpha, b+\beta\rho)$. Then

$$\begin{aligned} w(t|\rho,z) &= E^{\mu|\rho,z}[w(t|\mu,\rho)] = \int_0^\infty \text{Ex}(t|\mu(1-\rho))\cdot \text{Ga}(\mu|a+\alpha, b+\rho\beta)\,d\mu \\ &= \frac{(a+\alpha)(1-\rho)(b+\rho\beta)^{a+\alpha}}{[t(1-\rho)+(b+\rho\beta)]^{a+\alpha+1}} \\ &= (a+\alpha)\left[\frac{b(1+\rho B)}{(1-\rho)}\right]^{a+\alpha}\left[\frac{b(1+\rho B)}{(1-\rho)}+t\right]^{-(a+\alpha+1)}. \end{aligned} \tag{4.3}$$

Denote by $\text{Gg}(x|\alpha,\beta,k)$ with $\alpha>0, \beta>0, k>0$ the density of a Gamma–Gamma distribution with parameters α,β,k (see for instance, Ferrándiz and Sendra [18]) also called an inverse beta distribution (see Aitchison and Dunsmore [2]) as given by

$$\text{Gg}(x|\alpha,\beta,k) = \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+k)}{\Gamma(k)}\frac{x^{k-1}}{(\beta+x)^{k+\alpha}}, \quad x>0. \tag{4.4}$$

Notice that, if $X \sim \text{Gg}(x|\alpha,\beta,k)$ then X/β is distributed according to a standard form of Pearson Type VI distribution (Johnson and Kotz, [19]).

It can be seen from (4.3) and (4.4), that for any value of ρ ($\rho<1$), the conditional predictive distribution of T is a $\text{Gg}(t|a+\alpha, b(1+\rho B)/(1-\rho), 1)$. Its expected value is

$$E[T|\rho,z] = \frac{b(1+\rho B)}{(a+\alpha-1)(1-\rho)}, \tag{4.5}$$

which is finite for every value of ρ ($\rho<1$).

We compute now the predictive distribution of the waiting time in the system. From (4.3) and (2.12),

$$w(t|z) = E^{\rho|z}[w(t|\rho,z)] = C(a+\alpha)b^{a+\alpha}\int_0^1 \frac{\rho^{\alpha-1}(1-\rho)}{[(t+b)+\rho(Bb-t)]^{a+\alpha+1}}\,d\rho, \tag{4.6}$$

and using (2.11) and (2.13), we finally get

$$w(t|z) = \frac{a+\alpha}{b(\alpha+1)}\left[\frac{b}{(t+b)}\right]^{a+\alpha+1}\frac{F(a+\alpha+1,\alpha;\alpha+2;-\frac{bB-t}{b+t})}{F(a+\alpha,\alpha;\alpha+1;-B)}. \tag{4.7}$$

Here again, as in section 3, the predictive distribution $w(t|z)$ has no expectation even though the conditional predictive distribution $w(t|\rho,z)$ does have finite expectation (4.5) for every value $\rho<1$. To see this, notice that

$$E[T|z] = E^{\rho|z}E[T|\rho,z] = \frac{bC}{(a+\alpha-1)}\int_0^1 \frac{\rho^{\alpha-1}(1-\rho)^{-1}}{(1+B\rho)^{a+\alpha-1}}\,\mathrm{d}\rho, \tag{4.8}$$

which does not converge. Hence, the posterior predictive distribution (4.7) has no moments, no matter how many observations we get.

A waiting time of special interest is the time spent in the queue, T_q, whose distribution for any values of ρ ($\rho<1$) and μ ($\mu>0$) is such that assesses probability $1-\rho$ of no queueing at all ($T_q=0$), and distributes the remaining probability ρ on the positive real line according to a density $w_q(t)$ proportional to an $\mathrm{Ex}(\mu(1-\rho))$ density (see, for instance, Medhi [23, pp. 74–78]). That is,

$$\begin{aligned}&\Pr(T_q=0|\rho,\mu) = 1-\rho,\\ &w_q(t|\mu,\rho) = \mu\rho(1-\rho)\mathrm{e}^{-\mu(1-\rho)t}, \quad t>0.\end{aligned} \tag{4.9}$$

The predictive probability of not having to wait in line is simply

$$\Pr(T_q=0|z) = 1-E(\rho|z), \tag{4.10}$$

where $E(\rho|z)$ is given in (2.14). Obviously, this probability is equal to the probability $\Pr(N=0|z)$ that the system is idle.

Notice also from (4.9) and (4.1) that, for $t>0$,

$$w_q(t|\mu,\rho) = \rho w(t|\mu,\rho), \tag{4.11}$$

so that the conditional (on ρ) predictive density of T_q on $T_q>0$ can easily be deduced from the one for T and (4.11). Indeed,

$$w_q(t|\rho,z) = \rho w(t|\rho,z) = \rho\,\mathrm{Gg}\left(t|a+\alpha,\frac{b(1+\rho B)}{1-\rho},1\right). \tag{4.12}$$

Finally, the (marginal) predictive posterior density of T_q on $T_q>0$ is

$$\begin{aligned}w_q(t|z) &= \int_0^1 w_q(t|\rho,z)p(\rho|z)\,\mathrm{d}\rho\\ &= \frac{(a+\alpha)\alpha}{b(\alpha+1)(\alpha+2)}\left[\frac{b}{(t+b)}\right]^{a+\alpha+1}\frac{F(a+\alpha+1,\alpha+1;\alpha+3;-\frac{bB-t}{b+t})}{F(a+\alpha,\alpha;\alpha+1;-B)}.\end{aligned} \tag{4.13}$$

Again, since $E[T_q|\rho,z] = \rho b(1+\rho B)/[(a+\alpha-1)(1-\rho)]$, it follows that the predictive distribution of T_q has no moments.

5. Idle and busy periods

Sometimes, inferences concerning busy and/or idle periods are desired. In this section we find the predictive distributions on which these inferences are based.

Let T_d denote the length of an idle period, and let $d(t|\lambda)$ denote its density. Since $T_d \sim \text{Ex}(\lambda)$ (see, for instance, Medhi [23, pp. 126]), it follows that, for any values of ρ ($\rho<1$) and μ ($\mu>0$),

$$d(t|\rho,\mu) = \rho\mu e^{-\rho\mu t}, \quad t\geqslant 0. \tag{5.1}$$

Following a line of reasoning similar to that in section 4, we compute first the conditional predictive density

$$\begin{aligned} d(t|\rho,\mu) = E^{\mu|\rho,z}[d(t|\mu,\rho)] &= \int_0^\infty \text{Ex}(t|\mu\rho)\cdot \text{Ga}(\mu|a+\alpha, b+\rho\beta)\,d\mu \\ &= (a+\alpha)\left[\frac{b(1+\rho B)}{\rho}\right]^{a+\alpha}\left[t+\frac{b(1+\rho B)}{\rho}\right]^{-(a+\alpha+1)}, \end{aligned} \tag{5.2}$$

so that $T_d|\rho,\mu \sim \text{Gg}(a+\alpha, b(1+\rho B)/\rho, 1)$, and hence its expected value is

$$E[T_d|\rho,z] = \frac{b(1+\rho B)}{(a+\alpha-1)\rho}. \tag{5.3}$$

The predictive distribution of T_d, the length of an idle period, is

$$d(t|z) = E^{\rho|z}[d(t|z,\rho)] = C(a+\alpha)b^{(a+\alpha)}\int_0^1 \frac{\rho^\alpha}{[b+\rho(t+B)]^{a+\alpha+1}}\,d\rho, \tag{5.4}$$

or, using again (2.11) and (2.13), we can express it in terms of the hypergeometric function as

$$d(t|z) = \frac{(a+\alpha)\alpha}{b(\alpha+1)}\,\frac{F(a+\alpha+1,\alpha+1;\alpha+2;-(B+t/b))}{F(a+\alpha,\alpha;\alpha+1,-B)}. \tag{5.5}$$

It is noteworthy that this predictive does have expectation. Indeed

$$\begin{aligned} E[T_d|z] = E^{\rho|z}[E(T_d|\rho,z)] &= \frac{Cb}{(a+\alpha-1)}\int_0^1 \frac{\rho^{\alpha-2}}{(1+\rho B)^{a+\alpha-1}}\,d\rho \\ &= \frac{\alpha b}{(\alpha-1)(a+\alpha-1)}\,\frac{F(a+\alpha-1,\alpha-1;\alpha;-B)}{F(a+\alpha,\alpha;\alpha+1;-B)}. \end{aligned} \tag{5.6}$$

As a matter of fact, it can be shown that $E[T_d^k|z]$ exists for $k = 1,2,\ldots,\alpha-2$.

Finally, let's derive the posterior predictive distribution of the length of a busy period, T_b. For any values of ρ ($\rho>1$) and μ ($\mu>0$), the conditional distribution of T_b is given by the density

$$b(t|\rho,\mu) = \frac{e^{-\mu(1+\rho)t}}{t\rho^{1/2}} I_1(2t\mu\rho^{1/2}), \tag{5.7}$$

where $I_1(y)$ is a modified Bessel function of first order (Abramowitz and Stegun [1, chap. 9]), given by

$$I_1(y) = \sum_{k=0}^{\infty} \frac{(y/2)^{(2k+1)}}{k!\,(k+1)!}\,. \tag{5.8}$$

We shall again compute the predictive $b(t|z)$ by first computing $b(t|\rho, z) = E^{\mu|\rho,z}[b(t|\mu, \rho)]$ and then computing $b(t|z) = E^{\rho|z}[b(t|\rho, z)]$. For any value of ρ ($\rho<1$),

$$b(t|\rho, z) = \int_0^{\infty} b(t|\mu, \rho) \cdot \mathrm{Ga}(\mu|a+\alpha, b+\rho\beta)\,\mathrm{d}\mu\,. \tag{5.9}$$

But the modified Bessel function $I_1(y)$ can be expressed in terms of a confluent hypergeometric function $M(a, b, y)$ as follows:

$$I_1(y) = \frac{y}{2}\mathrm{e}^{-y} M(3/2, 3, 2y)\,, \tag{5.10}$$

where Kummer's function $M(a, b, y)$ (sometimes denoted ${}_1F_1(a; b; y)$ and also $\Phi(a; b; y)$) admits, for $b>a$, the following integral representation (see Abramowitz and Stegun [1, chap. 13]):

$$M(a, b, y) = \frac{\Gamma(b)}{\Gamma(b-a)\Gamma(a)} \int_0^1 \mathrm{e}^{yt} t^{a-1}(1-t)^{b-a-1}\,\mathrm{d}t\,. \tag{5.11}$$

Substituting (5.11), (5.10) and (5.7) into (5.9) and using the hypergeometric function $F(a, b; c; z)$ in (2.13) to integrate (5.9) gives

$$b(t|\rho, z) = \frac{(a+\alpha)(b+\rho\beta)^{a+\alpha}}{[A(t,\rho)]^{a+\alpha+1}} F\left(a+\alpha+1, 3/2; 3; \frac{4t\rho^{1/2}}{A(t,\rho)}\right), \tag{5.12}$$

where $A(t, \rho) = [(b+\rho\beta) + t(1+\rho^{1/2})^2]$.

Finally, we numerically compute

$$b(t|z) = \int_0^1 b(t|\rho, z) p(\rho|z)\,\mathrm{d}\rho\,, \tag{5.13}$$

which cannot be given a simplified, closed expression.

To compute the expected value of T_b, notice that, for any given values of μ and ρ ($\rho<1$), $E[T_b|\mu, \rho] = E[T|\mu, \rho] = 1/[\mu(1-\rho)]$, so that the conditional (on ρ) posterior expected length of the busy period, $E[T_b|\rho, z]$ is also equal to $E[T|\rho, z]$ as given in (4.5). It follows that the predictive distribution $b(t|z)$ has no moments either.

6. A numerical example

In this section we illustrate some of the results obtained in previous sections

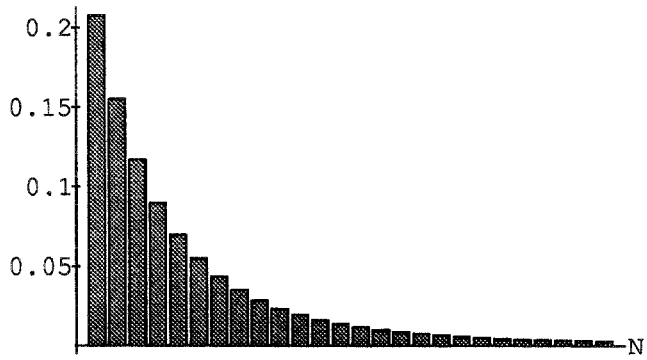

Fig. 1. Predictive distribution of N, number of customers in the system.

with a hypothetical queueing experiment in which $n_a = n_s = 100$, and the observed $\hat{\rho} = (n_a/t_a)/(n_s/t_s) = t_s/t_a$ equal 0.8. We use a non-informative prior.

Figure 1 shows the predictive distribution of N, the number of customers in the system. Probabilities of interest can easily be computed from this distribution; thus, the probability that the system is empty, which is also the probability of not having to wait in line, is $\Pr(N = 0|z) = 0.2076$ (and that the system is busy is 0.7924), and the probability of no customers in the queue is $\Pr(N = 0, 1|z) = 0.3625$. Some typical quantiles are given in table 1.

Table 1
Quantiles of $p(N|z)$.

Order	0.25	0.50	0.75	0.95
Quantiles	1	3	7	25

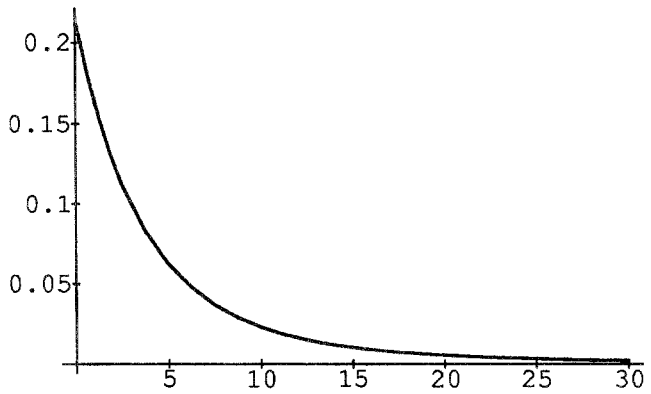

Fig. 2. Predictive distribution of standardized T, time spent in the system.

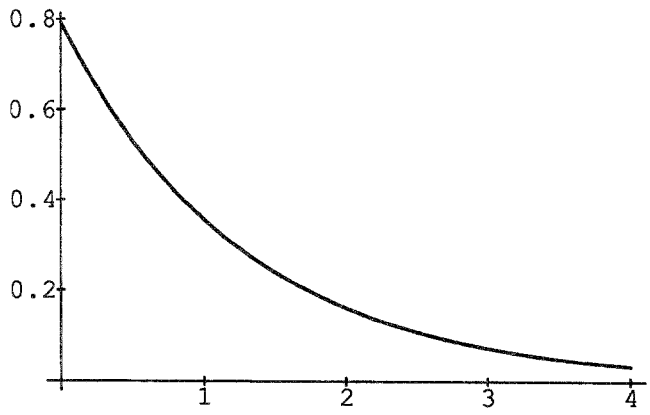

Fig. 3. Predictive distribution of standardized T_d, length of an idle period.

Figure 2 shows the predictive distribution of T, the time a customer spends in the system. The distribution that appears in not actually that of T, but the one of $U = T/t_s$, where t_s is the average time required to serve a customer in the conducted experiment. Hence, the units in the figure correspond to t_s each. We think this to be a better description of the way the queue behaves. (Beside, the distribution of U depends only on the samples sizes and $\hat{\rho}$.) Some quantiles of interest of the distribution of U are given in table 2.

We finally exemplify the only predictive distribution among the ones computed that has moments. Figure 3 shows the predictive distribution of the length of an idle period, T_d. Again, we have standardized the distribution in exactly the same way as the one for T; hence, the units correspond to t_s and the density shown is that of $U_d = T_d/t_s$. The mean can be computed to be $E[T_d|z] = 1.2725t_s$ and $\mathrm{Var}[T_d|z] = 1.6496t_s^2$. Table 3 shows some quantiles of interest.

Table 2
Quantiles of the standarized $p(T|z)$.

Order	0.25	0.50	0.75	0.95
Quantiles	$1.392t_s$	$3.583t_s$	$8.206t_s$	$29.10t_s$

Table 3
Quantiles of the standarized $p(T_d|z)$.

Order	0.25	0.50	0.75	0.95
Quantiles	$0.3632t_s$	$0.8767t_s$	$1.759t_s$	$3.8293t_s$

References

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (Dover Publ., New York, 1964).

[2] J. Aitchinson and I.R. Dunsmore, *Statistical Prediction Analysis* (Cambridge University Press, Cambridge, 1975).

[3] C. Armero, Bayesian Analysis of $M/M/1/\infty$/FIFO queues, in: *Bayesian Statistics 2*, eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith (North-Holland, Amsterdam, 1985) pp. 613–617.

[4] C. Armero, Análisis bayesiano de colas M/M (in Spanish), Ph.D. thesis, University of Valencia (1988).

[5] C. Armero, Bayesian inference in Markovian queues, Queueing Systems 15 (1994) 419–426.

[6] C. Armero and M.J. Bayarri, Prior assessments for prediction in queues, The Statistician (1994), in press.

[7] I.V. Basawa and B.L.S. Prakasa Rao, *Statistical Inference for Stochastic Processes* (Academic Press, New York, 1980).

[8] I.V. Basawa and N.U. Prabhu, Estimation in single server queues, Naval Res. Logist. Quart. 28 (1981) 475–487.

[9] I.V. Basawa and N.U. Prabhu, Large sample inference from single server queues, Queueing Systems 3 (1988) 289–304.

[10] V.E. Benes, A sufficient set of statistics for a simple telephone exchange model, Bell Syst. Tech. J. 36 (1957) 939–964.

[11] J.O. Berger, Objective Bayesian analysis: Development of reference noninformative priors, in: *Problemi di Recerca nella Statistica Bayesiana*, ed. W. Racugno (Societa Italiana di Statistica, Cagliari, 1992).

[12] J.O. Berger and R.L. Wolpert, *The Likelihood Principle*, 2nd ed. (Institute of Mathematical Statistics Monograph Series, Hayward, CA, 1988).

[13] J.O. Berger and J.M. Bernardo, On the development of the reference distributions, in: *Bayesian Statistics IV*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Oxford University Press, Oxford, 1992).

[14] U.N. Bhat and S.S. Rao, Statistical analysis of queueing systems, Queueing Systems 1 (1987) 217–247.

[15] A.B. Clarke, Maximum likelihood estimates in a simple queue, Ann. Math. Stat. 28 (1957) 1036–1040.

[16] D.R. Cox, Some problems of statistical analysis connected with congestion, in: *Proc. Symp. on Congestion Theory* (University of North Carolina Press, Chapel Hill, North Carolina, 1965) pp. 289–316.

[17] M.H. DeGroot, *Optimal Statistical Decision* (McGraw-Hill, New York, 1970).

[18] J.R. Ferrándiz and M. Sendra, *Tablas de Bioestadística Orientadas a la Metodología Bayesiana* (Gráficas Guada, Valencia, 1982).

[19] J.N.L. Johnson and S. Kotz, Continuous univariate distributions – 1 (Houghton Mifflin, Boston, 1970).

[20] N. Keiding, Maximum likelihood estimation in the birth-and-death process, Ann. Stat. 3 2 (1975) 363–372.

[21] M.F. McGrath, D. Gross and N.D. Singpurwalla, A subjective Bayesian approach to the theory of queues I – Modeling, Queueing Systems 1 (1987) 317–333.

[22] M.F. McGrath and N.D. Singpurwalla, A subjective Bayesian approach to the theory of queues II – Inference and information in $M/M/1$ queues, Queueing Systems 1 (1987) 335–353.

[23] J. Medhi, *Stochastic Models in Queueing Theory* (Academic Press, Boston, 1991).

[24] P.A.P. Moran, Estimation methods for evolutive processes, J. Roy. Stat. Soc. Ser. B13 (1951) 141–146.
[25] P.A.P. Moran, The estimation of parameters of a birth and death process, J. Roy. Stat. Soc. Ser. B15 (1953) 241–245.
[26] M.V. Muddapur, Bayesian estimates of parameters in some queueing models, Ann. Inst. Math. 24 (1972) 327–331.
[27] J.F. Reynolds, On estimating the parameters in some queueing models, Austral. J. Stat. 15 (1973) 35–43.
[28] D. Thiruvaiyaru and I.V. Basawa, Empirical Bayes estimation for queueing systems and networks, Queueing Systems 11 (1992) 179–202.