Estimation for queues from queue length data

J.V. Ross* T. Taimre† P.K. Pollett‡

September 15, 2006

Abstract

We consider the estimation of arrival and service rates for queues based on queue length data collected at successive, not necessarily equally spaced, time points. In particular, we consider the M/M/c queue, for c large, but application of the method to the repairman problem is almost identical, and the general approach presented should extend to other queue types. The estimation procedure makes use of an Ornstein-Uhlenbeck diffusion approximation to the Markov process description of the queue. We demonstrate the approach through simulation studies and discuss situations in which the approximation works best.

Keywords: Diffusion approximation, M/M/c queues, maximum likelihood, queue length, rate estimation, traffic intensity.

1 Introduction

While the literature on the stochastic modelling of queues is extensive, estimation and inference concerning the arrival and service rates has, in comparison, received little attention. Almost all the literature that addresses estimation issues describes methods that require continuous observation of the process over a fixed interval of time [12, 26, 10, 8, 3, 2, 11]. One exception is the work of Basawa *et al*. [9], who consider estimation for single server queues from waiting time data. Here we derive a method that requires substantially less information: simply the number in the queue at successive, not necessarily equally spaced, time points. The specific results we present apply to the M/M/c queue, in particular when $c > 40$ (as will be discussed), but the approach should extend to many other queue types. For example, our results extend almost immediately to the repairman problem [14].

Our method makes use of results of Kurtz [15, 16] and Barbour [4, 5, 6, 7] concerning density-dependent Markov processes. By taking the arrival rate of customers (packages) to be of the same order as the number of servers, we arrive at a Markov process with density-dependent transition rates. This allows us to apply the aforementioned results to derive an Ornstein-Uhlenbeck (OU) approximation to the queueing process. This OU

*Department of Mathematics, University of Queensland, QLD 4072, Australia. jvr@maths.uq.edu.au
†Department of Mathematics, University of Queensland, QLD 4072, Australia. ttaimre@maths.uq.edu.au
‡Department of Mathematics, University of Queensland, QLD 4072, Australia. pkp@maths.uq.edu.au

approximation provides us with an approximate likelihood for successive observations of the state of the queue, namely that of a multivariate Gaussian distribution, with explicit expressions for the mean, variance and covariance of the observations all in terms of the number of servers and the arrival and service rates. The use of diffusion approximations to estimate rates in this way has been suggested previously [13, 17], but its practical implementation has, to the best of our knowledge, not been undertaken until recently [19].

We demonstrate the approach by considering simulated examples of a queueing process corresponding to a large packet switching type telecommunication system, a smaller telecommunication system and to a simpler, shopping type queue. We discuss situations in which our method provides accurate estimates of rates, provide confidence regions of estimates and rules on how to achieve the best possible results.

2 Definitions

The M/M/c queue has Poisson arrivals at rate λ and independent exponentially distributed service times, and c servers each operating at rate μ. More formally, it is a Markov process $\{X(t),\, t \geq 0\}$ on the state space $S = \{0, 1, \ldots\}$ with non-zero transition rates

$$q(m, m+1) = \lambda, \qquad m \in S,$$

and

$$q(m, m-1) = \mu \min(m, c), \qquad m = \{1, 2, \ldots\},$$

where m is the state (number in the queue) of the process at time t.

Our approach to estimating λ and μ relies on results of Kurtz [15, 16] and Barbour [4, 5, 6, 7] concerning *density-dependent* processes. Pollett [18] extended the applicability of these results to *asymptotically* density-dependent processes: Let $\{m_c(\cdot)\}$ be a *family* of Markov processes indexed by $c > 0$, and suppose that $m_c(\cdot)$ takes values in S_c, which is contained in $\mathbb{Z}^k$, and has transition rates $Q_c = (q_c(m, n), m, n \in S_c)$. In practice, one has great freedom in identifying an index parameter. For definiteness, let us think of c as the number of servers in our queue.

Definition 2.1 *Suppose that there exists an open set $E \subseteq \mathbb{R}^k$ and a family $\{f_c, c > 0\}$ of continuous functions, with $f_c : E \times \mathbb{Z}^k \to \mathbb{R}$, such that*

$$q_c(m, m+l) = c f_c\left(\frac{m}{c}, l\right), \quad l \neq 0.$$

Then, the family of Markov chains is asymptotically density-dependent if, additionally, there exists a function $F : E \to \mathbb{R}$ such that $\{F_c\}$, given by $F_c(x) = \sum_l l f_c(x, l)$, $x \in E$, converges pointwise to F on E.

This definition of density-dependence is more general than that introduced in [15], which has f_c (and hence F_c) being the same for all c. Roughly speaking, the family is density-dependent if the transition rates of the corresponding "density process" $X_c(\cdot)$, defined by $X_c(t) = m_c(t)/c$, $t \geq 0$, depend on the present state m only through the "density" m/c, or, failing this, if this property is exhibited *asymptotically*, for large c.

3 Density dependence

Here we present results of Pollett [18], summarising and extending the remarkable work of Kurtz [15, 16] and Barbour [4, 5, 6, 7], that identify the limiting OU process we use as the basis for our estimation procedure.

Based on Definition 2.1, there appears to be a natural way to associate a density-dependent *deterministic* process with the Markov process, the intuition being that $X_c(\cdot)$ behaves more deterministically as c becomes large. Its trajectory is "tracked" by the process when c is large. The following (functional) law of large numbers establishes a deterministic approximation under appropriate conditions. It can be deduced immediately from Theorem 3.1 of [15].

Theorem 3.1 *Suppose that $f_c(\cdot, l)$ is bounded, for each l and c, that F is Lipschitz continuous on E and that $\{F_c\}$ converges uniformly to F on E. Then, if $\lim_{N\to\infty} X_c(0) = x_0$, the density process $X_c(\cdot)$ converges uniformly in probability on $[0,t]$ to $X(\cdot, x)$, the unique (deterministic) trajectory satisfying $X(0,x) = x$, $X(s,x) \in E$, $s \in [0,t]$, and*

$$\frac{\partial}{\partial s} X(s,x) = F(X(s,x)). \tag{1}$$

The following (functional) central limit law establishes that, for large c, the fluctuations about the deterministic path follow a Gaussian diffusion, provided that mild "second-order" conditions are satisfied. It can be deduced from Theorems 3.1 and 3.5 of [16].

Theorem 3.2 *Suppose $f_c(\cdot, l)$ is bounded, that F is Lipschitz continuous and has uniformly continuous first derivative on E, and that*

$$\lim_{c\to\infty} \sup_{x\in E} \sqrt{c}|F_c(x) - F(x)| = 0.$$

Suppose also that the sequence $\{G_c\}$, where

$$G_c(x) = \sum_l l^2 f_c(x,l), \quad x \in E,$$

converges uniformly to G, where G is uniformly continuous on E. Let $x_0 \in E$. Then, if

$$\lim_{c\to\infty} \sqrt{c}\,(X_c(0) - x_0) = z, \tag{2}$$

the family of processes $\{Z_c(\cdot)\}$, defined by $Z_c(s) = \sqrt{c}\,(X_c(s) - X(s,x_0))$, $0 \le s \le t$, converges weakly in $D[0,t]$ (the space of right-continuous, left-hand limit functions on $[0,t]$) to a Gaussian diffusion $Z(\cdot)$ with initial value $Z(0) = z$ and with mean and variance given by $\mu_s := \mathrm{E}(Z(s)) = M_s z$, where $M_s = \exp(\int_0^s B_u du)$ and $B_s = F'(X(s,x_0))$, and $\mathrm{Var}(Z(s)) = \sigma_s^2$, where $\sigma_s^2 = M_s^2 \int_0^s M_u^{-2} G(X(u,x_0))du$.

It follows that $X_c(s)$ has an approximate normal distribution with $\mathrm{Var}(X_c(s)) \simeq \sigma_s^2/c$. We would usually take $x_0 = X_c(0)$, thus giving $\mathrm{E}(X_c(s)) \simeq X(s,x_0)$.

In the important special case where x_0 is chosen as an equilibrium point of (1), we can be far more precise about the approximating diffusion.

Corollary 3.3 *If x_0 satisfies $F(x_0) = 0$ then, under the conditions of Theorem 3.2, the family $\{Z_c(\cdot)\}$, defined by $Z_c(s) = \sqrt{c}(X_c(s) - x_0)$, $0 \le s \le t$, converges weakly in $D[0,t]$ to an Ornstein-Uhlenbeck process $Z(\cdot)$ with initial value $Z(0) = z$, local drift $B = F'(x_0)$ and local variance $V = G(x_0)$. In particular, $Z(s)$ is normally distributed with mean $\mu_s = e^{Bs}z$ and variance $\sigma_s^2 = \frac{V}{2B}(e^{2Bs} - 1)$.*

We conclude that, for c large, $X_c(s)$ has an approximate normal distribution with $\mathrm{Var}(X_c(s)) \simeq \sigma_s^2/c$. A "working approximation" for the mean (that is, for a fixed value of c) is given by

$$\mathrm{E}(X_c(s)) \simeq x_0 + e^{Bs}(X_c(0) - x_0).$$

In the context of queueing models x_0 will usually be asymptotically stable, that is $B < 0$. However, it should be emphasised that it need not be for each of the above conclusions to hold. Indeed, the OU approximation is often very accurate in describing the fluctuations about centres and unstable equilibria (see Barbour [5]). We shall henceforth assume that $B < 0$, which is the case for our queueing model.

4 Method for estimation of rates

Our parameter estimation method makes use of the OU approximation outlined in the previous section. Firstly, the OU process is strongly stationary if we start it in equilibrium: $Z(0) \sim \text{Normal}\left(0, \sigma^2\right)$, where $\sigma^2 = V/(-2B)$. We therefore have (for large c) that $X_c(0) \sim \text{Normal}\left(x_0, \sigma^2/c\right)$. Hence, we may approximate $\mathrm{Cov}(X_c(s), X_c(s+t))$ by

$$c(t) := \frac{1}{c}\mathrm{Cov}(Z(s), Z(s+t)) = c(0)\exp(B|t|), \tag{3}$$

where $c(0) = \sigma^2/c$. Also, for large c, we know explicitly the correlation structure of the Gaussian vector $(X_c(t_1), X_c(t_2), \ldots, X_c(t_n))$, and hence its likelihood function:

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp\left[-\frac{1}{2}(x-m)C^{-1}(x-m)'\right], \tag{4}$$

where $m = (m_1, m_2, \ldots, m_n)$, $m_i = x_0$ for all $i = 1, 2, \ldots, n$, and

$$C = \begin{pmatrix} c_1 & c_{1,2} & c_{1,3} & \cdots & c_{1,n} \\ c_{1,2} & c_2 & c_{2,3} & \cdots & c_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{1,n} & \cdots & \cdots & \cdots & c_n \end{pmatrix}, \tag{5}$$

where $c_i = \sigma^2/c$ and $c_{i,i+s} = (\sigma^2/c)\exp(B|t_{i+s} - t_i|)$. The inversion of C, and calculation of its determinant, can be done explicitly, and this permits us to write down the (log-) likelihood explicitly (see Appendix I). The form that should be used in practice is given in (11) (or (12)).

We can therefore evaluate the (joint) maximum likelihood estimators of the parameters of the model, which are the values that maximise (4). Explicit calculation of the maximum likelihood estimators is not practical if the sample size is large. Therefore a numerical optimisation procedure will be required in practice to find the parameters which maximise the likelihood function. We believe the Cross-Entropy method (see [20])

to be an ideal approach, but many numerical optimisation procedures should be as effective. We illustrate this approach in Section 6 by using the Cross-Entropy method to estimate the parameters of three hypothetical M/M/c queueing examples. Appendix II also contains advice on selecting parameters when using the Cross-Entropy method.

It is pertinent to note that the OU approximation is achieved by letting the number of servers tend to infinity. Thus the OU approximation, and consequently the parameter estimation procedure presented, are best for queues with a *large* number of servers. It should also be noted that unequally spaced sampling of the process is not an obstacle to the method presented, as can be seen from the covariance structure (3).

Finally, we note that we can obtain estimates of relationships between parameters without resorting to numerically maximising the full likelihood. Under the assumptions used to obtain the OU approximation, we have

$$\mathrm{E}\,(\bar{m}) = cx_0, \quad \text{where} \quad \bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_c(t_i), \tag{6}$$

and

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}(m_c(t_i) - \bar{m})^2\right) = c\sigma^2. \tag{7}$$

Thus, the mean and variance of our data set provides some indication about the relationships between the parameters.

5 OU approximation for M/M/c queue

First, we derive the OU approximation for the M/M/c queue model using the results summarized in Section 3. We start by supposing that $\lambda = O(c)$, for c large, that is $\lambda \sim \alpha c$, where α is a constant for a particular queue. It is clear that the process is density-dependent with $F(x) = \alpha - \mu x$ and $G(x) = \alpha + \mu x$, and that $x_0 = \alpha/\mu$. Observe also that the traffic intensity $\lambda/(\mu c)$ tends to x_0 as $c \to \infty$, so we may interpret x_0 as the asymptotic traffic intensity.

We will consider the most interesting case $x_0 < 1$, when the traffic intensity is less than 1 and we consider the OU approximation about the stable equilibrium x_0. Since $F'(x) = -\mu$, we have local drift $B = F'(x_0) = -\mu$ and local variance $V = G(x_0) = 2\alpha$, being approximately $2\lambda/c$ when c is large. Thus, provided we arrange for (2) to hold, there will be a valid OU approximation. We conclude that, for c large, $X_c(t)$ has an approximate normal distribution with

$$\mathrm{E}(X_c(t)) \simeq x_0 + e^{-\mu t}(X_c(0) - x_0), \tag{8}$$

$$\mathrm{Var}(X_c(t)) \simeq \frac{\lambda}{\mu c^2}(1 - e^{-2\mu t}). \tag{9}$$

This diffusion approximation for M/M/c queues is not new [14]. However, we present the expressions in a form that explicitly shows the dependence on both λ and μ, which is required for estimation of these rates.

6 Estimation of rates

In this section we estimate the rates of M/M/c queues from simulated data. As discussed in Section 4, if we start the OU process in equilibrium, then $X_c(0) \sim \text{Normal}(\rho/c, \rho/c^2)$ for large c, where $\rho := \lambda/\mu$. Thus we have

$$\text{Cov}(X_c(s), X_c(s+t)) \simeq \frac{\rho}{c^2} \exp(-\mu|t|),$$

and hence the likelihood for the Gaussian vector $(X_c(t_1), X_c(t_2), \ldots, X_c(t_n))$ is given by equation (4), where $m_i = \rho/c$ for all $i \in \{1, 2, \ldots, n\}$, and C is given by (5) with $c_i = \rho/c^2$ and $c_{i,i+s} = \rho/c^2 \exp(-\mu|t_{i+s} - t_i|)$, for all $i, s \in \{1, 2, \ldots, n\}$.

We use the log-likelihood $l(x)$, given by

$$l(x) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|C|) - \frac{1}{2}(x-m)C^{-1}(x-m)'. \tag{10}$$

(Again, the form that should be used in practice is given in (12).)

For the M/M/c queue, the expected values of the mean and variance of our data set are (see (6) and (7))

$$\text{E}(\bar{m}) = \rho \quad \text{and} \quad \text{E}\left(\frac{1}{n}\sum_{i=1}^{n}(m_c(t_i) - \bar{m})^2\right) = \rho/c.$$

Hence, we have a quick estimate for ρ (the ratio of arrival rate to service rate) and the traffic intensity of the queue.

6.1 Large telecommunication system

We simulated the M/M/c queue with $c = 300$, $\lambda = 25$ and $\mu = 0.09$. This corresponds to a hypothetical system with 300 servers, calls (packets) arriving at rate 50 per second, each server being able to process 0.18 calls (packets) per second and being sampled every 0.5 of a second. The traffic intensity of this system is approximately 0.926. We simulated the queue to collect data corresponding to 5 minute's worth of observations (600 data points; 700 data points were collected, the first 100 discarded to reduce the influence of initial conditions). We produced 5 simulated data sets and ran the optimisation (Cross-Entropy) algorithm 5 times on each data set and reported the best, in terms of maximum likelihood, from each run. The Cross-Entropy parameters we use here are a sample size of 20,000, and an elite sample size of 1,000. We initialised the algorithm with pairs (λ, μ) drawn uniformly from the triangle $(0,0) - (100,0) - (100,100)$. The algorithm was stopped when the largest of the standard deviations of the sampling distributions dropped below the threshold $\varepsilon = 10^{-5}$. The table below reports the results and in addition the average (Avg.) estimates, and relative error (RE) in the averages.

Run	$\hat{\lambda}$	$\hat{\mu}$
1	24.6274	0.0863
2	25.6284	0.0930
3	24.0672	0.0886
4	23.9907	0.0875
5	23.4328	0.0854
Avg.	24.3493	0.08816
RE	0.026	0.02

It can be seen that our method of estimation works well in this situation. The relative errors in the average of our estimates for λ and μ are extremely small, being 2.6% and 2% respectively.

6.2 Small telecommunication system

We simulated the M/M/c queue with $c = 50$, $\lambda = 4250/600$ and $\mu = 1/6$. This corresponds to a hypothetical system with 50 servers, calls arriving at rate 4250/300 per minute, each server being able to process 1/3 of a call per minute and being sampled every 30 seconds. This system has a traffic intensity of 0.85. We simulated the queue to collect data corresponding to 2 hours worth of observations (240 data points; 440 data points were collected, the first 200 discarded to reduce the influence of initial conditions). We produced 5 simulated data sets and ran the optimisation (Cross-Entropy) algorithm 5 times on each data set and report the best, in terms of maximum likelihood, from each run. The same Cross-Entropy parameters were used and the same initialisation and stopping rule. The table below once again reports the results and in addition the average (Avg.) estimates, and relative error (RE) in the averages.

Run	$\hat{\lambda}$	$\hat{\mu}$
1	8.3436	0.1863
2	6.8675	0.1742
3	7.4461	0.1735
4	6.8598	0.1567
5	6.9116	0.1554
Avg.	7.2857	0.1692
RE	0.029	0.015

Once again, our method produces reasonably accurate estimates for both λ and μ, with the relative error in the averages in this situation being 2.9% and 1.5% respectively.

6.3 Shopping queue

We simulated the M/M/c queue with $c = 5$, $\lambda = 0.75$ and $\mu = 0.175$. This corresponds to a hypothetical system with 5 servers, customers arriving at rate 3 per minute, each server being able to process 0.7 customers per minute and being sampled every 15 seconds. These rates correspond to an express lane type queue often found at supermarkets, and has a traffic intensity of approximately 0.86. We simulated the queue to collect data corresponding to 1 hour worth of observations (240 data points; 440 data points were collected, the first 200 discarded to reduce the influence of initial conditions). We produced 5 simulated data sets and ran the optimisation (Cross-Entropy) algorithm 5 times on each data set and report the best, in terms of maximum likelihood, from each run. The same Cross-Entropy parameters were used and the same initialisation and stopping rule. The table below reports the results and in addition the average (Avg.) estimates, and relative error (RE) in the averages.

Run	$\hat{\lambda}$	$\hat{\mu}$
1	0.7567	0.0888
2	0.6615	0.1541
3	0.7468	0.0458
4	0.6436	0.1551
5	0.6586	0.1193
Avg.	0.69344	0.11262
RE	0.075	0.357

Our estimation procedure does not appear to work consistently in this situation. The relative error in the average estimate for λ is reasonable, at 7.5%, but the relative error in the average estimate for μ is abominable, at 35.7%.

6.4 Error bounds

Maximum likelihood estimators are asymptotically normally distributed with mean $\hat{\theta}$ and covariance matrix given by the inverse of the Fisher information matrix. This can be estimated by

$$\left(\mathscr{I}(\hat{\theta})\right)^{-1} = \left\{-\mathrm{E}\left[\frac{\partial^2 l(\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'}\right]\right\}^{-1}.$$

However, the second derivatives of the log-likelihood are often too complicated for their exact expected values to be calculated in practice. A second estimator widely used is

$$\left(\mathscr{I}(\hat{\theta})\right)^{-1} = \left(-\frac{\partial^2 l(\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'}\right)^{-1},$$

this being the inverse of (the negative of) the matrix of second derivatives of the log-likelihood, evaluated at the MLE.

For our, model we have

$$\left(\mathscr{I}((\hat{\lambda},\hat{\mu}))\right)^{-1} = -\begin{pmatrix} \frac{\partial^2 l}{\partial\hat{\lambda}^2} & \frac{\partial^2 l}{\partial\hat{\lambda}\partial\hat{\mu}} \\ \frac{\partial^2 l}{\partial\hat{\mu}\partial\hat{\lambda}} & \frac{\partial^2 l}{\partial\hat{\mu}^2} \end{pmatrix}^{-1} = \left(\frac{\partial^2 l}{\partial\hat{\lambda}^2}\frac{\partial^2 l}{\partial\hat{\mu}^2} - \frac{\partial^2 l}{\partial\hat{\lambda}\partial\hat{\mu}}\frac{\partial^2 l}{\partial\hat{\mu}\partial\hat{\lambda}}\right)^{-1}\begin{pmatrix} -\frac{\partial^2 l}{\partial\hat{\mu}^2} & \frac{\partial^2 l}{\partial\hat{\lambda}\partial\hat{\mu}} \\ \frac{\partial^2 l}{\partial\hat{\mu}\partial\hat{\lambda}} & -\frac{\partial^2 l}{\partial\hat{\lambda}^2} \end{pmatrix}.$$

Hence, we must compute the second derivatives of the log-likelihood given in Appendix I. This can be done exactly. However, the formulæ are rather cumbersome and will not be written out here. We use this approach to calculate the error bounds presented below.

Often it will be difficult to compute the matrix of second derivatives. In such cases a third estimator may be useful, which only requires first derivatives of the log-likelihood. This estimator is

$$\left(\mathscr{I}(\hat{\theta})\right)^{-1} = \left(\sum_{i=1}^{n}\widehat{g}_i\,\widehat{g}_i'\right)^{-1}$$

where

$$\widehat{g}_i = \frac{\partial l(x_i,\widehat{\theta})}{\partial\widehat{\theta}}.$$

We illustrate the error bounds by plotting in Figure 1 results for a set of simulated data (with 600 equally spaced observations, and parameters $(\lambda,\mu) = (25, 0.09)$ and $c = 300$).

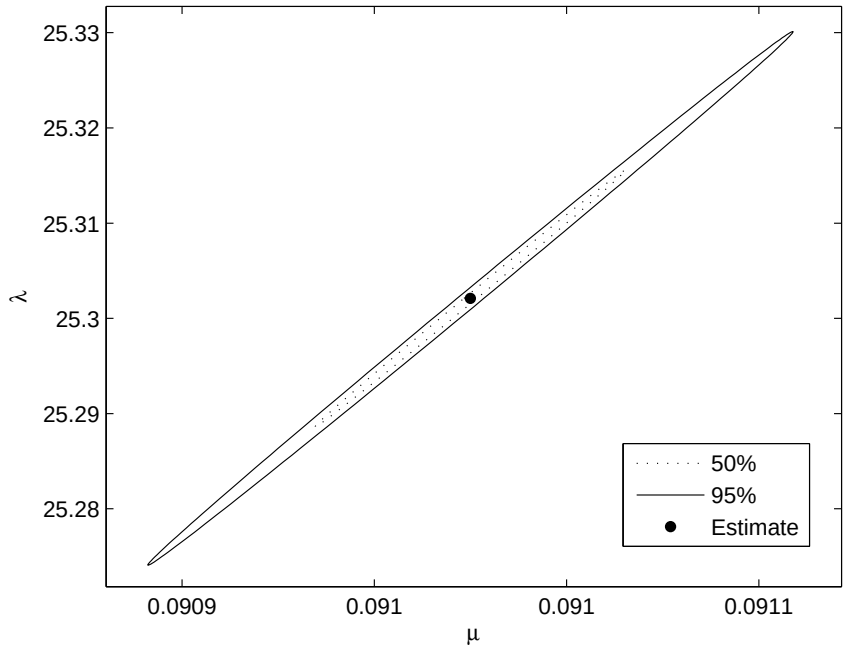

Figure 1: Confidence regions for simulated data of 600 observations, and parameters $\lambda = 25$, $\mu = 0.09$ and $c = 300$.

7 Discussion

It can be seen that in the large telecommunication system example, our methods produce extremely accurate estimates to the true parameter values from all 5 simulation runs. The largest relative error over these runs was 6.2%, and the relative error in our average estimates of λ and μ were 2.6% and 2% respectively. The parameter estimates reported were also consistently achieved by the Cross-Entropy algorithm in each run.

Similar comments apply to our small telecommunication system example, with the estimation procedure producing extremely accurate estimates to the true parameter values in 4 out of the 5 simulation runs, and a reasonable estimate in the other case, with relative errors of 17.8% and 11.8% for λ and μ respectively. The relative errors in our average estimates of λ and μ were once again impressive, being 2.9% and 1.5% respectively. The parameter estimates reported were once again consistently achieved over each of the 5 runs of the Cross-Entropy algorithm.

In our shopping queue example the quality of both estimates was poor. The estimates of μ consistently underestimated the true service rate and varied considerably over the 5 simulation runs. However, the estimates of λ appear to be reasonably close to the true values from our investigations for this (and other *small*) queues. The poor estimates in this case appear to be due to the small number of servers ($c = 5$), resulting in a poor approximation to the true likelihood of the queueing process observations. From our investigation it also appears that a larger sample size does not overcome this inadequacy. If in such situations this procedure must be used for parameter estimation, say due to lack of data or ability to acquire the data required to use other estimation methods, a

simulation study should be performed for the particular number of servers in the queue under consideration, the results of which can be used to correct for the bias in the procedure. An alternative approach may be to approximate the queue by a reflected Ornstein-Uhlenbeck process, a diffusion approximation that has been employed for many queue types [23, 24, 25].

From our investigations we have derived some rules that ensure our procedure works consistently and provides reasonably accurate estimates. The rules of thumb we provide below were derived from simulation studies, and for queueing processes with traffic intensities in the region 0.8 to 0.95, which is reasonable for many queueing systems. The first rule concerns the minimum number of servers in the queue. As already seen, if the number of servers is too small our procedure fails to produce accurate estimates. We found that the number of servers should be greater than 40 to provide accurate estimates. The other rule concerns the sampling interval that is required to produce accurate estimates. The sampling interval should be chosen so that the value of $\lambda+\mu$ for that interval is less than approximately 30. This requirement arises from the covariance structure (covariance of observations), which necessitates that the expected number of transitions between observations is not so large that the covariance between observations evanesces. Choosing such a sampling interval appears to require some prior information about the values of λ and μ, which in practice may often be only known approximately. However, as mentioned earlier, the estimates of λ produced by our method appear to be always reasonably close to the true parameter values, so one could estimate the rates from one sample of data, and based upon the estimate for λ, reduce the sampling interval if necessary.

Finally, we provided confidence contours for our estimates. The confidence region plotted corresponds to the large telecommunication system example, and we see that the confidence regions are very tight. However, it should be noted that they do not include the true parameter values. This is most likely a result of our estimates being biased due to inaccuracies in the Gaussian likelihood approximation for finite server size (see also Section 6.3 of [19]).

Results similar to those presented here should hold for other queue types, approximated by different diffusion approximations, such as the reflected Ornstein-Uhlenbeck approximations for queues with reneging or balking [23, 25]. As mentioned earlier, the use of such an approximation for the M/M/c queue may also increase the applicability of our method to cases when $c < 40$.

8 Summary

We have presented a method for estimating arrival and services rates for M/M/c queues from queue length data. The procedure makes use of an Ornstein-Uhlenbeck diffusion approximation to the original Markov process description of the queue. We demonstrated the applicability of the results through simulation studies and presented asymptotic confidence contours for the estimates. Rules on how to make best use of the procedure were also given. These concerned the minimum number of servers c (c greater than approximately 40) and the maximum sampling interval (sampling to ensure that $\lambda+\mu$ less than approximately 30 per sampling interval). The approach presented should extend to other queue types, in particular our results extend almost immediately to the repairman problem.

Acknowledgements. The authors thank the referee and Editors for their careful reading of the manuscript, and acknowledge the support of the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems.

Appendix I: The OU (log-) likelihood

We summarise results from [21] that allow us to invert explicitly, and calculate the determinant of, the time-dependent covariance matrix C of the OU process. This permits us to write the (log-) likelihood explicitly, and consequently note that only $O(n)$ operations are required to evaluate it.

Write the likelihood function (4) as

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp\left[-\tfrac{1}{2} y C^{-1} y'\right],$$

with $y = x - m$. It turns out ([22]) that

$$y C^{-1} y' = \frac{\sigma^2}{c} \sum_{i=1}^{n} \frac{(y_i - r_{i-1} y_{i-1})^2}{1 - r_{i-1}^2},$$

where $\sigma^2 = V/(-2B)$ and

$$\det(C) = \left(\frac{\sigma^2}{c}\right)^n \prod_{i=1}^{n-1} (1 - r_i^2),$$

where

$$r_k = \begin{cases} 0 & \text{if } k = 0, \\ \exp\left[B(t_{k+1} - t_k)\right] & \text{if } 1 \le k \le n-1. \end{cases}$$

This allows us to write the likelihood, and the log-likelihood, as

$$f(x) = \left(\frac{2\pi\sigma^2}{N}\right)^{-\frac{n}{2}} \left(\prod_{i=1}^{n-1} 1 - r_i^2\right)^{-\frac{1}{2}} \exp\left[-\frac{N}{2\sigma^2} \sum_{i=1}^{n} \frac{(y_i - r_{i-1} y_{i-1})^2}{1 - r_{i-1}^2}\right], \qquad (11)$$

and

$$l(x) = -\frac{n}{2} \log\left(\frac{2\pi\sigma^2}{c}\right) - \frac{1}{2} \sum_{i=1}^{n-1} \log\left(1 - r_i^2\right) - \frac{c}{2\sigma^2} \sum_{i=1}^{n} \frac{(y_i - r_{i-1} y_{i-1})^2}{1 - r_{i-1}^2}. \qquad (12)$$

We emphasise that the above formulæ hold in general, and should be used to evaluate the (log-) likelihood (instead of inverting C and calculating its determinant numerically).

Appendix II: Cross-Entropy Method

Here we describe the algorithm used to maximise the (log-) likelihood. This is followed by some general advice on choosing parameters when using the Cross-Entropy method. For further details and applications of the Cross-Entropy method, see [20].

Algorithm

1. Set $N_s = 20000$, $N_e = 1000$, $\varepsilon = 10^{-5}$, $a = 100$, maxits $= 5000$, $t = 0$.
2. Draw N_s pairs (λ_i, μ_i) uniformly from the triangle $(0,0) - (a,0) - (a,a)$.
3. Set $t = t + 1$. Calculate $l_i = l(x; \lambda_i, \mu_i)$.
4. Locate a set $\mathcal{E}$ of N_e indices i for which $l_k \geq l_j$ for all $k \in \mathcal{E}$, $j \notin \mathcal{E}$.
5. Calculate mean_λ, stddev_λ and mean_μ, stddev_μ as the means and standard deviations of λ_k and μ_k where $k \in \mathcal{E}$.
6. Draw N_s pairs of unknown parameters, from independent normal distributions with means and standard deviations calculated in the previous step.
7. If the largest σ is greater than ε, and $t <$ maxits then return to step 3; otherwise output $\widehat{\lambda}$ and $\widehat{\mu}$, where κ is an index such that $l_\kappa \geq l_j$ for all j.

In general, a user should trial parameter combinations on a sub-problem of the original, with initial "rule of thumb" settings of, say $N_e/N_s = \text{const}$ (0.01), $N_s = 5 \times d \times$ (20 to 200) (say, where d is the number of dimensions), maxits at 100 or 200 (so that it is usually never achieved). The suggestion is to subsequently refine these parameters if necessary. There is a "Fully Adaptive" version of the Cross-Entropy algorithm which limits the amount of tweaking required (see Chapter 5 of [20]).

References

[1] Acharya, S.K. (1999) On normal approximation for maximum likelihood estimation from single server queues. *Queuing Systems* 31, 207–216.

[2] Armero, C. (1994) Bayesian inference in Markovian queues. *Queuing Systems* 15, 419–426.

[3] Armero, C. and Bayarri, M.J. (1994) Bayesian prediction in M/M/1 queues. *Queuing Systems* 15, 401–417.

[4] Barbour, A.D. (1974) On a functional central limit theorem for Markov population processes. *Adv. Appl. Probab.* 6, 21–39.

[5] Barbour, A.D. (1976) Quasi-stationary distributions in Markov population processes. *Adv. Appl. Probab.* 8, 296–314.

[6] Barbour, A.D. (1980) Equilibrium distributions Markov population processes. *Adv. Appl. Probab.* 12, 591–614.

[7] Barbour, A.D. (1980) Density-dependent Markov population processes. In *Biological growth and spread.* Eds. W. Jäger, H. Rost and P. Tautu, Lecture Notes in Biomathematics 38, 36–49, Springer, Berlin.

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

[9] Basawa, I.V., Bhat, U.N. and Lund, R. (1997) Maximum likelihood estimation for single server queues from waiting time data. *Queuing Systems* 24, 155–167.

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

[11] Bingham, N.H. and Pitts, S.M. (1999) Non-parametric estimation for the $M/G/\infty$ queue. *Ann. Inst. Statist. Math.* 1, 71–97.

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

[13] Feigin, P.D. (1976) Maximum likelihood estimation for continuous-time stochastic processes. *Adv. Appl. Probab.* 8, 712–736.

[14] Iglehart, D.L. (1965) Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Probab.* 2, 429–441.

[15] Kurtz, T. (1970) Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* 7, 49–58.

[16] Kurtz, T. (1971) Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* 8, 344–356.

[17] McNeil, D.R. and Weiss, G.H. (1977) A large population approach to estimation of parameters in Markov population models. *Biometrika* 64, 553–558.

[18] Pollett, P.K. (1990) On a model for interference between searching insect parasites. *J. Austral. Math. Soc. Ser. B* 32, 133–150.

[19] Ross, J.V., Taimre, T. and Pollett, P.K. (2006) On parameter estimation in population models. *Theor. Popul. Biol. (to appear).*

[20] Rubinstein, R.Y. and Kroese, D.P. (2004) *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. Springer-Verlag, New York.

[21] Rybicki, G.B. (1994) Unpublished Notes: Notes on Gaussian Random Functions with Exponential Correlation Functions (Ornstein-Uhlenbeck Process) *from* `http://www.lanl.gov/DLDSTP/fast/`.

[22] Rybicki, G.B. and Press, W.H. (1995) A class of fast methods for processing irregularly sampled or otherwise inhomogeneous one-dimensional data. *Phys. Rev. Lett.* 74, 1060–1063.

[23] Ward, A.R. and Glynn, P.W. (2003) A diffusion approximation for a Markovian queue with reneging. *Queuing Systems* 43, 103–128.

[24] Ward, A.R. and Glynn, P.W. (2003) Properties of the reflected Ornstein-Uhlenbeck process. *Queuing Systems* 44, 109–123.

[25] Ward, A.R. and Glynn, P.W. (2005) A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queuing Systems* 50, 371–400.

[26] Wolff, R.W. (1965) Problems of statistical inference for birth and death queueing models. *Operat. Res.* 13, 343–357.