European Journal of Operational Research 279 (2019) 882–901

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Statistical inference for $M_t/G/Infinity$ queueing systems under incomplete observations

Dongmin Li[a,b], Qingpei Hu[a,b,*], Lujia Wang[c], Dan Yu[a,b]

[a] Center of Quality and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China
[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China
[c] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Phoenix, USA

ARTICLE INFO

Article history:
Received 30 October 2018
Accepted 30 June 2019
Available online 6 July 2019

Keywords:
Queueing
Interval censored data
Maximum-likelihood estimation (MLE)
Parametric bootstrap
Delta method

ABSTRACT

$M_t/G/Infinity$ queueing systems have been widely used to analyse complex systems, such as telephone call centres, software testing systems, and telecommunication systems. Statistical inferences of performance measures, such as the expected cumulative numbers of arrivals and departures, are indispensable for decision makers in analysing the current scenario, predicting future scenarios, and making cost-effective decisions. In most scenarios, we only obtain interval censored data, namely, counts in fixed time intervals, instead of complete data because we either do not want or are not able to monitor arrivals and departures. We provide a general framework for statistical inference in $M_t/G/Infinity$ queueing systems given interval censored data. A maximum-likelihood estimation (MLE) method is proposed for inferring the arrival rate and service duration. This method is applicable to general forms of the arrival rate functions and general service duration distributions. More importantly, we propose a combination of the bootstrap method and the delta method for inferring the expected cumulative numbers of arrivals and departures. The results of the simulation study demonstrate that the point and interval estimates of the proposed MLE method are satisfactory overall. As the number of intervals increases, the estimates based on the proposed MLE approach the estimates based on MLE with complete data. Our procedure enables estimates to be obtained without the need to keep track of each item, thereby substantially reducing resource consumption for monitoring items and storing data. An application in a software testing system demonstrates that the goodness-of-fit performance of the proposed MLE method is satisfactory.

1. Introduction

An $M_t/G/\infty$ queueing system is a relatively simple queueing system that has a nonhomogeneous Poisson arrival process with a time-dependent deterministic arrival rate function $\lambda_a \equiv \lambda_a(t)$, independent and identically distributed (i.i.d.) service durations that are independent of the arrival process, and infinitely many servers. Time-varying queueing models, including $M_t/G/\infty$ models, are standard models for describing the dynamics of large-scale service systems, such as telecommunication systems, call centres, and healthcare systems, e.g., hospitals (Pender, 2016). Researchers have applied $M_t/G/\infty$ models to service systems, such as telemarketing, police patrol, fire fighting, hospitals, copy machine repairs, and automatic teller machine operations. In these applications, an operating policy was to keep customer delays close to zero—a scenario that is consistent with the use of an infinite-server model (Green & Kolesar, 1998). $M_t/G/\infty$ models have been applied to storage systems to assess the day-by-day adequacy of stock (Crawford, 1977), to analyse stock requirements (Hillestad & Carrillo, 1980), and to evaluate war-readiness spare requirements for aircraft (Crawford, 1981). They also have been applied to software testing programs (Vizarreta et al., 2018; Yang, 1996) and internet traffic systems (Fay, Roueff & Soulier, 2007). The $M_t/G/\infty$ model is the offered load model for wireless and packet network systems, which describes the total packet carrying capacity of the channels or links in a packet network (Malhotra, Dey, van Doorn & Koonen, 2001; Palm, 1943; Singhai, Joshi & Bhatt, 2009). $M_t/G/\infty$ models have been used to describe the time-dependent variations in traffic at a base station in a nomadic computing, wireless environment (Malhotra et al., 2001). Risk measures have been studied to assess the performance of $M_t/G/\infty$ queueing systems and the measures can be used in staffing procedures, especially in healthcare systems (Pender, 2016). By considering the births and deaths of items as the arrival and departure processes, we can also analyse the births and deaths in a population with $M_t/G/\infty$ models.

[*] Corresponding author at: Center of Quality and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China.
E-mail address: qingpeihu@amss.ac.cn (Q. Hu).

https://doi.org/10.1016/j.ejor.2019.06.055
0377-2217/© 2019 Elsevier B.V. All rights reserved.

Although, in practice, systems do not have an infinitely many servers, various properties of an infinite server model are approximately (in an appropriate sense) true for finite-channel (and even single-channel) servers if the arrival rate is sufficiently low to yield a negligible probability that a customer will arrive to find the service full (Newell, 1966). The model also provides a reasonable description of congestion in actual operations under similar circumstances (Green & Kolesar, 1998). In addition, the number of busy servers in an infinite server model provides insight about the number of servers that are required in practical scenarios. Satisfactory analytical results are available for $M_t/G/\infty$ queueing systems. Compared with finite-server queueing systems, the number of busy servers, which is denoted as $N(t)$, and the departure rate, which is denoted as $\lambda_d(t)$, of an $M_t/G/\infty$ queueing system are relatively simple to determine. The theory for infinite-server models with time-dependent arrival rates is a useful frame of reference for examining more difficult finite-server models with time-dependent arrival rates (Massey & Whitt, 1993). Infinite-server models are of interest both in their own right (Eick, Massey & Whitt, 1993; Fay et al., 2007; Yang, 1996) and as approximations for lightly to moderately loaded multiserver models (Eick et al., 1993; Green & Kolesar, 1998; Massey & Whitt, 1993).

Research on queueing theory has provided important insights into the behavioural, operational, and statistical problems in queueing systems (Bhat, 1969). One key problem of queueing theory has been statistical inference given information regarding queueing systems such that we can analyse the current scenario, predict future scenarios, and make cost-effective decisions accordingly. In practical queueing scenarios, we often obtain large amounts of data regarding queueing systems through observation, such as the moments at which customers place calls to a call centre and the moments at which they hang up. Call centre service distribution behaviour has been analysed using Bayesian parametric and semi-parametric mixture models that can exhibit non-standard behaviour and are based on real call centre data (Aktekin, 2014). Inter-dependent, heterogeneous, and time-varying service-time distributions have been proposed that are based on a large-scale data-based investigation of service durations in a call centre with many heterogeneous agents and multiple call types (Ibrahim, L'Ecuyer, Shen & Thiongane, 2016). Information in the form of observed service durations was used to derive predictive probability results for the waiting times of customers in a queue. The results can be used in a multi-queue problem to assign arriving customers to queues with the objective of minimizing the waiting times (Coolen & Coolen-Schrijner, 2003). In $M_t/G/\infty$ queueing systems, we wish to infer the arrival rate function $\lambda_a(t)$; the cumulative distribution function (cdf), which is denoted by G, of the service duration S; and, consequently, the performance measures of interest, such as the expected cumulative numbers of arrivals and departures, which are denoted as $m_a(t)$ and $m_d(t)$, respectively, and the expected number of busy servers, which is denoted as $m(t)$. These measures give insight to service providers about the number of servers that are required. Given complete data on the queueing system, i.e., the arrival epoch and departure epoch of each item, it is simple to infer $\lambda_a(t)$ and G. However, in most scenarios, we can only obtain incomplete data since it is often difficult or impossible to keep track of each item from arrival to departure (Blanghaps, Nov & Weiss, 2013). Despite its substantial practical value in practical scenarios, statistical inference with incomplete data remains a major challenge, which is mainly due to the high complexity of time-dependent queueing systems with incomplete data.

Researchers have extensively studied statistical inference approaches given incomplete data in the $M/G/\infty$ queueing system, which is a simple and special case of the $M_t/G/\infty$ queueing system. An $M_t/G/\infty$ queueing system is an $M/G/\infty$ queueing system if the arrival process is a homogeneous Poisson process, i.e., if λ_a is a constant. Researchers mainly focus on three types of incomplete data: (1) the arrival and departure epochs without identification of items (Blanghaps et al., 2013; Brown, 1970; Goldenshluger, 2018); (2) the queue-length process $\{N(t)\}$ (Bingham & Pitts, 1999; Goldenshluger, 2016; Pickands & Stine, 1997); and (3) the "busy-period" process $\{I(N(t)) > 0\}$, which indicates only whether the system is empty or not (Bingham & Pitts, 1999; Hall & Park, 2004; Park, 2007). Due to the similarity between $M/G/\infty$ queueing systems and $M_t/G/\infty$ queueing systems, many approaches and results in $M/G/\infty$ queueing systems can be extended to $M_t/G/\infty$ queueing systems. However, in $M/G/\infty$ queueing systems, researchers focus on the steady state of the systems, in contrast to the transient behaviour in $M_t/G/\infty$ queueing systems. Common relations for steady-state queueing systems, such as Little's law, must be reformulated in $M_t/G/\infty$ queueing systems (Schwarz, Selinka & Stolletz, 2016). Since the arrival and departure rate functions are time-dependent in $M_t/G/\infty$ queueing systems, many properties in $M/G/\infty$ queueing systems do not hold in $M_t/G/\infty$ queueing systems. To examine the dynamics of real systems (Andersen, Nielsen, Reinhardt & Stidsen, 2019; Dhingra, Kumawat, Roy & de Koster, 2018; Liu, 2018; Massey, 2002, Parker; Pender, 2016; Schwarz et al., 2016), it is essential to study $M_t/G/\infty$ queueing systems.

We are interested in one type of incomplete data that is commonly used, interval censored data, which are also known as grouped data and panel count data and are specified as counts that occur in fixed time intervals, as opposed to the exact times because we cannot perform continuous observations in many cases. Interval censored data are similar but more complicated than data type (1) above since neither the correspondence between the arrivals and departures nor the exact arrival and departure times are known. Yang (1996) inferred the parameters regarding the arrival process and service duration separately: the arrival process parameters were estimated using the Laplace trend statistic and the service duration parameters were estimated based on the estimated arrival process parameters. The least-squares estimation (LSE) method is a convenient and efficient method for parameter estimation (Xie, Hu, Wu & Ng, 2007); however, the maximum-likelihood estimation (MLE) method is preferable due to its numerous asymptotic optimality properties. A likelihood function for the arrival and departure process was formulated with the hidden assumption that the combined distribution of any future cumulative numbers of arrivals and departures depends solely on the present values (Wu, Hu, Xie & Ng, 2007). Then, the hidden assumption was relaxed and an MLE method for the joint process was proposed, which can be applied to more general scenarios and can provide more accurate results (Wang, Hu & Liu, 2016). Since we occasionally encounter several difficulties in solving the likelihood function, an alternative parameter estimation algorithm that is based on the EM principle was developed (Wang, 2016). Algorithms that are based on the Bayesian framework were also presented (Wang, Hu & Xie, 2015). In all the research that is discussed above, point estimation and interval estimation are proposed for model parameters of the fault detection process and the correction process in software testing projects, which can be viewed as $M_t/G/\infty$ queueing systems. However, we are more concerned with the performance measures of queueing systems in practical queueing scenarios, such as the expected cumulative numbers of arrivals $m_a(t)$ and departures $m_d(t)$. Statistical inference on these measures has not yet been studied.

Our major contribution in this paper is that we provide a general framework for handling the statistical inference problem in $M_t/G/\infty$ queueing systems given interval censored data. In many cases, we may either not want or not be able to monitor arrivals and departures; in such cases, only the count of items in the queue is observed (Pickands & Stine, 1997). In addition, observers may only record data at fixed time points instead of observing continuously to maintain cost-effectiveness, among other reasons. As a result, interval censored data are commonly used in practical scenarios (Brown et al., 2005; Deng & Mark, 1993; Mandelbaum, Sakov & Zeltyn, 2000; Massey et al., 1996). Although extensive studies have been conducted on statistical inference approaches in similar scenarios for the homogeneous cases, $M/G/\infty$ queueing systems, the dynamics of real systems (Andersen et al., 2019; Dhingra et al., 2018; Liu, 2018; Massey, 2002; Massey et al., 1996; Pender, 2016; Schwarz et al., 2016) have not been considered. In this paper, we study the transient behaviour in the nonhomogeneous cases, the $M_t/G/\infty$ queueing systems, which better accord with practical scenarios. We provide a general MLE method for inferring the arrival rate and service duration in $M_t/G/\infty$ queueing models given interval censored data; this method is applicable to a general service duration distribution G. More importantly, we propose a combination of the bootstrap method and the delta method for inferring the expected cumulative numbers of arrivals and departures, which are indispensable for decision makers in determining the number of required servers. The remainder of the paper is organized as follows: In Section 2, we formulate the problem, describe the maximum-likelihood estimation method and propose a combination of the bootstrap method and the delta method for approximating the confidence intervals of $m_a(t)$ and $m_d(t)$. In Section 3, a simulation study is conducted to study the goodness-of-fit performance of our proposed MLE method. We study the impact of the number of intervals on $M_t/G/\infty$ queueing models with exponential and log-normal service durations. In Section 4, we apply our proposed MLE method on a software testing system. In Section 5, we present the conclusions of this work and discuss possible directions for future study.

2. Maximum-likelihood estimation in $M_t/G/\infty$ queueing systems given interval censored data

In $M_t/G/\infty$ queueing systems, we wish to infer the arrival rate function $\lambda_a(t)$; the cdf G of the service duration given interval censored data; and, consequently, other performance measures of interest, namely, the cumulative numbers of arrivals and departures. In Section 2.1, we review the essential properties of $M_t/G/\infty$ queueing systems, which are vital to the estimation of the performance measures in Section 2.3, and approaches for identifying these properties, which shed light on the formulation of the likelihood function in Section 2.2.

2.1. Performance analysis of $\boldsymbol{M_t/G}/\infty$ queueing systems

In $M_t/G/\infty$ queueing systems, customers do not interact. Because there are infinitely many servers, customers do not interfere with one another; due to the Poisson arrivals, the arrival time of a customer carries no information about the arrival time of the other customers (Eick et al., 1993). Thus, satisfactory analytical results are available for $M_t/G/\infty$ queueing systems. If a system is initially empty, the number of busy servers at time t, namely, $N(t)$, has a Poisson distribution with mean

$$m(t) = \int_0^t \lambda_a(u)[1 - G(t-u)]du \tag{2.1}$$

and the departure process is also a nonhomogeneous Poisson process with mean

$$m_d(t) = \int_0^t \lambda_a(u)G(t-u)du. \tag{2.2}$$

If the system is not initially empty, we can calculate the value of $m(t)$ via the probability generating function (pgf) of $N(t)$; see Keilson and Servi (1994).

The properties above can be obtained via several approaches. A simple approach is to take advantage of a property of Poisson processes: a censored Poisson process is Poisson and the sum of Poisson random variables is a Poisson random variable (Crawford, 1981). In addition, an item departs after it has arrived and its service has been completed; hence, the departure process can be modelled as a delayed arrival process (Xie et al., 2007). In addition to direct approaches for evaluating the probabilities, the relevant theory for infinite-server models is well established in the theory of stochastic point processes and random measures (Massey & Whitt, 1993). The arrival epoch and service duration generate a Poisson random measure on Euclidean space; see Daley and Vere-Jones (1988), Foley (1982), Foley (1986), Prekopa (1958), Rényi (1967) and pp. 26–31 of Serfozo (1990).

The most direct approach is to evaluate the probabilities for various events directly from the properties of the arrival and service duration distributions. This method sheds light on the formulation of the likelihood function given interval censored data. To obtain the distribution of $N(t)$, Hillestad and Carrillo (1980) initially evaluated the conditional probability. Denote by $N_a(t)$ the number of arrivals by time t. The conditional probability $P(N(t) = n|N_a(t) = a)$ can be viewed as the probability of n successes and $a-n$ failures in a independent trials with a success probability of p on each trial, i.e., $P(N(t) = n|N_a(t) = a)$ is a binomial distribution probability, where p denotes the probability that an arbitrary item that arrived prior to time t is still in the system at time t. By the total probability theorem,

$$p = \int_0^t [1 - G(t-u)]\frac{\lambda_a(u)}{m_a(u)}du, \tag{2.3}$$

where

$$m_a(t) = \int_0^t \lambda_a(\tau)d\tau \tag{2.4}$$

denotes the mean value function of the arrival process. Thus,

$$P(N(t) = n|N_a(t) = a) = \binom{a}{n}p^n(1-p)^{a-n} \qquad n = 0, 1, \ldots, a. \tag{2.5}$$

Unconditioning on $N_a(t)$ yields

$$\begin{aligned} P(N(t) = n) &= \sum_{k=n}^{\infty}\binom{a}{n}p^n(1-p)^{a-n}\frac{e^{-m_a(t)}[m_a(t)]^k}{k!} \\ &= \frac{e^{-pm_a(t)}[p \cdot m_a(t)]^n}{n!}. \end{aligned}$$

Therefore, formula (2.1), which is presented at the beginning of this section, is obtained: $N(t)$ has a Poisson distribution with mean

$$m(t) = pm_a(t) = \int_0^t \lambda_a(u)[1 - G(t-u)]du.$$

The property of the departure process can be obtained via a similar approach to the evaluation of the conditional probability (Yang, 1996). The main strategy of this approach is to evaluate the probability by first obtaining the conditional probability, which is a binomial distribution probability. In addition, the probability of success p is obtained via the total probability theorem. The approach of evaluating the conditional probability first can be extended to evaluate the probabilities of more complicated events. In the following section, in which we formulate the likelihood function given interval censored data, we obtain the likelihood function by partitioning it into a series of conditional probabilities, which are Poisson distribution probabilities and binomial distribution probabilities. In addition, the probabilities of successes can be obtained via similar approaches that utilize the total probability theorem.

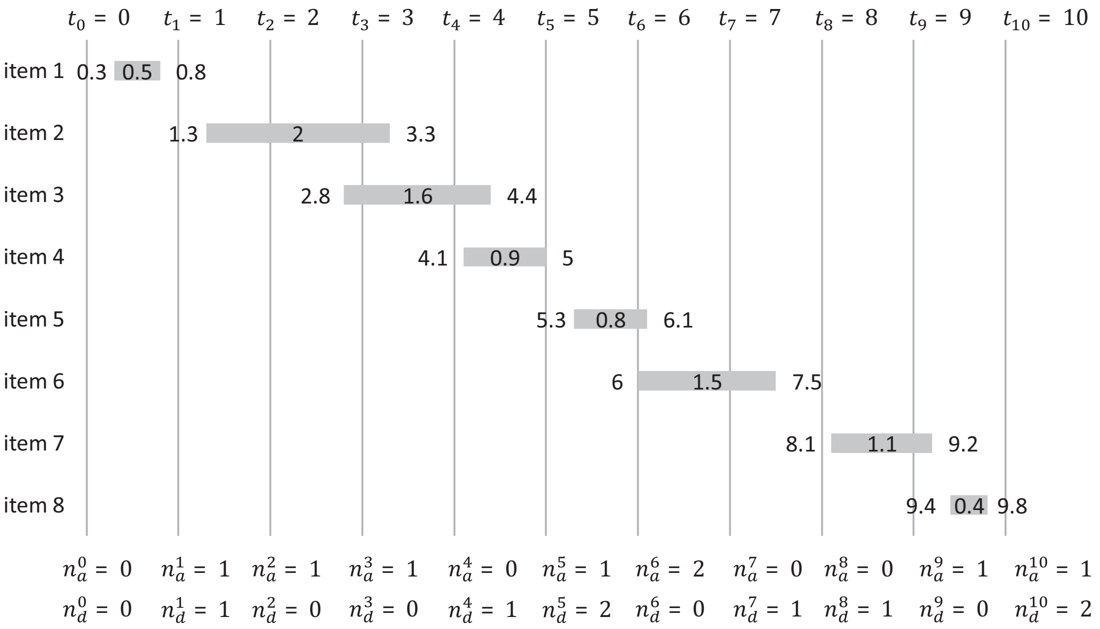
$t_0 = 0$
$t_1 = 1$
$t_2 = 2$
$t_3 = 3$
$t_4 = 4$
$t_5 = 5$
$t_6 = 6$
$t_7 = 7$
$t_8 = 8$
$t_9 = 9$
$t_{10} = 10$
item 1 0.3 0.5 0.8
item 2 1.3 2 3.3
item 3 2.8 1.6 4.4
item 4 4.1 0.9 5
item 5 5.3 0.8 6.1
item 6 6 1.5 7.5
item 7 8.1 1.1 9.2
item 8 9.4 0.4 9.8
$n_a^0 = 0$ $n_a^1 = 1$ $n_a^2 = 1$ $n_a^3 = 1$ $n_a^4 = 0$ $n_a^5 = 1$ $n_a^6 = 2$ $n_a^7 = 0$ $n_a^8 = 0$ $n_a^9 = 1$ $n_a^{10} = 1$
$n_d^0 = 0$ $n_d^1 = 1$ $n_d^2 = 0$ $n_d^3 = 0$ $n_d^4 = 1$ $n_d^5 = 2$ $n_d^6 = 0$ $n_d^7 = 1$ $n_d^8 = 1$ $n_d^9 = 0$ $n_d^{10} = 2$

Fig. 1. Illustration of interval censored data. We observe at time points $t_1, \dots, t_{10}$ and obtain interval censored data $n_1^a, \dots, n_{10}^a, n_1^d, \dots, n_{10}^d$. The start point and end point of each bar represent the arrival and departure epochs, respectively, of each item. The length of each bar represents the service duration of each item.

2.2. Problem formulation

Suppose the system is initially empty. We wish to infer the arrival rate function $\lambda_a(t)$; the cumulative distribution function (cdf) G of the service duration; and, consequently, the expected cumulative numbers of arrivals and departures, $m_a(t)$ and $m_d(t)$, in the $M_t/G/\infty$ queueing system given interval censored data. Suppose the form of the arrival rate function $\lambda_a(t)$, and the family of distributions of the service duration S are known. We wish to infer all the unknown parameters, denoted by parameter vector $\mathbf{\Psi} \in \Theta \subset R^m$, which contains all the parameters in $\lambda_a(t)$ and cdf G. We only observe at fixed time points, which are denoted as t_i $(i = 1, 2, \dots, k)$, and obtain the interval censored data $D = \{(t_i, n_i^a, n_i^d), i = 1, 2, \dots, k\}$, where n_i^a and n_i^d denote the numbers of arrivals and departures, respectively, in time interval $(t_{i-1}, t_i]$ for $i = 1, 2, \dots, k$. For simplicity, we define $t_0 = 0,\ n_0^a = n_0^d = 0$. Fig. 1 shows the interval censored data schematically.

Since both the arrival and departure processes are nonhomogeneous Poisson processes, we can infer $\lambda_a(t)$ from the arrival process and G from the departure process via the approaches of statistical inference for nonhomogeneous Poisson processes with interval censored data. We can also infer both $\lambda_a(t)$ and G from the departure process; however, we do not make full use of the data via this approach. Despite the complexity of the problem, by considering the arrival and departure processes together, we can make more effective use of all the interval censored data. Therefore, we formulate the likelihood function for the joint process of arrival and departure and apply the maximum-likelihood estimation method to infer model parameter vector $\mathbf{\Psi}$.

Similar to the approach that was reviewed in Section 2.1, we wish to obtain the joint likelihood function of the arrival and departure process via conditional probability density functions that are simpler to obtain. Inspired by the partition of a joint density function $f(x_1, x_2, \dots, x_K)$, which is expressed as

$$\begin{aligned} f(x_1, x_2, \dots, x_K) &= f(x_1) f(x_2|x_1) f(x_3|x_1, x_2) \dots f(x_K|x_1, \dots, x_{K-1}) \\ &= \prod_{i=1}^{K} f(x_i|x_1, \dots, x_{i-1}), \end{aligned} \tag{2.6}$$

we can partition the likelihood function in the same way by representing the joint density function as the product of a series of conditional density functions. Substituting $n_1^a, n_2^a, \dots, n_k^a, n_1^d, n_2^d, \dots, n_k^d$ for $x_1, x_2, \dots, x_K$ in formula (2.6) yields

$$\begin{aligned} L(D, \mathbf{\Psi}) &= f(n_1^a) f(n_2^a|n_1^a) f(n_3^a|n_1^a, n_2^a) \dots f\left(n_k^a|n_1^a, \dots, n_{k-1}^a\right) \\ &\quad \times f\left(n_1^d|n_1^a, \dots, n_{k-1}^a, n_k^a\right) \times f\left(n_2^d|n_1^a, \dots, n_k^a, n_1^d\right) \\ &\quad \times f\left(n_3^d|n_1^a, \dots, n_k^a, n_1^d, n_2^d\right) \dots \\ &\quad \times f\left(n_k^d|n_1^a, \dots, n_k^a, n_1^d, n_2^d, \dots, n_{k-1}^d\right) \\ &= \prod_{i=1}^{k} f\left(n_i^a|n_1^a, \dots, n_{i-1}^a\right) f\left(n_i^d|n_1^a, \dots, n_k^a, n_1^d, n_2^d, \dots, n_{i-1}^d\right). \end{aligned} \tag{2.7}$$

Since the arrival process is a nonhomogeneous Poisson process, the number of arrivals in time interval $(t_{i-1}, t_i]$ is independent of the number of arrivals prior to time t_{i-1}. Thus,

$$f\left(n_i^a|n_1^a, \dots, n_{i-1}^a\right) = f\left(n_i^a\right), \quad i = 1, 2, \dots, k. \tag{2.8}$$

Since n_i^d denotes the number of departures in time interval $(t_{i-1}, t_i]$, it is independent of the number of arrivals after time t_i. Thus,

$$\begin{aligned} & f\left(n_i^d|n_1^a, \dots, n_k^a, n_1^d, n_2^d, \dots, n_{i-1}^d\right) \\ &= f\left(n_i^d|n_1^a, \dots, n_i^a, n_1^d, n_2^d, \dots, n_{i-1}^d\right)\ , i = 1, 2, \dots, k. \end{aligned} \tag{2.9}$$

Therefore, the likelihood function can be simplified to

$$L(D, \mathbf{\Psi}) = \prod_{i=1}^{k} f\left(n_i^a\right) f\left(n_i^d|n_1^a, \dots, n_i^a, n_1^d, n_2^d, \dots, n_{i-1}^d\right). \tag{2.10}$$

For simplicity, we use $poi(\cdot; m)$ to represent a Poisson mass function with mean m:

$$poi(j\ ; m) = \frac{m^j}{j!} e^{-m}$$

From the properties of Poisson processes, we obtain

$$f\left(n_i^a\right) = poi\left(n_i^a\ ; m_a(t_i) - m_a(t_{i-1})\right) \quad , i = 1, 2, \dots, k. \tag{2.11}$$

We obtain $f(n_i^d|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d)$ next.

To obtain

$$f\left(n_i^d|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right),\quad i=1,2,\ldots,k,$$

we classify items that departed in time interval $(t_{i-1},t_i]$ into two types: Items of the first type are items that arrived prior to time t_{i-1}, the number of which is denoted as n_i^{d1}. Items of the second type are items that arrived in time interval $(t_{i-1},t_i]$, the number of which is denoted as n_i^{d2}. It follows that

$$n_i^d=n_i^{d1}+n_i^{d2}\quad ,i=1,2,\ldots,k. \tag{2.12}$$

The ranges of n_i^{d1} and n_i^{d2} are (Appendix A):

$$0\le n_i^{d1}\le\min\left(\sum_{l=1}^{i-1}n_l^a-\sum_{l=1}^{i-1}n_l^d,n_i^d\right)\ ,i=1,2,\ldots,k. \tag{2.13}$$

$$0\le n_i^{d2}\le\min\left(n_i^a,n_i^d\right)\qquad ,i=1,2,\ldots,k. \tag{2.14}$$

Based on the properties of Poisson processes, the items of these two types are independent. Since n_i^{d1} and n_i^{d2} are both unknown, we can partition $f(n_i^d|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d)$ as

$$\begin{aligned}
&f\left(n_i^d|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\\
&=f\left(n_i^{d1}+n_i^{d2}|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\\
&=\sum_{j=g_i}^{h_i}f\left(n_i^{d1}=n_i^d-j,n_i^{d2}=j|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\\
&=\sum_{j=g_i}^{h_i}\left[f\left(n_i^{d1}=n_i^d-j|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\right.\\
&\left.\times f\left(n_i^{d2}=j|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\right],i=1,2,\ldots,k.
\end{aligned} \tag{2.15}$$

where j represents all the possible values of n_i^{d2}, which range from g_i to h_i. From (2.13) and (2.14), which express the ranges of n_i^{d1} and n_i^{d2}, we obtain the minimum of j:

$$g_i=\max\left(0,n_i^d-\sum_{l=1}^{i-1}n_l^a+\sum_{l=1}^{i-1}n_l^d\right)\ ,i=1,2,\ldots,k \tag{2.16}$$

and the maximum of j:

$$h_i=\min\left(n_i^a,n_i^d\right)\ ,i=1,2,\ldots,k. \tag{2.17}$$

Whether an item that arrived in time interval $(t_{i-1},t_i]$ departed in time interval $(t_{i-1},t_i]$ depends entirely on its exact arrival epoch and its service duration; the arrivals and departures prior to time t_{i-1} are not relevant. Thus,

$$f\left(n_i^{d2}|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)=f\left(n_i^{d2}|n_i^a\right)\ ,i=1,2,\ldots,k. \tag{2.18}$$

Therefore, formula (2.15) simplifies to

$$\begin{aligned}
&f\left(n_i^d|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\\
&=\sum_{j=g_i}^{h_i}f\left(n_i^{d1}=n_i^d-j|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\\
&\times f\left(n_i^{d2}=j|n_i^a\right)\ ,i=1,2,\ldots,k.
\end{aligned} \tag{2.19}$$

First, we obtain $f(n_i^{d2}|n_i^a)$ $(i=1,2,\ldots,k)$. Similar to formula (2.5) in Section 2.1, $f(n_i^{d2}|n_i^a)$ are binomial distribution probabilities. In time interval $(t_{i-1},t_i]$, n_i^a items arrived and n_i^{d2} of these items departed. $f(n_i^{d2}|n_i^a)$ can be viewed as the probability of n_i^{d2} successes and $n_i^a-n_i^{d2}$ failures in n_i^a independent trials with a success probability of p_i^2 on each trial. Similar to formula (2.3) in Section 2.1, p_i^2 can be formulated using the total probability theorem (Appendix B):

$$p_i^2=[m_a(t_i)-m_a(t_{i-1})]^{-1}\int_{t_{i-1}}^{t_i}G(t_i-y)\lambda_a(y)dy\quad ,i=1,2,\ldots,k. \tag{2.20}$$

For simplicity, we use $bin(\cdot;p,n)$ to represent a binomial mass function with parameters p and n:

$$bin(j\ ;p,n)=\binom{n}{j}(p)^j(1-p)^{n-j}.$$

We have

$$f\left(n_i^{d2}=j|n_i^a\right)=bin\left(j\ ;p_i^2,n_i^a\right)\qquad ,i=1,2,\ldots,k. \tag{2.21}$$

Next, we obtain $f(n_i^{d1}|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d)$ $(i=1,2,\ldots,k)$, which are also binomial distribution probabilities, similar to formula (2.5) in Section 2.1. We define $t_{i0}=\max(t_j:\sum_{l=1}^{j}n_l^a=\sum_{l=1}^{j}n_l^d,\ j=1,\ldots,i-1)$. For time t_{i0}, which is before time t_{i-1}, the number of arrivals is equal to the number of departures prior to time t_{i0}; hence, items that arrived prior to time t_{i0} all departed prior to time t_{i0} and they will not depart in time interval $(t_{i-1},t_i]$. Therefore, we only need to consider items that arrived in time interval $(t_{i0},t_{i-1}]$. $\sum_{l=1}^{i-1}n_l^a-\sum_{l=1}^{i-1}n_l^d$ of them did not depart prior to t_{i-1} and n_1^{d1} of them departed in time interval $(t_{i-1},t_i]$. Thus, $f(n_i^{d1}|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d)$ is the probability of n_i^{d1}successes and $\sum_{l=1}^{i-1}n_l^a-\sum_{l=1}^{i-1}n_l^d-n_i^{d1}$ failures in $\sum_{l=1}^{i-1}n_l^a-\sum_{l=1}^{i-1}n_l^d$ independent trials with a success probability of p_i^1 on each trial. p_i^1 denotes the probability that an item departed in time interval $(t_{i-1},t_i]$ conditioned on its arrival in time interval $(t_{i0},t_{i-1}]$ and its departure after time t_{i-1}. Define p_i as the probability that an item departed prior to time t_{i-1} conditioned on its arrival in time interval $(t_{i0},t_{i-1}]$ and q_i as the probability that an item departed in time interval $(t_{i-1},t_i]$ conditioned on its arrival in time interval $(t_{i0},t_{i-1}]$. We have

$$p_i^1=\frac{q_i}{1-p_i}\quad ,i=1,2,\ldots,k. \tag{2.22}$$

We obtain via the total probability theorem (Appendix B)

$$p_i=[m_a(t_{i-1})-m_a(t_{i0})]^{-1}\int_{t_{i0}}^{t_{i-1}}G(t_{i-1}-y)\lambda_a(y)dy\ ,i=1,2,\ldots,k, \tag{2.23}$$

$$q_i=[m_a(t_{i-1})-m_a(t_{i0})]^{-1}\int_{t_{i0}}^{t_{i-1}}[G(t_i-y)-G(t_{i-1}-y)]\lambda_a(y)dy,\quad i=1,2,\ldots,k, \tag{2.24}$$

and, consequently, p_i^1. Thus,

$$\begin{aligned}
&f\left(n_i^{d1}=n_i^d-j|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d\right)\\
&\quad=bin\left(n_i^d-j\ ;p_i^1,\sum_{l=1}^{i-1}n_l^a-\sum_{l=1}^{i-1}n_l^d\right),i=1,2,\ldots,k.
\end{aligned} \tag{2.25}$$

When $i=1$, there was no arrival prior to time t_{i-1}. Thus, $n_1^{d1}=0$, $f(n_1^d|n_1^a,n_0^d)=f(n_1^{d2}=n_1^d|n_1^a)$. We can simply define $f(n_1^{d1}|n_1^a)=1$, which does not affect the value of the likelihood function. Similarly, if $t_{i0}=t_{i-1}$, then $\sum_{l=1}^{i-1}n_l^a=\sum_{l=1}^{i-1}n_l^d$. No item that arrived prior to t_{i-1} departed after t_{i-1}; hence, $n_i^{d1}=0$. Thus, $f(n_i^d|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d)=f(n_i^{d2}=n_i^d|n_i^a)$. For simplicity, we set $f(n_i^{d1}|n_1^a,\ldots,n_i^a,n_1^d,n_2^d,\ldots,n_{i-1}^d)=1$. Therefore, the joint likelihood function of the arrival and departure process is

$$L(D,\boldsymbol{\Psi}) = \prod_{i=1}^{k}\left\{\left[\sum_{j=g_i}^{h_i} bin(j\ ;p_i^2,n_i^a)\right.\right.$$
$$\left.\times bin\left(n_i^d-j\ ;p_i^1,\sum_{l=1}^{i-1}n_l^a-\sum_{l=1}^{i-1}n_l^d\right)\right]$$
$$\left.\times poi(n_i^a\ ;m_a(t_i)-m_a(t_{i-1}))\right\} \tag{2.26}$$

To obtain the maximum-likelihood estimate of parameter vector $\boldsymbol{\Psi}$,

$$\hat{\boldsymbol{\Psi}} = argmax_{\boldsymbol{\Psi}\in\Theta}L(D,\boldsymbol{\Psi}), \tag{2.27}$$

several optimization methods, such as the Newton–Raphson algorithm and the Nelder–Mead simplex algorithm, can be used. The maximum-likelihood estimate of the departure process may provide appropriate initial parameters in the iterative methods to ensure a globally optimum solution.

The MLE approach is applicable to the general form of the arrival rate function $\lambda_a(t)$, and the general service duration distribution G. The most frequently used parametric model of service is that of exponentially distributed durations (Bertsimas & Doan, 2010):

$$G(s) = 1-e^{-\upsilon s}, \tag{2.28}$$

where $\frac{1}{\upsilon}$ represents the mean of the exponential distribution. In practice, the main “theoretical” justification for its use has been analytical tractability, along with a lack of empirical evidence to the contrary (Gans, Koole & Mandelbaum, 2003). Given formulas (2.20) and (2.22)–(2.24), the expressions of p_i^1 and p_i^2 can be simplified if the forms of $\lambda_a(t)$ and G are simple. Since the exponential service duration has a cdf with a comparatively simple form, the expressions of p_i^1 and p_i^2 can be simplified for special forms of the arrival rate function $\lambda_a(t)$, such as the log-linear arrival rate. The log-linear arrival rate, which is expressed as

$$\lambda_a(t) = e^{\alpha_0+\alpha_1 t},$$

is commonly used in practical scenarios. The log-linear arrival rate represents a scenario in which the arrival rate is monotonically increasing or decreasing and is preferred over a linear model because it is positive for all values of α_0 and α_1, while the linear rate function can be kept positive only by imposing nonlinearity restrictions on α_0 and α_1 and leads to simple statistical procedure (Lewis & Shedler, 1976). When the arrival process has a log-linear arrival rate and the service duration is exponentially distributed, we can obtain the analytical solutions of the integrals in formulas (2.20), (2.23) & (2.24). Via formulas (2.23) & (2.24), p_i and q_i are obtained as

$$\begin{aligned}
p_i &= [m_a(t_{i-1})-m_a(t_{i0})]^{-1}\int_{t_{i0}}^{t_{i-1}} G(t_{i-1}-y)\lambda_a(y)dy\\
&= \left[\frac{e^{\alpha_0}}{\alpha_1}\left(e^{\alpha_1 t_{i-1}}-1\right)-\frac{e^{\alpha_0}}{\alpha_1}\left(e^{\alpha_1 t_{i0}}-1\right)\right]^{-1}\int_{t_{i0}}^{t_{i-1}}\left[1-e^{-\upsilon(t_{i-1}-y)}\right]e^{\alpha_0+\alpha_1 y}dy\\
&= 1+\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i0}-t_{i-1})}}{e^{\alpha_1(t_{i0}-t_{i-1})}-1}\quad (t_{i-1}\neq t_{i0})\quad ,i=1,2,\ldots,k.
\end{aligned}$$

$$\begin{aligned}
q_i &= [m_a(t_{i-1})-m_a(t_{i0})]^{-1}\int_{t_{i0}}^{t_{i-1}}[G(t_i-y)-G(t_{i-1}-y)]\lambda_a(y)dy\\
&= \left[\frac{e^{\alpha_0}}{\alpha_1}\left(e^{\alpha_1 t_{i-1}}-1\right)-\frac{e^{\alpha_0}}{\alpha_1}\left(e^{\alpha_1 t_{i0}}-1\right)\right]^{-1}\\
&\quad\times\int_{t_{i0}}^{t_{i-1}}\left\{\left[1-e^{-\upsilon(t_i-y)}\right]-\left[1-e^{-\upsilon(t_{i-1}-y)}\right]\right\}e^{\alpha_0+\alpha_1 y}dy\\
&= \left[1+\frac{\alpha_1 e^{\upsilon(t_{i-1}-t_i)}}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i0}-t_{i-1})}}{e^{\alpha_1(t_{i0}-t_{i-1})}-1}\right]\\
&\quad-\left[1+\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i0}-t_{i-1})}}{e^{\alpha_1(t_{i0}-t_{i-1})}-1}\right]\\
&= \left(e^{\upsilon(t_{i-1}-t_i)}-1\right)\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i0}-t_{i-1})}}{e^{\alpha_1(t_{i0}-t_{i-1})}-1}\quad (t_{i-1}\neq t_{i0}),
\end{aligned}$$
$i=1,2,\ldots,k.$

Therefore, via formula (2.22), p_i^2 are obtained as

$$\begin{aligned}
p_i^1 &= \frac{q_i}{1-p_i} = \left[\left(e^{\upsilon(t_{i-1}-t_i)}-1\right)\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i0}-t_{i-1})}}{e^{\alpha_1(t_{i0}-t_{i-1})}-1}\right]\\
&\quad\div\left\{1-\left[1+\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i0}-t_{i-1})}}{e^{\alpha_1(t_{i0}-t_{i-1})}-1}\right]\right\}\\
&= 1-e^{\upsilon(t_{i-1}-t_i)}\quad (t_{i-1}\neq t_{i0})\quad ,i=1,2,\ldots,k.
\end{aligned}$$

Similarly, via formula (2.20), p_i^2 are obtained as

$$\begin{aligned}
p_i^2 &= [m_a(t_i)-m_a(t_{i-1})]^{-1}\int_{t_{i-1}}^{t_i} G(t_i-y)\lambda_a(y)dy\\
&= \left[\frac{e^{\alpha_0}}{\alpha_1}\left(e^{\alpha_1 t_i}-1\right)-\frac{e^{\alpha_0}}{\alpha_1}\left(e^{\alpha_1 t_{i-1}}-1\right)\right]^{-1}\int_{t_{i-1}}^{t_i}\left[1-e^{-\upsilon(t_i-y)}\right]e^{\alpha_0+\alpha_1 y}dy\\
&= 1+\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i-1}-t_i)}}{e^{\alpha_1(t_{i-1}-t_i)}-1}\quad ,i=1,2,\ldots,k.
\end{aligned}$$

In this case (log-linear arrival rates paired with exponential service durations), p_i^1 and p_i^2 are converted to simpler forms:

$$p_i^1 = 1-e^{\upsilon(t_{i-1}-t_i)}\qquad (t_{i-1}\neq t_{i0}),\quad i=1,2,\ldots,k. \tag{2.29}$$

$$p_i^2 = 1+\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{(\alpha_1+\upsilon)(t_{i-1}-t_i)}}{e^{\alpha_1(t_{i-1}-t_i)}-1},\quad i=1,2,\ldots,k. \tag{2.30}$$

According to formulas (2.29) & (2.30), p_i^1 and p_i^2 do not depend on the starting point or the end point of each interval; they only depend on the length of each interval. When $t_{i-1}\neq t_{i0}$, p_i^1 does not depend on t_{i0} or parameters α_0 and α_1 in the arrival rate function; it only depends on parameter υ of the service duration. p_i^2 does not depend on parameter α_0; it depends only on parameter α_1 in the arrival rate function. When each interval has the same length Δt, p_i^1 and p_i^2 can be further simplified as:

$$p_i^1 = 1-e^{-\upsilon\Delta t}\quad (t_{i-1}\neq t_{i0})\quad ,i=1,2,\ldots,k. \tag{2.31}$$

$$p_i^2 = 1+\frac{\alpha_1}{\alpha_1+\upsilon}\frac{1-e^{-(\alpha_1+\upsilon)\Delta t}}{e^{-\alpha_1\Delta t}-1}\quad ,i=1,2,\ldots,k. \tag{2.32}$$

If $t_{i-1}\neq t_{i0}$, p_i^1 and p_i^2 do not depend on i, i.e., p_i^1 is the same in every interval, as is p_i^2.

In addition to the exponential distribution, the log-normal distribution has been shown to be a remarkably good fit for the service duration distribution (Brown et al., 2005). The cumulative distribution function (cdf) $G(s)$ of the log-normal distribution is

$$G(s) = \Phi\left(\frac{\ln(s)-\mu}{\sigma}\right), \tag{2.33}$$

where Φ denotes the cumulative distribution function of the standard normal distribution and μ and σ denote the mean and standard deviation of the variable’s natural logarithm. Note that μ denotes the mean of the variable’s natural logarithm; μ does not denote the mean of the service duration. The mean of the service duration is $e^{\mu+\frac{\sigma^2}{2}}$. Unlike the exponential distribution, the log-normal distribution does not have a closed-form expression for cdf G and there are no closed-form expressions for p_i^1 and p_i^2. If the arrival process has a log-linear arrival rate and the service duration has a log-normal distribution, p_i^1 and p_i^2 are expressed as

$$p_i^1 = \frac{q_i}{1-p_i},$$

$$p_i^2 = \frac{\alpha_1}{e^{\alpha_0}(e^{\alpha_1 t_i} - e^{\alpha_1 t_{i-1}})} \int_{t_{i-1}}^{t_i} \Phi\left(\frac{\ln(t_i - y) - \mu}{\sigma}\right) e^{\alpha_0 + \alpha_1 y} dy,$$
$$i = 1, 2, \ldots, k, \tag{2.34}$$

where

$$p_i = \frac{\alpha_1}{e^{\alpha_0}(e^{\alpha_1 t_{i-1}} - e^{\alpha_1 t_{i0}})} \int_{t_{i0}}^{t_{i-1}} \Phi\left(\frac{\ln(t_{i-1} - y) - \mu}{\sigma}\right) e^{\alpha_0 + \alpha_1 y} dy, \tag{2.35}$$

$$q_i = \frac{\alpha_1}{e^{\alpha_0}(e^{\alpha_1 t_{i-1}} - e^{\alpha_1 t_{i0}})} \int_{t_{i0}}^{t_{i-1}} \left[\Phi\left(\frac{\ln(t_i - y) - \mu}{\sigma}\right) - \Phi\left(\frac{\ln(t_{i-1} - y) - \mu}{\sigma}\right)\right] e^{\alpha_0 + \alpha_1 y} dy. \tag{2.36}$$

The analytical solutions of the integrals in formulas (2.34)–(2.36) cannot be obtained as in the exponential service duration case; however, the values of $\Phi(s)$ and $\lambda_a(y) = e^{\alpha_0 + \alpha_1 y}$ are available for all values of s and y. Therefore, the integrals can be solved via numerical integration methods and we can obtain the values of p_i^1 and p_i^2. In Sections 3 & 4, in which the simulation study and the application example will be presented, we will focus on exponentially and log-normally distributed service durations as representatives since these service durations are the two most commonly used service durations for modelling real-life queueing systems. However, our proposed MLE method is applicable to general parametric models and models with other service duration distributions can be analysed via this approach.

2.3. Statistical inference on $\mathbf{m_a}$ and $\mathbf{m_d}$

We consider the expected cumulative numbers of arrivals and departures, m_a and m_d, in addition to the parameter vector $\boldsymbol{\Psi}$, in practical scenarios since the performance measures m_a and m_d provide information about the current scenario. The mean number of busy servers at time t, which is denoted as $m(t)$ and provides information to service providers about the current server utilization and the number of required servers, can be obtained from $m_a(t)$ and $m_d(t)$:

$$m(t) = m_a(t) - m_d(t). \tag{2.37}$$

Point estimates of $m_a(t)$ and $m_d(t)$ at any fixed time t are obtained via formulas (2.4) & (2.2):

$$\hat{m}_a(t) = \int_0^t \hat{\lambda}_a(\tau) d\tau, \tag{2.38}$$

$$\hat{m}_d(t) = \int_0^t \hat{\lambda}_a(u) \hat{G}(t - u) du, \tag{2.39}$$

where estimates $\hat{\lambda}_a(u)$ and $\hat{G}(u)$ are the values of function $\lambda_a(u)$ and $G(u)$ that satisfy $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$. For instance, if the arrival process has a log-linear arrival rate with parameters α_0 and α_1 and the service duration has a log-normal distribution with parameters μ and σ, then the maximum-likelihood estimate is $\hat{\boldsymbol{\Psi}} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\mu}, \hat{\sigma})^T$. Point estimates of $m_a(t)$ and $m_d(t)$ are obtained as follows:

$$\hat{m}_a(t) = \int_0^t \hat{\lambda}_a(\tau) d\tau = \int_0^t e^{\hat{\alpha}_0 + \hat{\alpha}_1 \tau} d\tau = \frac{e^{\hat{\alpha}_0}}{\hat{\alpha}_1}\left(e^{\hat{\alpha}_1 t} - 1\right), \tag{2.40}$$

$$\hat{m}_d(t) = \int_0^t \hat{\lambda}_a(u) \hat{G}(t - u) du = \int_0^t e^{\hat{\alpha}_0 + \hat{\alpha}_1 u} \Phi\left(\frac{\ln(t - u) - \hat{\mu}}{\hat{\sigma}}\right) du. \tag{2.41}$$

However, to obtain the confidence intervals of $m_a(t)$ and $m_d(t)$, it is unreasonable to simply substitute the upper and lower confidence limits of each parameter. When $m_a(t)$ and $m_d(t)$ are not monotonic, the upper confidence limit may be even smaller than the lower confidence limit. Therefore, we propose a combination of the bootstrap method and the delta method for obtaining the approximate confidence intervals.

First, the covariance matrix, which is denoted by V, of $\hat{\boldsymbol{\Psi}}$ is needed. The asymptotic property of MLE has been used in previous studies to obtain V. The asymptotic property still holds under appropriate regularity conditions because $L(D, \boldsymbol{\Psi})$ is a special case of the marginal and conditional likelihood (Wu et al., 2007). Suppose $n \to \infty$. We have

$$\left(\hat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}\right) \to N\left(0, \frac{I(\boldsymbol{\Psi})^{-1}}{n}\right), \quad n \to \infty, \tag{2.42}$$

where $I(\boldsymbol{\Psi})$ is the Fisher information matrix of $\boldsymbol{\Psi}$. The covariance matrix V is obtained as $\frac{I(\boldsymbol{\Psi})^{-1}}{n}$. $I(\boldsymbol{\Psi})$ is typically estimated by the observed Fisher information matrix. However, the form of the likelihood function L is too complex for us to obtain the second derivative of $logL$ analytically. We could obtain the observed Fisher information matrix numerically; however, we would still face difficulties implementing numerical methods and obtaining the estimation error due to the high complexity of L. Therefore, we propose a parametric bootstrap method for estimating the covariance matrix V.

R sets of interval censored data are generated from estimate $\hat{\boldsymbol{\Psi}}$ and R maximum-likelihood estimates, which are denoted as $\boldsymbol{\Psi}_i^*$ $(i = 1, 2, \ldots, R)$, are obtained from each set of the generated interval censored data. The sample covariance matrix of the R estimates can be used as an estimate of V. The empirical approximation is justified by the law of large numbers. If confidence levels of 0.95 and 0.99 are to be used, then it is advisable to set $R = 999$ or higher (Davison & Hinkley, 1997). To generate interval censored data from $\hat{\boldsymbol{\Psi}}$, complete data are generated first. We generate the exact arrival epoch of each item and the service duration of each item separately and obtain the departure epoch of each item accordingly. The generation of the exact epochs from the nonhomogeneous Poisson arrival process in fixed interval $(0, t_k]$ can be reduced to the generation of a Poisson number of order statistics from a fixed density function according to the following result (Lewis & Shedler, 1976, 1979): If $T_1, \ldots, T_M$ denote the arrival epochs of the nonhomogeneous Poisson arrival process and if $m_a(t_k) = M$, then conditional on having observed $M(> 0)$ events in $(0, t_k]$, the T_is are distributed as the order statistics from a sample of size M from the distribution function

$$F(t) = \frac{m_a(t) - m_a(0)}{m_a(t_k) - m_a(0)} \qquad (0 \le t \le t_k). \tag{2.43}$$

The proof is presented in Chapter 2 of Cox and Lewis (1966). The steps that are carried out to estimate the covariance matrix V are as follows:

Step 1. Set $i = 1$ and $R = 1000$.

Step 2. Generate the total number of arrivals M in time interval $(0, t_k]$ based on mean $\hat{m}_a(t_k)$.

Step 3. Generate M order statistics $t_a^1, t_a^2, \ldots, t_a^M$ from cdf

$$F(t) = \frac{m_a(t) - m_a(0)}{m_a(t_k) - m_a(0)} \qquad (0 \le t \le t_k),$$

as the exact arrival epochs of the items in time interval $(0, t_k]$.

Step 4. Generate service durations $s_1, s_2, \ldots, s_M$ of the items based on cdf $\hat{G}(s)$ of the service duration.

Step 5. Obtain the departure epochs of the item $t_d^1, t_d^2, \ldots, t_d^M$, where

$$t_d^j = t_a^j + s_j, \qquad j = 1, 2, \ldots, M.$$

Step 6. Convert the complete data $t_a^1, t_a^2, \ldots, t_a^M, t_d^1, t_d^2, \ldots, t_d^M$ into interval censored data $n_1^a, n_2^a, \ldots, n_k^a, n_1^d, n_2^d, \ldots, n_k^d$, and obtain the MLE $\boldsymbol{\Psi}_i^*$ based on this interval censored data.

Step 7. Set $i = i+1$ and return to Step 2. When $i = R$, proceed to Step 8.

Step 8. Estimate V as

$$\hat{V} = \frac{1}{R-1}\sum_{i=1}^{R}\left(\boldsymbol{\Psi}_i^* - \bar{\boldsymbol{\Psi}}\right)\left(\boldsymbol{\Psi}_i^* - \bar{\boldsymbol{\Psi}}\right)^T, \tag{2.44}$$

where

$$\bar{\boldsymbol{\Psi}} = \frac{1}{R}\sum_{i=1}^{R}\boldsymbol{\Psi}_i^*. \tag{2.45}$$

Using the parametric bootstrap method, we obtain

$$\left(\hat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}\right) \to N\left(0, \hat{V}\right), \qquad n \to \infty. \tag{2.46}$$

Then, the delta method (Cramer, 1999; Davison, 2003) can be applied to obtain the approximate distributions of $m_a(t)$ and $m_d(t)$ at any fixed time t. The delta method is as follows: If

$$\left(\hat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}\right) \to N\left(0, \frac{\sum}{n}\right), \qquad n \to \infty, \tag{2.47}$$

where $\frac{\sum}{n}$ represents the variance of $\hat{\boldsymbol{\Psi}}_n$, then the function f of parameter vector $\boldsymbol{\Psi}$ has a similar asymptotic property,

$$\left[f\left(\hat{\boldsymbol{\Psi}}_n\right) - f(\boldsymbol{\Psi})\right] \to N\left(0, [\nabla f(\boldsymbol{\Psi})]^T \frac{\sum}{n} \nabla f(\boldsymbol{\Psi})\right), \qquad n \to \infty, \tag{2.48}$$

where $\nabla f(\boldsymbol{\Psi})$ denotes the gradient of f with respect to parameter vector $\boldsymbol{\Psi}$. Since $\hat{m}_a$ and $\hat{m}_d$ are the values of functions m_a and m_d given estimate $\hat{\boldsymbol{\Psi}}$, via formula (2.46), where $\hat{V}$ is equivalent to $\frac{\sum}{n}$ in formula (2.47), we can obtain

$$\hat{m}_a(t) - m_a(t) \to N\left(0, [\nabla m_a(t)]^T \hat{V} [\nabla m_a(t)]\right), \quad n \to \infty, \tag{2.49}$$

$$\hat{m}_d(t) - m_d(t) \to N\left(0, [\nabla m_d(t)]^T \hat{V} [\nabla m_d(t)]\right), \quad n \to \infty, \tag{2.50}$$

where $\nabla m_a(t)$ and $\nabla m_d(t)$ denote the gradients of m_a and m_d with respect to parameter vector $\boldsymbol{\Psi}$ at any fixed time t. Since the expressions of $\nabla m_a(t)$ and $\nabla m_d(t)$ still contain parameter vector $\boldsymbol{\Psi}$, when n is large, the approximate covariances can be estimated as $[\nabla \hat{m}_a(t)]^T \hat{V} [\nabla \hat{m}_a(t)]$ and $[\nabla \hat{m}_d(t)]^T \hat{V} [\nabla \hat{m}_d(t)]$, respectively, where $\nabla \hat{m}_a(t)$ and $\nabla \hat{m}_d(t)$ denote the values of $\nabla m_a(t)$ and $\nabla m_d(t)$ that satisfy $\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}$ at any fixed time t. Therefore, we can obtain the approximate $1-\alpha$ confidence intervals for $m_a(t)$ and $m_d(t)$:

$$\left[\hat{m}_a(t) - Z_{1-\frac{\alpha}{2}}\sqrt{\left[\nabla \hat{m}_a(t)\right]^T \hat{V}\left[\nabla \hat{m}_a(t)\right]}, \hat{m}_a(t) + Z_{1-\frac{\alpha}{2}}\sqrt{\left[\nabla \hat{m}_a(t)\right]^T \hat{V}\left[\nabla \hat{m}_a(t)\right]}\ \right], \tag{2.51}$$

$$\left[\hat{m}_d(t) - Z_{1-\frac{\alpha}{2}}\sqrt{\left[\nabla \hat{m}_d(t)\right]^T \hat{V}\left[\nabla \hat{m}_d(t)\right]}, \hat{m}_d(t) + Z_{1-\frac{\alpha}{2}}\sqrt{\left[\nabla \hat{m}_d(t)\right]^T \hat{V}\left[\nabla \hat{m}_d(t)\right]}\right], \tag{2.52}$$

where $Z_{1-\frac{\alpha}{2}}$ denotes quantile $1-\frac{\alpha}{2}$ of the standard normal distribution.

3. Simulation study

To study the goodness-of-fit performance of the proposed MLE method in queueing systems, a simulation study was conducted. We simulate cyclic arrivals since many service facilities, such as call centres and hospitals, follow periodic patterns. For simplicity, we consider the following arrival rate function:

$$\lambda_a(t) = \lambda + Asin\left(\frac{2\pi t}{T_0}\right), \tag{3.1}$$

where λ is the overall mean arrival rate, A is the amplitude of the arrival function and T_0 is its period. Motivated by the many practical cases in which a daily cycle is evident, we set $\lambda = 10$, $A = 5$ and $T_0 = 24$ hours. Our proposed MLE method is applicable to general service duration distributions; here, we study the exponential service duration distribution and the log-normal service duration distribution, which are the two most commonly used service duration distributions for modelling queueing systems in practice (Bertsimas & Doan, 2010; Brown et al., 2005; Gans et al., 2003). For an exponential service duration with cdf

$$G(s) = 1 - e^{-\upsilon s},$$

we set $\upsilon = 2$. For a log-normal service duration with cdf

$$G(s) = \Phi\left(\frac{\ln(s) - \mu}{\sigma}\right),$$

we set $\mu = -1.2$ and $\sigma = 1$. Note that μ here does not denote the mean of the service duration. For comparative purpose, we set the parameters as described above so that the exponential service duration and the log-normal service duration have the same mean of half an hour. We set the total time to $T = 48$ and simulate the complete data in time interval $(0, 48]$ via the same method as in Step 2-Step 5 in the bootstrap method that was proposed in Section 2.3. The exact arrival epochs of the items in time interval $(0, t_k]$, which are denoted as $t_a^1, t_a^2, \ldots, t_a^M$, are generated as M order statistics from cdf

$$F(t) = \frac{2\pi\lambda t + AT_0\left[1 - cos\left(\frac{2\pi t}{T_0}\right)\right]}{2\pi\lambda T + AT_0\left[1 - cos\left(\frac{2\pi T}{T_0}\right)\right]} \quad (0 \le t \le T). \tag{3.2}$$

$t_a^1, t_a^2, \ldots, t_a^M$ can be generated via inverse transform sampling. After obtaining the complete data, we convert them into interval censored data by dividing the total time T into N equal intervals. The impact of the number of intervals on the arrival process for interval censored data has been studied (Massey et al., 1996). We obtain results via point estimation and interval estimation for studying the impact of the number of intervals on the whole queueing system.

3.1. Point estimation

We compare the maximum-likelihood estimates of model parameters for various numbers of intervals with the same total time of $T = 48$. For both the exponential service duration distribution and the log-normal service duration distribution, we simulate 1000 sets of complete data. We divide each set of complete data into N equal intervals and obtain a single set of interval censored data.

Given the interval censored data, the MLE method that was proposed in Section 2.2 is used to obtain the maximum-likelihood estimates. The maximum-likelihood estimate $\hat{\boldsymbol{\Psi}}$ of the parameter vector can be obtained as the maximizer of likelihood function L, which is formulated as in formula (2.26). For the exponential service duration distribution, p_i^1 and p_i^2 in likelihood function L are obtained as

$$p_i^1 = \frac{q_i}{1-p_i},$$

$$p_i^2 = \left\{\lambda(t_i - t_{i-1}) + \frac{AT_0}{2\pi}\left[cos\left(\frac{2\pi t_{i-1}}{T_0}\right) - cos\left(\frac{2\pi t_i}{T_0}\right)\right]\right\}^{-1} \times \int_{t_{i-1}}^{t_i} \left(1 - e^{-\mu(t_i - y)}\right)\left[\lambda + Asin\left(\frac{2\pi y}{T_0}\right)\right]dy, \quad i = 1, 2, \ldots, k, \tag{3.3}$$

where

$$p_i = \left\{\lambda(t_{i-1} - t_{i0}) + \frac{AT_0}{2\pi}\left[cos\left(\frac{2\pi t_{i0}}{T_0}\right) - cos\left(\frac{2\pi t_{i-1}}{T_0}\right)\right]\right\}^{-1} \times \int_{t_{i0}}^{t_{i-1}} \left(1 - e^{-\mu(t_{i-1} - y)}\right)\left[\lambda + Asin\left(\frac{2\pi y}{T_0}\right)\right]dy, \tag{3.4}$$

$$q_i = \left\{\lambda(t_{i-1} - t_{i0}) + \frac{AT_0}{2\pi}\left[cos\left(\frac{2\pi t_{i0}}{T_0}\right) - cos\left(\frac{2\pi t_{i-1}}{T_0}\right)\right]\right\}^{-1} \times \int_{t_{i0}}^{t_{i-1}} \left(e^{-\mu(t_{i-1} - y)} - e^{-\mu(t_i - y)}\right)\left[\lambda + Asin\left(\frac{2\pi y}{T_0}\right)\right]dy. \tag{3.5}$$

For the log-normal service duration distribution, p_i^1 and p_i^2 are obtained as

$$p_i^1 = \frac{q_i}{1-p_i},$$

$$p_i^2 = \left\{\lambda(t_i - t_{i-1}) + \frac{AT_0}{2\pi}\left[cos\left(\frac{2\pi t_{i-1}}{T_0}\right) - cos\left(\frac{2\pi t_i}{T_0}\right)\right]\right\}^{-1} \times \int_{t_{i-1}}^{t_i} \Phi\left(\frac{\ln(t_i - y) - \mu}{\sigma}\right)\left[\lambda + Asin\left(\frac{2\pi y}{T_0}\right)\right]dy, \quad i = 1, 2, \ldots, k, \tag{3.6}$$

where

$$p_i = \left\{\lambda(t_{i-1} - t_{i0}) + \frac{AT_0}{2\pi}\left[cos\left(\frac{2\pi t_{i0}}{T_0}\right) - cos\left(\frac{2\pi t_{i-1}}{T_0}\right)\right]\right\}^{-1} \times \int_{t_{i0}}^{t_{i-1}} \Phi\left(\frac{\ln(t_{i-1} - y) - \mu}{\sigma}\right)\left[\lambda + Asin\left(\frac{2\pi y}{T_0}\right)\right]dy, \tag{3.7}$$

$$q_i = \left\{\lambda(t_{i-1} - t_{i0}) + \frac{AT_0}{2\pi}\left[cos\left(\frac{2\pi t_{i0}}{T_0}\right) - cos\left(\frac{2\pi t_{i-1}}{T_0}\right)\right]\right\}^{-1} \times \int_{t_{i0}}^{t_{i-1}} \left[\Phi\left(\frac{\ln(t_i - y) - \mu}{\sigma}\right) - \Phi\left(\frac{\ln(t_{i-1} - y) - \mu}{\sigma}\right)\right] \times \left[\lambda + Asin\left(\frac{2\pi y}{T_0}\right)\right]dy. \tag{3.8}$$

For computational efficiency, we calculate $\hat{\boldsymbol{\Psi}}$ as the minimizer of $-logL$. Since the form of $logL$ is complex, obtaining the derivative information of $logL$, either analytically or numerically, can be unreliable or time-consuming. Therefore, we apply the Nelder–Mead simplex algorithm (Lagarias, Reeds, Wright & Wright, 1998), which is a derivative-free method that does not use numerical or analytic gradients, to obtain the minimizer $\hat{\boldsymbol{\Psi}}$. In every iteration in the algorithm, the values of the integrals in p_i, q_i and p_i^2 are obtained via the adaptive numerical integration algorithm. The absolute error tolerance is set to 10^{-8} and the relative error tolerance 10^{-6}.

For the 1000 sets of interval censored data, 1000 estimates are obtained via the proposed MLE method. For any set of interval censored data, given estimate $\hat{\boldsymbol{\Psi}}$, values of p_i^1 and p_i^2 are between 0 and 1, corresponding with the definitions of p_i^1 and p_i^2—probabilities of events. In the case of log-normal service durations, the relative errors of the numerical integrations in formulas (3.6)–(3.8) are all smaller than 9×10^{-7}. The numbers of subintervals produced in the subdivision process range from 1 to 12. As $N \to \infty$, the interval censored data become complete data. Given the complete data, we can infer $\lambda_a(t)$ from the arrival process and G from the service duration separately via the MLE method. The sample mean and standard deviation of each parameter over 1000 maximum-likelihood estimates are listed in Table 1 (exponential service duration) and Table 2 (log-normal service duration). $N = \infty$ represents complete data. The means of relative errors (MREs), which are expressed as

$$MRE = \frac{1}{4}\left[\frac{\left|\hat{\lambda} - \lambda\right|}{\lambda} + \frac{\left|\hat{A} - A\right|}{A} + \frac{\left|\widehat{T_0} - T_0\right|}{T_0} + \frac{|\hat{\upsilon} - \upsilon|}{\mu}\right], \tag{3.9}$$

$$MRE = \frac{1}{5}\left[\frac{\left|\hat{\lambda} - \lambda\right|}{\lambda} + \frac{\left|\hat{A} - A\right|}{A} + \frac{\left|\widehat{T_0} - T_0\right|}{T_0} + \frac{|\hat{\mu} - \mu|}{\mu} + \frac{|\hat{\sigma} - \sigma|}{\sigma}\right], \tag{3.10}$$

are used to evaluate the goodness-of-fit performance for models with the exponential service duration distribution (formula (3.9)) and the log-normal service duration distribution (formula (3.10)), respectively. A lower value of MRE corresponds to higher goodness-of-fit performance.

Overall, the sample means of each parameter for various numbers of intervals are close to the true value of each parameter. All MREs are less than 5%; hence, the goodness-of-fit performance is satisfactory. The estimates become more accurate as N increases, especially for parameters that correspond to the service duration. The MREs are low (less than 1%) for models with exponential service durations, even when the number of intervals N is small (8). The log-normal service duration distribution has two parameters and, thus, a more complicated form than the exponential service duration distribution, which has only one parameter. The larger parameter space and the more complex form of the log-normal distribution leads to larger MREs compared to models with exponential service durations. Although the MRE for models with log-normal service durations given complete data is twice the MRE for models with exponential service durations given complete data, our proposed MLE method for models with log-normal service durations still performs well when N is large: almost all MREs for models with log-normal service durations are less than 1% when N is large. A small value of N corresponds to limited information. The smaller N is, the more information is lost. When $N = 8$, we only obtain data every 6 hours, which is a quarter of the period of 24 hours. A substantial amount of information cannot not be obtained from the data; hence, the comparatively high MRE (4.8%) for models with log-normal service durations is reasonable. Our proposed MLE method for models with exponential service durations still performs well despite the loss of information. The sample means of each parameter as N increases are shown in Fig. 2 (exponential service duration) and Fig. 3 (log-normal service duration). As N increases, the sample mean of each parameter approaches the sample mean given complete data.

3.2. Interval estimation

For $T = 48$, we simulate a set of complete data for exponential and log-normal service durations and compare the interval estimates of m_a and m_d for various numbers of intervals N. For each fixed value of N, a set of interval censored data is obtained from the complete data. From the interval censored data, the estimate $\hat{\boldsymbol{\Psi}}$ is obtained via the proposed MLE method. In the cases of exponential service duration distribution, point estimates of $m_a(t)$ and

Table 1
Comparison of parameter estimates using various numbers of intervals N (exponential service duration). SD represents the sample standard deviation.

Number of intervals N	$\hat{\lambda}$		$\hat{A}$		$\widehat{T}_0$		$\hat{\upsilon}$		MRE (%)
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
8	10.00	0.45	5.06	0.71	24.04	0.44	2.02	0.31	0.61
16	10.00	0.45	5.05	0.66	24.04	0.42	2.01	0.21	0.40
24	10.00	0.45	5.05	0.65	24.04	0.41	2.01	0.16	0.39
48	9.99	0.45	5.05	0.65	24.02	0.41	2.00	0.12	0.33
96	10.01	0.44	5.06	0.62	24.01	0.41	2.00	0.10	0.37
120	9.98	0.46	5.03	0.62	24.01	0.39	2.01	0.09	0.36
160	9.98	0.46	5.03	0.61	24.01	0.41	2.01	0.09	0.35
240	9.98	0.45	5.03	0.62	24.01	0.39	2.01	0.09	0.35
320	9.98	0.46	5.03	0.61	24.01	0.39	2.00	0.09	0.26
480	10.00	0.45	5.04	0.61	24.00	0.42	2.00	0.09	0.23
600	9.99	0.45	5.03	0.60	24.00	0.42	2.00	0.09	0.24
960	10.00	0.45	5.04	0.61	24.00	0.42	2.00	0.09	0.23
∞	10.00	0.45	5.04	0.61	24.00	0.42	2.00	0.09	0.24

Table 2
Comparison of parameter estimates using various numbers of intervals $\boldsymbol{N}$ (log-normal service duration). SD represents the sample standard deviation.

Number of intervals N	$\hat{\lambda}$		$\hat{A}$		$\widehat{T}_0$		$\hat{\mu}$		$\hat{\sigma}$		MRE (%)
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	
8	10.00	0.46	5.09	0.69	24.05	0.46	−1.44	0.90	1.02	0.71	4.78
16	10.00	0.46	5.07	0.62	24.04	0.43	−1.21	0.53	0.87	0.54	3.09
24	10.00	0.46	5.06	0.61	24.04	0.42	−1.16	0.41	0.84	0.48	4.20
48	10.01	0.48	5.05	0.61	24.03	0.42	−1.16	0.23	0.92	0.27	2.47
96	10.00	0.46	5.02	0.62	24.02	0.39	−1.19	0.14	0.98	0.14	0.62
120	9.99	0.45	5.02	0.63	24.02	0.40	−1.18	0.12	0.98	0.12	0.81
160	10.02	0.45	5.06	0.63	24.03	0.41	−1.19	0.11	0.99	0.11	0.81
240	9.99	0.45	5.02	0.62	24.02	0.40	−1.18	0.11	0.98	0.11	0.84
320	9.99	0.46	5.03	0.62	24.02	0.41	−1.18	0.12	0.98	0.13	0.88
480	10.01	0.45	5.09	0.62	24.02	0.42	−1.18	0.13	0.98	0.15	1.17
600	10.01	0.45	5.05	0.62	24.03	0.42	−1.18	0.13	0.98	0.16	0.98
960	10.01	0.45	5.09	0.62	24.02	0.42	−1.19	0.14	0.99	0.16	0.64
∞	10.01	0.45	5.09	0.62	24.02	0.42	−1.20	0.05	1.00	0.03	0.48

$m_d(t)$ at any fixed time t given estimate $\hat{\boldsymbol{\Psi}} = (\hat{\lambda}, \hat{A}, \hat{T}_0, \hat{\upsilon})^T$ are obtained as follows:

$$\hat{m}_a(t) = \hat{\lambda}t + \frac{\hat{A}\hat{T}_0}{2\pi}\left[1 - \cos\left(\frac{2\pi t}{\hat{T}_0}\right)\right], \tag{3.11}$$

$$\hat{m}_d(t) = \int_0^t \left[\hat{\lambda} + \hat{A}sin\left(\frac{2\pi u}{\hat{T}_0}\right)\right]\left[1 - e^{-\hat{\upsilon}(t-u)}\right]du. \tag{3.12}$$

Similarly, in the cases of log-normal service duration distribution, point estimates of $m_a(t)$ and $m_d(t)$ at any fixed time t given estimate $\hat{\boldsymbol{\Psi}} = (\hat{\lambda}, \hat{A}, \hat{T}_0, \hat{\mu}, \hat{\sigma})^T$ are obtained as follows:

$$\hat{m}_a(t) = \hat{\lambda}t + \frac{\hat{A}\hat{T}_0}{2\pi}\left[1 - \cos\left(\frac{2\pi t}{\hat{T}_0}\right)\right], \tag{3.13}$$

$$\hat{m}_d(t) = \int_0^t \left[\hat{\lambda} + \hat{A}sin\left(\frac{2\pi u}{\hat{T}_0}\right)\right]\left[\Phi\left(\frac{\ln(t-u) - \hat{\mu}}{\hat{\sigma}}\right)\right]du. \tag{3.14}$$

Given estimate $\hat{\boldsymbol{\Psi}}$, the parametric bootstrap method in Section 2.3 is applied to obtain the covariance matrix estimate $\hat{V}$, and we obtain the 95% confidence intervals for $m_a(t)$ and $m_d(t)$ at any fixed time t via formulas (2.51) & (2.52). In the cases of exponential service duration distribution, $\nabla\hat{m}_a(t)$ and $\nabla\hat{m}_d(t)$ at any fixed time t in formulas (2.51) & (2.52) are obtained as

$$\begin{aligned}\nabla\hat{m}_a(t) &= \left(\frac{\partial m_a(t)}{\partial\lambda}, \frac{\partial m_a(t)}{\partial A}, \frac{\partial m_a(t)}{\partial T_0}\right)^T\Bigg|_{\lambda=\hat{\lambda}, A=\widehat{A}, T_0=\hat{T}_0} \\ &= \left(\begin{matrix} t \\ \frac{T_0}{2\pi}\left[1-\cos\left(\frac{2\pi t}{T_0}\right)\right] \\ \frac{\partial}{\partial T_0}\left\{\lambda t + \frac{AT_0}{2\pi}\left[1-\cos\left(\frac{2\pi t}{T_0}\right)\right]\right\}\end{matrix}\right)\Bigg|_{\lambda=\hat{\lambda}, A=\widehat{A}, T_0=\hat{T}_0},\end{aligned} \tag{3.15}$$

$$\nabla\hat{m}_d(t) = \left(\frac{\partial m_d(t)}{\partial\lambda}, \frac{\partial m_d(t)}{\partial A}, \frac{\partial m_d(t)}{\partial T_0}, \frac{\partial m_d(t)}{\partial\upsilon}\right)^T\Bigg|_{\lambda=\hat{\lambda}, A=\widehat{A}, T_0=\hat{T}_0, \upsilon=\hat{\upsilon}}, \tag{3.16}$$

$\frac{\partial m_a(t)}{\partial T_0}$ can also be obtained analytically; we do not present the analytical result due to its length. The gradient of $m_d(t)$ is difficult to obtain since the form of $m_d(t)$ is complex. Therefore, $\nabla\hat{m}_d(t)$ is obtained via numerical differentiation. In the cases with log-normal service duration distribution, $\nabla\hat{m}_a(t)$ and $\nabla\hat{m}_d(t)$ are obtained via similar approaches.

We compare the differences between the upper and lower confidence limits at time $t = 48$, which are the lengths of the confidence intervals and are denoted as LCs, among various values of N; the results are listed in Table 3. For any value of N, the confidence interval of $m_a(48)$ is slightly longer than that of $m_d(48)$ for both exponential and log-normal service durations. In this simulation study, exponential and log-normal service durations with the same mean value are paired with the same arrival rate. The models with exponential service durations and log-normal service durations do not show substantial differences in terms of LC. The lengths of the confidence intervals as N increases are shown in

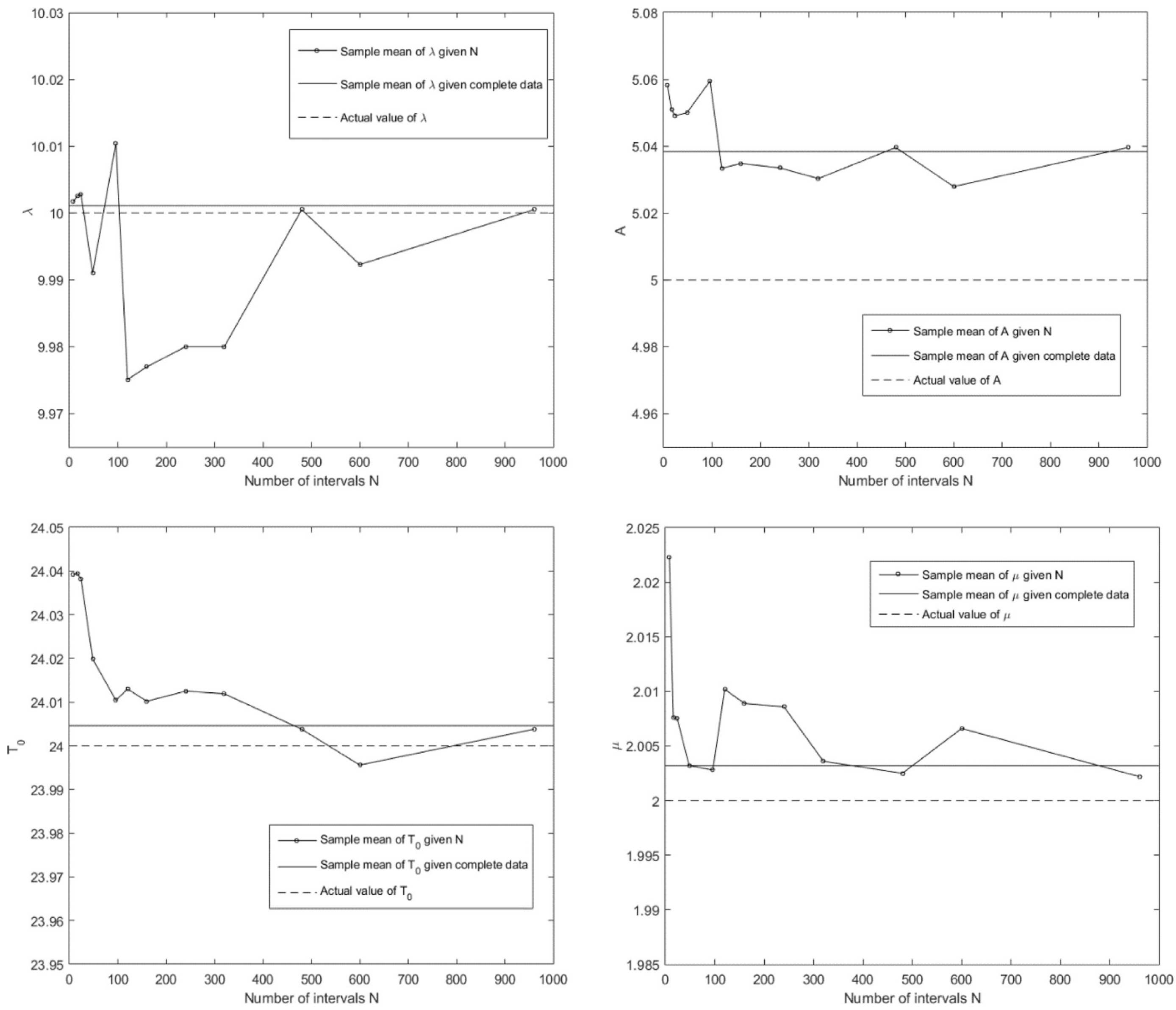

Fig. 2. Sample means of model parameters using various numbers of intervals ***N*** (exponential service duration).

Table 3
Lengths of confidence interval for various numbers of intervals N.

Number of intervals N		8	16	24	48	96	120	160
$m_a(48)$	exponential	85.39	83.91	88.36	86.39	82.21	86.28	86.15
	log-normal	90.61	87.74	87.01	90.92	88.38	84.94	84.86
$m_d(48)$	exponential	84.53	83.13	87.39	85.38	81.35	85.37	85.21
	log-normal	90.36	87.23	86.06	89.74	87.40	83.99	83.81
Number of intervals N		240	320	480	600	960	∞	
$m_a(48)$	exponential	86.24	88.74	86.60	85.70	86.60	86.56	
	log-normal	85.24	87.13	85.67	85.99	85.67	85.61	
$m_d(48)$	exponential	85.31	87.81	85.64	84.72	85.65	85.61	
	log-normal	84.28	86.09	84.76	85.03	84.75	84.67	

Fig. 4 (exponential service duration) and Fig. 5 (log-normal service duration). As N increases, the LC trends of $m_a(48)$ and $m_d(48)$ are the same, for both exponential and log-normal service durations. LC does not change substantially with N. Since no previous study is available for comparison, we compare the LCs for various numbers of intervals N with the LC for complete data. All LCs are in the 95%–105% range of the LC for complete data for the model with exponential service duration and in the 93%–107% range of the LC for complete data for the model with log-normal service duration. The LCs of both $m_a(48)$ and $m_d(48)$ approach the LC for complete data as N increases. These results demonstrate that the confidence intervals that are obtained via our procedure are close to those that are obtained via MLE for complete data, regardless of the number of intervals. Different service duration distributions that have the same mean result in similar confidence intervals; hence, we can obtain similar confidence intervals using

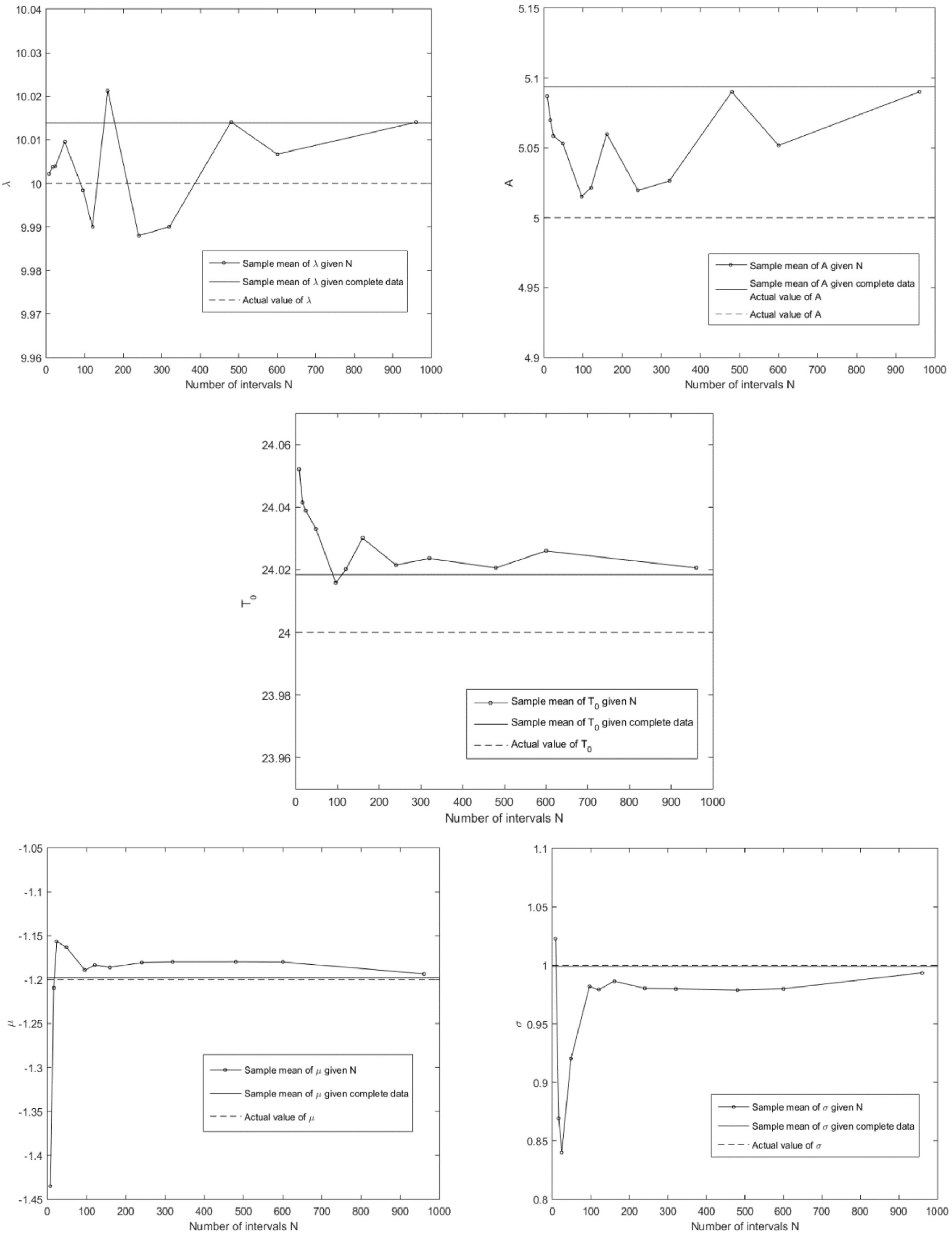

Fig. 3. Sample means of model parameters using various numbers of intervals ***N*** (log-normal service duration).

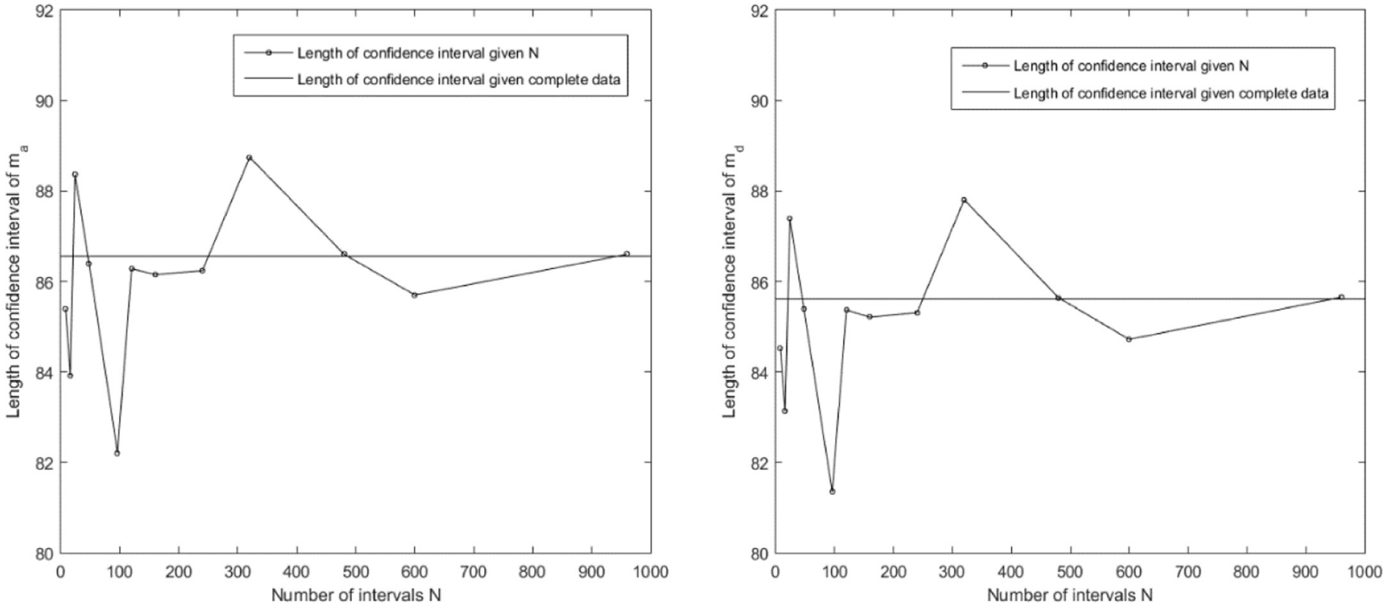

Fig. 4. Lengths of confidence intervals of $\boldsymbol{m_a}(48)$ and $\boldsymbol{m_d}(48)$ for various numbers of intervals $\boldsymbol{N}$ (exponential service duration).

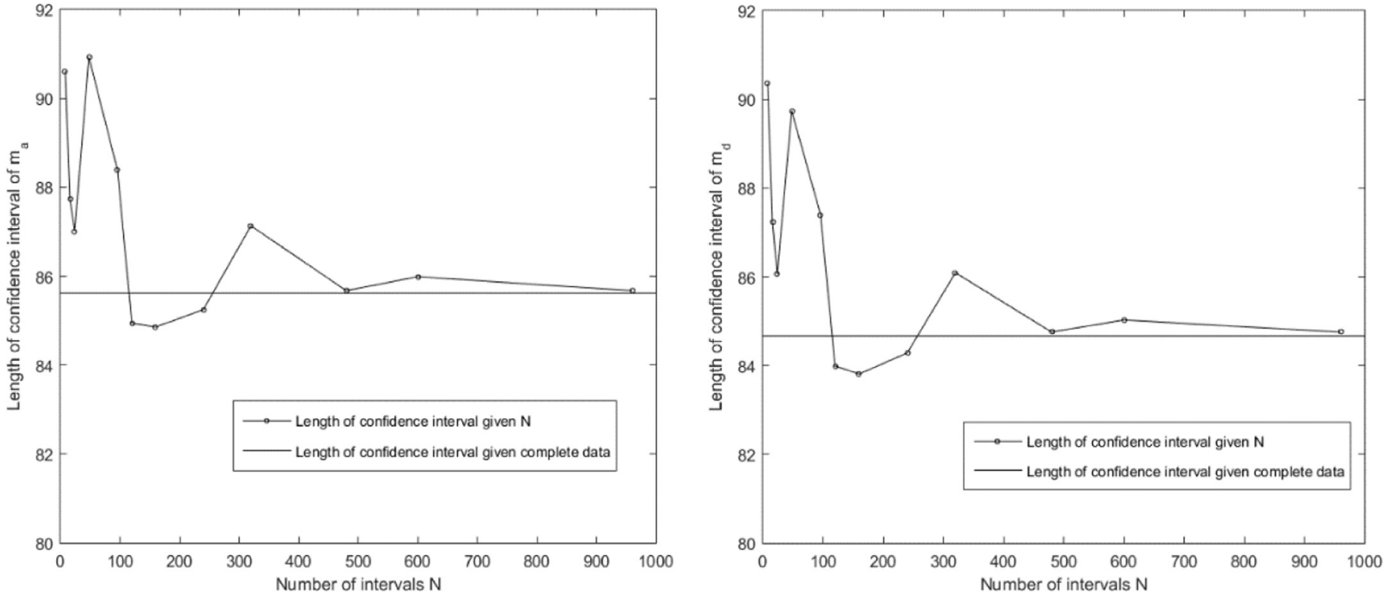

Fig. 5. Lengths of confidence intervals of $\boldsymbol{m_a}(48)$ and $\boldsymbol{m_d}(48)$ for various numbers of intervals $\boldsymbol{N}$ (log-normal service duration).

different service duration distributions to fit against data in practical scenarios.

Our proposed MLE method is applicable to general service duration distributions. The simulation results for models with both exponential and log-normal service durations demonstrate that our proposed MLE method realizes satisfactory goodness-of-fit performance. In the simulation study, the sample size is not large. We only generalize data in time interval $(0, 48]$, which is a realization of only two periods, and the results are already satisfactory; therefore, our proposed MLE method performs well even with a small sample that has limited information. Our procedure enables the estimation of model parameters m_a and m_d without having to keep track of each item from arrival to departure, which reduces the amount of resources that are expended monitoring items and storing data. The point and interval estimations of our proposed MLE method for models with exponential service durations perform extremely well with small *MRE*s and LCs that are similar to the LC for complete data, regardless of the number of intervals. For models with log-normal service durations, the *MRE*s are small overall and more accurate results can be obtained when the number of intervals is large. The LC is also similar to that for complete data, regardless of the number of intervals. These results demonstrate that in practical scenarios, we need not continuously monitor items to obtain complete data; we can observe at fixed time points and obtain interval censored data instead.

4. Application example

We apply our method to a large-scale software testing system (Yang, 1996) to evaluate the goodness-of-fit performance. In a software system, developers assess the system and detect and correct the faults inside before it can be released. Faults are detected when software is executed according to specified test cases. The fault detection process is assumed to be a nonhomogeneous Poisson

Table 4
Fault counts of the P1 system.

Time t_i	Number of detected faults n_i^a	Number of removed faults n_i^d	Time t_i	Number of detected faults n_i^a	Number of removed faults n_i^d
1	2	2	44	79	119
2	0	0	45	183	60
3	0	0	46	129	108
4	1	0	47	176	196
5	2	0	48	106	129
6	2	0	49	62	65
7	3	2	50	49	57
8	12	4	51	99	105
9	8	2	52	43	42
10	2	1	53	47	96
11	11	2	54	174	109
12	2	4	55	179	75
13	0	0	56	229	328
14	1	1	57	65	30
15	0	1	58	66	121
16	6	2	59	40	105
17	4	4	60	54	128
18	0	7	61	31	74
19	5	0	62	103	41
20	3	1	63	63	33
21	2	0	64	107	83
22	2	1	65	59	80
23	6	0	66	69	47
24	7	0	67	78	90
25	5	5	68	62	98
26	20	21	69	97	69
27	34	12	70	58	48
28	46	17	71	65	50
29	21	11	72	53	49
30	55	31	73	139	136
31	61	42	74	60	57
32	58	24	75	50	58
33	60	30	76	70	119
34	60	46	77	31	52
35	109	34	78	44	131
36	76	35	79	63	28
37	110	55	80	36	82
38	86	117	81	38	51
39	73	65	82	28	38
40	63	59	83	18	49
41	36	18	84	17	104
42	120	54	85	25	6
43	112	47	86	8	9

process, which is commonly used and has successfully measured the fault detection process (Okamura, Dohi & Osaki, 2013; Pham, 2000; Schneidewind, 2003; Xie, 1991). After each fault has been detected, a corrective action is performed immediately. Then, the detected fault is removed after developers spend time making the correction. The whole software testing system can be considered as an $M_t/G/\infty$ queueing system, where the fault detection process can be viewed as the arrival process and the fault removal process as the departure process. In practice, it is difficult to specify the exact number of servers for the fault detection and removal process in a large software development environment, where activities are usually carried out in parallel. For example, a member of a fault-removal team may be resolving several faults at the same time. Therefore, one team member can be counted as more than one server from a service perspective. This scenario is practically equivalent to an infinite-server scenario, especially for large systems that involve many team members (Yang, 1996). Although this software testing system does not have infinite servers, our fitting results below demonstrate that the $M_t/G/\infty$ queueing system serves as an efficient model.

In the software testing system P1, the number of faults that are detected and removed in each fixed time interval is available. The data set is listed in Table 4.

The cumulative numbers of arrivals and departures by time t_i are plotted in Fig. 6. Given the shape of the curve for the arrival process in Fig. 6, we apply the inflection S-shaped arrival rate (Ohba, 1984), which is expressed as

$$\lambda_a(t) = \frac{ab(1+c)e^{-bt}}{\left(1+ce^{-bt}\right)^2}, \tag{4.1}$$

paired with exponential service duration and log-normal service duration to fit against the real data.

The maximum-likelihood estimate $\hat{\boldsymbol{\Psi}}$ is obtained via the MLE method that was proposed above. In the case of an exponential service duration distribution, p_i^1 and p_i^2 in likelihood function L, which is presented as formula (2.26), are obtained as

$$p_i^1 = \frac{q_i}{1-p_i},\; p_i^2 = \left[\frac{a\left(1-e^{-bt_i}\right)}{1+ce^{-bt_i}} - \frac{a\left(1-e^{-bt_{i-1}}\right)}{1+ce^{-bt_{i-1}}}\right]^{-1} \times \int_{t_{i-1}}^{t_i} \left(1-e^{-\mu(t_i-y)}\right)\frac{ab(1+c)e^{-by}}{\left(1+ce^{-by}\right)^2}dy, \tag{4.2}$$

where

$$p_i = \left[\frac{a\left(1-e^{-bt_{i-1}}\right)}{1+ce^{-bt_{i-1}}} - \frac{a\left(1-e^{-bt_{i0}}\right)}{1+ce^{-bt_{i0}}}\right]^{-1} \times \int_{t_{i0}}^{t_{i-1}} \left(1-e^{-\mu(t_{i-1}-y)}\right)\frac{ab(1+c)e^{-by}}{\left(1+ce^{-by}\right)^2}\, dy, \tag{4.3}$$

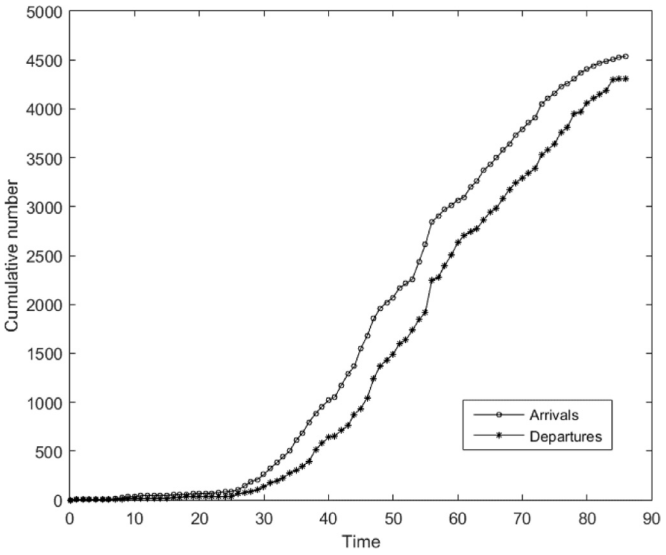

Fig. 6. Cumulative numbers of arrivals and departures in system P1.

$$q_i = \left[\frac{a\left(1-e^{-bt_{i-1}}\right)}{1+ce^{-bt_{i-1}}} - \frac{a\left(1-e^{-bt_{i0}}\right)}{1+ce^{-bt_{i0}}}\right]^{-1} \int_{t_{i0}}^{t_{i-1}} \left(e^{-\mu(t_{i-1}-y)} - e^{-\mu(t_i-y)}\right) \times \frac{ab(1+c)e^{-by}}{\left(1+ce^{-by}\right)^2} dy,$$
$$i = 1, 2, \ldots, k. \tag{4.4}$$

In the case of a log-normal service duration distribution, p_i^1 and p_i^2 are obtained as

$$p_i^1 = \frac{q_i}{1-p_i},$$
$$p_i^2 = \left[\frac{a\left(1-e^{-bt_i}\right)}{1+ce^{-bt_i}} - \frac{a\left(1-e^{-bt_{i-1}}\right)}{1+ce^{-bt_{i-1}}}\right]^{-1} \times \int_{t_{i-1}}^{t_i} \Phi\left(\frac{\ln(t_i-y)-\mu}{\sigma}\right) \frac{ab(1+c)e^{-by}}{\left(1+ce^{-by}\right)^2} dy. \tag{4.5}$$

where

$$p_i = \left[\frac{a\left(1-e^{-bt_{i-1}}\right)}{1+ce^{-bt_{i-1}}} - \frac{a\left(1-e^{-bt_{i0}}\right)}{1+ce^{-bt_{i0}}}\right]^{-1} \times \int_{t_{i0}}^{t_{i-1}} \Phi\left(\frac{\ln(t_{i-1}-y)-\mu}{\sigma}\right) \frac{ab(1+c)e^{-by}}{\left(1+ce^{-by}\right)^2} dy, \tag{4.6}$$

$$q_i = \left[\frac{a\left(1-e^{-bt_{i-1}}\right)}{1+ce^{-bt_{i-1}}} - \frac{a\left(1-e^{-bt_{i0}}\right)}{1+ce^{-bt_{i0}}}\right]^{-1} \times \int_{t_{i0}}^{t_{i-1}} \left[\Phi\left(\frac{\ln(t_i-y)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(t_{i-1}-y)-\mu}{\sigma}\right)\right] \times \frac{ab(1+c)e^{-by}}{\left(1+ce^{-by}\right)^2} dy,$$
$$i = 1, 2, \ldots, k. \tag{4.7}$$

Via the same approach as in Section 3.1, we obtain $\hat{\boldsymbol{\Psi}}$ as the minimizer of $-logL$ and apply the Nelder-Mead simplex algorithm to obtain minimizer $\hat{\boldsymbol{\Psi}}$ since the form of L is complex. The values of the integrals in p_i, q_i and p_i^2 are obtained via the adaptive numerical integration algorithm. The estimates of the model parameters that were obtained via the proposed MLE method are listed in Table 5. In the case of exponential service durations, given estimate $\hat{\boldsymbol{\Psi}}$, the values of p_i^1 are all 0.1563 for $i = 5, 6, \ldots, 86$, and the values of p_i^2 range from 0.0791 to 0.0816. In the case of log-normal service durations, given estimate $\hat{\boldsymbol{\Psi}}$, the values of p_i^1 range from 0.1091 to 0.2099, and the values of p_i^2 range from 0.0688 to 0.0718. The relative errors of the numerical integrations in formulas (4.5)-(4.7) are all smaller than 9×10^{-7}. The numbers of subintervals produced in the subdivision process range from 1 to 9.

In the case of the exponential service duration distribution, given estimate $\hat{\boldsymbol{\Psi}} = (\hat{a}, \hat{b}, \hat{c}, \hat{\upsilon})^T$, the point estimates of $m_a(t)$ and $m_d(t)$ at any fixed time t are obtained as

$$\hat{m}_a(t) = \frac{\hat{a}(1-e^{-\hat{b}t})}{1+\hat{c}e^{-\hat{b}t}}, \tag{4.8}$$

$$\hat{m}_d(t) = \int_0^t \left[\frac{\hat{a}\hat{b}\left(1+\hat{c}\right)e^{-\hat{b}u}}{\left(1+\hat{c}e^{-\hat{b}u}\right)^2}\right]\left[1-e^{-\hat{\upsilon}(t-u)}\right] du. \tag{4.9}$$

In the case of the log-normal service duration distribution, given estimate $\hat{\boldsymbol{\Psi}} = (\hat{a}, \hat{b}, \hat{c}, \hat{\mu}, \hat{\sigma})^T$, the point estimates of $m_a(t)$ and $m_d(t)$ at any fixed time t are obtained as

$$\hat{m}_a(t) = \frac{\hat{a}\left(1-e^{-\hat{b}t}\right)}{1+\hat{c}e^{-\hat{b}t}}, \tag{4.10}$$

$$\hat{m}_d(t) = \int_0^t \left[\frac{\hat{a}\hat{b}\left(1+\hat{c}\right)e^{-\hat{b}u}}{\left(1+\hat{c}e^{-\hat{b}u}\right)^2}\right]\left[\Phi\left(\frac{\ln(t-u)-\hat{\mu}}{\hat{\sigma}}\right)\right] du. \tag{4.11}$$

In both cases, given estimate $\hat{\boldsymbol{\Psi}}$, the parametric bootstrap method is applied to obtain the covariance matrix estimate $\hat{V}$, and we calculate the 95% confidence intervals for $m_a(t)$ and $m_d(t)$ at any fixed time t via formulas (2.51) & (2.52). In the case of the exponential service duration distribution, $\nabla \hat{m}_a(t)$ and $\nabla \hat{m}_d(t)$ at any fixed time t in formulas (2.51) & (2.52) are obtained as

$$\nabla \widehat{m_a}(t) = \left(\frac{\partial m_a(t)}{\partial a}, \frac{\partial m_a(t)}{\partial b}, \frac{\partial m_a(t)}{\partial c}\right)^T \Bigg|_{a=\hat{a}, b=\hat{b}, c=\hat{c}} = \left.\begin{pmatrix} \dfrac{1-e^{-bt}}{1+ce^{-bt}} \\ \dfrac{\partial}{\partial b}\left[\dfrac{a\left(1-e^{-bt}\right)}{1+ce^{-bt}}\right] \\ \dfrac{ae^{-bt}\left(e^{-bt}-1\right)}{\left(1+ce^{-bt}\right)^2} \end{pmatrix}\right|_{a=\hat{a}, b=\hat{b}, c=\hat{c}}, \tag{4.12}$$

$$\nabla \hat{m}_d(t) = \left(\frac{\partial m_d(t)}{\partial a}, \frac{\partial m_d(t)}{\partial b}, \frac{\partial m_d(t)}{\partial c}, \frac{\partial m_d(t)}{\partial \upsilon}\right)^T \Bigg|_{a=\hat{a}, b=\hat{b}, c=\hat{c}, \upsilon=\hat{\upsilon}}, \tag{4.13}$$

$\frac{\partial m_a(t)}{\partial b}$ can also be obtained analytically; we do not present the analytical result due to its length. As in the cases in the simulation study, $\nabla \hat{m}_d(t)$ is obtained via numerical differentiation since the form of $m_d(t)$ is complex. In the case of a log-normal service duration distribution, $\nabla \hat{m}_a(t)$ and $\nabla \hat{m}_d(t)$ are obtained via similar approaches. The point and interval estimates of $m_a(t_i)$ and $m_d(t_i)$ $(i = 1, 2, \ldots, 86)$ versus the observed

Table 5
Comparison of estimates and goodness-of-fit between the proposed MLE method and Yang's method.

Method	Parameter estimates	*MSE*
Yang Inflection S-shaped + Exponential service duration	$\hat{a} = 4713.33$ $\hat{b} = 0.10$ $\hat{c} = 210.26$ $\widehat{\upsilon} = 0.17$	$MSE_a = 10002$ $MSE_d = 9203$ $MSE = 9603$
Proposed MLE Inflection S-shaped + Exponential service duration	$\hat{a} = 4721.17$ $\hat{b} = 0.10$ $\hat{c} = 194.17$ $\widehat{\upsilon} = 0.17$	$MSE_a = 9648$ $MSE_d = 8866$ $MSE = 9257$
Proposed MLE Inflection S-shaped + Log-normal service duration	$\hat{a} = 4733.11$ $\hat{b} = 0.10$ $\hat{c} = 183.10$ $\widehat{\mu} = 1.16$ $\hat{\sigma} = 1.22$	$MSE_a = 9596$ $MSE_d = 7672$ $MSE = 8634$

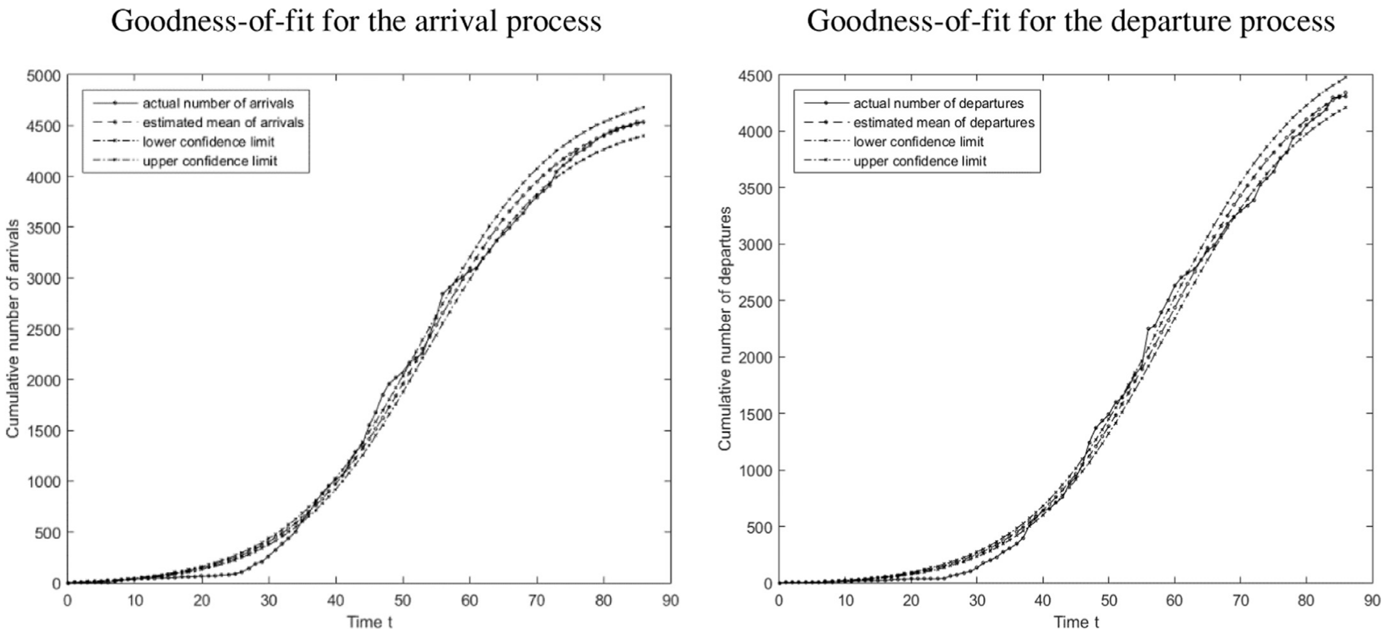

Fig. 7. Goodness-of-fit for inflection S-shaped arrival paired with exponential service duration.

data of both models are plotted in Fig. 7 (exponential service duration) and Fig. 8 (log-normal service duration). The proposed inflection S-shaped model, paired with both exponential service duration and log-normal service duration, fit the data set well for both the arrival process and the departure process. The confidence intervals of $m_a(86)$ and $m_d(86)$ are [4403.47,4672.53] and [4203.47,4484.26], respectively, for the model with exponential service duration and [4406.41,4668.38] and [4144.89,4399.48], respectively, for the model with log-normal service duration. The interval estimates for the two models do not differ significantly, which accords with the results and conclusions of the simulation study.

Since m_a and m_d are the means of the cumulative numbers of arrivals and departures, we use the mean-squared error between mean $m_a(t_i)$ and actual cumulative number of arrivals $\sum_{l=1}^{i} n_l^a$, which is expressed as

$$MSE_a = \frac{1}{k}\sum_{i=1}^{k}\left[\left(m_a(t_i) - \sum_{l=1}^{i} n_l^a\right)^2\right], \tag{4.14}$$

to evaluate the goodness-of-fit performance of the arrival process and

$$MSE_d = \frac{1}{k}\sum_{i=1}^{k}\left[\left(m_d(t_i) - \sum_{l=1}^{i} n_l^d\right)^2\right] \tag{4.15}$$

to evaluate the goodness-of-fit performance of the departure process. In this case, $k = 86$. We use the goodness-of-fit criterion,

$$MSE = \frac{1}{2k}\sum_{i=1}^{k}\left[\left(m_a(t_i) - \sum_{l=1}^{i} n_l^a\right)^2 + \left(m_d(t_i) - \sum_{l=1}^{i} n_l^d\right)^2\right], \tag{4.16}$$

to evaluate the performances of the proposed models, where a lower value of *MSE* corresponds to a higher goodness-of-fit performance. The estimates of the model parameters that were obtained via Yang's (1996) method and our proposed MLE method are listed in Table 5 and the *MSE*s are compared.

Yang's method is limited to models with exponential service durations, while our proposed MLE method is applicable to general service duration distributions and outperforms Yang's model in terms of goodness-of-fit. Under the same assumed model (inflection S-shaped arrival paired with exponential service duration), our proposed MLE method realizes higher goodness-of-fit performance than Yang's model for both the arrival and departure processes. *MSE*, MSE_a, and MSE_d of our proposed MLE method are all 4% smaller than those of Yang's method. In a software testing system, the remaining faults in the system are more difficult to detect and correct over time since the simple faults have already been detected and corrected at an early stage. At the late stage of fault detection and correction, substantial effort and expenditure are required for detecting or correcting even a single fault and a 4% increase in the goodness-of-fit performance can save a large

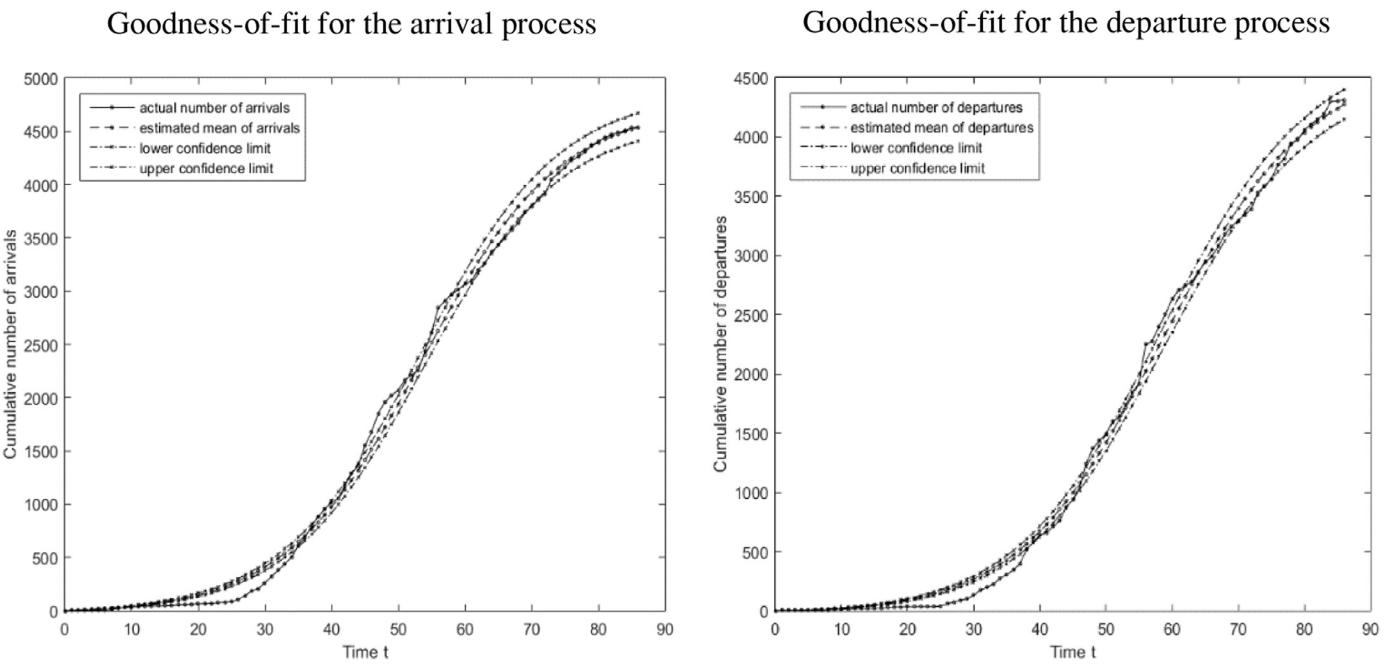

Fig. 8. Goodness-of-fit for inflection S-shaped arrival paired with log-normal service duration.

amount of money. The model with log-normal service duration fits the data set even better: the *MSE* is 10% smaller than that of Yang's method. Our proposed MLE method improves the goodness-of-fit performance; thus, the expenditure can be reduced. Since our proposed MLE method is applicable to general service duration distributions, more complex service duration distributions can be used to fit real data and higher goodness-of-fit performance can be realized. In this application example, the *MSE* of the model with log-normal service durations is 7% smaller than that of the model with exponential service durations. In contrast to the simulation study, where models with exponential service durations realize higher goodness-of-fit performance, the log-normal service duration distribution with 2 parameters is more flexible for fitting real data and realizes higher goodness-of-fit performance compared to the exponential service duration distribution.

The application example demonstrates that the goodness-of-fit performance of our proposed MLE method is satisfactory. We obtain the confidence intervals of m_a and m_d, which have not been obtained in other studies. Our proposed MLE method improves the goodness-of-fit performance and, more importantly, is applicable to general service duration distributions, including the exponential service duration distribution in the previous study (Yang, 1996). Models that have more complex service duration distributions can yield better goodness-of-fit performance than models with the exponential service duration distribution.

5. Conclusions and discussion

We provide a general framework for dealing with the statistical inference problem in $M_t/G/\infty$ queueing systems given interval censored data. We propose an MLE method for inferring model parameters. The method is applicable to a general service duration distribution G. More importantly, we propose a combination of the bootstrap method and the delta method for inferring the expected cumulative numbers of arrivals and departures, which facilitates cost-effective decision-making by service providers. We study exponential and log-normal service duration distributions in both a simulation study and an application example. These service duration distributions are the two most commonly used service durations for modelling queueing systems in practice and have been demonstrated to well fit the service duration distribution. The simulation results for models with both exponential and log-normal service durations demonstrate that our proposed MLE method realizes satisfactory goodness-of-fit performance. As the number of intervals increases, the estimates that are obtained via our proposed MLE approach the estimates that are obtained via MLE from complete data. Our procedure enables one to obtain estimates of model parameters m_a and m_d without having to keep track of each item, which reduces the amount of resources that are expended for monitoring items and storing data. The point and interval estimation approaches in our proposed MLE method for models with exponential service durations perform extremely well, regardless of the number of intervals. For models with log-normal service durations, the results are satisfactory overall and the point estimates are more accurate when the number of intervals is large. The application example demonstrates that the goodness-of-fit performance of our proposed MLE method is satisfactory. The model with a log-normal service duration distribution outperforms the model with an exponential service duration distribution in terms of goodness-of-fit in the application example when the actual family of service duration distributions is unknown. Our proposed MLE method enables more complex service duration distributions to be fit against real data and can yield higher goodness-of-fit performance.

However, we may encounter difficulties in obtaining the maximum-likelihood estimates. The form of the likelihood function is complex; hence, maximum-likelihood estimation is time-consuming in parameter estimation and the complement of the bootstrap method, especially when the number of parameters is large or the arrival rate and the service duration distributions are complex. In future studies, reducing the computational burden could be considered, e.g., by implementing the EM algorithm or utilizing simpler methods to obtain the estimates. Data imputation methods could be explored for coping with the interval censored data. A Bayesian framework could also be utilized to take experience into account.

Acknowledgment

The authors appreciate the reviewers' constructive suggestions and comments, which have helped the authors improve the quality of the revised paper. This work was partly supported by the National Key R&D Programs of the Ministry of Science and Technology of China (2018YFB0704304), National Center for Mathematics and Interdisciplinary Sciences, and Key Laboratory of Systems and Control (CAS).

Appendix A. Ranges of n_i^{d1} and n_i^{d2}

n_i^{d1} denotes the number of items that arrived prior to time t_{i-1} and departed in time interval $(t_{i-1}, t_i]$. It should not exceed the number of items that did not depart prior to time t_{i-1} or the number of items that departed in time interval $(t_{i-1}, t_i]$. Prior to time t_{i-1}, $\sum_{l=1}^{i-1} n_l^a$ items arrived and $\sum_{l=1}^{i-1} n_l^d$ items departed; hence, $\sum_{l=1}^{i-1} n_l^a - \sum_{l=1}^{i-1} n_l^d$ items did not depart prior to time t_{i-1}. Thus,

$$0 \le n_i^{d1} \le \min\left(\sum_{l=1}^{i-1} n_l^a - \sum_{l=1}^{i-1} n_l^d, n_i^d\right), i = 1, 2, \ldots, k. \tag{A.1}$$

Similarly, n_i^{d2} denotes the number of items that arrived in time interval $(t_{i-1}, t_i]$ and departed in time interval $(t_{i-1}, t_i]$, which should not exceed the number of items that arrived in time interval $(t_{i-1}, t_i]$ or the number of items that departed in time interval $(t_{i-1}, t_i]$. Thus,

$$0 \le n_i^{d2} \le \min\left(n_i^a, n_i^d\right), i = 1, 2, \ldots, k. \tag{A.2}$$

Appendix B. Formulations of p_i^1 and p_i^2

(a) Formulation of p_i^2

To obtain p_i^2 $(i = 1, 2, \ldots, k)$, according to the total probability theorem,

$$\begin{aligned} p_i^2 = {} & f(\text{departed in time interval } (t_{i-1}, t_i] | \text{arrived in time} \\ & \text{interval } (t_{i-1}, t_i]) \\ = {} & \int_{t_{i-1}}^{t_i} f(\text{departed in time interval } (t_{i-1}, t_i] | \text{arrived at time } y \\ & \text{in time interval } (t_{i-1}, t_i]) \\ & \times f(\text{arrived at time } y \text{ in time interval } (t_{i-1}, t_i] | \text{arrived in} \\ & \text{time interval } (t_{i-1}, t_i])dy. \end{aligned} \tag{B.1}$$

From the properties of Poisson processes, it follows that

$$\begin{aligned} & P(\text{arrived prior to time } y \text{ in time interval } (t_{i-1}, t_i] \\ & |\text{arrived in time interval } (t_{i-1}, t_i]) \\ & = \frac{P(\text{one arrival in time interval } (t_{i-1}, y], \text{ no arrival in time interval } (y, t_i])}{\mathrm{P}(\text{one arrival in time interval } (t_{i-1}, t_i])} \\ & = \frac{(m_a(y) - m_a(t_{i-1}))e^{-(m_a(y) - m_a(t_{i-1}))}e^{-(m_a(t_i) - m_a(y))}}{(m_a(t_i) - m_a(t_{i-1}))e^{-(m_a(t_i) - m_a(t_{i-1}))}} \\ & = \frac{m_a(y) - m_a(t_{i-1})}{m_a(t_i) - m_a(t_{i-1})}. \end{aligned} \tag{B.2}$$

Hence,

$$\begin{aligned} & f(\text{arrived at time } y \text{ in time interval } (t_{i-1}, t_i] | \text{arrived in time} \\ & \text{interval } (t_{i-1}, t_i]) = \frac{\lambda_a(y)}{m_a(t_i) - m_a(t_{i-1})}. \end{aligned} \tag{B.3}$$

To obtain $f(\text{departed in time interval } (t_{i-1}, t_i] | \text{arrived at time } y \text{ in time interval } (t_{i-1}, t_i])$, we consider an item that arrived at time y $(t_{i-1} < y \le t_i)$. If it departed in time interval $(t_{i-1}, t_i]$, its service duration should not exceed $t_i - y$, as shown in Fig. 9.

Thus,

$$\begin{aligned} & f(\text{departed in time interval } (t_{i-1}, t_i] | \text{arrived at time } y \text{ in time} \\ & \text{interval } (t_{i-1}, t_i]) = \mathrm{G}(t_i - y). \end{aligned} \tag{B.4}$$

It follows from (B.1), (B.3), and (B.4) that

$$p_i^2 = [m_a(t_i) - m_a(t_{i-1})]^{-1} \int_{t_{i-1}}^{t_i} G(t_i - y)\lambda_a(y)dy. \tag{B.5}$$

(a) Formulation of p_i^1

For $i = 1, 2, \ldots, k$, p_i^1 is obtained through p_i and q_i:

$$p_i^1 = \frac{q_i}{1 - p_i}. \tag{B.6}$$

By total probability theorem, we have

$$\begin{aligned} p_i = {} & \int_{t_{i0}}^{t_{i-1}} f(\text{departed prior to time } t_{i-1} | \text{arrived at time } y \\ & \text{in time interval } (t_{i0}, t_{i-1}]) \\ & \times f(\text{arrived at time } y \text{ in time interval } (t_{i0}, t_{i-1}] | \text{arrived} \\ & \text{in time interval } (t_{i0}, t_{i-1}])dy, \end{aligned} \tag{B.7}$$

$$\begin{aligned} q_i = {} & \int_{t_{i0}}^{t_{i-1}} f(\text{departed in time interval } (t_{i-1}, t_i] | \text{arrived at time } y \\ & \text{in time interval } (t_{i0}, t_{i-1}]) \\ & \times f(\text{arrived at time } y \text{ in time interval } (t_{i0}, t_{i-1}] | \text{arrived} \\ & \text{in time interval } (t_{i0}, t_{i-1}])dy. \end{aligned} \tag{B.8}$$

Similar as (B.3), we have

$$\begin{aligned} & f(\text{arrived at time } y \text{ in time interval } (t_{i0}, t_{i-1}] | \text{arrived in time} \\ & \text{interval } (t_{i0}, t_{i-1}]) = \frac{\lambda_a(y)}{m_a(t_{i-1}) - m_a(t_{i0})}. \end{aligned} \tag{B.9}$$

For an item which arrived at time y $(t_{i0} < y \le t_{i-1})$, if it departed prior to time t_{i-1}, its service duration should be not more than $t_{i-1} - y$, as is shown in Fig. 10.

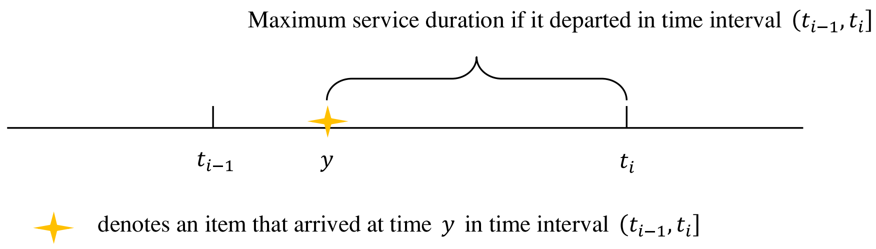
Maximum service duration if it departed in time interval $(t_{i-1}, t_i]$
t_{i-1}
y
t_i
denotes an item that arrived at time y in time interval $(t_{i-1}, t_i]$

Fig. 9. Range of the service duration.

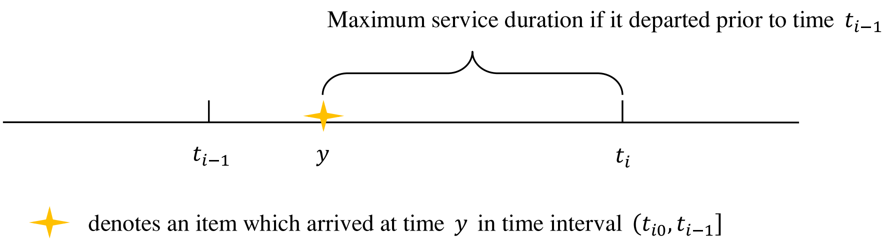
Maximum service duration if it departed prior to time t_{i-1}
t_{i-1}
y
t_i
denotes an item which arrived at time y in time interval $(t_{i0}, t_{i-1}]$

Fig. 10. The range of service duration.

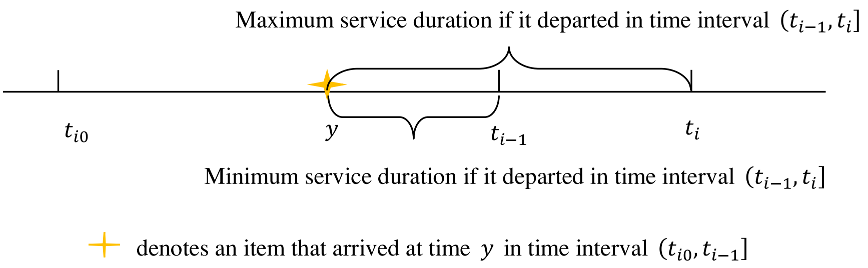
Maximum service duration if it departed in time interval $(t_{i-1}, t_i]$
t_{i0}
y
t_{i-1}
t_i
Minimum service duration if it departed in time interval $(t_{i-1}, t_i]$
denotes an item that arrived at time y in time interval $(t_{i0}, t_{i-1}]$

Fig. 11. Range of the service duration.

Thus,

$$f(\text{departed prior to time } t_{i-1}|\text{arrived at time } y \text{ in time interval } (t_{i0}, t_{i-1}]) = G(t_{i-1} - y). \tag{B.10}$$

If it departed in time interval $(t_{i-1}, t_i]$, its service duration should not exceed $t_i - y$ and should not be less than $t_{i-1} - y$, as shown in Fig. 11.

Thus,

$$f(\text{departed in time interval } (t_{i-1}, t_i]|\text{arrived at time } y \text{ in time interval } (t_{i0}, t_{i-1}]) = G(t_i - y) - G(t_{i-1} - y). \tag{B.11}$$

It follows from (B.7), (B.9), and (B.10) that

$$p_i = [m_a(t_{i-1}) - m_a(t_{i0})]^{-1} \int_{t_{i0}}^{t_{i-1}} G(t_{i-1} - y)\lambda_a(y)dy, \tag{B.12}$$

and according to (B.8), (B.9), and (B.11),

$$q_i = [m_a(t_{i-1}) - m_a(t_{i0})]^{-1} \int_{t_{i0}}^{t_{i-1}} [G(t_i - y) - G(t_{i-1} - y)]\lambda_a(y)dy, \tag{B.13}$$

and, consequently, p_i^1 by (B.6).

References

Aktekin, T. (2014). Call center service process analysis: Bayesian parametric and semi-parametric mixture modeling. *European Journal of Operational Research, 234*(3), 709–719.
Andersen, A. R., Nielsen, B. F., Reinhardt, L. B., & Stidsen, T. R. (2019). Staff optimization for time-dependent acute patient flow. *European Journal of Operational Research, 272*(1), 94–105.
Bertsimas, D., & Doan, X. V. (2010). Robust and data-driven approaches to call centers. *European Journal of Operational Research, 207*(2), 1072–1085.
Bhat, U. N. (1969). Sixty years of queueing theory. *Management Science, 15*(6), B280–B294.
Bingham, N. H., & Pitts, S. M. (1999). Non-parametric estimation for the M/G/∞ queue. *Annals of the Institute of Statistical Mathematics, 51*(1), 71–97.
Blanghaps, N., Nov, Y., & Weiss, G. (2013). Sojourn time estimation in an M/G/∞ queue with partial information. *Journal of Applied Probability, 50*(4), 1044–1056.
Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H. P., Zeltyn, S., et al. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association, 100*(469), 36–50.
Brown, M. (1970). An M/G/∞ estimation problem. *The Annals of Mathematical Statistics, 41*(2), 651–654.
Coolen, F. P. A., & Coolen-Schrijner, P. (2003). A nonparametric predictive method for queues. *European Journal of Operational Research, 145*(2), 425–442.
Cox, D. R., & Lewis, P. A. (1966). *Statistical analysis of series of events*. London: Methuen.
Cramer, H. (1999). *Mathematical methods of statistics*: 9. Princeton: University Press.
Crawford, G. (1981). *Palm's theorem for nonstationary processes*. Santa Monica, CA: The Rand Corporation.
Crawford, G.B. (1977). WRSK/BLSS analysis, the plans-oriented requirements model, Headquarters Pacific Air Forces/OA, March.
Daley, D. J., & Vere-Jones, D. (1988). *An introduction to the theory of point processes*. New York: Springer-Verlag.
Davison, A. C. (2003). *Statistical models*: 11. Cambridge University Press.
Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*: 1. Cambridge university Press.
Deng, L., & Mark, J. W. (1993). Parameter estimation for markov modulated poisson processes via the EM algorithm with time discretization. *Telecommunication Systems, 1*(1), 321–338.
Dhingra, V., Kumawat, G. L., Roy, D., & de Koster, R. (2018). Solving semi-open queuing networks with time-varying arrivals: An application in container terminal landside operations. *European Journal of Operational Research, 267*(3), 855–876.
Eick, S. G., Massey, W. A., & Whitt, W. (1993). The physics of the Mt/G/∞ queue. *Operations Research, 41*(4), 731–742.
Fay, G., Roueff, F., & Soulier, P. (2007). Estimation of the memory parameter of the infinite-source poisson process. *Bernoulli, 13*(2), 473–491.
Foley, R. D. (1982). The nonhomogeneous M/G/∞ queue. *Opsearch, 19*, 40–48.
Foley, R. D. (1986). Stationary poisson departure processes from non-stationary queues. *Journal of Applied Probability, 23*(1), 256–260.
Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call Centers: Tutorial, Review, and research **p**rospects. *Manufacturing & Service Operations Management, 5*(5), 79–141.
Goldenshluger, A. (2016). Nonparametric estimation of the service time distribution in the M/G/∞ queue. *Advances in Applied Probability, 48*(4), 1117–1138.
Goldenshluger, A. (2018). The M/G/∞ estimation problem revisited. *Bernoulli, 24*(4A), 2531–2568.
Green, L. V., & Kolesar, P. J. (1998). A note on approximating peak congestion in Mt/G/∞ queues with sinusoidal arrivals. *Management Science, 44*, S137–S144.
Hall, P., & Park, J. (2004). Nonparametric inference about service time distribution from indirect measurements. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 66*(4), 861–875.
Hillestad, R. J., & Carrillo, M. J. (1980). *Models and techniques for recoverable item stockage when demand and the repair process are nonstationary - Part I*. Santa Monica, CA: The Rand Corporation.
Ibrahim, R., L'Ecuyer, P., Shen, H., & Thiongane, M. (2016). Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *European Journal of Operational Research, 250*(2), 480–492.
Keilson, J., & Servi, L. D. (1994). Networks of non-homogeneous M/G/∞ systems. *Journal of Applied Probability, 31*(A), 151–168.
Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization, 9*(1), 112–147.
Lewis, P. A. W., & Shedler, G. S. (1976). Simulation of nonhomogeneous poisson processes with log linear rate function. *Biometrika, 63*(3), 501–505.
Lewis, P. A. W., & Shedler, G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics, 26*(3), 403–413.

Liu, Y. A. (2018). Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research, 66*(2), 514–534.

Malhotra, R., Dey, D., van Doorn, E. A., & Koonen, A. M. J. (2001). Traffic modeling in a reconfigurable broadband nomadic computing environment. *Internet Quality and Performance and Control of Network Systems, 47*(4), 255–267.

Mandelbaum, A., Sakov, A., & Zeltyn, S. (2000). Empirical analysis of a call center, Technical report, Technion, Israel Institute of Technology.

Massey, W. A. (2002). The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems, 21*(2), 173–204.

Massey, W. A., Parker, G. A., & Whitt, W. (1996). Estimating the parameters of a nonhomogeneous poisson process with linear rate. *Telecommunication Systems, 5*(2), 361–388.

Massey, W. A., & Whitt, W. (1993). Networks of infinite-server queues with nonstationary poisson input. *Queueing Systems, 13*(1), 183–250.

Newell, G. F. (1966). The M/G/∞ queue. *SIAM Journal on Applied Mathematics, 14*(1), 86–88.

Ohba, M. (1984). Software reliability analysis models. *IBM Journal of Research and Development, 28*(4), 428–443.

Okamura, H., Dohi, T., & Osaki, S. (2013). Software reliability growth models with normal failure time distributions. *Reliability Engineering & System Safety, 116*, 135–141.

Palm, C. (1943). Intensity variations in telephone traffic. *Ericsson Technics, 44*, 1–189.

Park, J. (2007). On the choice of an auxiliary function in the estimation. *Computational Statistics & Data Analysis, 51*(12), 5477–5482.

Pender, J. (2016). Risk measures and their application to staffing nonstationary service systems. *European Journal of Operational Research, 254*(1), 113–126.

Pham, H. (2000). *Software reliability*. New York: Springer-Verlag.

Pickands, J., & Stine, R. A. (1997). Estimation for an M/G/∞ queue with incomplete information. *Biometrika, 84*(2), 295–308.

Prekopa, A. (1958). On secondary processes generated by a random point distribution of poisson type. *Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae. Sectio Mathematic, 1*, 153–170.

Rényi, A. (1967). Remarks on the Poisson process. In *Proceedings of the symposium on probability methods in analysis* (pp. 280–286). Berlin Heidelberg: Springer.

Schneidewind, N. F. (2003). Fault correction profiles. In Proceedings of the *14th international symposiumon software reliability engineering*.

Schwarz, J. A., Selinka, G., & Stolletz, R. (2016). Performance analysis of time-dependent queueing systems: Survey and classification. *Omega, 63*, 170–189.

Serfozo, R. F. (1990). *Handbooks in OR and MS*: 2. Amsterdam: Elsevier Science Publishers.

Singhai, R., Joshi, S. D., & Bhatt, R. K. P. (2009). Offered-Load model for pareto inter-arrival network traffic. In *Proceedings of the 2009 IEEE 34th conference on local computer networks.*

Vizarreta, P., Trivedi, K., Helvik, B., Heegaard, P., Blenk, A., Kellerer, W., et al. (2018). Assessing the maturity of SDN controllers with software reliability growth models. *IEEE Transactions on Network and Service Management, 15*(3), 1090–1104.

Wang, L., Hu, Q. P., & Liu, J. (2016). Software reliability growth modelling and analysis with dual fault detection and correction processes. *IIE Transactions, 48*(4), 359–370.

Wang, L. J. (2016). *Studies over Statistical Inference Methods on Software Reliability* Master thesis. Beijing: University of Chinese Academy of Sciences.

Wang, L. J., Hu, Q. P., & Xie, M. (2015). Bayesian analysis for NHPP-based software fault detection and correction processes. In *Proceedings of the 2015 IEEE international conference on industrial engineering and engineering management (IEEM)*.

Wu, Y. P., Hu, Q. P., Xie, M., & Ng, S. H. (2007). Modeling and analysis of software fault detection and correction, process by considering time dependency. *IEEE Transactions on Reliability, 56*(4), 629–642.

Xie, M. (1991). *Software reliability modelling*. Singapore: World Scientific.

Xie, M., Hu, Q. P., Wu, Y. P., & Ng, S. H. (2007). A study of the modeling and analysis of software fault-detection and fault-correction processes. *Quality and Reliability Engineering International, 23*(4), 459–470.

Yang, K. Z. (1996). *An infinite server queueing model for software readiness assessment and related performance measures* (Doctor of Philosophy thesis). Syracuse University.