A Semi-parametric Bayesian Framework for Performance Analysis of Call Centers

Bangxian Wu and Xiaowei Zhang*

Department of Industrial Engineering and Logistics Management
Hong Kong University of Science and Technology

Abstract

Telephone call centers have become an integral part of todays economy. One of the most widely used approaches in practice to assess a call center's performance is the Erlang-A model, whose popularity stems from its analytical tractability. In particular, typical performance measures such as mean customer waiting time and the probability of customer waiting can be calculated explicitly in closed-form. However, recent empirical studies show that the Erlang-A model may be significantly off the chart. In this paper, we build a semi-parametric framework to assess the performance of a call center, which combines statistical learning techniques with the domain knowledge from the existing queueing theory. We also develop a Bayesian inference approach, which utilizes both "input data" (i.e. inter-arrival times, service times) and "output data" (observed performances). Empirical results show that our approach significantly reduces the error in performance assessment while retaining flexibility and analytical tractability.

Keywords: Call centers, Bayesian inference, Erlang-A model, semi-parametric

1 Introduction

Telephone call centers, as the primary contact interface between customers and their service providers, have become an integral part of todays economy and their importance is still growing. Since the labor cost accounts for up to 70% of the overall operating expense of a call center (Gans et al., 2003), from the managerial perspective it is essential to develop an well-designed staffing/scheduling scheme in order to balance agent efficiency and service quality. To that end, a widely used approach is to utilize the queueing-theoretical methods to analyze the queueing dynamics. The typical model primitives (i.e. "input") include the customer arrival process, the service requirement. The "output", on the other hand, is referred to the operational performance measures, such as the average customer waiting time and the fraction of the customers that experience delay before being answered.

A variety of queueing models have been proposed for performance assessment of call centers, among which the Erlang-A model is a very popular one due to its analytical tractability. The assumptions of this model include that the inter-arrival times, the service requirement as well as the patience time are independent exponential random variables. Moreover, the system has n identical agents and infinite capacity and operates with the first-in-first-out (FIFO) discipline. It is well

*Corresponding author. Email: xiaoweiz@ust.hk known that typical performance measures such as average waiting time and probability of waiting can be derived explicitly in closed-form under the Erlang-A model; see, for example, Mandelbaum and Zeltyn (2007) for an extensive survey of this model. There also has been significant effort for analyzing the extensions of the Erlang-A model by relaxing some of the above assumptions. For instance, Mandelbaum and Zeltyn (2004) allows the distribution of the patience time to be arbitrary and Iravani and Balcioğlu (2008) further allows the distribution of the service requirement to be arbitrary. In addition, Green et al. (2007) discusses the setting where both the customer arrival rate and the number of agents are time-varying. However, none of these extensions enjoy the same analytical tractability as the Erlang-A model and only approximation formulas are available for the performance measures of interest.

However, an essential assumption, that is the customers arrive independently, which underlies virtually all the queueing models including the aforementioned ones may not hold in practice as suggested by recent empirical studies; see Jongbloed and Koole (2001) and Avramidis et al. (2004). Hence, it is conceivable that the Erlang-A model and its extensions would yield an inaccurate assessment for the system performance; see for example Brown et al. (2005). One approach to address this issue is to build a more accurate model for the arrival process as in Jongbloed and Koole (2001), Avramidis et al. (2004), and Channouf and L'Ecuyer (2012). Nevertheless, this approach makes the analytical analysis of the queueing dynamics very challenging and the typical way to assess the system performance is system simulation.

Instead of viewing the traditional queueing models as inaccurate and attempting to completely revise them, we take a different perspective in this paper and treat them as building blocks that can *partially* explain the reality, with the discrepancy being interpreted as *modeling error* and modeled non-parametrically. More specifically, we build a non-parametric regression model, in which the explanatory variables are the variables of the queueing model whereas the response variable is the difference between the performance measure implied by the queueing model and that observed.

Besides the semi-parametric framework, our second contribution in this paper is to develop an associated Bayesian inference approach via Markov chain Monte Carlo (MCMC). An intuitive way for parameter estimation in our semi-parametric framework follows a two-stage procedure. Namely, one first estimates the parameters of the queueing model based on the "input data" and then estimates the coefficients of the regression model based on the parameters estimated from the first stage. Note that the "output data" (observed performances) also contains information which could be exploited for the inference of the queueing model. Therefore, we propose a statistical inference approach that can utilize both the input data and the output data in order to simultaneously estimate both the parameters of the queueing model and the coefficients of the regression model.

The rest of the paper is organized as follows. We briefly outline the widely adopted Erlang-A model in Section 2 and introduce our semi-parametric framework for performance analysis of call centers in Section 3, and propose our Bayesian inference approach in Section 4. We apply our framework for a case study in Section 5. Section 6 concludes the paper.

2 Erlang-A Model

Erlang-A is a popular model in call center management because it takes customer abandonment into account and it is analytical tractable. This model assumes that (i) the customer arrival process is Poisson process with constant rate λ; (ii) the service time of each customer has exponential distribution with rate μ; (iii) the time that a customer is willing to wait has exponential distribution with rate θ. Suppose the call center has k agents. Then, the typical performance measures of interest can be expressed in terms of λ, μ, and θ explicitly in closed-form. In particular, define $\rho = \frac{\lambda}{k\mu}$ and the incomplete Gamma function

$$\gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt,$$

then the probability of customer experiencing delay is

$$\mathrm{P}(W > 0) = \frac{A(\frac{n\mu}{\theta}, \frac{\lambda}{\theta})E(k)}{1 + (A(\frac{n\mu}{\theta}, \frac{\lambda}{\theta}) - 1)E(k)}, \tag{1}$$

and the average waiting time is

$$\mathrm{E}[W] = \frac{1}{\theta}\left(\frac{1}{\rho A(\frac{k\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}\right)\frac{A(\frac{k\mu}{\theta}, \frac{\lambda}{\theta})E(k)}{1 + (A(\frac{k\mu}{\theta}, \frac{\lambda}{\theta}) - 1)E(k)}, \tag{2}$$

where

$$A(x, y) = \frac{xe^y}{y^x}\gamma(x, y),$$

$$E(k) = \frac{\frac{(\lambda/\mu)^k}{k!}}{\sum_{i=0}^k \frac{(\lambda/\mu)^i}{i!}}.$$

3 Semi-parametric Framework

As seen in the previous section, typical modeling primitives of a queueing model include the probability distributions of the inter-arrival time of customers, the service requirements, and the patience time. Let Θ denote the set of the known parameters that characterize the above probability distributions. For instance, $\Theta = (\lambda, \mu, \theta)$ for the Erlang-A model. For a given queueing model, we use $g(\Theta)$ to denote the performance measure of interest. Note that the function $g(\cdot)$ can be either exact such as (1) or (2) for the Erlang-A model or approximate for other models. Let $Y = \{Y_i : i = 1, \ldots, n\}$ denote the observed performance measures in various time periods. Our semi-parametric framework for the performance assessment is as follows

$$Y_i = g(\Theta) + f(\Theta) + \epsilon_i, \tag{3}$$

for $i = 1, \ldots, n$, where f is an unknown smooth function and $\{\epsilon_i : i = 1, \ldots, k\}$ is a sequence of i.i.d. normal random variables with mean 0 and variance σ^2. Or similarly,

$$\log Y_i = \log g(\Theta) + f(\Theta) + \epsilon_i. \tag{4}$$

We call (3) the additive formulation and (4) the multiplicative formulation, because the non-parametric part $f(\Theta)$ characterizes the absolute discrepancy $Y - g(\Theta)$ between the queueing model and the observations in the former formulation whereas the relative discrepancy $Y/g(\Theta)$ in the latter.

One then has significant freedom in specifying the form of the non-parametric regression model. For example, $f(\Theta)$ can be modeled as linear regression

$$f(\Theta) = c + \alpha^{\mathsf{T}}\Theta, \tag{5}$$

or more general additive regression

$$f(\Theta) = \beta_0 + \sum_{j=1}^{J} \beta_j \phi(\Theta), \tag{6}$$

where $\{\phi_j : j = 1, \ldots, J\}$ is a family of basis functions and $\{\beta_j : j = 0, \ldots, J\}$ are the regression coefficients to be estimated. Examples of basis functions include polynomials, radial basis functions, or wavelet basis functions; see, for example, Kohn et al. (2001).

4 Bayesian Inference

Assume that in the i^{th} observation time interval, the number of customer arrivals is N_i, the the number of served customers is K_i, and the total agent service time is S_i. We are interested in estimating Θ, the unknown parameters of the queueing model, as well as the other unknown parameters of the regression model in (5) or (6) based on the combined data set $(N, K, S, Y) = \{(N_i, K_i, S_i, Y_i) : i = 1, \ldots, n\}$. With a bit abuse of notation, we still use Θ to denote all the unknown parameters to be estimated.

Clearly, the likelihood function $p(\Theta|N, K, S, Y)$ is not explicitly available so the maximum likelihood estimation approach is not appropriate in our setting. Instead, we follow a Bayesian inference approach to draw samples from the joint posterior distribution $p(\Theta|N, K, S, Y)$. Largely due to the complexity of the Erlang-A formula (1) or (2), it is prohibitively difficult to directly simulate from $p(\Theta|N, K, S, Y)$ so we use the Gibbs sampler to generate random samples the *full conditionals* $p(\Theta_j|\Theta_{-j}, N, K, S, Y)$, where Θ_{-j} is the collection of the unknown parameters except the j^{th} one; see, for example, Geman and Geman (1984) or Gelman et al. (2004) for an extensive treatment on the Gibbs sampler. It turns out that the full conditionals of some unknown parameters can be simulated by constructing conjugate priors, whereas others need to be simulated by the Metropolis-Hastings (see Metropolis and Ulam (1949) and Hastings (1970)) algorithm. We omit the details here.

5 Case Study: An Anonymous Bank in Israel

The data set we use is public and can be downloaded from the Service Enterprise Engineering (SEE) Center, Technion.[1] It contains the complete telephone records of a small telephone call center of an anonymous bank in Israel in 1999. Since there were no Israeli public holidays in November and December in 1999, we use only the data from these two months. Moreover, since the call arrival process on weekends has a completely different pattern, we focus on weekends only. Finally, there are 44 weekdays for us to study.

A well-known feature of the call arrivals is the so-called "time-of-day" effect, and the typical practice in call center management is to model the call arrival process as a inhomogeneous Poisson process with the arrival rate being a piecewise constant function. We consider five half hourly time intervals from 10 am to 14:30 am, each of which has a Poisson arrival rate λ_j, a service rate μ_j, and a patience rate θ_j. So there are 15 unknown parameters of the queuing model. We apply

[1] Website: http://ie.technion.ac.il/serveng

the multiplicative formulation (4) and linear regression (5). In Table 1, we compare our semi-parametric model against typical queueing models in terms of the discrepancy between the model and the observations

Model	Mean	S.D.	25%	50%	75%
Erlang-A	0.499	0.887	-0.017	0.629	1.103
Erlang-A approximation	0.656	0.900	0.105	0.782	1.260
M/GI/k + GI approximation	0.523	0.929	0.011	0.472	1.011
Erlang-A + linear regression	2.20e-4	0.816	-0.463	-0.172	0.602

Table 1: Comparison between the semi-parametric model with typical queuing models. Mean and S.D. refer to the mean and standard deviation of the model residuals. For typical queueing models (the first three models), the model residuals mean $\log Y_i - \log g(\Theta)$; whereas for the semi-parametric model (the last model), the model residuals mean $\log Y_i - \log g(\Theta) - f(\Theta)$. See (4). The last three columns are the quantiles of the residuals.

6 Conclusions

We proposed a robust semi-parametric framework to address the issue of modeling error in call center management and developed a Bayesian inference approach for parameter estimation with the aid of MCMC. We also applied our semi-parametric framework to a real call center and the numerical results showed that by adding a simple linear regression to the traditional queueing model can almost eliminate all the discrepancy between the model and the observations.

References

A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statisticala analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.*, 100 (1):36–50, 2005.

N. Channouf and P. L'Ecuyer. A normal copula model for the arrival process in a call center. *Intl. Trans. in Op. Res.*, 00:1–17, 2012.

N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall/CRC, 2 edition, 2004.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

F. Iravani and B. Balcioğlu. Approximations for the $M/GI/N + GI$ type call center. *QUESTA*, 58:137–153, 2008.

G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Appl. Stochastic Models Bus. Indust.*, 17:307–318, 2001.

R. Kohn, M. Smith, and D. Chan. Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11:313–322, 2001.

A. Mandelbaum and S. Zeltyn. The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the $M/M/n + G$ queue. *OR Spectrum*, 26:377–411, 2004.

A. Mandelbaum and S. Zeltyn. Service engineering in action: the Palm/Erlang-A queue, with applications to call centers. In D. Spath and K.-P. Fähnrich, editors, *Advances in Services Innovations*, pages 17–45. Springer, 2007.

N. Metropolis and S. Ulam. The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44(247):335–341, 1949.