
# MSCI 700

## Estimation of Parameters of Queueing Models

**Name:** Basil K. Ibrahim

**ID:** 20317669

**Semester:** Fall 2009

December 9, 2009

**"While the literature on the stochastic modelling of queues is extensive, estimation and inference concerning the arrival and service rates has, in comparison, received little attention."[3]**

# 1 Introduction: The Hole in the Queueing Literature

It may come to a surprise that queueing theory originated from the need to analyze existing queueing systems and find ways to maximize their functions. For instance, in a bus station, by knowing roughly how many people wait for a bus during rush hour and how often they come, the institution or company responsible for issuing them can determine the optimal number of buses required and their frequency of arrivals for taking all these people, minimizing the risk of having too few or too many passengers. Erlang, who is considered as the father of queueing theory, designed a simple queueing model as an attempt to reduce the waiting times of customers using telephones through Copenhagen Telephone Exchange; as it was in the old days, telephone calls (of an entire city possibly) were transferred by merely a handful of human operators. However, there is no indication (i.e. I have not been able to find any sources on the matter) that this simple model was successfully or even failingly applied.

In fact, as the literature grew and the number of complex queueing models have increased, enriching the field, only a small number of these models are or have been applied in practice. Although many papers boast about the application of their generalized queueing models in various fields such as epidemiology, trafficking and telecommunications, truth of the matter is that very few of these models end up being used by the experts of the field. And so, ironically, the field that was originally aimed to be used to help optimize real-life systems has been more or less, for the last century, oriented in the realm of theory and abstractness.

There are a couple of main reasons why this is the case. First of all, "one has the problem of deciding what model is appropriate from observing the behavior of a system." [5] With many models to choose from and "partly due to a lack of clear statistical procedures for fitting [these] models" [4], it is very difficult to determine the optimal queueing model for the given dataset. This is especially the case when attempting to derive/use goodness-of-fit tests on complex queueing models. In fact, even simple queueing models require some assumptions to be able to make statistical inferences, using the currently available methods in statistics.

The second reason why queueing theory is not as practical as it originally set out to be is because the estimation of the parameters of the queueing model is rarely straightforward and simple; even more problematic than this is calculating them computationally. With complex models, one deals with matrices of significant dimension which makes estimated parameters time consuming to calculate. Although computational power increases, the literature produces more complex models or sets out to make currently existing ones more realistic (e.g. increasing the buffer size); as a consequence, the computational complexity also increases.

This report shows some of the methodologies used for fitting, estimating and computing parameters of particular queueing models found in the literature. Except for one paper written in the sixties, this report is based on recent articles (between 2002 and 2009). I summarize the main findings of these papers in the literature in the report, starting with simple models and proceeding with more complex ones.

## 2 Parameter Estimation in Markovian Population Models [4]

From the start, [4] introduces Markovian populations under a biological context reflecting individuals of different types occupying different sites (e.g. ecosystems). The underlying process is assumed to be a Markov chain with three possible transitions: "the arrival of a new individual (birth or immigration), the departure of an existing individual from the system (death or emigration), or the transfer of an individual from one site to another (migration or predatorprey competition)." The authors then outline the methodology of estimating the parameter vectors under a general context and under a specific one: using density-dependent models (specific models that are indexed by a population ceiling).

In the general case, the authors use MLE (Maximum Likelihood Estimation) and argue that for computational optimization, one can use a method known as CE (Cross-Entropy). For density-dependent models, [4] shows that the process converges to a Verhulst model, which is a well established model in the literature, making estimation straightforward. For large values of the population ceiling, CLT begins to take effect and the authors argue that the process is asymptotically normally distributed converging to a one dimensional OU (Ornstein-Uhlenbeck) process. This, coupled with using the CE method for computational optimization makes parameter estimation quite feasible.

Feasibility, however, becomes a serious issue if one wishes to estimate the parameters under the general context as the dimensionality of the process increases. Yet, with increasing computer power and improvements in computational algorithms, one is optimistic that such models will become practically powerful enough to serve experts certain fields. Also, it is possible that the literature will produce a more efficient computational algorithm to reduce computational complexity, allowing models to be realistic.

## 3 Estimating Parameters for An M/M/c Model [3]

Written by the same authors in the previous section, [3] derives an OU approximation to estimate the parameters of the queue ($\lambda$ and $\mu$) for a large number of servers ($c > 40$). As mentioned in [4], for the density-dependent models case, one can show that the "density process" (the queueing process dependent on the number of servers) is asymptotically normally distributed approximately following an OU process. One then uses log-likelihood of this process and implements the CE algorithm to estimate $\lambda$ and $\mu$. As a consequence of using the OU approximation, one can also estimate the traffic intensity by calculating either the average number of customers at successive (not necessarily equal) time points in the sample or the variance (i.e. $\overline{m} = \hat{\rho}$ and $\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2 = \frac{\hat{\rho}}{c}$, where $\rho = \lambda/\mu$).

Through simulations the authors conclude that estimation accuracy is low for small values of $c$ large sample time intervals. Of the two mentioned examples in the paper, one of them yielded very accurate estimates while the other had poor estimates (they calculated the Fisher information matrix in order to analyze the error bounds of the estimates). In the latter example, the authors recommend using a reflected OU process instead. In general, the model will have greater practical use if the literature can produce methodologies for accurately estimating the model parameters using smaller values of $c$ or for finite queueing buffers. The next section generalizes this model but fixes $c$ to the value of one.

# 4 Birth-and-Death Queueing Models [5]

Although this paper was written in the sixties, I find it as the most attractive one I have read so far because of the explicit solutions it contains and because it is the only paper I have found that includes hypothesis tests of the parameters. However, the main reason behind this explicitness can be attributed to the simplicity of the queueing model: BD (birth-and-death) single server queues. Notwithstanding this, I consider this paper as a must-read for those who are interested in applying statistical inference on queueing models.

A model at state $E_j$ has an arrival rate $\lambda_j$ and a service rate $\mu_j$, where $j$ represents the number of customer in the system (including the one being served). As a result, the log-likelihood function is calculated to be:

$\ell(\underline{\theta}) = \sum_{j=0}^{\infty} u_j ln(\lambda_j) + \sum_{j=1}^{\infty} d_j ln(\mu_j) - \sum_{j=0}^{\infty} \gamma_j(\lambda_j + \mu_j)$, where

$\underline{\theta} = (\lambda_0, \lambda_1, \ldots, \mu_1, \mu_2, \ldots)$ (note that the queue can also be finite), $u_j$ is the number of upward transitions from state $E_j$ to $E_{j+1}$, $d_j$ is the number of downward transitions from $E_j$ to $E_{j-1}$ and $\gamma_j$ is the total time spent in state $E_j$. With $u_j$, $d_j$ and $\gamma_j$ given from empirical data, one simply uses MLE to estimate the elements of $\underline{\theta}$.

Next to estimating $\underline{\theta}$, the author derives estimate of the elements in the variance-covariance matrix of the parameters, denoted by $\sigma(\hat{\underline{\theta}})$. The author elaborates the importance of estimating the components of this matrix under three specific BD models. Basically, after estimating $\underline{\theta}$, one can conduct a goodness-of-fit of the parameters (see whether the $H_0$: $\theta = \theta_0$ is true), which requires the calculation of $\sigma(\hat{\underline{\theta}})$. One then calculates a log-likelihood ratio and compares it to a non-central chi-squared distribution. In all three models, the calculation of the parameters are straightforward and explicit.

The beauty of this BD queueing model is that it does not require a specific distribution of arrival or service, simply that the "mean recurrence time for each state is finite," and that there is only one server. The next section relaxes the BD assumption and uses QBD (quasi-BD) processes; specifically MAPs (Markov arrival processes) and BMAPs (Batch Markov Arrival Processes). Although the analysis assumes only one server, it is possible to extend it to multiple ones; however, the model will become more complex to analyze.

# 5 Estimating MAPs and BMAPs [1] and [2]

As mentioned in [2]: MAP "is one of the most general classes of stochastic counting processes that contains most of the commonly used arrival processes such as the Poisson process, the phase-type (PH) renewal process, and the Markov-modulated Poisson process (MMPP). Moreover, MAP is known to be dense[,] so it can approximate an arbitrary stochastic point process to a given degree of accuracy."

The main issue in estimating the MAP parameters is that for a MAP with $m$ phases, one has to estimate up to $2m^2 + m$ parameters, which leads to computational problems for large values of $m$. One can use two different approaches in estimating these parameters: moment-based and likelihood-based. The moment-based approach determines the parameters required to fit theoretical moments to the ones obtained from the sample; the likelihood-based approach, as shown previously, uses ML to estimate the parameters of the MAP. The main issues is that this involves large-scale matrix computations.

As a result, one uses an algorithm known as the EM (expectation-maximization) algorithm to compute the MLEs more effectively. As [2] states, the EM algorithm "is a statistical framework to compute MLEs under incomplete data and is particularly useful for stochastic models with many parameters." Note that we have incomplete data here because we only see the arrival times as empirical data, not the phases [1].

As mentioned in [1], the EM algorithm can used on a BMAP (Batch Markov Arrival Process) which is a superset of MAPs (as the arrival sizes can be greater than one now). An even simpler estimation procedure (labelled: "A Simpler Estimation Procedure" in [1]) can be used to estimate BMAPs, provided that the number of phases are known and are at least two. This procedure is as follows:

1. From the inter-arrival times, estimate $\hat{D}_0$ using the EM algorithm.
2. Estimate the probability distribution of every empirical arrival time being in a particular phase using discriminant analysis.
3. Calculate all $\hat{D}_n$'s, $n \geq 1$ from first and second steps.

The author of [1] even generalizes the use of this simpler procedure from BMAPs to HMMs (Hidden Markov Models), which are a very general class of Markov chains. As for [2], the authors outlining the EM algorithm for MAPs under the case of having group data.

This assumption has practical significance because often "in practice, only group data is available, as the exact arrival times may be unknown but are grouped into bins." The authors in [2] argue that the methods used for estimating the parameters of grouped data are completely different (from a statistical sense) than for non-grouped data, especially due to the fact that there is loss of information by not knowing the exact arrival times of customers.

The main idea of the EM algorithm for grouped data is to begin with an initial guess of the estimates and then feed these values into an improving function (similar to the Newton-Raphson

method) and iterate until an acceptable level of convergence. According to [2], an initial guess of the algorithm can be obtained using what is known as a $k$-means algorithm which divides all empirical arrivals into different $k$ classes based on inter-arrival times and assuming that the arrivals in each class occur in the same phase (i.e. one phase transition between arrivals). The authors also interestingly remark that one can also determine the optimal number of phases of a MAP by finding under which phase the MAP gives the smallest AIC (Akaike's Information Criterion) value.

A special case of MAPs, the authors use an approximate EM algorithm for MMPPs, which have less computational complexity than the MAP case. However, this method has its restrictions, such as losing accuracy the larger the inter-arrival times are, and assuming at "most one phase transition in each time interval."

Therefore, there has been some progress in MAPs and more general classes of it (BMAPs and HMMs), but the main issue lies in the computational complexity of calculating its parameter estimates. The literature on this process is quite recent and future improvements are expected. It may even be possible one day to conduct statistical inferences on such processes and determine when a simpler model can be used.

# 6 Conclusion

This report has attempted to give readers a main idea on how estimation is done in queueing theory. More importantly, one aims to expand upon this part of the literature, as without it, queueing theory will very seldom be of practical use, and will end up simply being an intellectual exercise. For estimation to become more established in queueing theory, there has to be a larger number of people interested and willing to expand upon it. Through mass interest and extensive practical experimentation (next to simulative studies), one can hope to bring queueing theory into new realms (even possibly changing its name to queueing studies or queue modelling for example). Only time will tell.

## References


[1] L. Breuer. "An EM Algorithm for Batch Markovian Arrival Processes And Its comparison to A Simpler Estimation Procedure." Annals of Operations Research, 112:123–138, 2002.

[2] H. Okamura, T. Dohi, K.S. Trivedi. "Markovian Arrival Process Parameter Estimation With Group Data," IEEE/ACM Transactions on Networking, Vol. 17, No. 4, August 2009

[3] J.V. Ross, T. Taimre, P.K. Pollett, "Estimation for Queues from Queue Length Data," Queueing Systems: Theory and Applications, v.55 n.2, p.131-138, February 2007

[4] J.V. Ross, T. Taimre, P.K. Pollett, "On Parameter Estimation in population models." Theor. Popul. Biol. 70 (2006) 498510.

[5] R.W. Wolff, "Problems of Statistical Inference for Birth And Death Queueing Models." Operat. Res. 13 (1965) 343357.