Advances on Methodological and Applied Aspects of Probability and Statistics

Edited by N. Balakrishnan

TAYLOR & FRANCIS
ALERE FLAMMAM
Founded 1798

Advances on Methodological and Applied Aspects of Probability and Statistics

Advances on Methodological and Applied Aspects of Probability and Statistics

Edited by

N. Balakrishnan

McMaster University
Hamilton, Canada

TAYLOR & FRANCIS
ALERE FLAMMAM
Founded 1798

USA	Publishing Office:	TAYLOR & FRANCIS 52 Vanderbilt Avenue, New York, NY 10017
UK		TAYLOR & FRANCIS 2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

ADVANCES ON METHODOLOGICAL AND APPLIED ASPECTS OF PROBABILITY AND STATISTICS

Cover design by Ellen Seguin.

A CIP catalog record for this book is available from the British Library.

Library of Congress Cataloging-in-Publication Data is available from the publisher.

ISBN 13: 978-1-56032-980-0 (hbk)

CONTENTS

Part II Models and Applications

Part III Estimation and Testing

PREFACE

This is one of two volumes consisting of 33 invited papers presented at the *International Indian Statistical Association Conference* held during October 10–11, 1998, at McMaster University, Hamilton, Ontario, Canada. This Second International Conference of IISA was attended by about 240 participants and included around 170 talks on many different areas of Probability and Statistics. All the papers submitted for publication in this volume were refereed rigorously. The help offered in this regard by the members of the Editorial Board listed earlier and numerous referees is kindly acknowledged. This volume, which includes 33 of the invited papers presented at the conference, focuses on *Advances on Methodological and Applied Aspects of Probability and Statistics.*

For the benefit of the readers, this volume has been divided into nine parts as follows:

Part I	Applied Probability
Part II	Models and Applications
Part III	Estimation and Testing
Part IV	Robust Inference
Part V	Regression and Design
Part VI	Sample Size Methodology
Part VII	Applications to Industry
Part VIII	Applications to Ecology, Biology and Health
Part IX	Applications to Economics and Management

I sincerely hope that the readers of this volume will find the papers to be useful and of interest. I thank all the authors for submitting their papers for publication in this volume.

Special thanks go to Ms. Arnella Moore and Ms. Concetta Seminara-Kennedy (both of Gordon and Breach) and Ms. Stephanie Weidel (of Taylor & Francis) for supporting this project and also for helping with the production of this volume. My final thanks go to Mrs. Debbie Iscoe for her fine typesetting of the entire volume.

I hope the readers of this volume enjoy it as much as I did putting it together!

N. BALAKRISHNAN

McMASTER UNIVERSITY
HAMILTON, ONTARIO, CANADA

LIST OF CONTRIBUTORS

Abraham, Bovas, IIQP, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
`babraham@setosa.uwaterloo.ca`

Agarwal, Manju, Department of Operations Research, University of Delhi, Delhi-110007, India

Aggarwala, Rita, Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4
`rita@math.ucalgary.ca`

Agrawal, Rehka, GE Corproate Research & Devleopment, Schenectady, NY 12065, U.S.A.
`rekha.agrawal@corporate.ge.com`

Balakrishnan, N., Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1
`bala@mcmail.cis.mcmaster.ca`

Basawa, Ishwar V., Department of Statistics, University of Georgia, Athens, GA 30602-1952, U.S.A.
`ishwar@stat.uga.edu`

Basu, Sanjib, Division of Statistics, Northern Illinois University, DeKalb, IL 60115, U.S.A.
`basu@niu.edu`

Basu, Sujit K., National Institute of Management, Calcutta 700027, India

Bera, Anil K., Department of Economics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, U.S.A.
`anil@fisher.econ.uiuc.edu`

Bhat, U. Narayan, Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0240, U.S.A.
nbhat@mail.smu.edu

Bhattacharjee, Manish C., Center for Applied Mathematics & Statistics, Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102-1982, U.S.A.
mabhat@chaos.njit.edu

Bhoj, Dinesh S., Department of Mathematical Sciences, Rutgers University, Camden, NJ 08102-1405, U.S.A.
dbhoj@crab.rutgers.edu

Billah, Md. Baki, Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh

Chaubey, Yogendra P., Department of Mathematics and Statistics, Concordia University, Montreal, Quebec, Canada H4B 1R6
chaubey@vax2.concordia.ca

Chaudhry, M. L., Department of Mathematics and Computer Science, Royal Military College of Canada, P.O. Box 17000, STN Forces, Kingston, Ontario, Canada K7K 7B4
chaudhry-ml@rmc.ca

Datta, Susmita, Department of Mathematics and Computer Science, Georgia State University, Atlanta, GA 30303-3083, U.S.A.
sdatta@cs.gsu.edu

Desu, M. M., Department of Statistics, State University of New York, Buffalo, NY 14214-3000, U.S.A.
desu@calcutta.med.buffalo.edu

Dhar, Sunil K., Department of Mathematical Sciences, New Jersey Institute of Technology, Newark, NJ 07102-1824, U.S.A.
sunidh@stat.njit.edu

Gadbury, Gary, Department of Mathematics, University of North Carolina at Greensboro, Greensboro, NC, U.S.A.

Gardiner, Joseph C., Department of Epidemiology, College of Human Medicine, Michigan State University, East Lansing, MI 48823, U.S.A.
gardine3@pilot.msu.edu

Geddes, R. R., Department of Economics, Fordham University, 441 East Fordham Road, Bronx, NY 10458-5158, U.S.A.

Ginebra, Josep, Departament d'Estadística, E.T.S.E.I.B., Universitat Politècnica de Catalunya, Avgda. Diagonal 647, 6[a] planta, 08028 Barcelona, Spain
ginebra@eio.upc.es

González, Enrique, Departamento de Estadística, Universidad de La Laguna, 38271 La Laguna, Spain
egonzale@ull.es

Govindarajulu, Z., Department of Statistics, University of Kentucky, Lexington, KY 40506, U.S.A.
raju@ms.uky.edu

Gupta, U. C., Department of Mathematics, Indian Institute of Technology, Kharagpur 721 302, India
umesh@maths.iitkgp.ernet.in

Huzurbazar, Aparna V., Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131-1141, U.S.A.
aparna@stat.unm.edu

Indurkhya, Alka, Department of Epidemiology, College of Human Medicine, Michigan State University, East Lansing, MI 48823, U.S.A.

Iyengar, Satish, Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A.
su@bacchus.stat.pitt.edu

Iyer, Hari, Department of Statistics, Colorado State University, Fort Collins, CO 80523, U.S.A.
hari@stat.colostat.edu

Jammalamadaka, S. Rao, Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, U.S.A.
rao@pstat.ucsb.edu

Kinateder, Kimberly K. K., Department of Mathematics and Statistics, Wright State University, Dayton, OH 45453, U.S.A.

Lu, Xuewen, Food Research Program, Sourthern Crop Protection and Food Research Centre, Agriculture and Agri-Food Canada, 43 McGilvray Street, Guelph, Ontario, Canada N1G 2W1
lux@em.agr.ca

Luo, Zhehui, Department of Epidemiology, College of Human Medicine, Michigan State University, East Lansing, MI 48823, U.S.A.

Mallick, Naresh C., Department of Economics and Finance, Alabama Agricultural and Mechanical University, Normal, AL, U.S.A.

Marchetti, Carol E., Department of Mathematics and Statistics, Rochester Institute of Technology, Rochester, NY 14623-5603, U.S.A.
cemsma@rit.edu

Mudholkar, Govind S., Department of Statistics, University of Rochester, Rochester, NY 14727, U.S.A.
govind@metro.bst.rochester.edu

Natarajan, Rajeshwari, Department of Statistics, University of Rochester, Rochester, NY 14727, U.S.A.
rajn@stat1.bst.rochester.edu

Patil, G. P., Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.
gpp@stat.psu.edu

Paul, Sudhir R., Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario, Canada N9B 3P4
smjp@uwindsor.ca

Prabhu, N. U., School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853-3801, U.S.A.
questa@orie.cornell.edu

Rao, J. N. K., School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6
jrao@math.carleton.ca

Saleh, A. K. Md. E., School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada K1S 5B6
esaleh@math.carleton.ca

Sarkar, Sanat K., Department of Statistics, Temple University, Philadelphia, PA 19122, U.S.A.
sanat@sbm.temple.edu

Sen, Kanwar, Department of Statistics, University of Delhi, Delhi-110007, India
ksen@ndb.vsnl.net.in

Serfling, Robert, Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75083-0688, U.S.A.
serfling@utdallas.edu

Singh, R. S., Department of Mathematics and Statistics, University of Guelph, Guelph, Ontario, Canada N1G 2W1
rsingh@msnet.mathstat.uoguelph.ca

Srivastava, Deo Kumar, Department of Biostatistics and Epidemiology, St. Jude Children's Research Hospital, 332 North Lauderdale St., Memphis, TN 38105-2794, U.S.A.
kumar.srivastava@stjude.org

Taillie, C., Department of Statistics, Pennsylvania State University, University Park, PA 16802, U.S.A.

Venkateswarlu, K., Department of Mathematics and Statistics, Concordia University, Montreal, Quebec, Canada H4B 1R6

Vinod, H. D., Department of Economics, Fordham University, 441 East Fordham Road, Bronx, NY 10458-5158, U.S.A.
vinod@murray.fordham.edu

Voss, Daniel T., Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, U.S.A.
dvoss@math.wright.edu

Wadhwa, Neerja, Card Services, GE Capital, Stamford, CT 06820, U.S.A.
pwadhwa@sprintmail.com

Wang, Weizhen, Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, U.S.A.

LIST OF TABLES

LIST OF FIGURES

Part I
Applied Probability

CHAPTER 1

FROM DAMS TO TELECOMMUNICATION – A SURVEY OF BASIC MODELS

N. U. PRABHU

Cornell University, Ithaca, NY

Abstract: In 1954 P.A.P. Moran formulated a simple discrete time model for a finite dam. This model was extended in several directions by J. Gani and the author during 1956–1963. The concepts underlying this model and the techniques used in its analysis are applicable in a wide variety of situations, as has already been demonstrated. Most recently, models for data communication systems have also been analyzed with these techniques. In this paper we survey some of this work.

Keywords and phrases: Buffer content, dam, data communication, idle time, input, fluid input, Lévy process, Markov chain, Markov-additive process, packets, Poisson arrivals, queues, subordinator, unmet demand, workload

1.1 INTRODUCTION

In 1954 P. A. P. Moran formulated a simple discrete time model for the finite dam. The basic components of this model are inputs that are independent and identically distributed random variables, a constant demand for water and the release policy "meet the demand if physically possible." During 1956–1963 J. Gani and the author extended this discrete time model to continuous time, where the input is described by a subordinator, the demand is at a unit rate and the release policy is the same as before. This continuous time model has several applications, in particular, to single server queues with Poisson arrivals and first come, first served discipline or priority discipline of the static or dynamic type. Because these models have several common features in regard to the underlying concepts and techniques of analysis, the author proposed the term *stochastic storage processes* to describe the processes that arise from the family of such models and presented a unified theory of these processes [see Prabhu (1998)]. The most recent extension of this theory is to models for transmission of telecommunication data. Here the input of data is characterized as a Markov-additive process, the desired transmission (demand) rate depends on the Markov component of the input and the actual transmission (release) policy is to "meet the demand if physically possible." The resulting theory may be viewed as the Markov-modulated version of the theory of dams.

In this paper we survey some of this work, emphasizing only the modeling aspects in order to point out the common features of the models considered. For detailed results and recent references see Prabhu (1998). For historical references see Prabhu (1965).

In Section 1.2 we describe Moran's discrete time model for the finite dam. The continuous time dam model is described in Section 1.3, and its extension to the data communication model in Section 1.4.

1.2 MORAN'S MODEL FOR THE FINITE DAM

Moran's discrete time model for a dam (water reservoir) is the following. A dam of finite capacity is designed to meet the demand for electric power (expressed in terms of the volume of water required to produce it) or for water to be supplied to a city. The demand for water at time n is m ($< c$) and this demand is met "if physically possible," that is, to the extent that this quantity is available in the dam at time n. The dam is fed by inputs of water such that if X_{n+1} denotes the input during the time interval $(n, n+1]$, then $\{X_n, n \geq 1\}$ is assumed to be a sequence of independent and identically distributed random variables. Because of this randomness the amount of water in the dam (the dam content) at time n is a random variable which we denote by Z_n ($n \geq 0$).

Since the capacity of the dam is finite there is a possibility of an overflow and the actual input during $(n, n+1]$ is therefore

$$\eta_{n+1} = \min(c - Z_n, X_{n+1}) \quad (n \geq 0). \tag{1.2.1}$$

The amount of water available for release at time $n+1$ is then $Z_n + \eta_{n+1}$ and the release policy implies that

$$Z_{n+1} = Z_n + \eta_{n+1} - \min(m, Z_n + \eta_{n+1}).$$

The sequence $\{Z_n, n \geq 0\}$ satisfies the relation

$$Z_{n+1} = (Z_n + \eta_{n+1} - m)^+ \quad (n \geq 0). \tag{1.2.2}$$

To see how the dam operates subject to these assumptions we note that during a time interval $(0, n]$ there is a certain amount F_n of overflow from the dam, and an amount D_n of the total demand nm that is not met. Easy calculations show that

$$Z_n = Z_0 + (S_n - nm) - F_n + D_n \quad (n \geq 0) \tag{1.2.3}$$

where $S_n = X_1 + X_2 + \cdots + X_n (n \geq 1), S_0 = 0$ and $S_n - nm$ is the net input (input minus demand) during $(0, n]$.

The assumption on the inputs X_n implies that $\{Z_n, n \geq 0\}$ is a time-homogeneous Markov chain on the state space $\mathbb{R}_+$. The problems of practical importance that arise in the analysis of the model are the derivation of (i) the steady state distribution of $\{Z_n\}$ and (ii) the distribution of the random variable

$$T(Z_0) = \min\{n \geq 1 : Z_n = 0\} \tag{1.2.4}$$

which is the duration of the wet period in the dam whose initial content is $Z_0 > 0$. Although these problems are standard in the theory of Markov chains, general solutions are not known because of the presence of the constant c $(< \infty)$ in (1.2.2). However, solutions are available for some important special cases of the input distributions [see Prabhu (1965)].

When $c = \infty$ (the case of the infinite dam) the equations (1.2.2) and (1.2.3) reduce to

$$Z_{n+1} = (Z_n + X_{n+1} - m)^+ \quad (n \geq 0) \tag{1.2.5}$$

and

$$Z_n = Z_0 + (S_n - m_n) + D_n \quad (n \geq 0). \tag{1.2.6}$$

These lead to the expressions

$$Z_n = \max\{Z_0 + S_n - nm, S_n - nm - m_n\} \tag{1.2.7}$$

$$D_n = (Z_0 + m_n)^- \tag{1.2.8}$$

where m_n is the minimum functional of the random walk $\{S_n - nm, n \geq 0\}$, namely

$$m_n = \min_{0 \leq k \leq n} (S_k - km) \quad (n \geq 0). \tag{1.2.9}$$

The equation (1.2.5) arises in queueing theory, specifically for waiting times Z_n in the single server queue with constant interarrival times $(= m)$ and general service times $X_n (n \geq 1)$. The quantity D_n in (1.2.6) is the total idle period during (0,n], while the random variable $T(Z_0)$ defined by (1.2.4) is the number of customers served during the busy period initiated by a waiting time $Z_0 > 0$. Thus the results for the infinite dam are applicable to queueing theory.

1.3 A CONTINUOUS TIME MODEL FOR THE DAM

In developing a continuous time model for the dam we first assume that its capacity is ∞. For the input we postulate a nonnegative process with stationary independent increments, that is, a Lévy process $\{X(t), t \geq 0\}$ with nondecreasing sample functions (also called a subordinator) and zero drift. The demand for water occurs at a rate $d \circ Z(t)$, where $Z(t)$ is the dam content at time $t \geq 0$. As in the discrete time case, this demand is met "if physically possible". These assumptions lead to the integral equation

$$Z(t) = Z(0) + X(t) - \int_0^t d \circ Z(s) 1_{\{Z(s)>0\}} ds. \tag{1.3.10}$$

We can rewrite this is

$$Z(t) = Z(0) + X(t) - \int_0^t d \circ Z(s) ds + \int_0^t d \circ Z(s) 1_{\{Z(s)=0\}} ds. \tag{1.3.11}$$

Here on the right side of (1.3.11) the first integral represents the total demand during $(0, t]$ and the second integral is the part of this demand that is not met. The equation (1.3.11) is the continuous time analogue of (1.2.6).

The most extensively studied special case of (1.3.10) is the one with unit demand rate (that is, $d(x) \equiv 1$), which arises also in the queueing system $M/G/1$ and single server queues with Poisson arrivals and static or dynamic priorities. In the queue $M/G/1$, the input $X(t)$ of workload is a compound Poisson process, and $Z(t)$ represents the remaining workload (virtual waiting time) at time t. In dam models the special cases of input include the gamma process, stable process with exponent 1/2 and the inverse Gaussian process. The integral equation (1.3.11) reduces in the case of unit demand rate to

$$Z(t) = Z(0) + Y(t) - \int_0^t 1_{\{Z(s)=0\}} ds. \tag{1.3.12}$$

where $Y(t) = X(t) - t$ (the net input) and the integral

$$I(t) = \int_0^t 1_{\{Z(s)=0\}} ds \tag{1.3.13}$$

represents the amount of unmet demand (dry period in a dam or idle time in the queue $M/G/1$).

As formulated above, the integral equation (1.3.12) does not have a unique nonnegative solution. However, if we modify it by writing

$$Z(t) = Z(0) + Y(t) - \int_0^t 1_{\{Z(s)\leq 0\}} ds \tag{1.3.14}$$

then the unique nonnegative solution of (1.3.14) is given by

$$Z(t) = \max\{Z(0) + Y(t), Y(t) - m(t)\} \tag{1.3.15}$$

where $m(t)$ is the minimum functional

$$m(t) = \inf_{0\leq s\leq t} Y(s). \tag{1.3.16}$$

It follows from (1.3.14) that

$$I(t) = \int_0^t 1_{\{Z(s)=0\}} ds = [Z(0) + m(t)]^- \tag{1.3.17}$$

on account of the nonnegativity of $Z(t)$. The results (1.3.15) and (1.3.17) are the continuous time analogues of (1.2.7) and (1.2.8) for the discrete time case.

Remarks.

1. When $Z(0) = 0$, the solution (1.3.15) reduces to

$$Z(t) = Y(t) - m(t). \tag{1.3.18}$$

In current literature (1.3.18) is referred to as reflection mapping. This term does not give credit to the pioneering 1958 paper by E. Reich, who derived (1.3.15) for the virtual waiting time in $M/G/1$. Furthermore, the identification of the idle time with the minimum functional does not follow from the reflection mapping.

2. The joint distribution of $Z(t)$ and $I(t)$ can be obtained directly from (1.3.12). For the compound Poisson input the older technique of analysis is based on the forward Kolmogorov integro-differential equation for the distribution of $Z(t)$. □

1.4 A MODEL FOR DATA COMMUNICATION SYSTEMS

A buffer of infinite capacity receives inputs of data represented as a Markov-additive process $\{X(t), J(t), t \geq 0\}$ on the state space $\mathbb{R}_+ \times \mathcal{E}$ in which the additive component is a compound Poisson process. Specifically

$$X(t) = X_0(t) + \int_0^t a \circ J(s)ds. \tag{1.4.19}$$

Here $X_0(t)$ is a compound Poisson process in which the rate at which jumps occur as well as the jump sizes depend on the state of the Markov process J on a countable state space $\mathcal{E}$, these jumps representing the arrivals of packets. In addition X has a drift that occurs at a rate $a(j)$ when J is in state j, and the integral in (1.4.19) represents the amount of data that arrive in a fluid fashion. The desired transmission (demand) rate is $d(j)$ when J is in state j and the transmission (release) policy is to meet the demand "if physically possible." Let $Z(t)$ denote the buffer content at time $t \geq 0$. The above assumptions lead to the integral equation

$$Z(t) = Z_0(t) + X(t) - \int_0^t r \circ (Z(s), J(s))ds \tag{1.4.20}$$

where the release rate r is given by

$$\begin{aligned} r(x,j) &= d(j) \text{ if } x > 0 \\ &= \min(d(j), a(j)) \text{ if } x = 0. \end{aligned} \tag{1.4.21}$$

Comparison with (1.3.10) show that (1.4.20) is indeed an extension of the (now classical) dam model. The presence of J is to be understood with reference to specific models. We first consider two special cases.

A fluid model for data communication. If the arrival of data is only in a fluid fashion, then $X_0(t) \equiv 0$ and the integral equation (1.4.20) reduces to

$$Z(t) = Z(0) + \int_0^t x \circ J(s)^+ ds - \int_0^t x \circ J(s)^- 1_{\{Z(s)>0\}} ds \tag{1.4.22}$$

where $x(j)$ is the net input rate

$$x(j) = a(j) - d(j).\square \tag{1.4.23}$$

A model with packet arrivals. In the presence of packet arrivals we need to assume that the desired transmission rate $d(j)$ exceeds the rate of fluid arrival $a(j)$. The integral equation (1.4.20) then reduces to

$$Z(t) = Z(0) + X_0(t) - \int_0^t d_1 \circ J(s) 1_{\{Z(s)>0\}} ds \tag{1.4.24}$$

where $d_1(j) = d(j) - a(j) > 0$. □

The integral equation that describes each of the above models is of the form

$$Z(t) = Z(0) + X(t) - \int_0^t r \circ (Z(s), J(s)) ds \tag{1.4.25}$$

where $\{X(t), J(t)\}$ is a Markov-additive process and

$$r(x, j) = d(j) 1_{\{x>0\}}. \tag{1.4.26}$$

Comparing (1.4.25) with the integral equation (1.3.10) we see that the data communication models described here are extensions of the continuous time dam model of Section (1.3). The unique nonnegative solution of (1.4.25), modified as in (1.3.14), is formally the same as (1.3.15), where the net input $Y(t)$ given by

$$Y(t) = X(t) - \int_0^t d \circ J(s) ds \tag{1.4.27}$$

and it should be noted that $\{Y(t), J(t)\}$ is a Markov-additive process.

The following are two fluid models that have been investigated in the literature. The presence of the Markov component J will be clear from these models.

a. A multiple source data handling system. There are N sources of messages, which are "on" or "off" from time to time. A switch receives messages at a unit rate from each source and transmits them at a fixed maximum rate c $(1 \le N < \infty, 0 < c < \infty)$. Messages that are not transmitted are stored in a buffer of infinite capacity (see Figure 1.1). Denoting by $J(t)$ the number of "on" sources at time $t \ge 0$, we assume that $\{J(t), t \ge 0\}$ is a birth and death process on the state space $\{0, 1, 2, \ldots, N\}$. Of interest is the buffer content $Z(t)$. It is seen that $Z(t)$ satisfies the integral equation

$$Z(t) = Z(0) + \int_0^t J(s) ds - \int_0^t r \circ (Z(s), J(s)) ds \tag{1.4.28}$$

where

$$\begin{aligned} r(x, j) &= c \text{ if } x > 0 \\ &= \min(j, c) \text{ if } x = 0. \end{aligned} \tag{1.4.29}$$

Clearly, this is a fluid model with $a(j) = j$ and $d(j) = c$. □

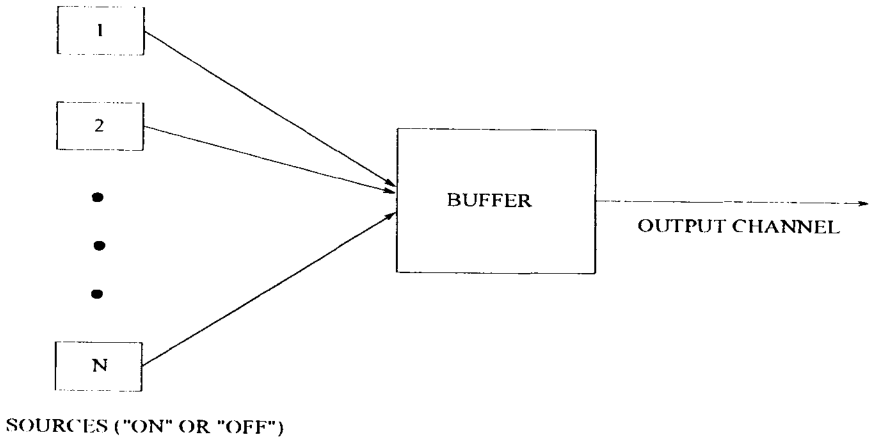

FIGURE 1.1 A buffer of infinite capacity for storage

b. An integrated circuit and packet switching multiplexer. A buffer of infinite capacity receives voice calls as well as data. There are $s+u$ output channels, of which u channels are reserved for data transmission, while the remaining s channels are shared by data and voice calls, with calls having preemptive priority over data and calls that find all s channels that serve them being lost (see Figure 1.2).

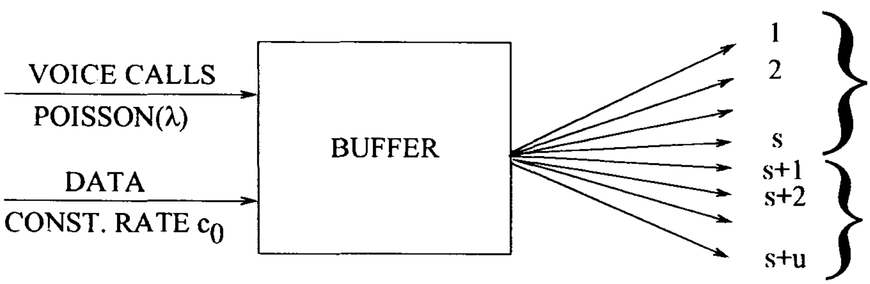

FIGURE 1.2 An integrated circuit and packet switching multiplexer

Voice calls arrive in a Poisson process and their service times have an exponential density. Data arrive continuously at a constant rate c_0 and are transmitted at a rate $c_1(< c_0)$. At time $t \geq 0$, let $Z(t)$ denote the amount of data in the buffer and $J(t)$ the number of channels available for data transmission. It is clear that $s + u - J(t)$ represents the queue length in an $M/M/s$ loss system, and $Z(t)$ satisfies the integral equation

$$Z(t) = Z(0) + \int_0^t c_0 ds - \int_0^t r \circ (Z(s), J(s))ds \tag{1.4.30}$$

where

$$\begin{aligned} r(x,j) &= c_1 j \text{ if } x > 0 \\ &= \min(c_1 j, c_0) \text{ if } x = 0. \end{aligned} \tag{1.4.31}$$

This is a fluid model with $a(j) = c_0$ and $d(j) = c_1 j$.

Remarks.

1. Some authors take (1.3.18) as the starting point of their investigation of data communication models. Such an approach neglects the modeling aspects that are important in any area of applied probability. In particular it does not emphasize the role of Markov-additive inputs.

2. The forward Kolmogorov equation (in the matrix form) can be used to derive the joint distribution of $Z(t)$ and $J(t)$. However, as in the case of the dam model it is much more straightforward to derive the joint distribution of $Z(t), I(t)$ and $J(t)$ directly from (1.4.25), $I(t)$ being the amount of the unmet demand.

3. It is hoped that this brief survey has made it clear that all of the models described in Sections (1.3) and (1.4) are indeed*storage models.* The use of the term *fluid queue*, currently in fashion, is obviously based on lack of familiarity with earlier literature in this subject area. This term is both unnecessary and unpleasant, and the author hopes that discriminating researchers will not use it in the future.

REFERENCES

Prabhu, N. U. (1965). *Queues and Inventories: A Study of Their Basic Stochastic Processes*, John Wiley & Sons, New York.

Prabhu, N. U. (1980). *Stochastic Storage Processes*, Springer-Verlag, New York. (1998): 2nd Edition.

CHAPTER 2

MAXIMUM LIKELIHOOD ESTIMATION IN QUEUEING SYSTEMS

U. NARAYAN BHAT

Southern Methodist University, Dallas, TX

ISHWAR V. BASAWA

University of Georgia, Athens, GA

Abstract: This paper provides an overview of the literature on the use of the maximum likelihood method for estimating parameters in queueing models. Two cases, one when the system elements are fully observable and the second when only a limited amount of information is available are considered. The paper also includes some new results in later sections.

Keywords and phrases: Parameter estimation, maximum likelihood, GI/G/1 queue, M/G/1 queue, GI/M/1 queue, waiting time, queue length

2.1 INTRODUCTION

There are two key steps in the use of the method of maximum likelihood estimation (m.l.e.): constructing the likelihood function and deriving estimators that maximize the function. It was Clarke (1957) who first demonstrated that the likelihood function can be constructed for the underlying queue length process in the queueing system M/M/1 (Poisson arrivals, exponential service times and single servers) if one can describe its sample path as a realization of random events that can be described in terms of distributions. The general maximum likelihood theory for Markov processes, of which M/M/1 is a simple example, has been given by Billingsley (1961). Since then, researchers have explored ways of using this method to non-Markovian systems as well.

In stochastic models, many times factors such as system structure and cost may prevent full observation. In such cases, inference on system parameters will have to be made using other system characteristics. For instance, in a queueing system where embedded Markov chains can be identified, observations relative to those Markov chains can be used to estimate parameters. Goyal and Harris (1972) provides one of the first examples of this procedure.

In this paper we provide an overview of the use of maximum likelihood estimation in queueing systems under both cases of complete and incomplete information. In addition to describing some of the basic work on Markovian systems, we review research on non-Markovian systems when the processes are fully observable and when information only on certain characteristics is available. In the latter case some new results are also presented. The paper is arranged in eight sections. Parameter estimation in Markovian and non-Markovian systems is described in Sections 2.2 and 2.3 respectively. These procedures assume the availability of complete information on the system, although in continuous time, discrete state Markovian systems the set of sufficient statistics used is smaller than that we normally require for non-Markovian systems. Section 2.4 deals with estimation using the embedded Markov chains for the waiting time process and in Section 2.5, the procedure described in Section 2.4 is modified for system time (waiting time plus service time) instead of only waiting time. In Sections 2.6 and 2.7 the process considered is the number of customers in the system and the two sections deal with the queues M/G/1 and GI/M/1 respectively. Finally, Section 2.8 provides some concluding observations. Also Sections 2.5 and 2.7 include new results.

We do not plan to provide a long bibliography in this overview. Only those papers with major influence in the course of research are cited. For the general theory of inference on Markov processes Billingsley's book (1961) is an excellent reference. Basawa and Prakasa Rao (1980) and Karr (1991) provide the theory of inference on stochastic processes, in general. For inference on queues, Bhat *et al.* (1997) is a good reference which includes an extensive bibliography.

2.2 M.L.E. IN MARKOVIAN SYSTEMS

Any discussion of m.l.e. in Markovian queues has to start with the paper by Clarke (1957). Even though two earlier papers by Moran (1951, 1953) described a procedure to estimate the birth and death parameters in the simple birth-and-death process, it was Clarke who used the complete description of the sample path to construct the likelihood function.

Let the system be observed for a length of time t such that the time spent in a busy state is a preassigned value t_b. Let n_a, n_s, t_e represent the number of arrivals, number of service completions, and the time spent in the empty state, respectively, during $[0, t]$. Furthermore, let n_0 be the initial queue length. Also assume that the system is in the steady state. The likelihood function can be written as

$$L(\lambda, \mu) = \left(\frac{\lambda}{\mu}\right)^{n_0} \left(1 - \frac{\lambda}{\mu}\right) \lambda^{n_a} \mu^{n_s} e^{-(\lambda+\mu)t_b} e^{-\lambda t_e}, \tag{2.2.1}$$

and the m.l.e.'s of λ and μ are found from the equations

$$\widehat{\lambda} = (\widehat{\mu} - \widehat{\lambda})(n_a + n_0 - \widehat{\lambda} t) \text{ and } \widehat{\lambda} = (\widehat{\lambda} - \widehat{\mu})(n_s - n_0 - \widehat{\mu} t_b). \tag{2.2.2}$$

Estimating $\widehat{\mu}$ from the second equation gives a quadratic in $\widehat{\lambda}$. Of the two solutions, any negative solution is rejected, and for the remaining values of $\widehat{\lambda}$, corresponding $\widehat{\mu}$ is obtained. Furthermore, any pair $(\widehat{\lambda}, \widehat{\mu})$ would be rejected for which $\widehat{\mu} \leq 0$ or $\widehat{\lambda}/\widehat{\mu} \geq 1$. If both solutions are valid, then the solution which maximizes the likelihood function is chosen.

For large $n_s - n_0$ Clarke gives a sample approximation for $\widehat{\lambda}$ and $\widehat{\mu}$ as

$$\widehat{\lambda} \simeq (n_a + n_0)/t, \ \widehat{\mu} \simeq (n_s - n_0)/t_b. \tag{2.2.3}$$

The consistency of $\widehat{\lambda}$ and $\widehat{\mu}$ has been examined by Samaan and Tracy (1978) who could establish only a weak consistency for $\widehat{\lambda}$. If we ignore the initial queue size, the estimates of λ and μ are, respectively, n_a/t and n_s/t_b.

As noted by Cox (1965), specializing Billingsley's (1961) results, this procedure can be extended to the generalized birth-and-death models. The conditional likelihood function (ignoring the contribution of the initial state) is of the form

$$e^{-\Sigma(\lambda_i+\mu_i)t_i} \Pi \lambda_i^{n_{a_i}} \mu_i^{n_{s_i}}, \tag{2.2.4}$$

where λ_i, μ_i are the rates of arrival and service compilations in state i, n_{a_i} and n_{s_i} are the numbers of arrivals and service completions in state i, and t_i is the total time spent in state i during the observation interval $(0, t]$.

For a finite state birth-death queue, ignoring the impact of the initial queue size, the m.l.e's of λ_i and μ_i are given by

$$\widehat{\lambda}_i = n_{a_i}/t_i \ (0 \leq i \leq M-1), \qquad \widehat{\mu}_i = n_{s_i}/t_i \ (1 \leq i \leq M). \qquad (2.2.5)$$

The above results and similar estimates for parameters in M/M/s, M/M/∞, and machine interference problem have been given by Wolff (1965), where many details are provided. For an extension of these methods to a simple Markovian queueing network, commonly known as the Jackson network, see Thiruvaiyaru *et al.* (1991), where joint asymptotic normality of the estimators is also established. Also see, Beneš (1957) for a discussion of the set of sufficient statistics in similar problems, and Cox (1965), and Lilliefors (1966) for confidence intervals for estimates.

2.3 M.L.E. IN NON-MARKOVIAN SYSTEMS

In Markovian systems, due to the memoryless property of the exponential distribution data-collection gets simplified because of our ability to pool observations without losing information. In non-Markovian systems this is not the case and therefore the two cases, one with complete information and the second with incomplete information (which arises when the system cannot be observed fully), become relevant. In this section we cover two important papers by Basawa and Prabhu (1981, 1988) which assume the availability of complete information. Research on cases with incomplete information is discussed in later sections.

Basawa and Prabhu (1981) obtain the m.l.e.'s of parameters of the arrival and service time distributions with continuous densities $f(u;\theta)$ and $g(v;\phi)$, respectively. The sampling scheme is to observe the queue until the first n customers have departed from the system and the service times of these n customers, say $(v_1, v_2, \ldots, v_n)$. Let the nth departure epoch be D_n and observe the interarrival times of all customers who arrive during $(0, D_n]$, giving the interarrival sequence $(u_1, u_2, \ldots, u_{N_A})$, where $N_A = N_A(D_n)$ =max$\{k : u_1 + u_2 + \cdots + u_k \leq D_n\}$. Under this sampling scheme, the likelihood function is

$$L_n(f,g) = \left\{\prod_{i=1}^{N_A} f(u_j;\theta)\right\}\left\{\prod_{i=1}^{n} g(v_j;\phi)\right\}[1 - F(x_n;\theta)], \qquad (2.3.6)$$

where

$$x_n = x_n(D_n) = D_n - \sum_{j=1}^{N_A} u_j \, .$$

Since the factor $[1 - F(x_n;\theta)]$ causes difficulty in obtaining simple estimates, consider the alternative approximate likelihood function obtained by dropping the last terms in (2.3.6):

$$L_n^a(f,g) = \left\{ \prod_{i=1}^{N_A} f(u_j;\theta) \right\} \left\{ \prod_{i=1}^{n} g(v_j;\phi) \right\}. \tag{2.3.7}$$

If $\widehat{\theta}_n^a, \widehat{\phi}_n^a$ are the m.l.e.'s of θ and ϕ based on $L_n^a(f,g)$, they are solutions of the equations

$$\sum_{j=1}^{N_A} \frac{\partial}{\partial\theta} \log f(u_j;\theta) = 0, \ \sum_{j=1}^{n} \frac{\partial}{\partial\phi} \log g(v_j;\phi) = 0. \tag{2.3.8}$$

Basawa and Prabhu prove that $\widehat{\theta}_n^a, \widehat{\phi}_n^a$ are consistent estimators of θ and ϕ and that

$$\begin{bmatrix} \sqrt{n}(\widehat{\theta}_n^a - \theta) \\ \sqrt{n}(\widehat{\phi}_n^a - \phi) \end{bmatrix} \xrightarrow{D} N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \left\{ \begin{matrix} \sigma_\theta^2/\eta & 0 \\ 0 & \sigma_\phi^2 \end{matrix} \right\} \right\}, \tag{2.3.9}$$

where N_2 represents a bivariate normal density with

$$\sigma_\theta^2 = \left[E\left(\frac{\partial}{\partial\theta} \log f \right)^2 \right]^{-1}, \sigma_\phi^2 = \left[E\left(\frac{\partial}{\partial\phi} \log g \right)^2 \right]^{-1}, \tag{2.3.10}$$

$\eta = \max(1,\rho)$, and ρ being the traffic intensity.

Let $\widehat{\theta}_n$ and $\widehat{\phi}_n$ be the estimators based on the full likelihood function (2.3.6). It is seen that $\widehat{\phi}_n = \widehat{\phi}_n^a$, and $\widehat{\theta}_n$ differs from $\widehat{\theta}_n^a$, but it can be shown that $\widehat{\theta}_n$ and $\widehat{\theta}_n^a$ have the same limiting distributions whenever

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial\theta} \log(1 - F(x_n;\theta)) \xrightarrow{P} 0. \tag{2.3.11}$$

This condition is satisfied for Erlangian arrivals. For large samples, estimators of θ and ϕ can be determined from (2.3.8) at least numerically, if not in closed form. Using (2.3.9) confidence intervals for θ and ϕ can also be constructed. From a practical point of view, it is significant to note that the limit properties of these statistics are obtained without the assumption on the existence of equilibrium. Basawa and Prabhu also consider m.l.e.'s for arrival and service rates in the M/M/1 queue based on a sample function observed over a fixed interval $(0,t]$, as done by Wolff (1965), and obtain limit distributions of the m.l.e.'s without any restrictions on ρ.

In a subsequent paper, Basawa and Prabhu (1988) have provided a unified framework for the estimation problem described above where the observation period is $(0, T]$, with a suitable stopping time T. Four different stopping rules are considered. It is shown that the limit distribution does not depend on the particular stopping rule if a random norming is used. They assume that the interarrival and service time distributions belong to the class of non-negative exponential families. Basawa and Prabhu also derive similar results using a generalized linear model for interarrival and service time distributions.

An extension of these procedures to Jackson-type queueing networks with arrivals at each node following a renewal process and service times being arbitrary has been carried out by Thiruvaiyaru and Basawa (1996). As an illustration, the inter-arrival time and service time distributions are assumed to belong to two separate exponential families of distributions. Two sampling plans, one based on a realization over a fixed interval and the second with observations over a certain random interval are used.

2.4 M.L.E. FOR SINGLE SERVER QUEUES USING WAITING TIME DATA

In Sections 2.4–2.7 m.l.e. procedures are described when complete information on the systems under consideration is not available. This section uses waiting time data, Section 2.6 employs system time (waiting time plus service time) data and the following two sections use queue length data for estimation.

A maximum likelihood procedure for the estimation of parameters in a single server queueing system GI/G/1 was presented in a recent paper by Basawa, Bhat and Lund (1996) using information on waiting times $\{W_t\}$, $t = 1, 2, \ldots, n$ of n successive customers. Information is collected from each of n successive customers on the amount of time spent by them in waiting for service. Let W_t denote the waiting time of the tth customer. The waiting time process $\{W_t,\, t = 1, 2, \ldots\}$ satisfies the following well known equation:

$$\begin{aligned} W_{t+1} &= \begin{cases} W_t + X_{t+1} & , \text{ if } W_t + X_{t+1} > 0 \\ 0 & , \text{ if } W_t + X_{t+1} \le 0 \end{cases} \\ &= \max(0, W_t + X_{t+1}) \end{aligned} \tag{2.4.12}$$

where $X_t = V_{t-1} - U_t$, with V_t and U_t denoting, respectively, the service and inter-arrival times corresponding to the tth customer. It should be noted that $\{X_t\}$ is a sequence of independent and identically distributed random variables and X_{t+1} is independent of W_t. It is clear that $\{W_t\}$ is a Markov chain and its transition distribution function can be written as

$$P(W_{t+1} \leq y | W_t = w) = \begin{cases} F_x(y - w) & , \quad y \geq 0 \\ 0 & , \quad y < 0 \end{cases} \tag{2.4.13}$$

where $F_x(\cdot)$ is the distribution of X_t. The transition distribution function has a discontinuity at 0. Define

$$\alpha(w) = 1 - F_x(-w). \tag{2.4.14}$$

Then, for the transition density we have

$$p(y|w) = \begin{cases} 1 - \alpha(w) & y = 0 \\ f_x(y - w) & y > 0 \\ 0 & y < 0 \end{cases}. \tag{2.4.15}$$

Define the indicator function

$$Z_{t+1} = \begin{cases} 0 & \text{if } W_{t+1} = 0 \\ 1 & \text{if } W_{t+1} > 0 \end{cases}. \tag{2.4.16}$$

Using Z_{t+1}, for the transition density of W_t, we can write

$$p(W_{t+1}|W_t) = [1 - \alpha(W_t)]^{1-Z_{t+1}} [f_x(W_{t+1} - W_t)]^{Z_{t+1}}. \tag{2.4.17}$$

The likelihood function based on the sample $(W_1, W_2, ..., W_n)$ is given by

$$L = p(W_1) \sum_{t=1}^{n-1} p(W_{t+1}|W_t). \tag{2.4.18}$$

Let $\theta = (\theta_1, \theta_2, ..., \theta_r)'$ be the unknown parameter vector corresponding to the distribution of X_t. Basawa *et al.* (1996) show that estimates for θ can in fact be determined using the likelihood function (2.4.18) following the standard procedure. Basawa *et al.* also have established the consistency and the asymptotic normality of the estimators, and discussed issues pertaining to their efficiency.

2.5 M.L.E. USING SYSTEM TIME

The sampling plan used in the last section requires the knowledge of the amount of time customers spend in waiting for service. In practice, in many instances, it may not be as easy to determine the actual waiting time as it is to determine the total time spent by customers in the system; i.e., the waiting time plus service time. We shall call this characteristic system time.

Let Y_t be the system time corresponding to the t^{th} customer. Based on its definition, we have

$$\begin{aligned} Y_{t+1} &= V_{t+1} + W_{t+1} \\ &= V_{t+1} + \text{ Max}(0,\, W_t + V_t - U_{t+1}) \\ &= V_{t+1} + \text{ Max}(0,\, Y_t - U_{t+1}) \end{aligned} \tag{2.5.19}$$

which can also be written in display form as

$$Y_{t+1} = \begin{cases} Y_t - U_{t+1} + V_{t+1} & \text{if } Y_t - U_{t+1} > 0 \\ V_{t+1} & \text{if } Y_t - U_{t+1} \leq 0 \end{cases} \tag{2.5.20}$$

Incidentally, the continuous time analog $\{Y(t),\, t \geq 0\}$ of the process $\{Y_t,\, t = 1, 2, 3 \ldots\}$ was originally introduced by Prabhu (1964) in the context of queue GI/M/1. The process $Y(t)$ exhibits properties of duality with the virtual waiting time process $W(t)$ as defined by Takács (1955) and the graph of $Y(t)$ can be looked upon as a mirror image of the graph of $W(t)$ [see, Prabhu (1965, p. 102)].

Equation (2.3.9) shows that $\{Y_t\}$, $t = 1, 2, \ldots.$ is a Markov process. We now proceed to derive the transition density corresponding to the Markov process $\{Y_t\}$. We have

$$\begin{aligned} &P(Y_{t+1} \leq y_{t+1} | Y_t = y_t) \\ &\quad = P(V_{t+1} \leq y_{t+1}) P(U_{t+1} \geq y_t) \\ &\quad\quad + \int_0^{y_t} P(V_{t+1} \leq y_{t+1} - y_t + u) a(u) du, \end{aligned} \tag{2.5.21}$$

where $a(u)$ is the inter-arrival time density. The result in (2.5.21) follows readily from (2.5.20), considering the two possibilities: $U_{t+1} \geq Y_t$ and $U_{t+1} < Y_t$ and applying the addition law. The transition probability of Y_{t+1} given Y_t is then obtained by differentiating (2.5.21) with respect to y_{t+1}:

$$\begin{aligned} P(y_{t+1} | y_t) &= b(y_{t+1})(1 - A(y_t)) \\ &\quad + \int_0^{y_t} b(y_{t+1} - y_t + u) a(u) du, \end{aligned} \tag{2.5.22}$$

where $b(\cdot)$ and $A(\cdot)$ denote the density of service time V and the distribution function of inter-arrival time, U, respectively. The likelihood function based on $(Y_1, \ldots, Y_n)$ is then given by

$$L(\theta) = p(Y_1;\theta) \prod_{j=1}^{n-1} p(Y_{j+1}|Y_j;\theta), \tag{2.5.23}$$

where θ is the parameter of interest, and $p(Y_1;\theta)$ is the initial density of Y_1. The ML estimator $\widehat{\theta}$ is obtained as a solution of the equation

$$\frac{d \log L(\theta)}{d\theta} = 0. \tag{2.5.24}$$

Since $\{W_t\}$ is an ergodic process (assuming that the traffic intensity $\rho < 1$), it follows that $\{Y_t\}$, $Y_t = W_t + V_t$, is also ergodic. The consistency and the asymptotic normality of the MLE, $\widehat{\theta}$, can therefore be deduced as in Basawa *et al.* (1996).

2.6 M.L.E. IN M/G/1 USING QUEUE LENGTH DATA

In this and the next section, the sampling scheme used for collecting data includes only observing the number of customers in the system for a fixed length of time or some variation of it.

Consider the embedded Markov chain of the queue length in M/G/1, defined at departure epochs. Let Q_t be the number of customers in the system immediately after the tth departure. The process $\{Q_t,\ t = 0, 1, 2, \ldots\}$ is a Markov chain. Let $B(\cdot)$ be the service time distribution and the Poisson arrival rate be λ. If we denote by A_t, the number of arriving customers during the service period, we get the distribution of A_t as

$$P(A_t = j) = k_j = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} dB(x), \qquad j = 0, 1, 2, \ldots \tag{2.6.25}$$

It is well known that Q_t satisfies the relation

$$\begin{aligned} Q_{t+1} &= \begin{cases} Q_t - 1 + A_{t+1} & \text{if } Q_t > 0 \\ A_{t+1} & \text{if } Q_t = 0 \end{cases} \\ &= \begin{cases} Q_t - 1 + A_{t+1} & \text{if } Q_t - 1 > 0 \\ A_{t+1} & \text{if } Q_t - 1 \leq 0 \end{cases} \end{aligned} \tag{2.6.26}$$

which is similar in structure to Eq. (2.5.20). For the transition probabilities of $\{Q_t\}$, we have

$$P(Q_{t+1} = i_{t+1}|Q_t = i_t) = \begin{cases} P(A_{t+1} = i_{t+1}) & \text{if } i_t = 0 \\ P(A_{t+1} = i_{t+1} - i_t + 1) & \text{if } i_t \geq 1. \end{cases} \tag{2.6.27}$$

Suppose the process is observed until the number of departures reaches a fixed value n. Now tracing the sample path of the process we may write down the likelihood function as

$$L(\theta) = p(Q_1;\theta) \prod_{t=1}^{n-1} p(Q_{t+1}|Q_t;\theta)\,. \tag{2.6.28}$$

Let n_{ij} be the number of transitions of Q_t from i to j on the sample path, and θ, the vector of parameters for which estimators are being sought. We get

$$\begin{aligned} \log L(\theta) &= \log P(Q_0 = i_0) \\ &\quad + \sum_{j=0}^{\infty} (n_{0j} + n_{1j}) \log k_j \\ &\quad + \sum_{i=2}^{\infty} \sum_{j=i-1}^{\infty} n_{ij} \log k_{j-i+1}\,. \end{aligned} \tag{2.6.29}$$

Depending on the form of the service time distribution, an explicit expression for the likelihood function can be written down and maximized in the usual manner to determine maximum likelihood estimates. The same general formulation holds when the service times are dependent. Goyal and Harris (1972) consider two such systems: (i) service times are exponential but with different means when the queue size is 1 and when it is > 1, (ii) service times are exponential with means linearly dependent on the number of customers in the system ($\mu_t = t\mu$). They derive m.l.e.'s for utilization factors (arrival rate/service rate) in the case of these two systems when the effect of the initial queue length can or cannot be ignored. Depending on the complexity of likelihood functions to be maximized, some equations will have to be solved using numerical approximation methods.

Another approach to maximum likelihood estimation using embedded Markov chains is to observe only the number of arrivals during successive service periods. In particular, when the arrivals are Poisson and the service times are Erlangian, Harishchandra and Rao (1984) have constructed the likelihood function using the number of arrivals during successive service periods as the sample. In an M/E$_k$/1 queue, in which k is the shape parameter of the Erlangian distribution and ρ is the traffic intensity, let A_t denote the number of arrivals during the service of the $(t+1)$th customer. Then A_t has the negative binomial distribution given by

$$P(A_t = x) = f(x, \rho) = \binom{x+k-1}{x} \left(\frac{\rho}{\rho+k}\right)^x \left(\frac{k}{\rho+k}\right)^k, \quad x = 0, 1, 2, \ldots . \tag{2.6.30}$$

Suppose the system is observed only at departure epochs. Using equation (2.6.26), the queue length data can be easily converted into arrival data. Let $x_1, x_2, \ldots, x_n$ be the number of arrivals during the first n service times, respectively. The likelihood function for this sample is then

$$L(x_1, x_2, \ldots, x_n; \rho) = \prod_{i=1}^{n} \binom{x_i+k-1}{x_i} \left(\frac{\rho}{\rho+k}\right)^{x_i} \left(\frac{k}{\rho+k}\right)^k . \tag{2.6.31}$$

The maximum likelihood estimate of ρ is found to be $\widehat{\rho} = \Sigma x_i/n$. This estimator is unbiased and consistent, since $E(\widehat{\rho}) = \rho$ and $\text{Var}(\widehat{\rho}) = \rho(\rho + k)/(kn)$. Furthermore, it turns out that $\widehat{\rho}$ is also the minimum variance bound (MVB) estimator and therefore uniformly minimum variance unbiased estimator (UMVUE) of ρ. It can be shown that the probability distribution of X belongs to the one-parameter exponential family and hence $T = \Sigma x_i$ is a sufficient statistic for ρ. Finally, for large values of n,

$$\frac{1}{\sigma}\sqrt{n}(\widehat{\rho} - \rho) \xrightarrow{D} N(0, 1), \tag{2.6.32}$$

where

$$\sigma^2 = \left[E\left(\frac{\partial}{\partial\rho}\log f(x, \rho)\right)^2\right]^{-1} = \frac{\rho(\rho+k)}{k}. \tag{2.6.33}$$

Even though, conceptually, estimating k using the likelihood function (2.6.31) is only a mathematical problem, due to the complexities of the expressions, the procedure does not become tractable. The results derived by Miller and Bhat (1997) overcome this problem by using a different approach.

Miller and Bhat use the number of customers served while the system has been busy for a specific length of time as the data element. In this formulation the service process, after eliminating idle times, resembles a renewal process. Consider the following two sampling plans for this renewal process.

Sampling Plan I: Assuming that the first observation period begins at time zero, observe the renewal process at time τ and record the number of renewals in $(0, \tau)$. To assure independent observations, the next observation period will begin when the next renewal occurs. Then after a period of τ time units, the number of renewals occurring in this second period is recorded. Wait until the next renewal occurs and the renewal epoch begins the following observation period, etc.

Sampling Plan II: Assuming the first observation period begins at time zero, observe the renewal process at time τ and record the number of renewals in $(0, \tau)$. Also record the time until the next renewal following time τ which will signal the start of a new observation period. Then after a period of τ time units, the number of renewals occurring in this second period is recorded. Record the time elapsed until the next renewal and the renewal epoch begins the following observations period, etc.

The second sampling plan uses the additional information on the waiting time to start the next observation.

Let $N_1^\tau, N_2^\tau, \ldots.$ denote the number of renewals (service completions) occurring in the observation periods, $1, 2, 3, \ldots$, respectively. In the second sampling plan the observations will be bivariate $\{(N_i^\tau, Y_i(\tau)), i = 1, 2, \ldots\}$ where $Y_i(\tau)$ is the excess life of the renewal period encountered at the ith observation. Using these observations, $\{N_i^\tau, i = 1, 2, \ldots\}$ with Sampling Plan I and $\{(N_i^\tau, Y_i(\tau))\}, i = 1, 2, \ldots\}$ with Sampling Plan II, Miller and Bhat construct likelihood functions which can be used to derive m.l.e. for k either assuming k to be continuous first and determining the best integer k from that result, or using the method of integer maximum likelihood estimation. As one would expect Sampling Plan II leads to better results in estimation.

2.7 M.L.E. IN GI/M/1 USING QUEUE LENGTH DATA

Consider the imbedded Markov chain $\{Q_t, t = 0, 1, 2, \ldots\}$ in a GI/M/1 queue in which arrivals from a renewal process and service times are exponential. Let Q_t represent the number of customers in the system just before the tth arrival. Let $A(\cdot)$ be the inter-arrival time distribution function and μ be the service rate so that the exponential service time density is given by $\mu e^{-\mu x}(x > 0)$. Define D_t as the number of potential departures during an inter-arrival period if an unlimited number of customers are available for service. The random variable D_t has the distribution

$$P(D_t = j) = \delta(j) = \int_0^\infty e^{-\mu x} \frac{(\mu x)^j}{j!} dA(x), \; j = 0, 1, 2, \ldots \qquad (2.7.34)$$

It is well known that Q_t satisfies the relation

$$Q_{t+1} = \begin{cases} Q_{t+1} - D_t & , \ textif \ Q_t + 1 - D_t > 0 \\ 0 & , \ \text{if} \ \ Q_t + 1 - D_t \leq 0 \end{cases} . \tag{2.7.35}$$

Let

$$X_{t+1} = 1 - D_t.$$

Then, (2.7.35) can be re-written in the form

$$Q_{t+1} = \begin{cases} Q_t + X_{t+1} & , \ \text{if} \ \ Q_t + X_{t+1} > 0 \\ 0 & , \ \text{if} \ \ Q_t + X_{t+1} \leq 0 \end{cases} \tag{2.7.36}$$

which is similar in structure to Eq. (2.4.12).

From equation (2.7.36) we get

$$\begin{aligned} P(Q_{t+1)} = 0 | Q_t = i) &= P(i + X_{t+1} \leq 0) \\ &= P(D_t \geq i + 1) \\ &= \sum_{r=i+1}^{\infty} \delta(r) \\ &= 1 - \alpha(i) \end{aligned} \tag{2.7.37}$$

where we have written $\sum_{r=0}^{i} \delta(r) = \alpha(i)$. Also

$$\begin{aligned} P(Q_{t+1} = j | Q_t = i) &= P(i + X_{t+1} = j) \\ &= P(D_t = i + 1 - j) \\ &= \delta(i + 1 - j), \ (j > 0). \end{aligned} \tag{2.7.38}$$

Using the indicator function Z_t defined in (2.4.16), with W_t replaced by Q_t, we may write the transition probability as

$$p(Q_{t+1} | Q_t) = [1 - \alpha(Q_t)]^{1 - Z_{t+1}} [\delta(Q_t + 1 - Q_{t+1})]^{Z_{t+1}} \tag{2.7.39}$$

and the likelihood function as

$$L(\theta) = p(Q_1; \theta) \prod_{t=1}^{n-1} p(Q_{t+1} | Q_t; \theta). \tag{2.7.40}$$

It should be noted that when estimating θ using maximization of (2.7.40), numerical methods maybe needed. For instance, when the inter-arrival time distribution is Erlangian with

$$\begin{aligned} dA(x) &= e^{-k\lambda x}\frac{(k\lambda x)^{k-1}}{(k-1)!}k\lambda dx \qquad (x>0) \\ \delta(r) &= \int_0^\infty e^{-\mu x}\frac{(\mu x)^r}{r!}e^{-k\lambda x}\frac{(k\lambda x)^{k-1}}{(k-1)!}k\lambda dx \\ &= \binom{r+k-1}{r}\left(\frac{\mu}{\mu+k\lambda}\right)^r\left(\frac{k\lambda}{\mu+k\lambda}\right)^k. \end{aligned} \tag{2.7.41}$$

Even though $\delta(r)$ lends itself convenient for taking logarithms and differentiating, $\alpha(i)=\sum_{r=0}^{i}\delta(r)$ is not easily tractable in such operations. Then, direct maximization using numerical techniques is recommended.

If k is also an unknown parameter, methods using integer-maximum likelihood estimation will have to be incorporated in the process [see, Dahiya (1986) and Miller (1997)]. Another approach is to follow the procedure of Miller and Bhat (1977) described in Section 2.6. The arrival process is a renewal process and the estimation procedure proposed by Miller and Bhat gives m.l.e. for Erlang k of the arrival distribution.

In deriving Eq. (2.7.35), we note that D_t has been defined as the number of potential departures during an inter-arrival period. (It is the actual number when the system is busy throughout the period; otherwise it is the number of departures if there are an unlimited number of customers in the system). Consequently, the information available on $\{Q_t\}$ cannot be transformed into information on D_t completely as done for Eq. (2.6.30) in Section 2.6. Therefore, if one has to carry out inference based solely on queue length, the maximum likelihood method described above seems to be the best approach.

2.8 SOME OBSERVATIONS

From a review of research papers on the use of m.l.e. to estimate parameters of queueing models, it is clear that if one is interested in deriving simple readily usable results, a Markovian model is almost a necessity. Even when using information from an embedded chain in the queue M/G/1, the procedure leads to closed-form solutions only when the service time distribution is Erlangian. When likelihood function becomes complex, maximization can be accomplished only through numerical approximation methods. Therefore, in applications with non-Markovian models where easy numerical results are needed, regardless of the sophistication of the maximum likelihood procedure and the desirable properties possessed by the estimators resulting from it, we may not have any recourse but to use moment estimators.