Journal of

Applied Mathematics and Decision Sciences

Special Issue
Statistics and Applied Probability: A Tribute to Jeffrey J. Hunter

Hindawi Publishing Corporation
http://www.hindawi.com

Volume 2007

Journal of Applied Mathematics and Decision Sciences

Olivier Thas
Ghent University, Belgium
olivier.thas@ugent.be

Wing-Keung Wong
National University of Singapore, Singapore
ecswwk@nus.edu.sg

Graham Raymond Wood
Macquarie University, Australia
gwood@efs.mq.edu.au

Journal of Applied Mathematics and Decision Sciences
Volume 2007

Special Issue

Statistics and Applied Probability: A Tribute to Jeffrey J. Hunter
Guest Editors: Graeme Charles Wake and Paul Cowpertwait

Contents

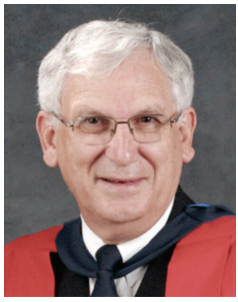
Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 57619, 4 pages
doi:10.1155/2007/57619

Editorial

Statistics and Applied Probability: A Tribute to Jeffrey J. Hunter

Paul Cowpertwait, Graeme Wake, Robert D. Anderson,
Howard Edwards, and Shayle Searle

Received 16 September 2007; Accepted 16 September 2007



We are delighted to be associated as guest editors with the publication of this special issue of the Journal of Applied Mathematics and Decision Sciences, in honour of our esteemed colleague Professor Jeffrey J. Hunter. This special issue marks Jeff's retirement from Massey University, although Jeff plans to continue his research. It also acknowledges Jeff's many contributions to probability and statistics over a long career of 40-plus years, during which he has served his colleagues and university exceedingly well, with the utmost integrity, and has served and promoted his academic discipline in a consistent and highly professional manner, representing the interests of the statistics discipline in many capacities and roles. We have been honoured to have him as a colleague, and given the impressive response of the authors of this issue, we believe this feeling is shared by many in the academic community. Some personal notes from Robert Anderson (Pro Vice-Chancellor of the College of Sciences, Massey University), Howard Edwards (Programmes Director in the College of Sciences), and Emeritus Professor Shayle Searle are included in this editorial.

We are indebted to the Hindawi Publishing Corporation for supporting this initiative, and for their efficient handling of the manuscripts. We thank Tony Norris (Head of our Institute) and Robert McKibbin (former Head) for their helpful suggestions and moral support, and are grateful to Tony and Gaven Martin, for providing financial support via

the institute and personal research funds, respectively, without which this publication would not have been possible. Helpful comments from Stephen Haslett are acknowledged. We thank the many referees who helped tremendously with the reviewing process; in particular, we acknowledge the following referees for their helpful comments and/or detailed reviews: Tasos Tsoularis, Martin Hazelton, Valerie Isham, Beatrix Jones, Claire Jordan, Dominic Lee, Michael Leonard, Barry McDonald, Andrew Metcalfe, Christian Onof, Marco Reale, Danny Walsh, and Thomas Yee. Finally, we thank the authors, many of whom know Jeff well and are his friends; we are very grateful to them for their impressive response to this special issue and for their comments on each others work, which has helped ensure that this publication is of high standard.

Paul Cowpertwait and Graeme Wake
Guest Co-Editors
July, 2007

Professor Jeffrey J. Hunter, FNZMS by Robert D. Anderson, Pro Vice-Chancellor

It is an honour to have been asked to contribute some thoughts on Professor Jeffrey Hunter to this special publication.

I first met Jeff at the time of his interview for the Chair in Statistics at Massey University, in 1990. In the subsequent 17 years, we have developed a strong personal and professional friendship. Such is his instinctively generous and helpful nature, it is impossible not to enjoy, and value, his friendship and professional interaction. He has been an outstanding colleague and mentor to many.

When one reflects on Jeff's career and achievements, a remarkable aspect that stands out, and that others rarely emulate, is that at varying times he has held some high-level managerial posts, the Foundation Deanship of the Faculty of Information and Mathematical Sciences at Massey University for example, yet he still managed to sustain an enviable personal record of research and advanced scholarship-compared to most other university administrators. The award of a Doctor of Science degree by Massey University in 2005 is compelling evidence of enduring scholarship at the highest level. The New Zealand Statistical Association Campbell Award (its highest) in 2006 for "contributions to statistical research and education, and services to the profession of statistics" further reflects his ability to keep his managerial responsibilities-without exception, successfully discharged-in perspective in order that he could find time to concentrate on what actually counts in any university of substance, namely; advanced education and scholarship.

In the early stages of our association, Jeff and I were fellow Deans. Later when the Institute of Information and Mathematical Sciences, based at the Auckland Campus of Massey University, was founded, the wider College of Sciences needed to find a leader with an ability to inspire, and to organise, others to commit to a shared vision for a newly created multidisciplinary entity. Fortunately, Jeff was available for the role and (with high probability of success-that particular word needs to appear in any statement about him) he was appointed Foundation Head as from 1st January, 1998. It is important for the formal record that appreciation be expressed for his excellent leadership of the institute over those formative years.

It is entirely apt that there is a dedicated publication honouring Professor Jeffrey Hunter's outstanding career of advanced scholarship in statistics and mathematics.

Robert D. Anderson, ONZM
Pro Vice-Chancellor (Sciences)
July, 2007

Jeff Hunter's Contribution to Statistics at Massey University by Howard Edwards

It has been my great pleasure to work under and with Jeff Hunter in his variety of roles at Massey University: Professor and Head of the Statistics Department at Palmerston North, Dean of the Faculty of Information and Mathematical Sciences which spanned both the Palmerston North campus and the new Albany campus, and Foundation Head of the Institute of Information and Mathematical Sciences based on the new Albany campus.

Jeff arrived at Massey University in 1990 after an already distinguished career at the University of Auckland. At the time, the Statistics Group within the combined Department of Mathematics and Statistics at Massey had recently tried to break away and form a separate department but had been unsuccessful in this regard. Jeff recognised that Statistics as a discipline within the university would only receive its proper recognition and grow if it were a separate department and argued accordingly for this, within a larger grouping of departments, including that of Mathematics. Under his leadership a de facto department was effectively created, and the university recognised this in 1991 with the establishment of the Department of Statistics.

Jeff's background in applied probability led some to worry that he might alter the applied focus of statistics at Massey. However, Jeff always put the interests of the statisticians as a group first, and the subject flourished as a result of judicious staff appointments through the early 1990s.

In 1993, Massey University opened its new Albany campus at Auckland and Jeff was fully supportive of the small Statistics group (consisting of Denny Meyer, Barry McDonald, and myself) that was set up there in 1994 to support the offerings in Business and in Information Sciences. The campus grew swiftly and so did the group, and when the Faculty of Information and Mathematical Sciences went out of existence as part of university restructuring in 1997 Jeff was appointed as Foundation Head of the Institute of Information and Mathematical Sciences which was (and continued to be) based on the Albany campus. Both Statistics in particular and Information Sciences in general flourished under Jeff's leadership, and when Jeff stepped down as Head in 2002, he continued to lead the Statistics group at Albany.

Those of us who have been part of Statistics at Massey University have been fortunate to benefit from Jeff's leadership, vision, and collegiality over the past 17 years, and we owe him a huge debt of thanks in making Statistics the success it is at Massey University.

Howard Edwards
Programmes Director for Information Sciences
July, 2007

A Personal Note to Jeff J. Hunter on His Retirement by Shayle Searle

What would mathematically bring together a Markov chain enthusiast and a linear models devotee? The link for Jeff and me is generalized inverse matrices. And it is of no small importance that we are also both New Zealanders each with an off-shore Ph.D., both enjoy travel and for more than twenty years have had much pleasure in being guests in each other's homes. For a variety of reasons, some legitimate and some patently transparent, I have been fairly inactive statistically since retiring twelve years ago. Hence, I cannot, in the time made available to me, offer anything new, as much as I would like to be in a position to do so. Retirement for me has been, and is, a delight, especially for relaxingly doing things other than statistics. I sincerely hope that my good friend Jeff and his wife Hazel get the same pleasure from it as I do. All good wishes to you both.

Emeritus Professor Shayle Searle
April, 2007

Paul Cowpertwait: Institute of Information and Mathematical Sciences, College of Sciences, Massey University, Auckland, New Zealand
Email address: p.s.cowpertwait@massey.ac.nz

Graeme Wake: Institute of Information and Mathematical Sciences, College of Sciences, Massey University, Auckland, New Zealand
Email address: g.c.wake@massey.ac.nz

Robert D. Anderson: College of Sciences, Massey University, Auckland, New Zealand
Email address: r.d.anderson@massey.ac.nz

Howard Edwards: Institute of Information and Mathematical Sciences, College of Sciences, Massey University, Auckland, New Zealand
Email address: h.edwards@massey.ac.nz

Shayle Searle: Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 13101, USA
Email address: srs24@cornell.edu

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 48138, 14 pages
doi:10.1155/2007/48138

Research Article

Marked Continuous-Time Markov Chain Modelling of Burst Behaviour for Single Ion Channels

Frank G. Ball, Robin K. Milne, and Geoffrey F. Yeo

Received 2 May 2007; Accepted 8 August 2007

Recommended by Graeme Charles Wake

Patch clamp recordings from ion channels often show *bursting* behaviour, that is, periods of repetitive activity, which are noticeably separated from each other by periods of inactivity. A number of authors have obtained results for important properties of theoretical and empirical bursts when channel gating is modelled by a continuous-time Markov chain with a finite-state space. We show how the use of marked continuous-time Markov chains can simplify the derivation of (i) the distributions of several burst properties, including the total open time, the total charge transfer, and the number of openings in a burst, and (ii) the form of these distributions when the underlying gating process is time reversible and in equilibrium.

1. Introduction

Movement of ions across biological membranes is selectively controlled by specialised protein molecules, called ion channels, which thereby regulate many aspects of cell function. The many kinds of ion channels vary in location, size, chemical structure and function; see, for example, Sakmann and Neher [1]. Usually, ion conduction occurs through a single aqueous pore having a gate that is controlled, for example, by a neurotransmitter, voltage, or membrane tension. Understanding the behaviour of ion channels is important in the study of cell regulation and its pathologies; certain diseases and drugs may affect behaviour of particular channels, and consequently cell functioning. Recordings of the ion flux (tiny current of the order of a few picoamperes) from a single channel are possible through the patch clamp technique (Hamill et al. [2]). At typical recording time resolution, channel gating appears instantaneous, and at any particular time the channel is in one of its stable conductance levels; the simplest channel types exhibit just two,

commonly termed *open* (conducting) and *closed* (nonconducting), though some have multiple conductance levels.

Gating behaviour of a single ion channel is usually modelled by a continuous-time homogeneous finite-state Markov chain; see Colquhoun and Hawkes [3]. (Other background and references can be gleaned from the work of Sakmann and Neher [1].) Two complications often need to be addressed in such modelling: because each conductance level may arise from several states, there may be *aggregation* of states into conductance classes which partition the state space into open and closed states in the case of just two conductance levels; also, because of inherent limitations of the recording procedure, very brief sojourns in a class may not be observable (see, e.g., Ball et al. [4] and Hawkes et al. [5]).

Periods of repetitive open channel activity known as *bursts* are often present in a single channel record, and these are noticeably separated from each other by periods of inactivity. Essentially, a burst is a sequence of periods during which the channel is open together with the intervening short closed times, commonly called *gaps*; neighbouring bursts are separated by much longer closed times, termed *interburst sojourns*. Two types of burst have been studied: *theoretical bursts* depend on a partitioning of the closed states into short-lived and long-lived states; *empirical bursts* depend on closed-times being classified as short or long according to whether they do not or do exceed some specified critical time t_c. In practice, from a single channel record only empirical bursts can be determined, and some of their global properties (such as total charge transfer—see Section 4.2) may be less sensitive to problems caused by missed brief sojourns than individual open and closed sojourns. Furthermore, activity *within* a burst is likely to come from only *one* channel even when there are several channels in the patch; consequently data from within empirical bursts are often used for statistical analyses (see, e.g., Colquhoun et al. [6] and Beato et al. [7]). Ball et al. [8, 9] have discussed other reasons for studying bursts.

For a channel with two conductance levels, Colquhoun and Hawkes [3] showed, under diagonalisability assumptions, that the distributions of the duration, total open time, and number of openings in a theoretical burst are each linear combinations of (resp.) exponential or geometric distributions, and that the numbers of these components can be related to the structure of the underlying gating process. Empirical bursts were first considered by Colquhoun and Sakmann [10]; later studies include Ball [11], Li et al. [12], and Yeo et al. [13].

Ball et al. [8, 9] developed a multivariate semi-Markov framework for analysing burst properties of multiconductance channels, that encompassed both theoretical and empirical bursts in a unified fashion, and investigated the form of distributions of burst properties when the underlying channel is in equilibrium and time reversible. (In the absence of a free energy source, any plausible model of channel gating should be time reversible, see Laüger [14].) The aim of the present paper is to show how the results in Ball et al. [8, 9] can be accessed more easily through a marked continuous-time Markov chain (cf. He and Neuts [15]) which is derived from the underlying continuous-time Markov model describing the channel gating behaviour by deleting closed sojourns and concatenating the open sojourns; the marks allow transitions corresponding to the deleted closed sojourns to be labelled according to whether they are gaps or interburst periods. Concatenated

processes have been used previously to explain some burst properties; see, for example, Colquhoun and Hawkes [3, pages 20–22] and Ball et al. [8, pages 192-193], [9, page 217]. However, they have not been used previously to provide a systematic approach like that developed in the present paper for derivation of burst properties.

Some background and basic notation for Markov modelling of a single ion channel is given in Section 2, along with definitions of bursts and the key marked continuous-time Markov chain. Section 3 develops some fundamental structural properties of transition-rate matrices and equilibrium distributions relevant for study of bursts, and shows that the key marked process inherits time reversibility from the underlying process. Section 4 then presents derivations for some particular burst properties, the total open time, total charge transfer, and number of openings during a burst. In addition, it summarizes results for other properties, such as the time spent in and the number of visits to a subclass of the open states during a burst. Section 5 makes concluding remarks about some extensions, the advantages and disadvantages of the present approach relative to previous ones, and other applications.

Throughout this paper, vectors and matrices are rendered in bold, all vectors are column vectors, and "$^\top$" denotes transpose, which is used to express row vectors. Furthermore, $\boldsymbol{I}$ denotes an identity matrix, $\mathbf{1}$ a column vector of ones, and $\mathbf{0}$ a matrix (vector) of zeros, dimensions of these being clear from their context.

2. Background and notation

We assume that the gating mechanism of a single ion channel is modelled by an irreducible homogeneous continuous-time Markov chain $\{X(t)\} = \{X(t) : t \geq 0\}$, with finite-state space $E = \{1,2,\ldots,n\}$, transition-rate matrix $\boldsymbol{Q} = [q_{ij}]$, and equilibrium distribution $\boldsymbol{\pi} = [\pi_1,\pi_2,\ldots,\pi_n]^\top$. The state space is partitioned as $E = O \cup C$, where $O = \{1,2,\ldots,n_O\}$ and $C = \{n_O+1, n_O+2,\ldots,n\}$ correspond to the channel being open and closed, respectively. The closed states are further partitioned as $C = S \cup L$, where $S = \{n_O+1, n_O+2,\ldots,n_O+n_S\}$ and $L = \{n_O+n_S+1, n_O+n_S+2,\ldots,n\}$ are the short-lived and long-lived closed states, respectively. Let $n_C = n - n_O$ be the number of closed states and $n_L = n - n_O - n_S = n_C - n_S$ be the number of long-lived closed states.

The transition-rate matrix $\boldsymbol{Q}$ may be partitioned in various ways according to the problem under consideration, for example, by the open and closed classes O and C, or by the open, short-lived closed and long-lived closed classes O, S, and L, giving, respectively,

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_{OO} & \boldsymbol{Q}_{OC} \\ \boldsymbol{Q}_{CO} & \boldsymbol{Q}_{CC} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{OO} & \boldsymbol{Q}_{OS} & \boldsymbol{Q}_{OL} \\ \boldsymbol{Q}_{SO} & \boldsymbol{Q}_{SS} & \boldsymbol{Q}_{SL} \\ \boldsymbol{Q}_{LO} & \boldsymbol{Q}_{LS} & \boldsymbol{Q}_{LL} \end{bmatrix}. \tag{2.1}$$

Corresponding partitions are used for the equilibrium distribution $\boldsymbol{\pi}$, that is, $\boldsymbol{\pi}^\top = [\boldsymbol{\pi}_O^\top, \boldsymbol{\pi}_C^\top] = [\boldsymbol{\pi}_O^\top, \boldsymbol{\pi}_S^\top, \boldsymbol{\pi}_L^\top]$.

We now give formal definitions of the two types of burst. For a *theoretical burst*, a sojourn of $\{X(t)\}$ in the class C is classified as an *interburst sojourn* if it contains a visit to L, and is classified as a *gap* if it is purely within the class S. The interburst sojourns are used to partition the channel record into bursts. Thus, a given burst begins at the start of the first O sojourn following an interburst sojourn, and ends at the start of the

subsequent interburst sojourn. An *empirical burst* is defined by specifying a *critical time* $t_c > 0$ and classifying sojourns in C of duration $> t_c$ as interburst sojourns and those of duration $\leq t_c$ as gaps. A given burst is then defined as for a theoretical burst but with this new definition of interburst sojourns and gaps.

Some basic results for aggregated continuous-time Markov chains, required in the sequel, are now summarized. For $t \geq 0$, let $\boldsymbol{P}_C(t) = [p_{ij}^C(t)]$, where

$$p_{ij}^C(t) = P\big(X(t) = j,\, X(u) \in C \text{ for } 0 \leq u \leq t \mid X(0) = i\big) \quad (i, j \in C). \tag{2.2}$$

Then, a standard forward argument (see, e.g., Colquhoun and Hawkes [3, pages 9, 10]) shows that

$$\boldsymbol{P}_C(t) = \exp\big(\boldsymbol{Q}_{CC}t\big) \quad (t \geq 0), \tag{2.3}$$

where $\exp(\boldsymbol{Q}_{CC}t) = \sum_{k=0}^{\infty} t^k \boldsymbol{Q}^k/k!$ denotes the usual matrix exponential.

Suppose that $X(0) \in C$ and let $T_C = \inf\{t > 0 : X(t) \in O\}$ denote the time elapsing until the channel enters an open state. Then T_C has (matrix) probability density function given by

$$\boldsymbol{f}_{T_C}(t) = \exp\big(\boldsymbol{Q}_{CC}t\big)\boldsymbol{Q}_{CO} \quad (t > 0), \tag{2.4}$$

where $\boldsymbol{f}_{T_C}(t) = [f_{ij}^{T_C}(t)]$ with

$$f_{ij}^{T_C}(t) = \frac{d}{dt}P\big(T_C \leq t,\, X(T_C) = j \mid X(0) = i\big) \quad (i \in C,\ j \in O). \tag{2.5}$$

Hence, if $\boldsymbol{P}_{CO} = [p_{ij}^{CO}]$, where $p_{ij}^{CO} = P(X(T_C) = j \mid X(0) = i)$ $(i \in C,\ j \in O)$, then

$$\boldsymbol{P}_{CO} = \int_0^{\infty} \boldsymbol{f}_{T_C}(t)\,dt = \int_0^{\infty} \exp\big(\boldsymbol{Q}_{CC}t\big)\boldsymbol{Q}_{CO}\,dt = \big(-\boldsymbol{Q}_{CC}^{-1}\big)\boldsymbol{Q}_{CO}. \tag{2.6}$$

Note that $\boldsymbol{Q}_{CC}$ is nonsingular since C is a transient class (as $\{X(t)\}$ is irreducible), and hence by Asmussen [16, page 77] all the eigenvalues of $\boldsymbol{Q}_{CC}$ have strictly negative real parts.

Let $\{\widetilde{X}(t)\}$ be the process obtained from $\{X(t)\}$ by deleting all closed sojourns and concatenating the open sojourns; see Figures 2.1(a) and 2.1(b). The process $\{\widetilde{X}(t)\}$ is a continuous-time Markov chain, with state space O. Let $\boldsymbol{Q}_{OO}^{\text{cat}} = [q_{ij}^{\text{cat}}]$ denote the transition-rate matrix for concatenated open-to-open transitions; that is, for $i, j \in O$, q_{ij}^{cat} is the rate that, given the channel is in state i, it leaves the open states and subsequently reenters the open states via state j. Then it follows from (2.6) that

$$\boldsymbol{Q}_{OO}^{\text{cat}} = \boldsymbol{Q}_{OC}\big(-\boldsymbol{Q}_{CC}^{-1}\big)\boldsymbol{Q}_{CO}. \tag{2.7}$$

Thus $\{\widetilde{X}(t)\}$ has $n_O \times n_O$ transition-rate matrix, $\widetilde{\mathbf{Q}}$ say, given by

$$\widetilde{\mathbf{Q}} = \boldsymbol{Q}_{OO} + \boldsymbol{Q}_{OC}\big(-\boldsymbol{Q}_{CC}^{-1}\big)\boldsymbol{Q}_{CO}. \tag{2.8}$$

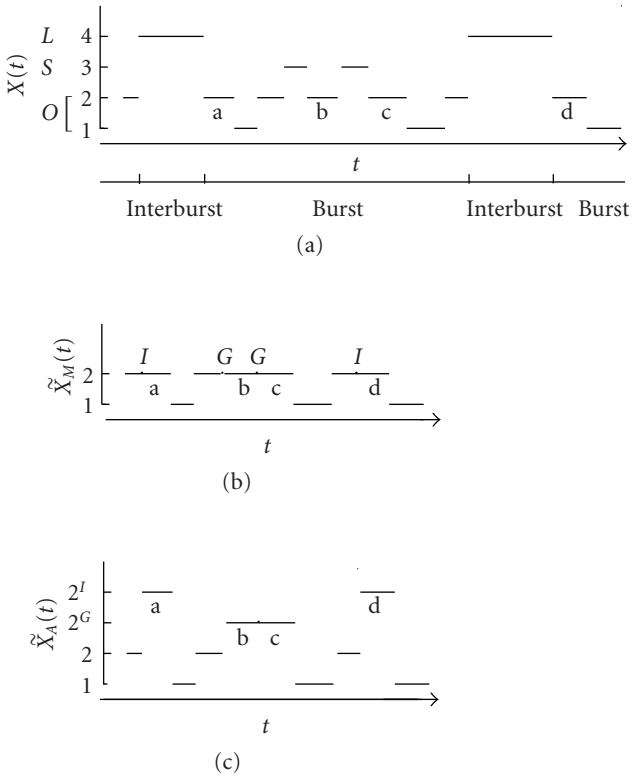


FIGURE 2.1. (a) Partial realization of $\{X(t)\}$ based on $O = \{1,2\}$, $S = \{3\}$, and $L = \{4\}$, with $q_{13} = q_{31} = q_{14} = q_{41} = 0$ and all other entries of $\boldsymbol{Q}$ nonzero. (For corresponding state space graph see Figure 3.1(a).) Labels a, b, c, d indicate open sojourns immediately preceded by a closed sojourn (allowing these to be tracked in parts (b) and (c)). Also indicated are (theoretical) burst and interburst periods. (b) Corresponding realization of $\{\widetilde{X}_M(t)\}$, that is, realization from (a) after omission of gaps and interburst sojourns, concatenation of neighbouring (open) sojourns, and addition of marks I and G to indicate (preceding omitted) interburst sojourn or gap. Corresponding realization of $\{\widetilde{X}(t)\}$ is obtained by omitting marks. (c) Corresponding realization of augmented process $\{\widetilde{X}_A(t)\}$ (as introduced following proof of Theorem 3.2). This requires two states, denoted 2^G and 2^I, additional to states 1 and 2; these carry information previously indicated by marks I and G in (b).

It is easily verified that $\widetilde{\boldsymbol{Q}}$ satisfies $\widetilde{\boldsymbol{Q}}\mathbf{1} = \mathbf{0}$, so it is a proper transition-rate matrix. To see this, start with $\widetilde{\boldsymbol{Q}}\mathbf{1} = \boldsymbol{Q}_{OO}\mathbf{1} + \boldsymbol{Q}_{OC}(-\boldsymbol{Q}_{CC}^{-1})\boldsymbol{Q}_{CO}\mathbf{1}$. Expanding $\boldsymbol{Q}\mathbf{1} = \mathbf{0}$ in partitioned form yields $\boldsymbol{Q}_{OO}\mathbf{1} + \boldsymbol{Q}_{OC}\mathbf{1} = \mathbf{0}$ and $\boldsymbol{Q}_{CO}\mathbf{1} + \boldsymbol{Q}_{CC}\mathbf{1} = \mathbf{0}$. The latter implies that $(-\boldsymbol{Q}_{CC}^{-1})\boldsymbol{Q}_{CO}\mathbf{1} = \mathbf{1}$, whence $\widetilde{\boldsymbol{Q}}\mathbf{1} = \mathbf{0}$ using the former.

Finally, let $\{\widetilde{X}_M(t)\}$ be defined analogously to $\{\widetilde{X}(t)\}$ except that whenever a closed sojourn of $\{X(t)\}$ is deleted, the corresponding transition of $\{\widetilde{X}(t)\}$ (which may not involve a change of state) is marked G or I, according to whether the closed sojourn of $\{X(t)\}$ is a gap or an interburst sojourn; see Figure 2.1(b).

3. Basic results

The transition-rate matrix, $\widetilde{\boldsymbol{Q}}$, of $\{\widetilde{X}(t)\}$ can be decomposed as $\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}_O + \boldsymbol{Q}_G + \boldsymbol{Q}_I$, where $\boldsymbol{Q}_O$ corresponds to transitions of $\{X(t)\}$ purely within O (i.e., without any deletion and concatenation), and $\boldsymbol{Q}_G$ and $\boldsymbol{Q}_I$ to transitions which result from deleted sojourns which were gaps and interburst sojourns, respectively.

Theorem 3.1. *For both types of burst,*

$$\boldsymbol{Q}_O = \boldsymbol{Q}_{OO}. \tag{3.1}$$

For a theoretical burst,

$$\boldsymbol{Q}_G = \boldsymbol{Q}_{OS}(-\boldsymbol{Q}_{SS}^{-1})\boldsymbol{Q}_{SO}, \tag{3.2}$$

$$\boldsymbol{Q}_I = -[\boldsymbol{Q}_{OL} + \boldsymbol{Q}_{OS}(-\boldsymbol{Q}_{SS}^{-1})\boldsymbol{Q}_{SL}][\boldsymbol{Q}_{LL} + \boldsymbol{Q}_{LS}(-\boldsymbol{Q}_{SS}^{-1})\boldsymbol{Q}_{SL}]^{-1}[\boldsymbol{Q}_{LO} + \boldsymbol{Q}_{LS}(-\boldsymbol{Q}_{SS}^{-1})\boldsymbol{Q}_{SO}]. \tag{3.3}$$

For an empirical burst,

$$\boldsymbol{Q}_G = \boldsymbol{Q}_{OC}(-\boldsymbol{Q}_{CC}^{-1})(\boldsymbol{I} - \mathrm{e}^{\boldsymbol{Q}_{CC}t_c})\boldsymbol{Q}_{CO}, \tag{3.4}$$

$$\boldsymbol{Q}_I = \boldsymbol{Q}_{OC}(-\boldsymbol{Q}_{CC}^{-1})\,\mathrm{e}^{\boldsymbol{Q}_{CC}t_c}\boldsymbol{Q}_{CO}. \tag{3.5}$$

Proof. The off-diagonal elements of (3.1) are clear; since $\widetilde{\boldsymbol{Q}}$ is a proper transition matrix, the diagonal elements of (3.1) follow, respectively, for each type of burst once (3.2) and (3.3), or (3.4) and (3.5), have been established.

For a theoretical burst, (3.2) follows from (2.7) with C replaced by S. To prove (3.3), consider an alternative concatenation of $\{X(t)\}$ in which sojourns in S are deleted unless they are gaps. This yields a continuous-time Markov chain $\{X'(t)\}$ say, with transition-rate matrix $\boldsymbol{Q}'$ having partitioned form

$$\boldsymbol{Q}' = \begin{bmatrix} \boldsymbol{Q}'_{OO} & \boldsymbol{Q}'_{OS} & \boldsymbol{Q}'_{OL} \\ \boldsymbol{Q}'_{SO} & \boldsymbol{Q}'_{SS} & \boldsymbol{0} \\ \boldsymbol{Q}'_{LO} & \boldsymbol{0} & \boldsymbol{Q}'_{LL} \end{bmatrix}. \tag{3.6}$$

Now, arguing as for (2.8), $\boldsymbol{Q}'_{LL} = \boldsymbol{Q}_{LL} + \boldsymbol{Q}_{LS}(-\boldsymbol{Q}_{SS}^{-1})\boldsymbol{Q}_{SL}$. Also $\boldsymbol{Q}'_{OL} = \boldsymbol{Q}_{OL} + \boldsymbol{Q}_{OS}(-\boldsymbol{Q}_{SS}^{-1})\boldsymbol{Q}_{SL}$, where the first term corresponds to transitions directly from O to L and the second to transitions that involve an intervening sojourn in S. Similarly, $\boldsymbol{Q}'_{LO} = \boldsymbol{Q}_{LO} + \boldsymbol{Q}_{LS}(-\boldsymbol{Q}_{SS}^{-1})\boldsymbol{Q}_{SO}$. It then follows as in (2.7), with C replaced by L, that $\boldsymbol{Q}_I = \boldsymbol{Q}'_{OL}(-\boldsymbol{Q}'_{LL})^{-1}\boldsymbol{Q}'_{LO}$, yielding (3.3).

For an empirical burst, using (2.4), $\boldsymbol{Q}_G = \int_0^{t_c} \boldsymbol{Q}_{OC}\mathrm{e}^{\boldsymbol{Q}_{CC}t}\boldsymbol{Q}_{CO}\,dt$ and $\boldsymbol{Q}_I = \int_{t_c}^{\infty} \boldsymbol{Q}_{OC}\mathrm{e}^{\boldsymbol{Q}_{CC}t}\boldsymbol{Q}_{CO}\,dt$; hence (3.4) and (3.5) follow. □

The process $\{\widetilde{X}(t)\}$ inherits irreducibility from $\{X(t)\}$, so $\{\widetilde{X}(t)\}$ possesses an equilibrium distribution, $\widetilde{\boldsymbol{\pi}}$ say (of dimension n_O). It is intuitively clear that

$$\widetilde{\boldsymbol{\pi}} = (\boldsymbol{\pi}_O^{\top}\boldsymbol{1})^{-1}\boldsymbol{\pi}_O, \tag{3.7}$$

since concatenating closed sojourns does not affect the long-term relative proportions of time that $\{X(t)\}$ spends in the different open states. More formally, it is easily verified that $\widetilde{\boldsymbol{\pi}}^{\top}\widetilde{\boldsymbol{Q}} = \mathbf{0}$. For example, for empirical bursts, $\boldsymbol{\pi}_O^{\top}\widetilde{\boldsymbol{Q}} = \boldsymbol{\pi}_O^{\top}\boldsymbol{Q}_{OO} + \boldsymbol{\pi}_O^{\top}\boldsymbol{Q}_{OC}(-\boldsymbol{Q}_{CC}^{-1})\boldsymbol{Q}_{CO}$. Now, $\boldsymbol{\pi}^{\top}\boldsymbol{Q} = \mathbf{0}$, since $\boldsymbol{\pi}$ is the equilibrium distribution of $\{X(t)\}$, and expanding this in partitioned form yields $\boldsymbol{\pi}_O^{\top}\boldsymbol{Q}_{OC} + \boldsymbol{\pi}_C^{\top}\boldsymbol{Q}_{CC} = \mathbf{0}$, so $\boldsymbol{\pi}_O^{\top}\boldsymbol{Q}_{OC}(-\boldsymbol{Q}_{CC}^{-1})\boldsymbol{Q}_{CO} = \boldsymbol{\pi}_C^{\top}\boldsymbol{Q}_{CO}$. Hence, $\boldsymbol{\pi}_O^{\top}\widetilde{\boldsymbol{Q}} = \boldsymbol{\pi}_O^{\top}\boldsymbol{Q}_{OO} + \boldsymbol{\pi}_C^{\top}\boldsymbol{Q}_{CO} = \mathbf{0}$, since $\boldsymbol{\pi}^{\top}\boldsymbol{Q} = \mathbf{0}$. Thus $\widetilde{\boldsymbol{\pi}}^{\top}\widetilde{\boldsymbol{Q}} = \mathbf{0}$, as required. A similar argument holds for theoretical bursts.

Recall that $\{X(t)\}$ is reversible if and only if the detailed balance conditions

$$\pi_i q_{ij} = \pi_j q_{ji} \quad (i, j \in E) \tag{3.8}$$

are satisfied. Let $\boldsymbol{W} = \operatorname{diag}(\boldsymbol{\pi})$ be the diagonal matrix whose entries on the diagonal are those of $\boldsymbol{\pi}$. Then (3.8) can be written as

$$\boldsymbol{W}^{1/2}\boldsymbol{Q}\boldsymbol{W}^{-1/2} = (\boldsymbol{W}^{1/2}\boldsymbol{Q}\boldsymbol{W}^{-1/2})^{\top}. \tag{3.9}$$

Expanding (3.9) in partitioned form yields (cf. Fredkin et al. [17]) that if $A \subseteq E$ and $\boldsymbol{W}_A = \operatorname{diag}(\boldsymbol{\pi}_A)$ then

$$\boldsymbol{W}_A^{1/2}\boldsymbol{Q}_{AA}\boldsymbol{W}_A^{-1/2} = (\boldsymbol{W}_A^{1/2}\boldsymbol{Q}_{AA}\boldsymbol{W}_A^{-1/2})^{\top}, \tag{3.10}$$

while if $A, B \subset E$ are disjoint then

$$\boldsymbol{W}_A^{1/2}\boldsymbol{Q}_{AB}\boldsymbol{W}_B^{-1/2} = (\boldsymbol{W}_B^{1/2}\boldsymbol{Q}_{BA}\boldsymbol{W}_A^{-1/2})^{\top}. \tag{3.11}$$

Theorem 3.2. *For both theoretical and empirical bursts, if $\{X(t)\}$ is reversible, then so are $\{\widetilde{X}(t)\}$ and $\{\widetilde{X}_M(t)\}$.*

Proof. Again this is clear on intuitive grounds. For a formal proof we show that detailed balance holds for the three types of transition in $\{\widetilde{X}_M(t)\}$, that is, that $\widetilde{\boldsymbol{W}}^{1/2}\boldsymbol{Q}_O\widetilde{\boldsymbol{W}}^{-1/2}$, $\widetilde{\boldsymbol{W}}^{1/2}\boldsymbol{Q}_G\widetilde{\boldsymbol{W}}^{-1/2}$, and $\widetilde{\boldsymbol{W}}^{1/2}\boldsymbol{Q}_I\widetilde{\boldsymbol{W}}^{-1/2}$ are all symmetric, where $\widetilde{\boldsymbol{W}} = \operatorname{diag}(\widetilde{\boldsymbol{\pi}})$. Note that, because of (3.7), it is sufficient to show that $\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_O\boldsymbol{W}_O^{-1/2}$, $\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_G\boldsymbol{W}_O^{-1/2}$, and $\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}$ are all symmetric. Setting $A = O$ in (3.10) and recalling (3.1) shows that $\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_O\boldsymbol{W}_O^{-1/2}$ is symmetric for both types of burst.

For theoretical bursts, using (3.2),

$$\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_G\boldsymbol{W}_O^{-1/2} = \boldsymbol{W}_O^{1/2}\boldsymbol{Q}_{OS}\boldsymbol{W}_S^{-1/2}[\boldsymbol{W}_S^{1/2}(-\boldsymbol{Q}_{SS})\boldsymbol{W}_S^{-1/2}]^{-1}\boldsymbol{W}_S^{1/2}\boldsymbol{Q}_{SO}\boldsymbol{W}_O^{-1/2}, \tag{3.12}$$

which is symmetric, because of (3.10) with $A = S$ and (3.11) with $A = O$ and $B = S$. A similar argument shows that $\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}$ is symmetric.

For empirical bursts, noting that $-\boldsymbol{Q}_{CC}^{-1}(\boldsymbol{I} - \mathrm{e}^{\boldsymbol{Q}_{CC}t_c}) = \sum_{k=1}^{\infty}\boldsymbol{Q}_{CC}^{k-1}t_c^k/k!$,

$$\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_G\boldsymbol{W}_O^{-1/2} = \boldsymbol{W}_O^{1/2}\boldsymbol{Q}_{OC}\boldsymbol{W}_C^{-1/2}\left[\sum_{k=1}^{\infty}(\boldsymbol{W}_C^{1/2}\boldsymbol{Q}_{CC}\boldsymbol{W}_C^{-1/2})^{k-1}t_c^k/k!\right]\boldsymbol{W}_C^{1/2}\boldsymbol{Q}_{CO}\boldsymbol{W}_O^{-1/2}, \tag{3.13}$$

which is symmetric, because of (3.10) with $A = C$ and (3.11) with $A = O$ and $B = C$. Similarly, $\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}$ is symmetric. □

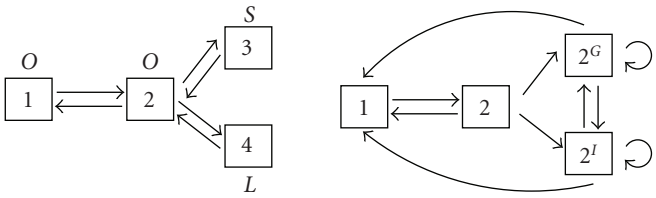


FIGURE 3.1. State space graphs based on $O = \{1,2\}$, $S = \{3\}$, and $L = \{4\}$, with $q_{13} = q_{31} = q_{14} = q_{41} = 0$ and all other entries of $\boldsymbol{Q}$ nonzero. (a) State space graph for underlying process. (b) State space graph for augmented process corresponding to (a); note that only states 1, 2, and additional states 2^G and 2^I are required in this case. The augmented process is clearly nonreversible; for example, state 1 can be reached from, but not followed by, 2^G or 2^I.

The marked process $\{\widetilde{X}_M(t)\}$ could in principle be modelled by augmenting the state space of $\{\widetilde{X}(t)\}$ to indicate whether the current state was immediately preceded by another open state, a deleted gap, or a deleted interburst sojourn. This augmented process, $\{\widetilde{X}_A(t)\}$ say, is a continuous-time Markov chain. Suppose that $\{X(t)\}$ has state space graph as in Figure 3.1(a). In this example, since state 1 cannot be reached directly from either state 3 or state 4, only states 1 and 2, and two additional states, 2^G and 2^I (say), are required for the augmented process. Figure 3.1(b) gives the state space graph and shows the nonreversibility of this augmented process; see Figure 2.1(c) for a typical (partial) realization of $\{\widetilde{X}_A(t)\}$, corresponding to those for $\{X(t)\}$ and $\{\widetilde{X}_M(t)\}$ in Figures 2.1(a) and 2.1(b). In general, the augmented process requires a state space which is up to three times the size of that of the marked process: $\{1,2,\ldots,n_O,\ 1^G,2^G,\ldots,n_O^G,\ 1^I,2^I,\ldots,n_O^I\}$ (say). Hence, this approach would not be so useful because, as well as increasing the size of the state space, such an augmented process need not be reversible.

Let $\{J_k\}$ be the discrete-time Markov chain that records the entry state of successive bursts, that is, the state of $\{\widetilde{X}_M(t)\}$ immediately following successive I-marked transitions. The transition matrix of $\{J_k\}$ is $\boldsymbol{P}_B = -\widetilde{\boldsymbol{Q}}_O^{-1}\boldsymbol{Q}_I$, where $\widetilde{\boldsymbol{Q}}_O = \boldsymbol{Q}_O + \boldsymbol{Q}_G$. (By analogy with (2.3), the (matrix) probability that $\{\widetilde{X}_M(t)\}$ does not have an I-transition in $(0,t]$ is $\exp(\widetilde{\boldsymbol{Q}}_O t)$, so $\boldsymbol{P}_B = \int_0^\infty \exp(\widetilde{\boldsymbol{Q}}_O t)\boldsymbol{Q}_I\,dt = -\widetilde{\boldsymbol{Q}}_O^{-1}\boldsymbol{Q}_I$. The matrix $\widetilde{\boldsymbol{Q}}_O$ is nonsingular because its eigenvalues have strictly negative real parts, since $\exp(\widetilde{\boldsymbol{Q}}_O t) \to \boldsymbol{0}$ as $t \to \infty$). Note that $\{J_k\}$ also inherits irreducibility from $\{X(t)\}$, though the state space of $\{J_k\}$ may be a proper subset of O, for example, if there are open states which cannot be entered directly from C. If $\{J_k\}$ is also aperiodic, as is necessarily the case when $\boldsymbol{Q}$ is such that $q_{ij} > 0$ if and only if $q_{ji} > 0$ (a condition that is satisfied by most physically plausible channel gating models and by all time reversible models), then $\{J_k\}$ possesses an equilibrium distribution, $\boldsymbol{\psi} = [\psi_1, \psi_2, \ldots, \psi_{n_O}]^\top$ say, where ψ_i is the equilibrium probability that a burst begins in state i. (If the state space of $\{J_k\}$ is a proper subset of O, then some of the elements of $\boldsymbol{\psi}$ are zero.)

LEMMA 3.3. *The equilibrium distribution $\boldsymbol{\psi}$ of $\{J_k\}$ is given by* $\boldsymbol{\psi}^\top = \boldsymbol{\pi}_O^\top \boldsymbol{Q}_I / (\boldsymbol{\pi}_O^\top \boldsymbol{Q}_I \boldsymbol{1})$.

Proof. Recall that $\widetilde{\boldsymbol{Q}} = \widetilde{\boldsymbol{Q}}_O + \boldsymbol{Q}_I$ and, using (3.7), that $\boldsymbol{\pi}_O^\top \widetilde{\boldsymbol{Q}} = \boldsymbol{0}$. Thus $\boldsymbol{\pi}_O^\top \boldsymbol{Q}_I = -\boldsymbol{\pi}_O^\top \widetilde{\boldsymbol{Q}}_O$, so using $\boldsymbol{P}_B = -\widetilde{\boldsymbol{Q}}_O^{-1} \boldsymbol{Q}_I$ gives $\boldsymbol{\pi}_O^\top \boldsymbol{Q}_I \boldsymbol{P}_B = \boldsymbol{\pi}_O^\top \boldsymbol{Q}_I$. Hence $\boldsymbol{\psi}^\top \boldsymbol{P}_B = \boldsymbol{\psi}^\top$, as required. □

The equilibrium distribution in Lemma 3.3 is intuitively clear in view of (3.7) and the fact that a burst is immediately preceded by an I-transition of $\{\widetilde{X}_M(t)\}$. Alternative expressions for $\boldsymbol{\psi}$ have been given by, for example, Colquhoun and Hawkes [3, Equation (3.2)] for theoretical bursts, and Ball [11, Equation (3.9)] and Li et al. [12, Equation (2.10)] for empirical bursts.

4. Properties of bursts

4.1. Total open time during a burst. Suppose that $\{\widetilde{X}_M(t)\}$ is in equilibrium. Then the times of I-transitions of $\{\widetilde{X}_M(t)\}$ form a stationary point process. Let T_O denote the length of a typical interval in this point process (i.e., the time between two successive I-transitions) and let U_O denote a typical excess lifetime (i.e., the time from an arbitrary time point until the next I-transition of $\{\widetilde{X}_M(t)\}$). Note that, because in $\{\widetilde{X}_M(t)\}$ all closed sojourns have been omitted and the open sojourns concatenated, T_O gives the total open time during a typical burst. Since $\{\widetilde{X}_M(t)\}$ is in equilibrium, the survivor function, $\overline{F}_{U_O}(t)$ say, of U_O is given by

$$\overline{F}_{U_O}(t) = \widetilde{\boldsymbol{\pi}}^\top \mathrm{e}^{\widetilde{\boldsymbol{Q}}_O t} \mathbf{1} \quad (t > 0). \tag{4.1}$$

Thus, by the standard relationship between the distributions of a typical lifetime and a typical excess lifetime of a stationary point process, the pdf of T_O, $f_{T_O}(t)$ say, is given by

$$f_{T_O}(t) = \mu_{T_O} \overline{F}''_{U_O}(t) = \mu_{T_O} \widetilde{\boldsymbol{\pi}}^\top \widetilde{\boldsymbol{Q}}_O^2 \mathrm{e}^{\widetilde{\boldsymbol{Q}}_O t} \mathbf{1} \quad (t > 0), \tag{4.2}$$

where, with D_+ denoting right-hand derivative, $\mu_{T_O} = E[T_O] = [-D_+ \overline{F}_{U_O}(0)]^{-1} = (-\widetilde{\boldsymbol{\pi}}^\top \widetilde{\boldsymbol{Q}}_O \mathbf{1})^{-1}$; cf. Ball and Milne [18].

Now, suppose that $\{X(t)\}$, and hence $\{\widetilde{X}(t)\}$, is time reversible. Then

$$\overline{F}_{U_O}(t) = \mathbf{1}^\top \widetilde{\boldsymbol{W}} \mathrm{e}^{\widetilde{\boldsymbol{Q}}_O t} \mathbf{1} = \mathbf{1}^\top \widetilde{\boldsymbol{W}}^{1/2} \widetilde{\boldsymbol{W}}^{1/2} \mathrm{e}^{\widetilde{\boldsymbol{Q}}_O t} \widetilde{\boldsymbol{W}}^{-1/2} \widetilde{\boldsymbol{W}}^{1/2} \mathbf{1} \quad (t > 0). \tag{4.3}$$

Now, using the series expression for the matrix exponential, $\widetilde{\boldsymbol{W}}^{1/2} \mathrm{e}^{\widetilde{\boldsymbol{Q}}_O t} \widetilde{\boldsymbol{W}}^{-1/2} = \exp(\widetilde{\boldsymbol{W}}^{1/2} \widetilde{\boldsymbol{Q}}_O \widetilde{\boldsymbol{W}}^{-1/2} t)$. Further, $\widetilde{\boldsymbol{W}}^{1/2} \widetilde{\boldsymbol{Q}}_O \widetilde{\boldsymbol{W}}^{-1/2} = \widetilde{\boldsymbol{W}}^{1/2} (\boldsymbol{Q}_O + \boldsymbol{Q}_G) \widetilde{\boldsymbol{W}}^{-1/2}$ is symmetric as $\{\widetilde{X}_M(t)\}$ is time reversible. Hence, $\widetilde{\boldsymbol{W}}^{1/2} \widetilde{\boldsymbol{Q}}_O \widetilde{\boldsymbol{W}}^{-1/2}$ admits the spectral representation

$$\widetilde{\boldsymbol{W}}^{1/2} \widetilde{\boldsymbol{Q}}_O \widetilde{\boldsymbol{W}}^{-1/2} = \sum_{i=1}^{n_O} \lambda_i \boldsymbol{x}_i \boldsymbol{x}_i^\top, \tag{4.4}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_{n_O}$ are the eigenvalues of $\widetilde{\boldsymbol{Q}}_O$, which are all real (as $\widetilde{\boldsymbol{W}}^{1/2} \widetilde{\boldsymbol{Q}}_O \widetilde{\boldsymbol{W}}^{-1/2}$ is symmetric) and strictly negative, and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_O}$ is a corresponding orthonormal set of right eigenvectors.

Substituting (4.4) into (4.3) yields

$$\overline{F}_{U_O}(t) = \sum_{i=1}^{n_O} \alpha_i \mathrm{e}^{\lambda_i t} \quad (t > 0), \tag{4.5}$$

where, for $i = 1,2,\ldots,n_O$, $\alpha_i = \mathbf{1}^\top \widetilde{\boldsymbol{W}}^{1/2} \boldsymbol{x}_i \boldsymbol{x}_i^\top \widetilde{\boldsymbol{W}}^{1/2} \mathbf{1} = (\boldsymbol{x}_i^\top \widetilde{\boldsymbol{W}}^{1/2} \mathbf{1})^2 \geq 0$. Thus, if $\{X(t)\}$ is time-reversible and in equilibrium, then T_O is distributed as a mixture of at most n_O negative exponential random variables; this distribution is obtained from [9, Equation (3.17)] by Ball et al. (by taking their $\boldsymbol{c} = \mathbf{1}$).

4.2. Total charge transfer during a burst. For $i \in O$, let c_i denote the current when $X(t) = i$, that is, when the channel is in open state i. The total charge transfer during a burst is the integral of the current over the burst, which is given by $\int_0^{T_O} c_{\widetilde{X}(t)}\, dt$, assuming that the burst starts at $t = 0$ and the current is zero when $X(t) \in C$. Suppose that $c_i > 0$ $(i \in O)$. Let $\{\widehat{X}(t)\}$ and $\{\widehat{X}_M(t)\}$ denote the random time-changed versions of $\{\widetilde{X}(t)\}$ and $\{\widetilde{X}_M(t)\}$, respectively, obtained by running the clock at rate c_i^{-1} when $\widetilde{X}(t) = i$ $(i \in O)$. Let $C = \mathrm{diag}(\boldsymbol{c})$, where $\boldsymbol{c} = (c_1, c_2, \ldots, c_{n_O})^\top$. The transition-rate matrix, $\widehat{\boldsymbol{Q}}$ say, of $\{\widehat{X}(t)\}$ admits the decomposition $\widehat{\boldsymbol{Q}} = \widehat{\boldsymbol{Q}}_O + \widehat{\boldsymbol{Q}}_G + \widehat{\boldsymbol{Q}}_I$, where $\widehat{\boldsymbol{Q}}_O = \boldsymbol{C}^{-1}\boldsymbol{Q}_O$, $\widehat{\boldsymbol{Q}}_G = \boldsymbol{C}^{-1}\boldsymbol{Q}_G$, and $\widehat{\boldsymbol{Q}}_I = \boldsymbol{C}^{-1}\boldsymbol{Q}_I$. It is easily verified that $\{\widehat{X}(t)\}$ has equilibrium distribution, $\widehat{\boldsymbol{\pi}}$ say, given by $\widehat{\boldsymbol{\pi}}^\top = (\widetilde{\boldsymbol{\pi}}^\top \boldsymbol{C}\mathbf{1})^{-1}\widetilde{\boldsymbol{\pi}}^\top \boldsymbol{C}$, and that $\{\widehat{X}_M(t)\}$ is time reversible if and only if $\{\widetilde{X}_M(t)\}$ is time reversible.

Let $\widehat{T}_O$ be the time elapsing between two successive I-transitions of $\{\widehat{X}_M(t)\}$, that is, the total charge transfer over a typical burst (since all closed sojourns have been omitted and the open sojourns concatenated). Then, in equilibrium, the distribution of $\widehat{T}_O$ is given by (4.2), with $\widetilde{\boldsymbol{\pi}}$ replaced by $\widehat{\boldsymbol{\pi}}$, $\widetilde{\mathbf{Q}}_O$ replaced by $\widetilde{\widehat{\mathbf{Q}}}_O = \widehat{\mathbf{Q}}_O + \widehat{\mathbf{Q}}_G$, and μ_{T_O} replaced by $\widehat{\mu}_{T_O} = -(\widehat{\boldsymbol{\pi}}_O^\top \widetilde{\widehat{\mathbf{Q}}}_O \mathbf{1})^{-1}$. Further, it follows as in Section 4.1 that if $\{X(t)\}$ is time-reversible, then, in equilibrium, $\widehat{T}_O$ is distributed as a mixture of at most n_O negative exponential random variables; see [9, Equation (3.17)] by Ball et al.

4.3. Number of openings during a burst. Let N_O be the number of openings in a burst. Note that $N_O = k$ if and only if, in $\{\widetilde{X}_M(t)\}$, the number of G-marks between two successive I-marks is $k-1$. The (substochastic) transition matrix between two successive marks in $\{\widehat{X}_M(t)\}$ is $-\boldsymbol{Q}_O^{-1}\boldsymbol{Q}_G$ if the second mark is a G, and $-\boldsymbol{Q}_O^{-1}\boldsymbol{Q}_I$ if the second mark is an I. Thus, in equilibrium, and using Lemma 3.3,

$$P(N_O = k) = (\boldsymbol{\pi}_O^\top \boldsymbol{Q}_I \mathbf{1})^{-1} \boldsymbol{\pi}_O^\top \boldsymbol{Q}_I (-\boldsymbol{Q}_O^{-1}\boldsymbol{Q}_G)^{k-1} (-\boldsymbol{Q}_O^{-1}\boldsymbol{Q}_I)\mathbf{1} \quad (k = 1,2,\ldots). \tag{4.6}$$

Suppose that $\{X(t)\}$, and hence $\{\widetilde{X}_M(t)\}$, is time reversible. The strictly positive definite matrix $-\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_O\boldsymbol{W}_O^{-1/2}$ is then symmetric, so $(-\boldsymbol{Q}_O)^{-1/2}$ exists and $\boldsymbol{A}_O$ defined by

$\boldsymbol{A}_O = \boldsymbol{W}_O^{1/2}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2}$ is symmetric. Thus,

$$\begin{aligned}
&P(N_O = k)\\
&= (\boldsymbol{\pi}_O^\top \boldsymbol{Q}_I \mathbf{1})^{-1}\mathbf{1}^\top \boldsymbol{W}_O \boldsymbol{Q}_I (-\boldsymbol{Q}_O)^{-1/2}\left[-(\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_G(-\boldsymbol{Q}_O)^{-1/2}\right]^{k-1}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_I\mathbf{1}\\
&= (\boldsymbol{\pi}_O^\top \boldsymbol{Q}_I \mathbf{1})^{-1}\mathbf{1}^\top \boldsymbol{W}_O^{1/2}\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I (-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2}\left[\boldsymbol{W}_O^{1/2}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_G(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2}\right]^{k-1}\\
&\qquad \times \boldsymbol{W}_O^{1/2}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}\boldsymbol{W}_O^{1/2}\mathbf{1}.
\end{aligned} \tag{4.7}$$

The matrix $\boldsymbol{W}_O^{1/2}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_G(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2} = \boldsymbol{A}_O(\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_G\boldsymbol{W}_O^{-1/2})\boldsymbol{A}_O$ is symmetric and positive (semi-) definite (noting that the eigenvalues of $\boldsymbol{Q}_G$ are nonnegative for both types of burst) and hence admits the spectral representation

$$\boldsymbol{W}_O^{1/2}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_G(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2} = \sum_{i=1}^{n_O}\rho_i \boldsymbol{y}_i\boldsymbol{y}_i^\top. \tag{4.8}$$

Further, the eigenvalues satisfy $0 \le \rho_i < 1$ $(i = 1,2,\ldots,n_O)$ as the matrix on the left in (4.8) is similar to the substochastic matrix $-\boldsymbol{Q}_O^{-1}\boldsymbol{Q}_G$. Substituting (4.8) into (4.7) yields

$$P(N_O = k) = \sum_{i=1}^{n_O}\beta_i\rho_i^{k-1} \quad (k = 1,2,\ldots), \tag{4.9}$$

where $\beta_i = (\boldsymbol{\pi}_O^\top \boldsymbol{Q}_I\mathbf{1})^{-1}\mathbf{1}^\top\boldsymbol{W}_O^{1/2}\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2}\boldsymbol{y}_i\boldsymbol{y}_i^\top\boldsymbol{W}_O^{1/2}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}\boldsymbol{W}_O^{1/2}\mathbf{1}$. Now, as $\{\widetilde{X}_M(t)\}$ is reversible,

$$\begin{aligned}
\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2} &= \boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}\boldsymbol{A}_O\\
&= \left[\boldsymbol{A}_O\boldsymbol{W}_O^{1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}\right]^\top\\
&= \left[\boldsymbol{W}_O^{1/2}(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{Q}_I\boldsymbol{W}_O^{-1/2}\right]^\top.
\end{aligned} \tag{4.10}$$

Thus,

$$\beta_i = (\boldsymbol{\pi}_O^\top\boldsymbol{Q}_I\mathbf{1})^{-1}\left[\mathbf{1}^\top\boldsymbol{W}_O\boldsymbol{Q}_I(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2}\boldsymbol{y}_i\right]\left[\mathbf{1}^\top\boldsymbol{W}_O\boldsymbol{Q}_I(-\boldsymbol{Q}_O)^{-1/2}\boldsymbol{W}_O^{-1/2}\boldsymbol{y}_i\right]^\top \ge 0, \tag{4.11}$$

so if $\{X(t)\}$ is time-reversible and in equilibrium then N_O is distributed as a mixture of at most rank$(\boldsymbol{Q}_G)$ geometric random variables; cf. [9, Equation (3.28)] by Ball et al.

4.4. Other properties. Various other properties of bursts may be obtained by using appropriate marked processes. For example, suppose that $n_O > 1$ and consider a proper subset $A \subset O$ of the class O of open states. Let T_A be denoting the time $\{\widetilde{X}(t)\}$ spends in A between two successive I-transitions. (e.g., if A denotes the open states that have a specified conductance level, then T_A is the total time the channel spends at that conductance

level during a typical burst.) Let $B = O\backslash A$. Now partition $\widetilde{\mathbf{Q}}_O$ and $\boldsymbol{\psi}$, giving, respectively,

$$\widetilde{\mathbf{Q}}_O = \begin{bmatrix} \widetilde{\mathbf{Q}}_{AA} & \widetilde{\mathbf{Q}}_{AB} \\ \widetilde{\mathbf{Q}}_{BA} & \widetilde{\mathbf{Q}}_{BB} \end{bmatrix}, \qquad \boldsymbol{\psi}^\top = [\boldsymbol{\psi}_A^\top, \boldsymbol{\psi}_B^\top]. \tag{4.12}$$

Then, in equilibrium, the probability that the channel visits A during a burst is given by $P(T_A > 0) = \boldsymbol{\psi}_A^\top \mathbf{1} + \boldsymbol{\psi}_B^\top (-\widetilde{\mathbf{Q}}_{BB})^{-1} \widetilde{\mathbf{Q}}_{BA} \mathbf{1}$. Note that it is possible for this probability to be strictly less than one, in which case $P(T_A = 0) > 0$.

The distribution of $T_A \mid T_A > 0$ can be obtained by a further concatenation of $\{\widetilde{X}(t)\}$, in which sojourns in B are deleted. Denote the resulting process by $\{X^*(t)\}$ and the corresponding marked process by $\{X_M^*(t)\}$. The transition-rate matrix of $\{X^*(t)\}$, $\mathbf{Q}^*$ say, admits the decomposition $\mathbf{Q}^* = \mathbf{Q}_A^* + \mathbf{Q}_G^* + \mathbf{Q}_I^*$. Moreover, if $\widetilde{\mathbf{Q}}_A^* = \mathbf{Q}_A^* + \mathbf{Q}_G^*$, then arguing as for (2.8) yields $\widetilde{\mathbf{Q}}_A^* = \widetilde{\mathbf{Q}}_{AA} + \widetilde{\mathbf{Q}}_{AB}(-\widetilde{\mathbf{Q}}_{BB}^{-1})\widetilde{\mathbf{Q}}_{BA}$.

Suppose that $\{X_M^*(t)\}$ is in equilibrium and let U_A denote a typical excess lifetime from an arbitrary time until the next I-transition of $\{X_M^*(t)\}$. Then U_A has survivor function given by $\overline{F}_{U_A}(t) = (\boldsymbol{\pi}^*)^\top \mathrm{e}^{\widetilde{\mathbf{Q}}_A^* t} \mathbf{1}$ $(t > 0)$, where $\boldsymbol{\pi}^* = (\boldsymbol{\pi}_A^\top \mathbf{1})^{-1} \boldsymbol{\pi}_A$ is the equilibrium distribution of $\{X^*(t)\}$. Arguing as in Section 4.1 now shows that, if $\{X(t)\}$ is time reversible and in equilibrium, then $T_A \mid T_A > 0$ is distributed as a mixture of at most n_A negative exponential random variables, where n_A is the number of states in A. The distribution of the total charge transfer whilst in A during a burst can be determined using a random time transformation of $\{X_M^*(t)\}$, as in Section 4.2; details concerning this distribution are given in Ball et al. [9, Equations (3.13) and (3.14)].

Let N_A denote the number of visits to A during a burst. The distribution of N_A comprises a point mass at zero, given by $P(N_A = 0) = P(T_A = 0)$, and a (possibly defective) distribution on the positive integers. Moreover, similar arguments to those used in Section 4.3 show that if $\{X(t)\}$ is time-reversible and in equilibrium, then $N_A \mid N_A > 0$ is distributed as a mixture of at most rank$(\boldsymbol{P}_A)$ geometric random variables, where $\boldsymbol{P}_A = (-\widetilde{\mathbf{Q}}_{AA}^{-1})\widetilde{\mathbf{Q}}_{AB}(-\widetilde{\mathbf{Q}}_{BB}^{-1})\widetilde{\mathbf{Q}}_{BA}$ is the (substochastic) transition matrix for entry states of two successive visits to A during a burst; for details see [9, Equation (3.32)] by Ball et al.

5. Concluding remarks

In previous papers, notably Ball et al. [4, 8, 9], we have derived results about ion channel gating behaviour by exploiting structure arising from relevant Markov renewal processes that are embedded in the underlying Markov or Markov renewal process which describes the channel gating. Especially, in [8, 9] the focus was on derivation of burst properties. The present paper has shown that many of the results of those two papers can be obtained much more simply using a suitably marked continuous-time Markov chain which is derived from the assumed underlying continuous-time Markov chain by deleting closed sojourns and concatenating the open sojourns. Other results in those papers, such as the form of autocorrelation functions of burst properties, can also be obtained using the present framework but details are omitted owing to space restrictions. The clarity of the derivations appears to result from them accessing precisely the details of structure which are relevant in each situation, and from exploiting two other aspects. First, the use of excess lifetimes simplifies the derivation of properties of sojourn time pdfs, since they avoid

the use of burst entry process equilibrium distributions and consequently lead more directly to mixtures of exponentials. Second, the expression for the burst entry process equilibrium distribution given in Lemma 3.3 (that arises naturally in the present setting) leads to more efficient derivations of mixture properties than in [9]. The approach of the present paper is not readily applicable when knowledge of the time spent in the deleted closed sojourns is required (e.g., in determining the distribution of burst duration). Also, the method is generally less useful in cases where the channel gating behaviour is modelled by a Markov renewal process that is not a continuous-time Markov chain; the concatenated process is a Markov renewal process but its semi-Markov kernel usually does not take a simple form. Concatenated processes may also prove useful in other areas of application of aggregated processes, such as system reliability (cf. Csenki [19]).

Acknowledgments

Frank Ball thanks the School of Mathematics and Statistics, The University of Western Australia, for hospitality and support during short visits in 2004 and 2006 when some of this research was started. The work was completed whilst Frank Ball held a Leverhulme Research Fellowship, which also is gratefully acknowledged. The authors thank Jeff Hunter for his friendship and salute him on the occasion of his retirement, in particular for his many scholarly contributions to Markov chains, Markov renewal processes, and more widely in applied probability. We are especially grateful for his interest over many years in our own work on Markov renewal processes and their applications in modelling the gating behaviour of ion channels.

References

[1] B. Sakmann and E. Neher, Eds., *Single-Channel Recording*, Plenum Press, New York, NY, USA, 2nd edition, 1995.

[2] O. P. Hamill, A. Marty, E. Neher, B. Sakmann, and F. J. Sigworth, "Improved patch-clamp techniques for high-resolution current recording from cells and cell-free membrane patches," *Pflügers Archiv*, vol. 391, no. 2, pp. 85–100, 1981.

[3] D. Colquhoun and A. G. Hawkes, "On the stochastic properties of bursts of single ion channel openings and of clusters of bursts," *Philosophical Transactions of the Royal Society of London Series B*, vol. 300, no. 1098, pp. 1–59, 1982.

[4] F. G. Ball, R. K. Milne, and G. F. Yeo, "Aggregated semi-Markov processes incorporating time interval omission," *Advances in Applied Probability*, vol. 23, no. 4, pp. 772–797, 1991.

[5] A. G. Hawkes, A. Jalali, and D. Colquhoun, "Asymptotic distributions of apparent open times and shut times in a single channel record allowing for the omission of brief events," *Philosophical Transactions of the Royal Society of London Series B*, vol. 337, no. 1282, pp. 383–404, 1992.

[6] D. Colquhoun, C. J. Hatton, and A. G. Hawkes, "The quality of maximum likelihood estimates of ion channel rate constants," *Journal of Physiology*, vol. 547, no. 3, pp. 699–728, 2003.

[7] M. Beato, P. J. Groot-Kormelink, D. Colquhoun, and L. G. Sivilotti, "The activation mechanism of $\alpha 1$ homomeric glycine receptors," *Journal of Neuroscience*, vol. 24, no. 4, pp. 895–906, 2004.

[8] F. G. Ball, R. K. Milne, and G. F. Yeo, "Multivariate semi-Markov analysis of burst properties of multiconductance single ion channels," *Journal of Applied Probability*, vol. 39, no. 1, pp. 179–196, 2002.

[9] F. G. Ball, R. K. Milne, and G. F. Yeo, "A unified approach to burst properties of multiconductance single ion channels," *Mathematical Medicine and Biology*, vol. 21, no. 3, pp. 205–245, 2004.

[10] D. Colquhoun and B. Sakmann, "Fluctuations in the microsecond time range of the current through single acetylcholine receptor ion channels," *Nature*, vol. 294, no. 5840, pp. 464–466, 1981.

[11] F. G. Ball, "Empirical clustering of bursts of openings in Markov and semi-Markov models of single channel gating incorporating time interval omission," *Advances in Applied Probability*, vol. 29, no. 4, pp. 909–946, 1997.

[12] Y. Li, G. F. Yeo, R. K. Milne, B. W. Madsen, and R. O. Edeson, "Burst properties of a supergated double-barrelled chloride ion channel," *Mathematical Biosciences*, vol. 166, no. 1, pp. 23–44, 2000.

[13] G. F. Yeo, R. K. Milne, B. W. Madsen, Y. Li, and R. O. Edeson, "Markov modelling of burst behaviour in ion channels," in *Stochastic Processes: Modelling and Simulation*, D. N. Shanbhag and C. R. Rao, Eds., vol. 21 of *Handbook of Statistics*, pp. 931–968, North-Holland, Amsterdam, The Netherlands, 2003.

[14] P. Laüger, "Conformational transitions of ionic channels," in *Single-Channel Recording*, B. Sakmann and E. Neher, Eds., pp. 651–662, Plenum Press, New York, NY, USA, 2nd edition, 1995.

[15] Q.-M. He and M. F. Neuts, "Markov chains with marked transitions," *Stochastic Processes and Their Applications*, vol. 74, no. 1, pp. 37–52, 1998.

[16] S. Asmussen, *Applied Probability and Queues*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Chichester, UK, 1987.

[17] D. R. Fredkin, M. Montal, and J. A. Rice, "Identification of aggregated Markovian models: application to the nicotinic acetylcholine receptor," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. I (Berkeley, Calif, 1983)*, L. Le Cam and R. A. Olshen, Eds., Wadsworth Statistics/Probability Series, pp. 269–289, Wadsworth, Belmont, Calif, USA, 1985.

[18] F. G. Ball and R. K. Milne, "Applications of simple point process methods to superpositions of aggregated stationary processes," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 1, pp. 181–196, 2004.

[19] A. Csenki, *Dependability for Systems with a Partitioned State Space. Markov and Semi-Markov Theory and Computational Implementation*, vol. 90 of *Lecture Notes in Statistics*, Springer, New York, NY, USA, 1994.

Frank G. Ball: School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham NG7 2RD, UK
Email address: frank.ball@nottingham.ac.uk

Robin K. Milne: School of Mathematics and Statistics, The University of Western Australia, Crawley, WA 6009, Australia
Email address: milne@maths.uwa.edu.au

Geoffrey F. Yeo: School of Mathematics and Statistics, The University of Western Australia, Crawley, WA 6009, Australia
Email address: yeo@maths.uwa.edu.au

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 98086, 13 pages
doi:10.1155/2007/98086

Review Article

Putting Markov Chains Back into Markov Chain Monte Carlo

Richard J. Barker and Matthew R. Schofield

Received 7 May 2007; Accepted 8 August 2007

Recommended by Graeme Charles Wake

Markov chain theory plays an important role in statistical inference both in the formulation of models for data and in the construction of efficient algorithms for inference. The use of Markov chains in modeling data has a long history, however the use of Markov chain theory in developing algorithms for statistical inference has only become popular recently. Using mark-recapture models as an illustration, we show how Markov chains can be used for developing demographic models and also in developing efficient algorithms for inference. We anticipate that a major area of future research involving mark-recapture data will be the development of hierarchical models that lead to better demographic models that account for all uncertainties in the analysis. A key issue is determining when the chains produced by Markov chain Monte Carlo sampling have converged.

1. Introduction

Markov chains, and related stochastic models, have long played an important role in helping ecologists understand population dynamics. In the main, this has been through the application of probability models to the problem of predicting the realized dynamics of plant and animal populations over time. In this context, the challenge is to construct models that are relatively simple, in terms of the numbers of parameters and the relationships between (and among) parameters and state variables, and yet that are able to provide a reasonable approximation to the sorts of dynamics typically observed in nature.

The development of these models has been driven by ecologists motivated by an interest in underlying theory, and ecologists with applications in mind. These applications are generally based on inference about the long-term dynamics of the study population. Examples include predicting the time to possible extinction for threatened populations or establishing safe levels of harvest for exploited populations.

Complementary to the problem of developing theoretical models that predict the temporal dynamics of populations has been the development of a body of theory for estimating population parameters. This estimation theory is the main emphasis in the remainder of this contribution. Almost universally, these models treat quantities such as population size at time t, N_t, or the numbers of individuals born between t and $t + \Delta t$, as fixed quantities to be predicted (here we use prediction in the sense of predicting the value of an unobserved realization of a random variable as well as future realizations). In contrast, the population models discussed above treat these same quantities as random variables whose behavior we are interested in describing.

A particularly important class of estimation models for population dynamics is the mark-recapture model [1]. Mark-recapture data comprises repeated measures on individual animals in the population obtained through samples of animals drawn from the population (usually) at discrete sample times $t_1, \ldots, t_k$. Because the capture process is imperfect, not all animals in the population at time t_j are captured, and a model is required to describe this process of repeated imperfect captures.

In the simplest case, the population is regarded as closed to births and deaths and so the population size at time t_j is the same for all k sample occasions. Studies of the dynamics of the population can then be based on repeated experiments generating a sequence of abundance estimates.

A more interesting class of models is the open population models in which individuals may enter (through birth or immigration) or leave (through death or emigration) during the interval (t_1, t_k). Here the challenge is to model the sequence of captures of animals in terms of parameters of demographic interest. These are usually survival probabilities S_j, and parameters that describe the birth process. Quantities such as N_j, the abundance of animals at the time of sample j, and B_j, the number of animals born between samples j and $j + 1$, are then predicted from the model.

Traditionally, inference for mark-recapture models has been based on maximum likelihood. Importantly, demographic models for $\{N_t\}$ have played no role in these estimation models. Inference about the $\{N_t\}$ process has instead been based on ad hoc methods for summarizing sequences of estimates such as $\{\hat{N}_t\}$. Recently there have been many developments that apply Bayesian inference methods to mark recapture models based on Markov chain Monte Carlo (MCMC). Here the utility of Markov chain theory appears in a fundamentally different context to that described above for population modeling; it arises as a tool for inference. An exciting feature of Bayesian inference methods based on MCMC is that fitting complicated hierarchical models has become feasible. Hierarchical models provide a link between the demographic models and estimation models in a way that should lead to better and more relevant inference. It is recent developments in Markov chain theory, in particular Gibbs sampling [2] and reversible jump Markov chain Monte Carlo [3], that allow this link to take place.

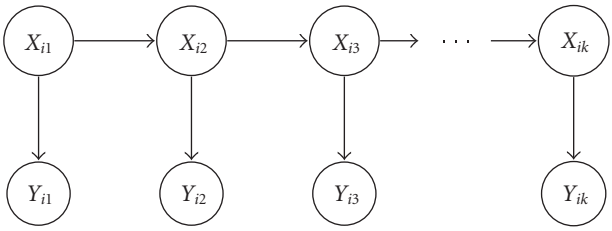


FIGURE 2.1. Directed acyclic graph for the hidden Markov model for mark-recapture data.

2. Open population capture-recapture models

Define ν as the total number of animals alive on at least one of the sampling occasions t_j ($j = 1,\ldots,k$). Let Y_{ij} be the indicator that animal i ($i = 1,\ldots,\nu$) is caught in sample j, which takes place at time t_j. Any animal that is caught for the first time in a particular sample is marked with some form of unique tag and then released, with any recaptures noted. The data are of the form $y = \{y_{ij}\}$ for $i = 1,\ldots,u.$ and $j = 1,\ldots,k$, where u_j is the number of unmarked animals caught in sample j and $u. = \sum_j u_j$ is the total number of animals caught at least once.

For animal i, the set of values $y_i = \{y_{i1},\ldots,y_{ik}\}$ is called the capture history and these provide censored information on the time of birth and death. We know that the time of birth occurred before the sample time corresponding to the first nonzero value of y_i and the time of death occurred sometime after the sample time corresponding to the last nonzero value of y_i.

2.1. Hidden Markov model. The problem with capture-recapture methods is that not all animals in the population are caught; this causes the censoring of the birth and death intervals. Define X_{ij} as the indicator of the event that animal i is alive and in the study population at time t_j. If $X = \{x_{ij}\}$, $i = 1,\ldots,\nu$, $j = 1,\ldots,k$, were observed for an animal population in which ν animals were ever alive and available for capture on at least one of the k sample occasions, then inference about the parameters $\{S_j\}$ and $\{\beta_j\}$ would be straight-forward. It would also be straight-forward to calculate observed values of the random variables $N_j = \sum_i x_{ij}$.

Instead of observing x_{ij}, we observe y_{ij} with the joint distribution of the sequence of pairs $\{X_{ij}, Y_{ij}\}$ described by a hidden Markov model (Figure 2.1). Importantly, y does not include the history given by the vector of k zeros associated with animals that were available for capture at sometime during the experiment but never caught.

Conditional on the event that animal i is alive (and available for capture) on at least one of the capture occasions t_j, the sequence $(X_{i1},\ldots,X_{ik})$ can then be modeled as a three-state Markov chain. $X_{ij} = 0$ corresponds to the event that animal i has not yet entered the population. $X_{ij} = 1$ corresponds to the event that animal i is in the population and alive, and $X_{ij} = 2$ corresponds to animal i being dead. The possible transitions are $0 \to 0$, $0 \to 1$,

$1 \to 1$, $1 \to 2$, and $2 \to 2$. Define:

β_0 as the probability that an animal is born before the start of the experiment, given that it is alive at some time during the experiment,

β_j as the probability that an animal is born between the times of sample j and $j+1$, given that it was available for capture at sometime during the experiment,

S_j as the probability that an animal alive at the time of sample j is still alive at the start of sample $j+1$.

Conditioning on the ν individuals that are alive and available for capture at sometime during the experiment means that $\beta_0 + \beta_1 + \ldots + \beta_{k-1} = 1$. If $\mathcal{A}$ is used to denote the event that animal i is alive and available for capture at sometime during the experiment, then

$$\Pr\left(X_{i1} = x_{i1} \mid \mathcal{A}\right) = \begin{cases} \beta_0, & x_{i1} = 1, \\ 1 - \beta_0, & x_{i1} = 0, \end{cases} \tag{2.1}$$

and for $j = 2, \ldots, k-1$,

$$\Pr\left(X_{ij} = x_{ij} \mid x_{ij-1}, \mathcal{A}\right) = \begin{cases} 1, & x_{ij} = 2, x_{ij-1} = 2, \\ 0, & x_{ij} = 1, x_{ij-1} = 2, \\ 0, & x_{ij} = 0, x_{ij-1} = 2, \\ 1 - S_j, & x_{ij} = 2, x_{ij-1} = 1, \\ S_j, & x_{ij} = 1, x_{ij-1} = 1, \\ 0, & x_{ij} = 0, x_{ij-1} = 1, \\ 0, & x_{ij} = 2, x_{ij-1} = 1, \\ \beta'_{j-1}, & x_{ij} = 1, x_{ij-1} = 0, \\ 1 - \beta'_{j-1}, & x_{ij} = 0, x_{ij-1} = 0, \end{cases} \tag{2.2}$$

where $\beta'_j = \beta_j / \sum_{h=j}^{k-1} \beta_h$. Note that we exclude from the study population individuals that are born and then die during (t_j, t_{j+1}). Clearly these individuals are invisible to the mark-recapture experiment.

2.2. Observed data likelihood. A common assumption in mark recapture models is that $y_{ij} | x_{ij}$ is the outcome of a Bernoulli trial with probability p_j if $x_{ij} = 1$ or 0 otherwise. That is, only those animals alive and in the population at the time of sample j are at risk of being caught, which happens with probability p_j. The standard approach to fitting this model is to derive an observed data likelihood from the marginal distribution with pdf $[y \mid \nu, S, \beta, p]$ which is described by summing up across all possible sample paths for X that are compatible with the data (here we use the notation $[y|x]$ to denote the pdf of the distribution for the random variable Y, evaluated at y, conditional on X, evaluated at x).

Using the notation $y = \{y_{ij}\}$ $(i = 1, \ldots, u.;\ j = 1, \ldots, k)$ and $u = (u_1, \ldots, u_k)'$ the likelihood can be expressed as

$$[y \mid \nu, S, \beta, p] = [u. \mid \nu][u \mid u., \beta, S][y \mid u, S, p], \tag{2.3}$$

where $[u.|\nu]$ describes a binomial distribution with index ν and parameter π_0 and $[u|u.]$ a multinomial distribution with index $u.$ and parameter vector ξ.

Both π_0 and ξ are complicated functions of the parameters S and β that account for the censoring of the birth times in y. If we let ψ_j denote the probability that an animal is available for capture in sample j but has not yet been caught, then

$$\psi_{j+1} = \psi_j(1-p_j)S_j + \beta_j \tag{2.4}$$

for $j = 2,\ldots,k-1$, with $\psi_1 = \beta_0$. Such an animal is first caught in sample j with probability $\psi_j p_j$ hence

$$\begin{aligned} \pi_0 &= 1 - \sum_{j=1}^{t} \psi_j p_j, \\ \xi_j &= \frac{\psi_j p_j}{1-\pi_0}. \end{aligned} \tag{2.5}$$

The term $[y \mid u,S,p]$ describes the celebrated Cormack-Jolly-Seber model (see [4–6]). If we index the first sample in which an animal i was caught by r_i and the last sample in which animal i was seen by l_i, then we can write

$$[y \mid u,S,p] = \prod_{i=1}^{u.} \prod_{j=r_i}^{l_i} p_j^{y_{ij}} (1-p_j)^{1-y_{ij}} \chi_{l_i}, \tag{2.6}$$

where the term χ_j, which accounts for the censoring of time of death, can be defined recursively as

$$\chi_j = \begin{cases} 1, & j = k, \\ 1 - S_j + S_j(1-p_{j+1})\chi_{j+1}, & j = 1,\ldots,k-1. \end{cases} \tag{2.7}$$

A nice feature of the observed data likelihood (2.3) is that it is straightforward to find maximum likelihood estimators. [7] showed that the parameter ν is not identifiable, however the partial likelihood $[u \mid u.,\beta,S][y \mid u,S,p]$ obtained by conditioning on $u.$ contains all practically useful information on the identifiable parameters in the model. Closed-form solutions to the ML equations exist for all identifiable parameters in this partial-likelihood.

The model can also be easily generalized to allow parameters to be individual-specific, by introducing covariates, or to allow captures to depend on the earlier capture history of the animal. The model (2.3) has also been extended by reparameterizing the model in terms of per capita birth rate f_j [7, 8] and an index to population growth rate [8]

$$\begin{aligned} \lambda_j &= S_j + f_j \\ &= \frac{E[N_{j+1} \mid N_j,\nu]}{E[N_j \mid \nu]}. \end{aligned} \tag{2.8}$$

The complete model (2.3), which we refer to as the Crosbie-Manly-Arnason-Schwarz model, was first described in [9] building on earlier work of [10]. It differs from the well-known Jolly-Seber (JS) model ([5, 6]) in the way that captures of unmarked animals are modeled. In the JS model, the terms

$$[u. \mid \nu][u \mid u.] \tag{2.9}$$

are replaced by

$$[u \mid U] \tag{2.10}$$

and the elements of $U = (U_1, \ldots, U_k)'$, where U_j is the number of unmarked animals in the population at the time of sample j, are treated as unknown parameters to be estimated. While historically important, this approach does not allow the reparameterizations in terms of f_j and λ_j described above as f_j and λ_j cannot be written as deterministic functions of the parameters S_j and β_j. Also, the U_j parameters are of little demographic interest although predictors of N_j and B_j exist.

2.3. Complete data likelihood. An alternative to using the observed data likelihood is to describe the model in terms of the complete data likelihood (CDL) [11]. The CDL is the likelihood of *all* data, both missing and observed. The observed data are

(i) the values y_{ij} for all $u.$ individuals that we observed at least once,

(ii) the censored information about X that we obtain from y.

The missing data are

(i) the unknown number of individuals that were available for capture but not caught (we include this by specifying ν as a parameter),

(ii) the realized but unknown values of X_{ij}, x_{ij}.

Including both the observed and missing data gives the CDL that we can express as

$$[x, y \mid S, \beta, p, \nu] = [y \mid p, x, \nu][x \mid S, \beta, \nu]. \tag{2.11}$$

The CDL has been naturally factored into a term that describes how the data were corrupted, $[y \mid p, x, \nu]$, and a term that models the underlying birth and death processes of interest, $[x \mid S, \beta, \nu]$. As for the observed data likelihood section, we model the corruption by assuming that an individual that was alive in sample j ($x_{ij} = 1$) was observed in sample j ($y_{ij} = 1$) with probability p_j. We model the birth and death process as in the Markov chain model for X described in Section 2.1.

Even though the CDL provides a natural approach to looking at the problem, we must still integrate over the missing data in order to obtain valid inference. Computational techniques such as Markov chain Monte Carlo (MCMC, discussed below) can be used that iteratively integrate out all missing data, allowing models to be specified in terms of the CDL. This means that in each iteration of the MCMC chain we need values for the missing quantities x and ν (and all other parameters) that were obtained from the

posterior distribution of all unknowns. One such MCMC algorithm we can use is Gibbs sampling, described in the following section.

2.4. Markov chain Monte Carlo. The Gibbs sampler, also known as alternating conditional sampling, is a remarkable algorithm for efficiently constructing a Markov chain for complex joint probability distributions $[Z_1,\ldots,Z_k \mid \theta]$ by sampling from the full conditional distributions $[Z_i \mid \{Z_j\}_{j\neq i},\theta]$ of each component. The stationary distribution of this Markov chain has the target density $[Z_1,\ldots,Z_k \mid \theta]$. A particularly useful feature of the Gibbs sampler is that the Markov chain can be constructed even though the target joint probability density may only be known up to the normalizing constant. This has led to a resurgence of interest in Bayesian inference which, historically, has been held back by the need for high-dimension integration needed to normalize posterior probability densities.

Here, we outline the construction of a Gibbs sampler whose target density is the CDL described in the previous section. Once we have collected our mark-recapture data we have data y, that is known; and the unknown are any parameters and any unknown realized values of random variables of interest. For the Crosbie-Manly-Arnason-Schwarz model (2.3), the parameters are $p_1,p_2,\ldots,p_k$; $S_1,\ldots,S_{k-1}$, $\beta_0,\beta_1,\ldots,\beta_{k-1}$, and ν. In addition unknown realized random variables of interest (might) include $N_1,\ldots,N_k$, $B_0,B_1,\ldots,B_{k-1}$.

Starting with (2.3) and specifying independent beta $Be(\alpha_p,\gamma_p)$ prior distributions for the parameters p_j, we can show that the full conditional distribution for p_j is a beta $Be(n_j+\alpha_p,N_j-n_j+\gamma_p)$ for $j=1,\ldots,k$, and where n_j is the total number of individuals caught at time of sample j.

If we specify independent beta $\mathrm{Be}(\alpha_S,\gamma_S)$ prior distributions for the parameters S_j, we obtain beta $\mathrm{Be}(N_j-D_j+\alpha_S,D_j+\gamma_S)$ full conditional distributions for $j=1,\ldots,k-1$.

If we specify independent beta $\mathrm{Be}(\alpha_{\beta'},\gamma_{\beta'})$ prior distributions for the parameters β'_j, we obtain beta $\mathrm{Be}(B_j+\alpha_{\beta'},N-\sum_{h=0}^{j}B_h+\gamma_{\beta'})$ full conditional distributions for $j=1,\ldots,k-2$ (note that $\beta'_{k-1}=1$). If desired, we can transform the generated values of β'_j to any other birth parameter, such as β_j or η_j. For example, β_j, $j=0,\ldots,k-1$, is obtained by taking

$$\beta_j=\begin{cases}\beta'_j, & j=0,\\ \beta'_j\displaystyle\prod_{h=0}^{j-1}(1-\beta'_h), & j=1,\ldots,k-2.\end{cases}\tag{2.12}$$

We also need to calculate the full conditional probability of the missing values of x. This can be done a number of ways, but we choose to represent the information in x by matrices that give the interval censored times of birth and death, denoted by b and d, respectively. The value $b_{ij}=1$ means individual i was born between sample j and $j+1$ with $b_{ij}=0$ otherwise. The value $b_{i0}=1$ means the individual was born before the study began. The value $d_{ij}=1$ means that individual i died between sample j and $j+1$ with $d_{ij}=0$ otherwise. The value $d_{ik}=1$ means the individual was still alive at the end of the study. The assumptions about the underlying birth and death processes impose

restrictions on the values of b and d,

$$\sum_{j=0}^{k-1} b_{ij} = 1, \quad \sum_{j=1}^{k} d_{ij} = 1, \quad d_{ij} - b_{ij} \geq 0 \ \forall i, j. \tag{2.13}$$

The matrices b and d are censored. For the $u.$ individuals that were ever observed we know that they were not born after first capture, $b_{ij} = 0$, $j = r_i, \ldots, k$, where r_i indexes the sample in which animal i first appeared. Likewise we know that they did not die before the last capture, $d_{ij} = 0$, $j = 1, \ldots, l_i$, where l_i indexes the sample in which animal i last appeared. For the $\nu - u.$ individuals that were never observed we have no information about either b or d.

The full conditional distribution of the unknown values of b for individual i is a multinomial distribution with probability vector γ^b, where

$$\begin{aligned} \gamma_j^b &= \frac{\zeta_j^b}{\sum_{h=0}^{\lambda_i^b - 1} \zeta_h^b}, \\ \zeta_j^b &= \beta_j \prod_{h=j}^{\lambda_i^b - 1} S_h (1 - p_h), \end{aligned} \tag{2.14}$$

and λ_i^b is the time of first capture r_i for all individuals observed, and it is the last period in which the individual was alive (obtained from d) for individuals $i = u. + 1, \ldots, \nu$ that were not observed.

The full conditional distribution of the unknown values of d for individual i is a multinomial distribution with probability vector γ^d, where

$$\begin{aligned} \gamma_j^d &= \frac{\zeta_j^d}{\sum_{h=\lambda_i^d}^{t} \zeta_h^d}, \\ \zeta_j^d &= \begin{cases} \displaystyle\prod_{h=\lambda_i^d}^{j-1} S_h \prod_{h=\lambda_i^d+1}^{j} (1 - p_h)(1 - S_j), & j < k, \\ \displaystyle\prod_{h=\lambda_i^d}^{j-1} S_h \prod_{h=\lambda_i^d+1}^{j} (1 - p_h), & j = k, \end{cases} \end{aligned} \tag{2.15}$$

and λ_i^d is the time of last capture l_i for all individuals observed, and it is the first period in which the individual was alive (obtained from b) for individuals $i = u. + 1, \ldots, \nu$ that were not observed.

For the parameter ν we specify a discrete uniform prior distribution $DU(0, \kappa_\nu)$. Obtaining a sample from the full conditional distribution for ν has two problems:

(1) the full conditional distribution is only known to a proportionality constant,

(2) the value of the parameter changes the dimension of other unknowns in the model.

Table 2.1. Summary statistics for fitting the Crosbie-Manly-Arnason-Schwarz model to the meadow vole data. The statistic u_j is the number of individuals caught in sample j, R_j is the number of individuals released following sample j, q_j is the number of the R_j that were ever recaptured, and m_j is the number of marked animals caught in sample j.

Month	$u_j = R_j$	q_j	m_j
1	96	81	—
2	42	74	76
3	27	65	71
4	30	61	70
5	38	82	59
6	61	—	87

Table 2.2. Maximum likelihood estimates of identifiable parameters in the Crosbie-Manly-Arnason-Schwarz model fitted to the meadow vole data.

			95% C. I.	
Parameter	Estimate	SE	lower	upper
S_1	0.875	0.041	0.770	0.936
S_2	0.659	0.049	0.559	0.747
S_3	0.681	0.050	0.576	0.770
S_4	0.619	0.050	0.518	0.710
p_2	0.905	0.040	0.792	0.960
p_3	0.855	0.047	0.737	0.925
p_4	0.934	0.036	0.816	0.979
p_5	0.909	0.039	0.800	0.961
f_2	0.220	0.056	0.130	0.347
f_3	0.253	0.062	0.152	0.391
f_4	0.378	0.078	0.242	0.538

To overcome the first problem we can use a sampling scheme, such as the Metropolis-Hastings algorithm, that allows us to sample from a distribution that we only know up to the proportionality constant. The second problem requires an extension of the Metropolis-Hastings algorithm called reversible jump Markov chain Monte Carlo [3] where there is dimension matching to ensure reversibility of the Markov chain. Details of the reversible-jump sampler are given in [11].

2.5. Example. We illustrate the use of the MCMC algorithm described in the previous section by fitting the model (2.3) to meadow vole (*Microtus pensylvannicus*) data collected at the Patuxent Wildlife Research Center, Laurel, Md, USA [12]. The meadow vole population was trapped at one-month intervals; untagged animals were tagged and released, tagged animals had their identity recorded and were then released. The model (2.3) can be fitted using sufficient statistics (Table 2.1) and closed-form solutions to the ML equations [7]. The ML estimates (Table 2.2) and the posterior summaries from fitting the model using the Gibbs sample algorithm (Table 2.3) of the previous section are

Table 2.3. Posterior distribution summary statistics for the Gibbs sampler posterior simulations and for the identifiable parameters in the Crosbie-Manly-Arnason-Schwarz model fitted to the meadow vole data. Included are predictions of the abundances N_j at the time of each sample and B_j, the number of individuals born in (t_j, t_{j+1}).

				Quantile	
Parameters	Mean	Median	0.025	0.5	0.975
S_1	0.874	0.042	0.788	0.876	0.953
S_2	0.659	0.049	0.563	0.659	0.755
S_3	0.681	0.051	0.580	0.682	0.779
S_4	0.616	0.050	0.517	0.616	0.711
f_2	0.222	0.058	0.123	0.218	0.348
f_3	0.257	0.064	0.147	0.252	0.397
f_4	0.377	0.079	0.241	0.370	0.549
p_2	0.886	0.043	0.791	0.891	0.956
p_3	0.839	0.048	0.733	0.843	0.921
p_4	0.913	0.041	0.819	0.919	0.975
p_5	0.891	0.041	0.799	0.896	0.958
Realized random variables					
N_2	132.4	5.1	125	132	145
N_3	116.3	5.1	108	116	128
N_4	108.8	3.9	103	108	118
N_5	108.1	3.7	103	107	117
B_2	28.7	4.1	21	29	37
B_3	29.3	3.4	22	29	36
B_4	41.0	3.6	35	41	49

in close agreement. For the Gibbs sampler we used beta Be(1, 1) priors for S_i, p_i, and β_i' with the f_i parameters found deterministically as a function of the β_i and S_i parameters [7]. We used a discrete uniform DU(200000) prior for ν.

A nice feature of the MCMC approach is that it is relatively simple to obtain a posterior distribution for unobserved random variables of interest such as N_i and B_i, also reported in Table 2.3. Predictions for these in a likelihood analysis must be based on method of moment-type estimators as the N_i and B_i parameters do not explicitly appear in the likelihood function. Obtaining predictions using MCMC is straightforward with values sampled from the posterior predictive distribution. In practice, we simply use the set of b and d values obtained in the Markov chain and for each iteration compute the current value of the variable of interest. For example, [11] show that

$$N_j = \sum_{i=1}^{\nu} \left(\sum_{h=0}^{j-1} b_{ih} - \sum_{h=1}^{j-1} d_{ih} \right). \tag{2.16}$$

Progress of the Markov chains is instructive (Figure 2.2). Convergence of the Markov chains for the identifiable parameters $S_1, \ldots, S_{k-2}$; $p_2, \ldots, p_{k-1}$; $f_2, \ldots, f_{k-2}$ is rapid and the

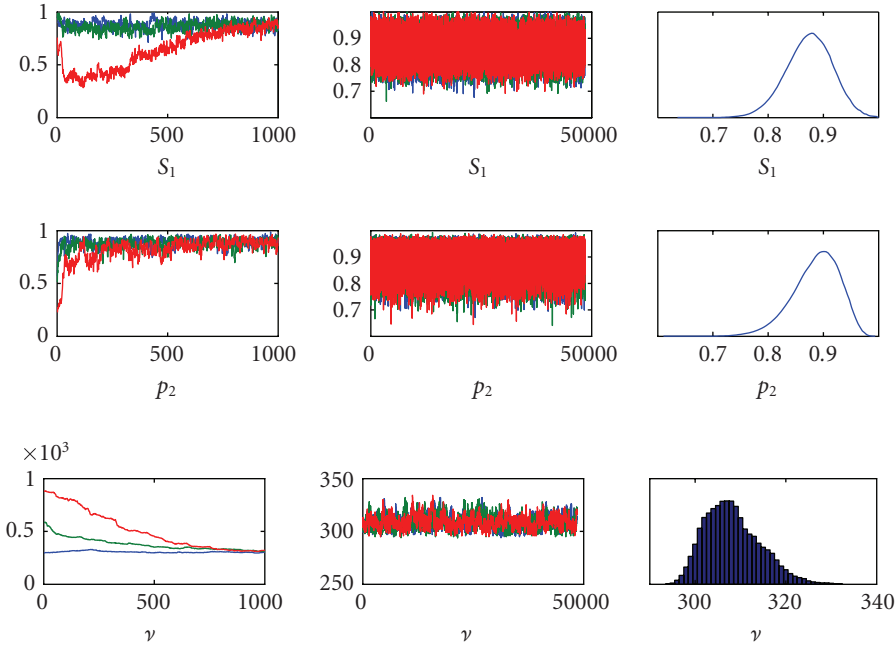


FIGURE 2.2. Gibbs sampler Markov chains for the first 1000 iterations (column of plots on left-hand side), for 50 000 iterations following a discarded burn-in of 10 000 iterations (middle column of plots) and posterior density plots (right-hand column of plots) for the identifiable parameters S_1 and p_1 and the nonidentifiable parameter ν, for the Gibbs sampler posterior simulations from fitting the Crosbie-Manly-Arnason-Schwarz model fitted to the meadow vole data.

chains appear well-mixed after a few thousand iterations. Even chains for nonidentifiable parameters such as ν appear well mixed after a few thousand iterations.

3. Discussion

Markov chain theory plays an important role in statistical inference both in the formulation of models for data and in the construction of efficient algorithms for inference. The ability to compactly represent stochastic processes that evolve over time has meant that Markov chain theory has played an important role in describing animal and plant population dynamics. The fact that each generation gives rise to the next makes the Markov property a natural starting point in developing population models.

The role of the Markov chain in statistical inference is a much more recent development. Although the Metropolis-Hastings algorithm first appeared in 1953 [13], with roots back to the Manhattan project, and the Gibbs sampler in 1984 [2] MCMC entered into widespread use in statistics only in the last 15 years. This followed [14] who illustrated the application of Gibbs sampling to a variety of statistical problems.

MCMC is particularly useful for fitting hierarchical models using Bayesian inference methods. A simple example of a hierarchical model is one where we have two components

to the model $[Y \mid \theta][\theta \mid \gamma]$. Here, conditional on parameters θ, the random variable Y has density $[Y \mid \theta]$. The parameters θ are themselves modeled as random variables and have distribution with pdf $[\theta \mid \gamma]$. In order to fit this model by a method such as maximum likelihood (ML), we must find the marginal likelihood: $\mathcal{L}(\theta \mid y] = \int_\theta [y \mid \theta][\theta \mid \gamma] d\theta$. In practice this integral is very difficult, if not practically impossible, in most cases of interest. The equivalent Bayesian problem is to find the posterior marginal density $[\gamma \mid y] \propto [Y \mid \theta][\theta \mid \gamma][\gamma]$ which involves essentially the same integral. However, MCMC allows us to approximate a sample from the required density but we must specify a prior distribution from γ. Thus, Bayesian inference methods based on MCMC allows inference for the model $[Y \mid \theta][\theta \mid \gamma]$ but the same problem may be difficult or impossible using standard ML theory.

In the context of mark-recapture modeling, the advent of MCMC has meant that it is now possible to fit complex hierarchical models. Starting with a model such as (2.3), it is relatively easy to add hierarchical structure. In the context of the hidden Markov model described in Figure 2.1, we mean that it is relatively straightforward to add in components that allow us to further model $[X \mid \beta, S, \nu]$. These issues have been explored by [7, 11]. [7] showed how the model (2.3) could be used to explore relationships among parameters and in particular considered a hierarchical model in which it was assumed that survival probabilities were related to per capita birth rates (functions of the β parameters). Such a model would make sense when there were common environmental influences on both survival probabilities and birth rates. Such an analysis is virtually impossible using ML methods. Similarly, [11] showed how the CDL discussed above could be used in hierarchical modeling, and in particular fit models in which there is feedback between the unknown random variables N_j and survival S_j or per capita birth rates f_j. The practical benefit of hierarchical modeling is that a much richer class of models is available for the ecologist to explore, and the advent of MCMC means that methods of inference are fully able to incorporate all different sources of uncertainty in the analysis.

A particularly important technical issue with MCMC is that of convergence. A typical problem in Markov chain theory is the determination of a stationary distribution given the transition kernel. In MCMC, the problem is reversed, and is one of constructing a transition kernel that has, as its stationary distribution, the target distribution of interest. Under mild conditions, the Metropolis-Hastings algorithm and the Gibbs sampler as a special case are methods of constructing transition kernels that have the required property. An important practical problem is the development of rules for helping determine when the chain has converged. Experience indicates that for some problems, convergence can be rapid and in others it can be slow. Rapid convergence appears to be an associated with likelihood functions that are well behaved. Methods for assessing convergence are ad hoc, and generally based on informal graphical assessment (Figure 2.2) or computation of simple statistics, such as the Brooks-Gelman-Rubin diagnostic statistic. There is surprisingly little underlying theory about the rate at which Markov chains constructed in MCMC should converge to the stationary distribution that can be applied to the practical problems of MCMC. Clearly, time to stationarity will be a property of the model. Again, observation indicates that problems for which the likelihood function are well behaved tend to converge rapidly. The work in [15, 16] is important in this regard.

References

[1] G. A. F. Seber, *The Estimation of Animal Abundance and Related Parameters*, Macmillan, New York, NY, USA, 2nd edition, 1982.

[2] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[3] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[4] R. M. Cormack, "Estimates of survival from the sighting of marked animals," *Biometrika*, vol. 51, pp. 429–438, 1964.

[5] G. M. Jolly, "Explicit estimates from capture-recapture data with both death and immigration-stochastic model," *Biometrika*, vol. 52, pp. 225–247, 1965.

[6] G. A. F. Seber, "A note on the multiple-recapture census," *Biometrika*, vol. 52, pp. 249–259, 1965.

[7] W. A. Link and R. J. Barker, "Modeling association among demographic parameters in analysis of open population capture-recapture data," *Biometrics*, vol. 61, no. 1, pp. 46–54, 2005.

[8] R. Pradel, "Utilization of capture-mark-recapture for the study of recruitment and population growth rate," *Biometrics*, vol. 52, no. 2, pp. 703–709, 1996.

[9] C. J. Schwarz and A. N. Arnason, "A general methodology for the analysis of capture-recapture experiments in open populations," *Biometrics*, vol. 52, no. 3, pp. 860–873, 1996.

[10] S. F. Crosbie and B. F. J. Manly, "Parsimonious modelling of capture-mark-recapture studies," *Biometrics*, vol. 41, no. 2, pp. 385–398, 1985.

[11] M. R. Schofield and R. J. Barker, "A unified capture-recapture framework," to appear in *Journal of Agricultural, Biological and Ecological Statistics*.

[12] B. K. Williams, J. D. Nichols, and M. J. Conroy, *Analysis and Management of Animal Populations*, Academic Press, London, UK, 2002.

[13] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[14] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990.

[15] J. J. Hunter, "Mixing times with applications to perturbed Markov chains," *Linear Algebra and Its Applications*, vol. 417, no. 1, pp. 108–123, 2006.

[16] J. J. Hunter, "Coupling and mixing times in a Markov chains," *Research Letters in Information and Mathematical Science*, vol. 11, pp. 1–22, 2007.

Richard J. Barker: Department of Mathematics and Statistics, University of Otago,
P.O. Box 56, Dunedin 9010, New Zealand
Email address: rbarker@maths.otago.ac.nz

Matthew R. Schofield: Department of Mathematics and Statistics, University of Otago,
P.O. Box 56, Dunedin 9010, New Zealand
Email address: mschofield@maths.otago.ac.nz

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 43565, 18 pages
doi:10.1155/2007/43565

Research Article

Reliability of Modules with Load-Sharing Components

Mark Bebbington, Chin-Diew Lai, and Ričardas Zitikis

Received 17 April 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

To increase the reliability of modules, and thus of systems assembled from them, they are frequently constructed using parallel load-sharing components. Examples include jet engines, electrical power networks, and telecommunications networks. We consider the situation when the components operate independently, but when any one of them fails, the load of the failed component is instantaneously distributed among the working components. The entire module fails when the last working component fails. We analyze the survival probability and residual life expectancy of such modules. An obvious application is to the case of the 1998 Auckland power supply failure in New Zealand.

1. Introduction and motivation

Reliability systems often consist of several subsystems, which may be called modules. In practical reliability analysis, one often considers first the reliability of each module, and derives the reliability of the system as a whole. A classical example of such a system is a combination of data transmission routers. Since, in many areas, the continuity of data flow is of utmost importance, the system's reliability is increased by incorporating redundancy in the form of parallel components or subsystems. For instance, data transfers between two points may be accomplished by multiple (identical or not) parallel routers, with electricity supplied to each of the routers by several (identical or not) power units.

In general, we are interested in a module consisting $K \geq 2$ parallel components. We denote the lifetimes of the components by T_k, $1 \leq k \leq K$, with survival functions $S_k(t) = \mathbf{P}\{T_k > t\}$, and hazard rate (HR) functions $h_k(t) = -S_k'(t)/S_k(t)$, respectively. When one

of the components fails, its load is distributed among the working components. The entire module fails when the last working component fails; denote the module's failure time by $M(K)$. The corresponding survival and mean residual life (MRL) functions are, respectively,

$$S_{M(K)}(t) = \mathbf{P}\{M(K) > t\},$$
$$\mu_{M(K)}(t) = \frac{IS_{M(K)}(t)}{S_{M(K)}(t)}, \tag{1.1}$$

where $IS_{M(K)}(t) = \int_t^\infty S_{M(K)}(x)dx$. We next give a couple of illustrative examples, where the need to estimate the above two functions is a natural one.

> "Consider jet engines functioning under full load on a commercial airplane. One functioning jet engine is enough for a small airplane, while 2 engines are necessary for a big airplane. But for higher reliability, 2 engines are functioning for the small airplane and 4 for the big airplane. An engine controller manages the load sharing. When 2 engines function in a small airplane, the load on each is much less than when they function alone. From the test data, the failure rate of the engines is reduced to 45% under half load. Similarly, if 4 engines are functioning for a big airplane, the failure rate for each engine is reduced to 45%, while if three engines are functioning, the failure rate is reduced to 75% … for how long can the small and big airplanes fly before the reliability drops below 0.9?" [1].

We see from this excerpt that it is natural to aim at estimating the airplane's survival function $S_{M(K)}(t)$. We may also want to know for how long, on average, the airplane can still stay in the air, for which we need to estimate the MRL function $\mu_{M(K)}(t)$. Of course, the above questions are more mathematical idealizations than reflections of reality, but they serve as conceptual examples of some of the types of problems in the area. In practice, even large jets can land relatively safely without a single functioning engine [2, 3].

> "The 1998 Auckland power crisis was an event that occurred in the Auckland, New Zealand, Central Business District. The area suffered a five-week-long power outage in 1998. At the beginning of 1998, almost all of downtown Auckland received electricity from the supplier Mercury Energy via only four power cables, two of them were 40-year-old oil-filled cables past their replacement date. One of the cables failed on 20 January, possibly due to the unusually hot and dry conditions, another on 9 February, and due to the increased load from the failure of the first cables, the remaining two failed on 19 and 20 February, leaving the central business district (except parts of a few streets) without power" [4].

For a detailed account and analysis of the power crisis, see [5]. In this case, estimation of the mean residual life is of utmost importance in deciding what emergency repair or replacement activities may be (more) effective.

To get an initial feel about the module's survival, HR, and MRL functions, we note that if the failure of any one of the K components does not influence the HR functionsof the

functioning components, then the module's survival function $S_{M(K)}(t)$ can be written in terms of the individual survival functions as $1 - \prod_{k=1}^{K}(1 - S_k(t))$. The individual survival functions $S_k(t)$ can in turn be expressed using the corresponding HR functions $h_k(t)$ as $S_k(t) = \exp\{-\int_0^t h_k(y)dy\}$. In the context of the present paper, due to the load-sharing scenario, the dynamics of the entire module and thus of its survival and MRL functions are quite different from those in the case of non-interacting parallel components.

There are a few closely related references on this topic. The reliability of load sharing systems may be studied through positively dependent multivariate life distributions [6]; for positively dependent bivariate life distributions, we refer to [7, Section 9.2]. Another approach of studying dependency among parallel components is by using interaction schemes. For example, Murthy and Nguyen [8], and Murthy and Wilson [9] propose and analyze an interaction scheme where, in a two-component system, the failure of one component provokes the failure of another component with probability p, and thus does not provoke with $1 - p$. Another failure interaction scheme in various generalities—we follow a similar line of thought in the present paper—is where the failure of a component modifies the HR function of the other components by not provoking its failure instantaneously but modifying its conditional time to failure [10–13]. These papers assume piecewise constant failure rates, or various degrees of interchangeability and symmetry in their components and/or redistribution schemes, whereas our results are presented in complete generality, and include estimators for the MRL. Perhaps more importantly, our work starts with the notion that there might be too few observations of failing entire modules in order to derive desired statistical inferential results, but failure times of individual module's components might be more readily available (e.g., from laboratory-type testing). Hence assuming the availability of such data, we then aim at deriving formulae for the survival function—and thus, in turn, failure, MRL, and other functions—of the entire module. In contrast, the aforementioned papers are concerned with estimating the component failure rate function given the observed failure times of entire systems. Note also that this problem can be considered [1, 12] in the context of a more general system, the k-out-of-K:G, which, by definition, functions as long as there are at least k $(1 \le k \le K)$ components working. These papers consider specific distributions and load sharing rules, with less generality than our results.

The remainder of the paper is organized as follows. In Section 2, we present the model, notational conventions, and other mathematical formalities. Section 5 contains expressions for the survival and MRL functions, $S_{M(K)}(t)$ and $\mu_{M(K)}(t)$, in terms of individual components that work under the original or increased loads. The general results, Theorems 5.1 and 5.2, are preceded in Section 3 by a detailed analysis of the case $K = 2$, which is of interest in its own right, as well as for a more easily comprehended example of the general theorems. Explicit examples of the $K = 2$ case are given in Section 4, where the performance of parametric and nonparametric estimators of the survival and MRL functions are examined.

Two of us (M.B. and C.D.L.) were fortunate enough to be colleagues of Jeff Hunter when he occupied the Chair in Statistics at Massey University. Jeff's inaugural address was on the subject of reliability and warranty analysis, and we hope he enjoys this sequel. The many visits of the third author (R.Z.) to Massey University in PalmerstonNorth did

not pass by without Jeff flying in from Auckland either to give an inspiring seminar on Generalized Inverses and Stochastic Processes, or to enliven morning and afternoon teas.

2. Mathematical formalism

We assume that the failure times $T_1,\dots,T_K$ are independent, though not necessarily identically distributed, random variables. We work with continuous life-time distributions, and hence assume that there are no multiple failures at any time as multiple failures can occur only with zero probabilities. The first failure occurs at the time $T_{1:K} = \min_{1\le k\le K} T_k$, which is the first order statistic of $T_1,\dots,T_K$. Let D be the first antirank of $T_1,\dots,T_K$, which is (uniquely) defined by $T_D = T_{1:K}$. Hence the pair $(D, T_{1:K})$ tells us which of the components $\{1,\dots,K\}$ fails first and at what time.

At the time $T_{1:K}$, the load of the failed component D is instantaneously distributed among the remaining $K-1$ components, whose set we denote by $\Delta^{(1)} = \{1,\dots,K\} \setminus \{D\}$. Specifically, for every $k \in \Delta^{(1)}$, the failure of the Dth component increases the HR function $h_k(t)$ of the kth (working) component by a function $a^{(1)}_{D,k}(t)$, where the superscript (1) indicates that the redistribution has occurred (immediately) after the 1st failure. Hence for every $k \in \Delta^{(1)}$, we have the conditional-on-$\{T_1,\dots,T_K\}$ HR function $h^{(1)}_k(t) = (h_k(t) + a^{(1)}_{D,k}(t))\mathbf{1}_{\{T_{1:K}\le t\}}$, where the indicator $\mathbf{1}_{\{T_{1:K}\le t\}}$ is equal to 1 when the statement $T_{1:K} \le t$ is true and is 0 otherwise. Let $T^{(1)}_k$, $k \in \Delta^{(1)}$ be conditionally-on-$\{T_1,\dots,T_K\}$ independent random variables whose conditional-on-$\{T_1,\dots,T_K\}$ distributions have the HR functions $h^{(1)}_k(t)$.

Before proceeding further, let us discuss intuitively what we have introduced so far. First, note that $h^{(1)}_k(t) = 0$ for all $t < T_{1:K}$, which implies that the random variables $T^{(1)}_k$, $k \in \Delta^{(1)}$ do not take on any value in the interval $[0, T_{1:K}]$. Hence in addition to the 'original' situation with K random variables $T_1,\dots,T_K$, we have constructed an "artifact" with $K-1$ random variables $T^{(1)}_k$, $k \in \Delta^{(1)}$, which are "activated" at the moment $t = T_{1:K}$ and governed by the HR functions $h_k(t) + a^{(1)}_{D,k}(t)$. When one of the $\Delta^{(1)}$ components fails, we create new $K-2$ "artificial" components. Proceeding in a similar fashion, we specify the mechanism that governs the life of the entire module and allows us, via a conditioning technique, to determine its survival, HR, and MRL functions. We next describe this procedure rigorously and also introduce additional notation to be used throughout the rest of the paper.

To begin, we find it convenient to use the notation $T^{(0)}_1,\dots,T^{(0)}_K$ instead of $T_1,\dots,T_K$, respectively, and $D^{(0)}$ instead of D. Next, starting with the "initial" random variables $T^{(0)}_k$, we recursively, for all $i = 1,\dots,K-2$, define the following quantities.

(i) The random variables $D^{(i)}$ and $T^{(i)}_{1:(K-i)}$, which respectively specify the $(i+1)$st failed component and its failure time, which are related via (or defined by) the equations $T^{(i)}_{D^{(i)}} = T^{(i)}_{1:(K-i)} \equiv \min_{k\in\Delta^{(i)}} T^{(i)}_k$, where $\Delta^{(0)} = \{1,\dots,K\}$ and, for any $i \ge 1$, the set $\Delta^{(i)} = \Delta^{(i-1)} \setminus \{D^{(i-1)}\}$ consists of all working components immediately before the $(i+1)$st failure.

(ii) Conditionally-on-$\{D^{(0)},\dots,D^{(i)}, T^{(i)}_{1:(K-i)}\}$ independent random variables $T^{(i+1)}_k$, $k \in \Delta^{(i)}$, whose conditional-on-$\{D^{(0)},\dots,D^{(i)}, T^{(i)}_{1:(K-i)}\}$ distributions have the HR

functions

$$h_k^{(i+1)}(t) = \left(h_k(t) + \sum_{m=1}^{i+1} a^{(m)}_{D^{(m-1)},k}(t)\right)\mathbf{1}_{\{T^{(i)}_{1:(K-i)} \le t\}}. \tag{2.1}$$

Hence, $T_k^{(i+1)}$ is the lifetime of the kth component after $i+1$ failed components, which are $D^{(0)},\ldots,D^{(i)}$. The random variable $T_k^{(i+1)}$ starts its life at the time $t = T^{(i)}_{1:(K-i)}$.

Note that, since there are K components in the module, the largest value of i is $K-1$ as there are no functioning components after the Kth failure. When $i = K-2$, then there is only one "surviving" random variable $T_k^{(K-1)}$, whose index k is the (only) member of the singleton set $\{1,\ldots,K\} \setminus \{D^{(0)},\ldots,D^{(K-2)}\}$; denote the member by $\kappa(K-1)$. Hence we have $M(K) = T^{(K-1)}_{\kappa(K-1)}$, and so the module's survival function $S_{M(K)}(t)$ can be written as $S_{M(K)}(t) = \mathbf{P}\{T^{(K-1)}_{\kappa(K-1)} > t\}$. With the help of the latter equation, the corresponding formula for the MRL function $\mu_{M(K)}(t)$ can be expressed in terms of the survival function of the random variable $T^{(K-1)}_{\kappa(K-1)}$ using (1.1). Of course, one can also derive an analogous expression for the HR function via the equation $h_{M(K)}(t) = -S'_{M(K)}(t)/S_{M(K)}(t)$. Section 3 provides a detailed analysis of the survival and MRL functions when $K = 2$.

3. Survival and MRL functions for two components

In this section, we give a detailed analysis of the survival function $S_{M(2)}(t)$ of a module with *two* (possibly different) components whose independent lifetime variables are T_1 and T_2 with (possibly different) survival functions $S_1(t)$ and $S_2(t)$, respectively. At the time $T_{1:2} = \min(T_1, T_2)$, one of the two components fails; let it be i. As a result of the failure, the HR function of the working component $k = \text{NOT}(i)$ increases by a function $a^{(1)}_{i,k}(t)$, for all $t \ge T_{1:2}$. (Note that $\text{NOT}(i) = 3 - i$ as we consider the $K = 2$ case.) Let $S_k^{+i}(t)$ be the survival function of the component k when it is working under its own load plus the load of the failed component i, which in our current two-component situation means that the component k takes on the whole module's load.

There is a possibility that we might have a sufficiently large number of failure times of such modules, in which case we estimate $S_{M(2)}(t)$ using the empirical survival function, or fit a parametric distribution to the failure times. Failing a sufficiently large number of modules may not, however, be feasible, due to time and/or cost considerations. However, assessing the reliability of individual components under normal and/or increased loads can be quite a feasible task, say, in a laboratory environment. Quantitative accelerated life testing techniques can be used to speed up the process (cf., e.g., Nelson [14]). For the reasons noted above, in the next theorem, we express $S_{M(2)}(t)$ in terms of the "individual" survival functions $S_i(t)$ and $S^{+i}_{\text{NOT}(i)}(t)$, for $i = 1$ and 2.

Theorem 3.1. *We have that*

$$S_{M(2)}(t) = -\sum_{i=1}^{2} S^{+i}_{\text{NOT}(i)}(t) \int \mathbf{1}_{\{y \le t\}} \frac{S_{\text{NOT}(i)}(y)}{S^{+i}_{\text{NOT}(i)}(y)} dS_i(y) + S_1(t)S_2(t). \tag{3.1}$$

We can estimate the survival functions $S_1(t)$ and $S_2(t)$ on the right-hand side of (3.1) by exposing (e.g., in a laboratory environment) the two components to their "normal"

loads, and we can also estimate the survival functions $S_1^{+2}(t)$ and $S_2^{+1}(t)$ by exposing the corresponding components to the load of the entire module. In the nonparametric approach, we estimate the survival functions $S_i(t)$, $i = 1,2$ as $\widehat{S}_i(t) = (1/n_i)\sum_{\ell=1}^{n_i} \mathbf{1}_{\{T_i(\ell)>t\}}$, where $T_i(1),\ldots,T_i(n_i)$ are independent copies of the random variable $T_i \sim S_i$. (For a given random variable X, it is customary to use the notation $X_1,\ldots,X_n$ for copies of X. Since we already use subscripts for other good reasons, throughout the paper, we use $X(1),\ldots,X(n)$ to denote copies of X.) Next, we use independent copies $T_j^{+i}(1),\ldots,T_j^{+i}(m_j)$ of the random variable $T_j^{+i} \sim S_j^{+i}$ to construct an estimator for $S_j^{+i}(t)$, which is $\widehat{S}_j^{+i}(t) = (1/m_j)\sum_{\ell=1}^{m_j} \mathbf{1}_{\{T_j^{+i}(\ell)>t\}}$. Thus, we have the nonparametric estimator of the module's survival function

$$\widehat{S}_{M(2)}(t) = \sum_{i=1}^{2} \widehat{S}^{+i}_{\mathrm{NOT}(i)}(t)\frac{1}{n_i}\sum_{\ell=1}^{n_i} \mathbf{1}_{\{T_i(\ell)\le t\}}\frac{\widehat{S}_{\mathrm{NOT}(i)}(T_i(\ell))}{\widehat{S}^{+i}_{\mathrm{NOT}(i)}(T_i(\ell))} + \widehat{S}_1(t)\widehat{S}_2(t). \tag{3.2}$$

To derive an analogous expression for the MRL function $\mu_{M(2)}(t)$ in terms of the four "individual" survival functions, we need to derive an analogous expression for the integral $IS_{M(2)}(t)$, which can be done by either integrating the right-hand side of (3.1) or by using general Theorem 5.2 with $K = 2$. This gives us the following corollary.

Corollary 3.2. *We have that*

$$\begin{aligned} IS_{M(2)}(t) = &\sum_{i=1}^{2}\iint (x-\max(y,t))_+ dS^{+i}_{\mathrm{NOT}(i)}(x)\frac{S_{\mathrm{NOT}(i)}(y)}{S^{+i}_{\mathrm{NOT}(i)}(y)}dS_i(y) \\ &-\sum_{i=1}^{2}\int (y-t)_+ S_{\mathrm{NOT}(i)}(y)dS_i(y), \end{aligned} \tag{3.3}$$

where $c_+ = c$ if $c > 0$ and $c_+ = 0$ otherwise.

Equations (3.1) and (3.3) can be used for constructing parametric estimators for the MRL function $\mu_{M(2)}(t)$. If, however, we want to use a nonparametric estimator, then we can construct it with the help of the non-parametric estimator for the integral $IS_{M(2)}(t)$,

$$\begin{aligned} \widehat{IS}_{M(2)}(t) = &\sum_{i=1}^{2}\frac{1}{n_i m_{\mathrm{NOT}(i)}}\sum_{\ell=1}^{n_i}\sum_{\nu=1}^{m_{\mathrm{NOT}(i)}} \left(T^{+i}_{\mathrm{NOT}(i)}(\nu) - \max\left(T_i(\ell),t\right)\right)_+ \frac{\widehat{S}_{\mathrm{NOT}(i)}(T_i(\ell))}{\widehat{S}^{+i}_{\mathrm{NOT}(i)}(T_i(\ell))} \\ &+\sum_{i=1}^{2}\frac{1}{n_i}\sum_{\ell=1}^{n_i}(T_i(\ell)-t)_+\widehat{S}_{\mathrm{NOT}(i)}(T_i(\ell)). \end{aligned} \tag{3.4}$$

We now define a nonparametric estimator for the MRL function $\mu_{M(2)}(t)$ as

$$\widehat{\mu}_{M(2)}(t) = \frac{\widehat{IS}_{M(2)}(t)}{\widehat{S}_{M(2)}(t)}. \tag{3.5}$$

The above expressions for the module's survival and MRL functions are based on the survival functions of individual components under their original and increased loads. If desired, however (and we will find it convenient in Section 4), the expressions can easily

be rewritten in terms of the corresponding HR functions. This can be done using the equations $S_k(t) = \exp\{-\int_0^t h_k(x)dx\}$, $S_k^{+i}(t) = \exp\{-\int_0^t h_k(x) + a_{i,k}^{(1)}(x)dx\}$, and so forth, or simply using (A.6) derived in the appendix. (Indeed, the proof of general Theorem 5.1 is based on HR functions.) Clearly now, we have $S_k(t)/S_k^{+i}(t) = \exp\{\int_0^t a_{i,k}^{(1)}(x)dx\}$, which is convenient when dealing with the right-hand sides of (3.1) and (3.3). (Of course, we have $i \neq k$.)

4. Examples

As an example, consider the simple but important case when the module's two components have *exponential* lifetimes. (For a recent discussion of tests for exponentiality, we refer to Mimoto and Zitikis [15] and references therein.) That is, we assume the survival function $S_k(t) = \exp(-\lambda_k t)$ and, consequently, the HR function $h_k(t) = \lambda_k$. (We will later find it also convenient to use the notation $S(t;\lambda_k)$ instead of $S_k(t)$, and the notation $f(t;\lambda_k)$ for the corresponding density function.) Since the exponential HR function is constant, it leaps to mind to choose the redistribution function also as a constant; hence we assume that $a_{i,k}^{(1)}(t) \equiv \alpha_{i,k}$. Under this assumption and using (3.1), we obtain the survival function

$$S_{M(2)}(t) = \left(1 + t\sum_{i=1}^{2} \lambda_i \Delta\left(t;\lambda_i - \alpha_{i,k}\right) e^{(\lambda_i - \alpha_{i,k})t}\right) e^{-(\lambda_1+\lambda_2)t}, \tag{4.1}$$

where

$$\Delta(t;c) = \begin{cases} \dfrac{1}{ct}\left(1 - e^{-ct}\right) & \text{if } c \neq 0, \\ 1 & \text{if } c = 0. \end{cases} \tag{4.2}$$

Irrespectively of the sign of c, the quantity $\Delta(t;c)$ is nonnegative, and so we have the bound $S_{M(2)}(t) \geq e^{-(\lambda_1+\lambda_2)t}$, which can be rewritten as $S_{M(2)}(t) \geq \mathbf{P}\{\min(T_1, T_2) > t\}$; hence the obvious fact is that the module functions at least until the time of the first failure.

We next derive the HR function, which is

$$h_{M(2)}(t) = \frac{(\lambda_1+\lambda_2) - \sum_{i=1}^{2} \lambda_i e^{(\lambda_i - \alpha_{i,k})t} + t(\lambda_1+\lambda_2)\sum_{i=1}^{2} \lambda_i \Delta\left(t;\lambda_i - \alpha_{i,k}\right) e^{(\lambda_i - \alpha_{i,k})t}}{1 + t\sum_{i=1}^{2} \lambda_i \Delta\left(t;\lambda_i - \alpha_{i,k}\right) e^{(\lambda_i - \alpha_{i,k})t}}. \tag{4.3}$$

Integrating (4.1), we obtain an expression for $IS_{M(2)}(t)$ and, in turn, for the MRL function:

$$\mu_{M(2)}(t) = \frac{1 + \sum_{i=1}^{2} \lambda_i / (\lambda_k + \alpha_{i,k}) e^{(\lambda_i - \alpha_{i,k})t} + t\sum_{i=1}^{2} \lambda_i \Delta\left(t;\lambda_i - \alpha_{i,k}\right) e^{(\lambda_i - \alpha_{i,k})t}}{\left(1 + t\sum_{i=1}^{2} \lambda_i \Delta\left(t;\lambda_i - \alpha_{i,k}\right) e^{(\lambda_i - \alpha_{i,k})t}\right)(\lambda_1+\lambda_2)}. \tag{4.4}$$

We will next further examine two special cases.

4.1. Scenario A. If we suppose that the components are functionally identical but the HR functions differ because the load is shared unequally, then we can have $a_{i,k}^{(1)}(t) \equiv \lambda_i$.

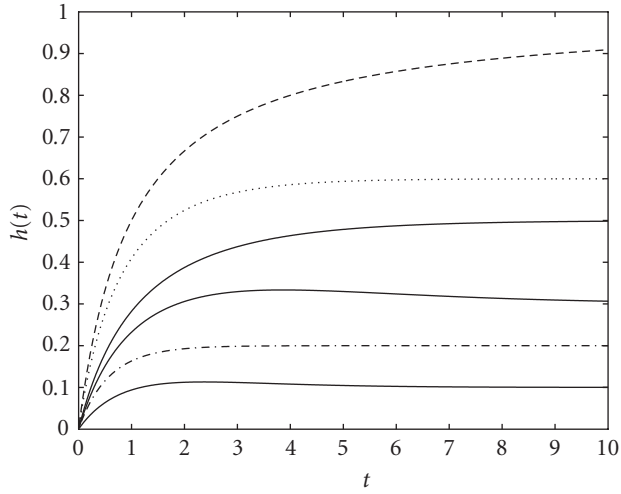


FIGURE 4.1. Representative shapes of the failure rate of a module consisting of two exponential components in parallel with failure rates $\lambda_1, \lambda_2 : \lambda_1 + \lambda_2 = 1$. The solid lines are for the independent case [16, Equation 2.2] with the upper, middle, and lower curves having $\lambda_1 = 0.5, 0.3, 0.1$, respectively. The dashed curve is Scenario A ($\lambda_1 = 0.5, 0.3, 0.1$) and Scenario B with $\lambda_1 = 0.5$, the dotted curve is Scenario B with $\lambda_1 = 0.3$, and the dot-dashed curve is Scenario B with $\lambda_1 = 0.1$.

(As a special case, we may have $\lambda_1 = \lambda_2 =$, say, λ.) Equations (4.1) and (4.3) yield the survival and HR functions

$$S_{M(2)}(t) = \big(1 + t(\lambda_1 + \lambda_2)\big)e^{-(\lambda_1+\lambda_2)t}, \qquad h_{M(2)}(t) = \frac{t(\lambda_1 + \lambda_2)^2}{1 + t(\lambda_1 + \lambda_2)}, \tag{4.5}$$

while (4.4) gives the MRL function

$$\mu_{M(2)}(t) = \frac{2 + t(\lambda_1 + \lambda_2)}{\big(1 + t(\lambda_1 + \lambda_2)\big)(\lambda_1 + \lambda_2)}. \tag{4.6}$$

4.2. Scenario B. As an alternative to Scenario A, we might suppose that the components are sharing the load equally but the component reliabilities differ. In this case, we set $a_{i,k}^{(1)}(t) \equiv \lambda_k$, assuming without loss of generality that $\lambda_1 \neq \lambda_2$, as the case of equality (i.e., $\lambda_1 = \lambda_2 =$, say, λ) is covered by Scenario A. As before, (4.1) and (4.3) give

$$S_{M(2)}(t) = \frac{\lambda_1 e^{-2\lambda_2 t} - \lambda_2 e^{-2\lambda_1 t}}{\lambda_1 - \lambda_2}, \qquad h_{M(2)}(t) = \frac{2\lambda_1\lambda_2\big(e^{-2\lambda_2 t} - e^{-2\lambda_1 t}\big)}{\lambda_1 e^{-2\lambda_2 t} - \lambda_2 e^{-2\lambda_1 t}}. \tag{4.7}$$

Finally, from (4.4), we have the MRL function

$$\mu_{M(2)}(t) = \frac{\lambda_2^2 e^{-2\lambda_1 t} - \lambda_1^2 e^{-2\lambda_2 t}}{2\lambda_1\lambda_2\big(\lambda_2 e^{-2\lambda_1 t} - \lambda_1 e^{-2\lambda_2 t}\big)}. \tag{4.8}$$

Figure 4.1 shows the behaviour of the HR function for various combinations of λ_1, λ_2, normalized so that $\lambda_1 + \lambda_2 = 1$.

In both scenarios, the survival, HR, and MRL functions depend only on λ_1 and λ_2, which are parameters of individual components and can, therefore, be estimated by failing the components under, for example, their "usual" loads a number of times in a laboratory environment. Assuming that we have such data

$$\begin{aligned} t_1(1),\ldots,t_1(n_1) - &\text{ observations of } T_1 \sim S(\bullet;\lambda_1), \\ t_2(1),\ldots,t_2(n_2) - &\text{ observations of } T_2 \sim S(\bullet;\lambda_2), \end{aligned} \tag{4.9}$$

the MLEs of λ_i, $i = 1,2$ are the standard ones: $\widehat{\lambda}_i = n_i/s_i$, where $s_i = \sum_{\ell=1}^{n_i} t_i(\ell)$. However, we may have more information about failures under the original and redistributed load.

First, consider the case of individual components. Suppose that the reliability of individual components can be determined in a laboratory environment, providing n_i observations of T_i, and m_i observations of $T^{+i}_{\text{NOT}(i)}$. Hence in addition to data (4.9), we now also have

$$\begin{aligned} t_1^+(1),\ldots,t_1^+(m_1) - &\text{ observations of } T_1^{+2} \sim \begin{cases} S(\bullet;\lambda_1+\lambda_2) & \text{Scenario A,} \\ S(\bullet;2\lambda_1) & \text{Scenario B,} \end{cases} \\ t_2^+(1),\ldots,t_2^+(m_2) - &\text{ observations of } T_2^{+1} \sim \begin{cases} S(\bullet;\lambda_1+\lambda_2) & \text{Scenario A,} \\ S(\bullet;2\lambda_2) & \text{Scenario B.} \end{cases} \end{aligned} \tag{4.10}$$

(It would be more precise to write $t_i^{+\text{not}(i)}(\ell)$ instead of $t_i^+(\ell)$, but the latter is simpler and we expect no confusion.) The likelihood is the product of the $n_1+n_2+m_1+m_2$ individual likelihoods. Denote $s_i^+ = \sum_{\ell=1}^{m_i} t_i^+(\ell)$. Then in Scenario A, the loglikelihood function is

$$\begin{aligned} \log L(\boldsymbol{\lambda}) = {} & n_1\log\lambda_1 - \lambda_1 s_1 + n_2\log\lambda_2 - \lambda_2 s_2 \\ & + (m_1+m_2)\log(\lambda_1+\lambda_2) - (\lambda_1+\lambda_2)(s_1^+ + s_2^+). \end{aligned} \tag{4.11}$$

Solving the system of equations $(\partial/\partial\lambda_i)\log L(\boldsymbol{\lambda}) = 0$, $i = 1,2$ yields the MLEs for $i = 1,2$,

$$\widehat{\lambda}_i = \frac{b \pm \sqrt{b^2 - 4ac}}{2a}, \tag{4.12}$$

where $a = (s_i - s_{3-i})(s_i + s_1^+ + s_2^+)$, $b = (s_i - s_{3-i})(n_i + m_1 + m_2) + (n_1+n_2)(s_i + s_1^+ + s_2^+)$, and $c = n_i(n_1+n_2+m_1+m_2)$. In Scenario B, we have the loglikelihood function

$$\begin{aligned} \log L(\boldsymbol{\lambda}) = {} & n_1\log\lambda_1 - \lambda_1 s_1 + n_2\log\lambda_2 - \lambda_2 s_2 \\ & + m_1\log(2\lambda_1) + m_2\log(2\lambda_2) - 2\lambda_1 s_1^+ - 2\lambda_2 s_2^+, \end{aligned} \tag{4.13}$$

which yields the MLEs for $i = 1,2$,

$$\widehat{\lambda}_i = \frac{n_i + m_i}{s_i + 2s_i^+}. \tag{4.14}$$

Now, consider the case where we have data on failures of the entire module. We have already noted the "trivial" situation when the module's survival, HR, and MRL function

can be estimated using modules' observed failures, provided that the number of such observations is sufficiently large. If, however, the sample size is not large, then in order to increase the reliability of statistical inference, we want to use every possible bit of information. Hence assume that we have n independent observations of the random vector $(D, T_{1:2}, T^{+D}_{\mathrm{not}(D)})$, where D is the first failed component, $T_{1:2}$ is the time of the first failure, and $T^{+D}_{\mathrm{not}(D)}$ is the time of module's failure. (Note that $\mathrm{not}(D) = 3 - D$.) Our data are the three-dimensional vectors $(d(\ell), t(\ell), t^+(\ell))$, $\ell = 1, \ldots, n$, which are independent observations of the random vector $(D, T_{1:2}, T^{+D}_{\mathrm{not}(D)})$. (It would be more precise to write $t^{+d(\ell)}_{\mathrm{not}(d(\ell))}(\ell)$ instead of $t^+(\ell)$, but the latter is less cumbersome and we expect no confusion.) In addition, we assume that we also know $n_1 = \sum_{\ell=1}^{n} \mathbf{1}_{\{d(\ell)=1\}}$, the number of times component 1 has failed first. The frequency of component 2 failing first is, therefore, $n_2 = n - n_1$. Whether we are dealing with Scenario A or B, the (unknown) parameter is $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, and we need to estimate it. In Scenario A, we have the likelihood function

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \prod_{\ell=1}^{n} f\left(t(\ell); \lambda_{d(\ell)}\right) S\left(t(\ell); \lambda_{3-d(\ell)}\right) f\left(t^+(\ell) - t(\ell); \lambda_1 + \lambda_2\right) \\ &= (\lambda_1 + \lambda_2)^n \lambda_1^{n_1} \lambda_2^{n_2} \exp\left\{ -(\lambda_1 + \lambda_2) \sum_{\ell=1}^{n} t^+(\ell) \right\}. \end{aligned} \tag{4.15}$$

Solving the system of equations $(\partial/\partial\lambda_i) \log L(\boldsymbol{\lambda}) = 0$, $i = 1, 2$ yields the MLEs for $i = 1, 2$, $\widehat{\lambda}_i = 2n_i / \sum_{\ell=1}^{n} t^+(\ell)$. In Scenario B, the likelihood function is

$$\begin{aligned} L(\boldsymbol{\lambda}) &= \prod_{\ell=1}^{n} f\left(t(\ell); \lambda_{d(\ell)}\right) S\left(t(\ell); \lambda_{3-d(\ell)}\right) f\left(t^+(\ell) - t(\ell); 2\lambda_{3-d(\ell)}\right) \\ &= (2\lambda_1\lambda_2)^n \exp\left\{ -(\lambda_1 + \lambda_2) \sum_{\ell=1}^{n} t(\ell) \right\} \exp\left\{ -2 \sum_{\ell=1}^{n} \lambda_{3-d(\ell)} \left(t^+(\ell) - t(\ell)\right) \right\}, \end{aligned} \tag{4.16}$$

which gives the MLEs, for $i = 1, 2$,

$$\widehat{\lambda}_i = \frac{n}{\sum_{\ell=1}^{n} t(\ell) + 2 \sum_{\ell=1}^{n} \mathbf{1}_{\{d(\ell)=3-i\}} \left(t^+(\ell) - t(\ell)\right)}. \tag{4.17}$$

We are now able to compare the performance of the parametric estimators obtained from (3.1) and (3.3), and the nonparametric estimators (3.2) and (3.5), using a small simulation study. We suppose that $\lambda_1 = 0.001$, $\lambda_2 = 0.002$, and that we have $n_1 = n_2 = m_1 = m_2$ observations of failure times of individual components in a laboratory setting, allowing us to estimate the parameters from (4.12) or (4.14). Figure 4.2 compares the estimated survival and MRL functions for Scenario A, while Figure 4.3 shows the same for Scenario B. We can see in both examples that the estimators appear to be unbiased, except, possibly, the nonparametric estimator of the MRL, where there may be underestimation. The variation is larger, as expected, for the nonparametric estimators, and increases over time, except in the case of the parametric estimate of the MRL, where the 90-percentile band appears to be of approximately constant width.

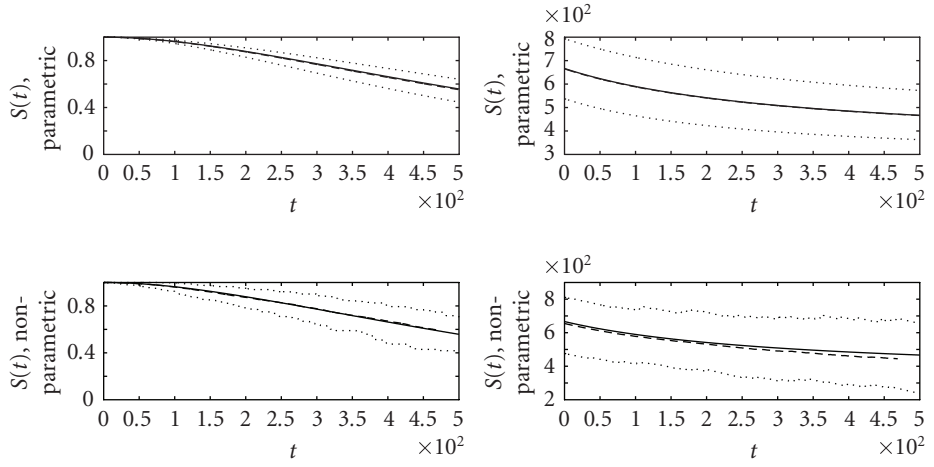


FIGURE 4.2. The estimated survival (left) and MRL (right) functions for Scenario A. Parametric estimates are shown in the top panel, nonparametric in the bottom. The true curve is a solid line. The mean of 100 repetitions is shown as a dashed line, while the dotted lines are the 5th and 95th percentiles.

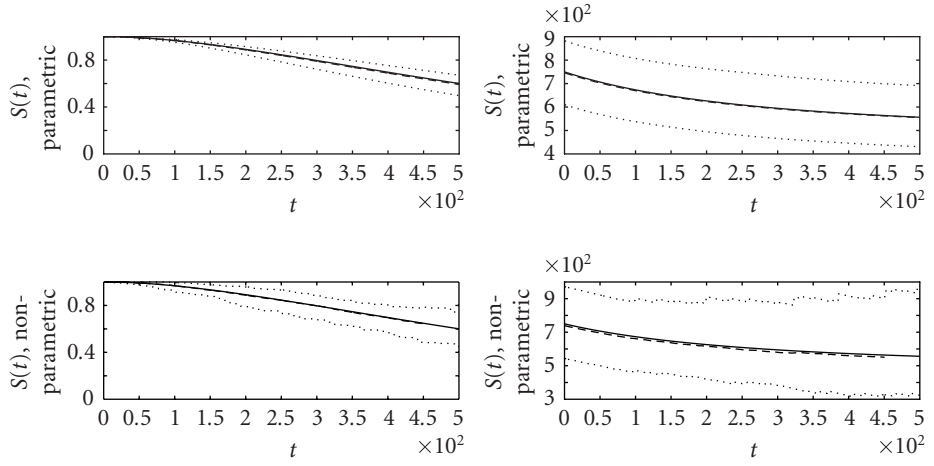


FIGURE 4.3. The estimated survival (left) and MRL (right) functions for Scenario B. Parametric estimates are shown in the top panel, nonparametric in the bottom. The true curve is a solid line. The mean of 100 repetitions is shown as a dashed line, while the dotted lines are the 5th and 95th percentiles.

5. Survival and MRL functions for more than two components

In this section, we consider the survival and MRL functions of modules with arbitrarily, $K \geq 2$, many components. We will need additional notation. Let $S_k^{+(i,j)}(t)$ denote the survival function of a working component k when two other components, i and j, have

failed. Likewise, we interpret the survival functions $S_{i_{K-1}}^{+(i_1,\ldots,i_{K-2})}(t)$, $S_{\mathrm{NOT}(i_1,\ldots,i_{K-1})}^{+(i_1,\ldots,i_{K-1})}(t)$, and so forth.

Theorem 5.1. *For every $K \geq 2$, we have $S_{M(K)}(t) = S_{M(K)}^{*}(t) + S_{M(K)}^{**}(t)$, where*

$$\begin{aligned}
S_{M(K)}^{*}(t) = (-1)^{K-1} &\sum_{i_1 \in \{1,\ldots,K\}} \cdots \sum_{i_{K-1} \in \{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-2}\}} S_{\mathrm{NOT}(i_1,\ldots,i_{K-1})}^{+(i_1,\ldots,i_{K-1})}(t) \int \cdots \int \mathbf{1}_{\{y_{K-1} \leq t\}} \\
&\times \prod_{q \in \{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-1}\}} \frac{S_q^{+(i_1,\ldots,i_{K-2})}(y_{K-1})}{S_q^{+(i_1,\ldots,i_{K-1})}(y_{K-1})} \mathbf{1}_{\{y_{K-1} > y_{K-2}\}} dS_{i_{K-1}}^{+(i_1,\ldots,i_{K-2})}(y_{K-1}) \\
&\cdots \\
&\times \prod_{q \in \{1,\ldots,K\} \setminus \{i_1\}} \frac{S_q(y_1)}{S_q^{+i_1}(y_1)} \mathbf{1}_{\{y_1 > 0\}} dS_{i_1}(y_1), \\
S_{M(K)}^{**}(t) = (-1)^{K-1} &\sum_{i_1 \in \{1,\ldots,K\}} \cdots \sum_{i_{K-1} \in \{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-2}\}} \int \cdots \int \mathbf{1}_{\{y_{K-1} > \max(y_{K-2},t)\}} \\
&\times S_{\mathrm{NOT}(i_1,\ldots,i_{K-1})}^{+(i_1,\ldots,i_{K-2})}(y_{K-1}) \mathbf{1}_{\{y_{K-1} > y_{K-2}\}} dS_{i_{K-1}}^{+(i_1,\ldots,i_{K-2})}(y_{K-1}) \\
&\times \prod_{q \in \{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-2}\}} \frac{S_q^{+(i_1,\ldots,i_{K-3})}(y_{K-2})}{S_q^{+(i_1,\ldots,i_{K-2})}(y_{K-2})} \mathbf{1}_{\{y_{K-2} > y_{K-3}\}} dS_{i_{K-1}}^{+(i_1,\ldots,i_{K-3})}(y_{K-2}) \\
&\cdots \\
&\times \prod_{q \in \{1,\ldots,K\} \setminus \{i_1\}} \frac{S_q(y_1)}{S_q^{+i_1}(y_1)} \mathbf{1}_{\{y_1 > 0\}} dS_{i_1}(y_1).
\end{aligned} \tag{5.1}$$

The proof of Theorem 5.1 is deferred from the appendix.

In the following theorem, we consider the integral $IS_{M(K)}(t)$ for arbitrary $K \geq 2$, from which we can arrive at the MRL function $\mu_{M(K)}(t)$ via the equation $\mu_{M(K)}(t) = IS_{M(K)}(t)/S_{M(K)}(t)$.

Theorem 5.2. *For every $K \geq 2$, we have $IS_{M(K)}(t) = IS_{M(K)}^{*}(t) + IS_{M(K)}^{**}(t)$, where*

$$\begin{aligned}
IS_{M(K)}^{*}(t) = (-1)^{K} &\sum_{i_1 \in \{1,\ldots,K\}} \cdots \sum_{i_{K-1} \in \{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-2}\}} \\
&\int \cdots \int (x - \max(y_{K-1}, t))_{+} dS_{\mathrm{NOT}(i_1,\ldots,i_{K-1})}^{+(i_1,\ldots,i_{K-1})}(x) \\
&\times \prod_{q \in \{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-1}\}} \frac{S_q^{+(i_1,\ldots,i_{K-2})}(y_{K-1})}{S_q^{+(i_1,\ldots,i_{K-1})}(y_{K-1})} \mathbf{1}_{\{y_{K-1} > y_{K-2}\}} dS_{i_{K-1}}^{+(i_1,\ldots,i_{K-2})}(y_{K-1}) \\
&\cdots \\
&\times \prod_{q \in \{1,\ldots,K\} \setminus \{i_1\}} \frac{S_q(y_1)}{S_q^{+i_1}(y_1)} \mathbf{1}_{\{y_1 > 0\}} dS_{i_1}(y_1),
\end{aligned}$$

$$IS^{**}_{M(K)}(t) = (-1)^{K-1} \sum_{i_1\in\{1,\ldots,K\}} \cdots \sum_{i_{K-1}\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}}$$
$$\int\cdots\int (y_{K-1}-t)_+ S^{+(i_1,\ldots,i_{K-2})}_{\mathrm{NOT}(i_1,\ldots,i_{K-1})}(y_{K-1})\,\mathbf{1}_{\{y_{K-1}>y_{K-2}\}}\,dS^{+(i_1,\ldots,i_{K-2})}_{i_{K-1}}(y_{K-1})$$
$$\times \prod_{q\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}} \frac{S^{+(i_1,\ldots,i_{K-3})}_q(y_{K-2})}{S^{+(i_1,\ldots,i_{K-2})}_q(y_{K-2})}\,\mathbf{1}_{\{y_{K-2}>y_{K-3}\}}\,dS^{+(i_1,\ldots,i_{K-3})}_{i_{K-1}}(y_{K-2})$$
$$\cdots$$
$$\times \prod_{q\in\{1,\ldots,K\}\setminus\{i_1\}} \frac{S_q(y_1)}{S^{+i_1}_q(y_1)}\,\mathbf{1}_{\{y_1>0\}}\,dS_{i_1}(y_1). \tag{5.2}$$

The proof of Theorem 5.2 is again deferred to the appendix. We have by now established all the necessary formulas to derive the MRL function $\mu_{M(K)}(t)$ via original and increased loads of individual components.

Explicit formulae for Theorems 5.1 and 5.2 in the case of three and four components are available from the authors. The case $K = 4$ features prominently in our motivating examples in Section 1.

6. Summary

In this paper, we argue that reliability of modules with load-sharing components can be expressed in terms of the reliabilities of individual components exposed to various levels of load (normal and increased). This is of practical interest since the reliability of individual components can be conveniently estimated in a laboratory environment using either a natural aging regime (if time permits) or employing, for example, a quantitative accelerated life testing technique (cf., e.g., Nelson [14]). Hence we have derived equations expressing the module's survival, and thus HR and MRL, functions in terms of the survival functions of individual components. We have also discussed parametric and nonparametric inference for the latter functions, or their parameters if a parametric model has been assumed, under various load-sharing scenarios and data gathering regimes.

Appendix

A. Proofs

Proof of Theorem 5.1. We start calculating the survival function $S_{M(K)}(t)$ using first conditioning and then the formula of total probability. Hence

$$S_{M(K)}(t) = \mathbf{E}\left[\mathbf{P}\left\{T^{(K-1)}_{\kappa(K-1)} > t \mid D^{(0)},\ldots,D^{(K-2)},T^{(K-2)}_{1:2}\right\}\right]$$
$$= \mathbf{E}\left[\exp\left\{-\mathbf{1}_{\{T^{(K-2)}_{D^{(K-2)}}\le t\}}\int_{T^{(K-2)}_{D^{(K-2)}}}^{t}\left(h_{\kappa(K-1)}(x)+\sum_{m=1}^{K-1}a^{(m)}_{D^{(m-1)},\kappa(K-1)}(x)\right)dx\right\}\right]$$

$$= \sum_{i_1 \in \{1,\ldots,K\}} \sum_{i_2 \in \{1,\ldots,K\}\setminus\{i_1\}} \cdots \sum_{i_{K-1} \in \{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}} \mathbf{E}\Bigg[\exp\left\{ -\mathbf{1}_{\{T_{i_{K-1}}^{(K-2)} \le t\}} \int_{T_{i_{K-1}}^{(K-2)}}^{t} \left(h_{i_K}(x) + \sum_{m=1}^{K-1} a_{i_m,i_K}^{(m)}(x) \right) dx \right\} \times \mathbf{1}_{\{D^{(0)}=i_1\}} \cdots \mathbf{1}_{\{D^{(K-3)}=i_{K-2}\}} \mathbf{1}_{\{D^{(K-2)}=i_{K-1}\}} \Bigg], \tag{A.1}$$

where i_K is the (only) member of the singleton set $\{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-1}\}$. Given $D^{(0)} = i_1,\ldots,D^{(K-3)} = i_{K-2}$, the event $D^{(K-2)} = i_{K-1}$ is equivalent to $T_{i_{K-1}}^{(K-2)} < T_{i_K}^{(K-2)}$. By construction, the latter two random variables are independent. Hence we calculate the conditional expectation of $\mathbf{1}_{\{D^{(K-2)}=i_{K-1}\}}$ by first writing

$$\mathbf{P}\left\{ T_{i_K}^{(K-2)} > t \mid D^{(0)} = i_1,\ldots,D^{(K-3)} = i_{K-2} \right\} = \exp\left\{ -\mathbf{1}_{\{T_{i_{K-2}}^{(K-3)} \le t\}} \int_{T_{i_{K-2}}^{(K-3)}}^{t} \left(h_{i_K}(x) + \sum_{m=1}^{K-2} a_{i_m,i_K}^{(m)}(x) \right) dx \right\}. \tag{A.2}$$

Next, we use (A.2) with $t = T_{i_{K-1}}^{(K-2)}$ to get the desired probability of the event $T_{i_{K-1}}^{(K-2)} < T_{i_K}^{(K-2)}$. This, together with (A.1), gives

$$S_{M(K)}(t) = \sum_{i_1 \in \{1,\ldots,K\}} \sum_{i_2 \in \{1,\ldots,K\}\setminus\{i_1\}} \cdots \sum_{i_{K-1} \in \{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}} \mathbf{E}\Bigg[\exp\left\{ -\mathbf{1}_{\{T_{i_{K-1}}^{(K-2)} \le t\}} \int_{T_{i_{K-1}}^{(K-2)}}^{t} \left(h_{i_K}(x) + \sum_{m=1}^{K-1} a_{i_m,i_K}^{(m)}(x) \right) dx \right\} \times \exp\left\{ -\mathbf{1}_{\{T_{i_{K-2}}^{(K-3)} \le T_{i_{K-1}}^{(K-2)}\}} \int_{T_{i_{K-2}}^{(K-3)}}^{T_{i_{K-1}}^{(K-2)}} \left(h_{i_K}(x) + \sum_{m=1}^{K-2} a_{i_m,i_K}^{(m)}(x) \right) dx \right\} \times \mathbf{1}_{\{D^{(0)}=i_1\}} \cdots \mathbf{1}_{\{D^{(K-3)}=i_{K-2}\}} \Bigg]. \tag{A.3}$$

Our next step is to integrate the expression inside $\mathbf{E}[\cdots]$ on the right-hand side of (A.3) with respect to the random variable $T_{i_{K-1}}^{(K-2)}$, for which we need to derive the survival function. Analogously to (A.2), we have that

$$\mathbf{P}\left\{ T_{i_{K-1}}^{(K-2)} > t \mid D^{(0)} = i_1,\ldots,D^{(K-3)} = i_{K-2} \right\} = \exp\left\{ -\int_0^t \left(h_{i_{K-1}}(x) + \sum_{m=1}^{K-2} a_{i_m,i_{K-1}}^{(m)}(x) \right) \mathbf{1}_{\{T_{i_{K-2}}^{(K-3)} \le x\}} dx \right\}. \tag{A.4}$$

Using the latter equation on the right-hand side of (A.3), we have that

$$
\begin{aligned}
S_{M(K)}(t) = & \sum_{i_1\in\{1,\ldots,K\}} \sum_{i_2\in\{1,\ldots,K\}\setminus\{i_1\}} \cdots \sum_{i_{K-1}\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}} \\
& \mathbf{E}\Bigg[\int_{T_{i_{K-2}}^{(K-3)}}^{\infty} \exp\Bigg\{ -\mathbf{1}_{\{y_{K-1}\le t\}} \int_{y_{K-1}}^{t} \left(h_{i_K}(x) + \sum_{m=1}^{K-1} a_{i_m,i_K}^{(m)}(x) \right) dx \Bigg\} \\
& \quad \times \exp\Bigg\{ - \sum_{q\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}} \int_{T_{i_{K-2}}^{(K-3)}}^{y_{K-1}} \left(h_q(x) + \sum_{m=1}^{K-2} a_{i_m,q}^{(m)}(x) \right) dx \Bigg\} \\
& \quad \times \left(h_{i_{K-1}}(y_{K-1}) + \sum_{m=1}^{K-2} a_{i_m,i_{K-1}}^{(m)}(y_{K-1}) \right) dy_{K-1} \mathbf{1}_{\{D^{(0)}=i_1\}} \cdots \mathbf{1}_{\{D^{(K-3)}=i_{K-2}\}} \Bigg].
\end{aligned}
\tag{A.5}
$$

Comparing the latter equation with (A.1), we see that we have "eliminated" the indicator $\mathbf{1}_{\{D^{(K-2)}=i_{K-1}\}}$. Continuing the above arguments until the last indicator $\mathbf{1}_{\{D^{(0)}=i_1\}}$ is "eliminated," we arrive at

$$
\begin{aligned}
S_{M(K)}(t) = & \sum_{i_1\in\{1,\ldots,K\}} \sum_{i_2\in\{1,\ldots,K\}\setminus\{i_1\}} \cdots \sum_{i_{K-1}\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}} \int_0^{\infty} \int_{y_1}^{\infty} \cdots \int_{y_{K-2}}^{\infty} \\
& \exp\Bigg\{ -\mathbf{1}_{\{y_{K-1}\le t\}} \sum_{q\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-1}\}} \int_{y_{K-1}}^{t} \left(h_q(x) + \sum_{m=1}^{K-1} a_{i_m,q}^{(m)}(x) \right) dx \Bigg\} \\
& \times \exp\Bigg\{ - \sum_{q\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}} \int_{y_{K-2}}^{y_{K-1}} \left(h_q(x) + \sum_{m=1}^{K-2} a_{i_m,q}^{(m)}(x) \right) dx \Bigg\} \\
& \times \left(h_{i_{K-1}}(y_{K-1}) + \sum_{m=1}^{K-2} a_{i_m,i_{K-1}}^{(m)}(y_{K-1}) \right) dy_{K-1} \\
& \cdots \\
& \times \exp\Bigg\{ - \sum_{q\in\{1,\ldots,K\}\setminus\{i_1\}} \int_{y_1}^{y_2} \left(h_q(x) + a_{i_1,q}^{(1)}(x) \right) dx \Bigg\} \left(h_{i_2}(y_2) + a_{i_1,i_2}^{(1)}(y_2) \right) dy_2 \\
& \times \exp\Bigg\{ - \sum_{q\in\{1,\ldots,K\}} \int_0^{y_1} h_q(x)dx \Bigg\} h_{i_1}(y_1)dy_1.
\end{aligned}
\tag{A.6}
$$

We will next modify the last $K-1$ exponents in (A.6). We start with

$$
\exp\Bigg\{ - \sum_{q\in\{1,\ldots,K\}} \int_0^{y_1} h_q(x)dx \Bigg\} h_{i_1}(y_1)dy_1 = -\exp\Bigg\{ - \sum_{q\in\{1,\ldots,K\}\setminus\{i_1\}} \int_0^{y_1} h_q(x)dx \Bigg\} dS_{i_1}(y_1).
\tag{A.7}
$$

We now combine the exponent on the right-hand side of (A.7) with the penultimate exponent in (A.6). The last two lines of (A.6) become

$$\begin{aligned}\cdots\times\exp\left\{-\sum_{q\in\{1,\ldots,K\}\setminus\{i_1\}}\int_0^{y_2}\left(h_q(x)+a_{i_1,q}^{(1)}(x)\right)dx\right\}\left(h_{i_2}(y_2)+a_{i_1,i_2}^{(1)}(y_2)\right)dy_2\\ \times(-1)\exp\left\{\sum_{q\in\{1,\ldots,K\}\setminus\{i_1\}}\int_0^{y_1}a_{i_1,q}^{(1)}(x)dx\right\}dS_{i_1}(y_1),\end{aligned}\tag{A.8}$$

which can be rewritten as

$$\begin{aligned}\cdots\times(-1)\exp\left\{-\sum_{q\in\{1,\ldots,K\}\setminus\{i_1,i_2\}}\int_0^{y_2}\left(h_q(x)+a_{i_1,q}^{(1)}(x)\right)dx\right\}dS_{i_2}^{+i_1}(y_2)\\ \times(-1)\exp\left\{\sum_{q\in\{1,\ldots,K\}\setminus\{i_1\}}\int_0^{y_1}a_{i_1,q}^{(1)}(x)dx\right\}dS_{i_1}(y_1).\end{aligned}\tag{A.9}$$

We continue with these arguments and arrive at

$$\begin{aligned}S_{M(K)}(t)=&\sum_{i_1\in\{1,\ldots,K\}}\cdots\sum_{i_{K-1}\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}}\int_0^{\infty}\cdots\int_{y_{K-2}}^{\infty}\\ &\exp\left\{-\mathbf{1}_{\{y_{K-1}\le t\}}\sum_{q\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-1}\}}\int_{y_{K-1}}^{t}\left(h_q(x)+\sum_{m=1}^{K-1}a_{i_m,q}^{(m)}(x)\right)dx\right\}\\ &\times(-1)\exp\left\{-\sum_{q\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-1}\}}\int_0^{y_{K-1}}\left(h_q(x)+\sum_{m=1}^{K-2}a_{i_m,q}^{(m)}(x)\right)dx\right\}\\ &\times dS_{i_{K-1}}^{+(i_1,\ldots,i_{K-2})}(y_{K-1})\\ &\cdots\\ &\times(-1)\exp\left\{\sum_{q\in\{1,\ldots,K\}\setminus\{i_1\}}\int_0^{y_1}a_{i_1,q}^{(1)}(x)dx\right\}dS_{i_1}(y_1).\end{aligned}\tag{A.10}$$

Next, we write $S_{M(K)}(t)=S_{M(K)}^{*}(t)+S_{M(K)}^{**}(t)$, where

$$\begin{aligned}S_{M(K)}^{*}(t)=(-1)^{K-1}&\sum_{i_1\in\{1,\ldots,K\}}\cdots\sum_{i_{K-1}\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-2}\}}S_{\mathrm{NOT}(i_1,\ldots,i_{K-1})}^{+(i_1,\ldots,i_{K-1})}(t)\int_0^{\infty}\cdots\int_{y_{K-2}}^{\infty}\mathbf{1}_{\{y_{K-1}\le t\}}\\ &\times\exp\left\{\sum_{q\in\{1,\ldots,K\}\setminus\{i_1,\ldots,i_{K-1}\}}\int_0^{y_{K-1}}a_{i_m,q}^{(K-1)}(x)dx\right\}dS_{i_{K-1}}^{+(i_1,\ldots,i_{K-2})}(y_{K-1})\\ &\cdots\\ &\times\exp\left\{\sum_{q\in\{1,\ldots,K\}\setminus\{i_1\}}\int_0^{y_1}a_{i_1,q}^{(1)}(x)dx\right\}dS_{i_1}(y_1)\end{aligned}$$

$$S^{**}_{M(K)}(t) = (-1)^{K-1} \sum_{i_1 \in \{1,\ldots,K\}} \cdots \sum_{i_{K-1} \in \{1,\ldots,K\} \setminus \{i_1,\ldots,i_{K-2}\}} \int_0^\infty \int_{y_1}^\infty \cdots \int_{y_{K-2}}^\infty \mathbf{1}_{\{y_{K-1}>t\}} S^{+(i_1,\ldots,i_{K-2})}_{\mathrm{NOT}(i_1,\ldots,i_{K-1})}(y_{K-1}) dS^{+(i_1,\ldots,i_{K-2})}_{i_{K-1}}(y_{K-1}) \cdots \times \exp\left\{ \sum_{q \in \{1,\ldots,K\} \setminus \{i_1\}} \int_0^{y_1} a^{(1)}_{i_1,q}(x)dx \right\} dS_{i_1}(y_1). \tag{A.11}$$

Write $a^{(1)}_{i_1,q}(x)$ as the sum of $h_q(x) + a^{(1)}_{i_1,q}(x)$ and $-h_g(x)$, which shows that the rightmost exponent in (A.11) can be written as the ratio $S_q(y_1)/S^{+i_1}_q(y_1)$. Similarly, we have the equations

$$\exp\left\{ \int_0^{y_2} a^{(2)}_{i_2,q}(x)dx \right\} = \frac{S^{+i_1}_q(y_2)}{S^{+(i_1,i_2)}_q(y_2)}, \ldots, \exp\left\{ \int_0^{y_{K-1}} a^{(K-1)}_{i_2,q}(x)dx \right\} = \frac{S^{+(i_1,\ldots,i_{K-2})}_q(y_{K-1})}{S^{+(i_1,\ldots,i_{K-1})}_q(y_{K-1})}. \tag{A.12}$$

Theorem 5.1 follows. □

Proof of Theorem 3.1. This is a consequence of Theorem 5.1 and the observation that the product $S_1(t)S_2(t)$ is equal to $-\sum_{i=1}^2 \int_t^\infty S_{\mathrm{NOT}(i)}(y)dS_i(y)$, which appears in the result of Theorem 5.1 when $K = 2$. □

Proof of Theorem 5.2. For any random variable X, whose survival function we denote by $S_X(t)$, the integral $\int_t^\infty S_X(x)\mathbf{1}_{\{z \le x\}}dx$ is equal to the expectation $\mathbf{E}[(X - \max(z,t))_+]$, which is of course equal to $-\int_0^\infty (x - \max(z,t))_+ dS_X(x)$. Furthermore, $\int_t^\infty \mathbf{1}_{\{y>x\}}dx$ is equal to $(y-t)_+$. These observations and (3.1) complete the proof of Theorem 5.2. □

References

[1] J. Shao and L. R. Lamberson, "Modeling a shared-load k-out-of-n:G system," *IEEE Transactions on Reliability*, vol. 40, no. 2, pp. 205–209, 1991.

[2] Flight Saftey Australia, "Deadstick in an Airbus 330," 2001, http://www.casa.gov.au/fsa/2001/sep/10-13.pdf.

[3] Wikipedia, "Gimli Glider," 2007, http://en.wikipedia.org/w/index.php?title Gimli_Glider&oldid = 140801954.

[4] Wikipedia, "1998 Auckland Power Crisis," 2006, http://en.wikipedia.org/wiki/1998\Auckland\power\crisis.

[5] Ministry of Economic Development, "Auckland Power Supply Failure 1998: The Report of the Ministerial Inquiry into the Auckland Power Supply Failure," 1998, Ministry of Economic Development, Wellington, New Zealand, http://www.med.govt.nz/templates/ContentTopicSummary.aspx?id 12143.

[6] W. Q. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data*, John Wiley & Sons, New York, NY, USA, 1998.

[7] C.-D. Lai and M. Xie, *Stochastic Ageing and Dependence for Reliability*, Springer, New York, NY, USA, 2006.

[8] D. N. P. Murthy and D. G. Nguyen, "Study of two-component system with failure interaction," *Naval Research Logistics Quarterly*, vol. 32, no. 2, pp. 239–247, 1985.

[9] D. N. P. Murthy and R. J. Wilson, "Parameter estimation in multi-component systems with failure interaction," *Applied Stochastic Models and Data Analysis*, vol. 10, no. 1, pp. 47–60, 1994.
[10] H. Kim and P. H. Kvam, "Reliability estimation based on system data with an unknown load share rule," *Lifetime Data Analysis*, vol. 10, no. 1, pp. 83–94, 2004.
[11] P. H. Kvam and E. A. Peña, "Estimating load-sharing properties in a dynamic reliability system," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 262–272, 2005.
[12] H. Liu, "Reliability of a load-sharing k-out-of-n:G system: non-iid components with arbitrary distributions," *IEEE Transactions on Reliability*, vol. 47, no. 3, part 1, pp. 279–284, 1998.
[13] R. I. Zequeira and C. Bérenguer, "On the inspection policy of a two-component parallel system with failure interaction," *Reliability Engineering & System Safety*, vol. 88, no. 1, pp. 99–107, 2005.
[14] W. B. Nelson, *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*, John Wiley & Sons, New York, NY, USA, 2004.
[15] N. Mimoto and R. Zitikis, "The Atkinson index, the Moran statistic, and testing exponentiality," submitted to *Journal of the Japan Statistical Society*.
[16] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, To Begin With, Silver Spring, Md, USA, 1981.

Mark Bebbington: Institute of Information Sciences and Technology, Massey University, Private Bag 11222, Palmerston North, New Zealand
Email address: m.Bebbington@massey.ac.nz

Chin-Diew Lai: Institute of Information Sciences and Technology, Massey University, Private Bag 11222, Palmerston North, New Zealand
Email address: c.lai@massey.ac.nz

Ričardas Zitikis: Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7
Email address: zitikis@stats.uwo.ca

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 83852, 15 pages
doi:10.1155/2007/83852

Research Article

Long-Range Dependence in a Cox Process Directed by a Markov Renewal Process

D. J. Daley, T. Rolski, and R. Vesilo

Received 19 June 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

A Cox process N_{Cox} directed by a stationary random measure ξ has second moment $\operatorname{var} N_{\mathrm{Cox}}(0,t] = \mathrm{E}(\xi(0,t]) + \operatorname{var}\xi(0,t]$, where by stationarity $\mathrm{E}(\xi(0,t]) = (\text{const.})t = \mathrm{E}(N_{\mathrm{Cox}}(0,t])$, so long-range dependence (LRD) properties of N_{Cox} coincide with LRD properties of the random measure ξ. When $\xi(A) = \int_A \nu_{J(u)}\mathrm{d}u$ is determined by a density that depends on rate parameters ν_i ($i \in \mathbb{X}$) and the current state $J(\cdot)$ of an $\mathbb{X}$-valued stationary irreducible Markov renewal process (MRP) for some countable state space $\mathbb{X}$ (so $J(t)$ is a stationary semi-Markov process on $\mathbb{X}$), the random measure is LRD if and only if each (and then by irreducibility, every) generic return time Y_{jj} ($j \in \mathbb{X}$) of the process for entries to state j has infinite second moment, for which a necessary and sufficient condition when $\mathbb{X}$ is finite is that at least one generic holding time X_j in state j, with distribution function (DF) H_j, say, has infinite second moment (a simple example shows that this condition is not necessary when $\mathbb{X}$ is countably infinite). Then, N_{Cox} has the same Hurst index as the MRP N_{MRP} that counts the jumps of $J(\cdot)$, while as $t \to \infty$, for finite $\mathbb{X}$, $\operatorname{var} N_{\mathrm{MRP}}(0,t] \sim 2\lambda^2 \int_0^t \mathcal{G}(u)\mathrm{d}u$, $\operatorname{var} N_{\mathrm{Cox}}(0,t] \sim 2\int_0^t \sum_{i\in\mathbb{X}}(\nu_i - \overline{\nu})^2 \varpi_i \mathcal{H}_i(t)\mathrm{d}u$, where $\overline{\nu} = \sum_i \varpi_i \nu_i = \mathrm{E}[\xi(0,1]]$, $\varpi_j = \Pr\{J(t) = j\}$, $1/\lambda = \sum_j \check{p}_j \mu_j$, $\mu_j = \mathrm{E}(X_j)$, $\{\check{p}_j\}$ is the stationary distribution for the embedded jump process of the MRP, $\mathcal{H}_j(t) = \mu_i^{-1}\int_0^\infty \min(u,t)[1 - H_j(u)]\mathrm{d}u$, and $\mathcal{G}(t) \sim \int_0^t \min(u,t)[1 - G_{jj}(u)]\mathrm{d}u/m_{jj} \sim \sum_i \varpi_i \mathcal{H}_i(t)$ where G_{jj} is the DF and m_{jj} the mean of the generic return time Y_{jj} of the MRP between successive entries to the state j. These two variances are of similar order for $t \to \infty$ only when each $\mathcal{H}_i(t)/\mathcal{G}(t)$ converges to some $[0,\infty]$-valued constant, say, γ_i, for $t \to \infty$.

1. Introduction

This paper is a sequel to Daley [1] which arose from wanting to decide whether the detailed long-range dependent (LRD) behavior of a Cox process N_{Cox} directed by the ON phases of a stationary ON/OFF alternating renewal process N is the same as the LRD behavior of N. It was shown that both processes have the same Hurst index but that the ratio $\operatorname{var} N_{\mathrm{Cox}}(0,t]/\operatorname{var} N(0,t]$ need not converge for $t \to \infty$.

Here, we examine the nature of these two variance functions for the case of a Cox process whose instantaneous rate ν_i is determined by the state $i \in \mathbb{X}$, with $\mathbb{X}$ being countable (sometimes it must be finite), of a LRD stationary Markov renewal process (MRP), of which our earlier example of an alternating renewal process (ARP) is the simplest. MRPs have long been an interest of Jeff Hunter (e.g., Hunter [2]), and it is a pleasure to contribute this paper to a volume that marks his contributions to the academic community both inside New Zealand and further afield where D. Daley in particular has enjoyed his company many times since first meeting him in Chapel Hill, NC, and T. Rolski at Cornell University.

In Section 2, we introduce the necessary notation and recall known results that are relevant to the problem at hand. Section 3 develops formulae for univariate and bivariate marginal probabilities for MRPs that take us into the realm of Markov renewal equations which enable us to address the questions we raise when $\mathbb{X}$ is finite. We conclude in Section 4 with remarks on the case where $\mathbb{X}$ is countably infinite. In the appendix, we prove an asymptotic convergence result due originally, we believe, to Sgibnev [3].

2. The setting and known results

A Cox process N_{Cox} driven by the random measure ξ is a point process which, conditional on the realization ξ, is a Poisson process with parameter measure ξ (e.g., Daley and Vere-Jones [4, Section 6.2]). Then, when N_{Cox} and ξ are located in the half-line $\mathbb{R}_+$, for Borel subsets A of $\mathbb{R}_+$,

$$\mathrm{E}[N_{\mathrm{Cox}}(A)] = \mathrm{E}[\xi(A)], \qquad \operatorname{var} N_{\mathrm{Cox}}(A) = \mathrm{E}[\xi(A)] + \operatorname{var} \xi(A) \tag{2.1}$$

[4, Proposition 6.2.II]. A stationary point process or random measure ξ on $\mathbb{R}$ is LRD when

$$\limsup_{t\to\infty} \frac{\operatorname{var} \xi(0,t]}{t} = \infty \tag{2.2}$$

[4, Section 12.7], and its Hurst index H is defined by

$$H = \inf\left\{h : \limsup_{t\to\infty} \frac{\operatorname{var} \xi(0,t]}{t^{2h}} < \infty\right\}. \tag{2.3}$$

It follows from (2.1) and (2.2) that a Cox process is LRD if and only if the random measure driving it is LRD, and that they both have the same Hurst index (this is Daley [1, Proposition 1]).

We choose to describe a Markov renewal process (see, e.g., Çinlar [5] or Kulkarni [6] for a textbook account) both in terms of the sequence $\{(X_n, J_n)\}$ of successive intervals

X_n between jumps of a Markov chain $\{J_n\}$ on a countable state space $\mathbb{X}$ with one-step transition probabilities $(p_{ij},\ i,j \in \mathbb{X})$, and the $\mathbb{X}$-valued semi-Markov process $\{J(t) : t \in \mathbb{R}\}$ which can be related via the time epochs $T_n = T_0 + X_1 + \cdots + X_n$ subsequent to some initial epoch T_0, as $J_{n+1} = J(T_n+)$ and

$$\begin{aligned} J(t) &= J_n \quad (T_{n-1} \le t < T_n,\ n = 1,2,\dots) \\ &= \sum_{n=1}^{\infty} J_n I_{T_{n-1} \le t < T_n}(t). \end{aligned} \tag{2.4}$$

We use the random measure

$$\xi(A) \equiv \int_A \nu_{J(u)} \mathrm{d}u, \tag{2.5}$$

where $\{\nu_i\}$ is a family of nonnegative constants defined over $\mathbb{X}$, as the driving measure of the Cox process N_{Cox} that we consider. This means that if $\sigma_i(0,t]$ is the (Lebesgue) measure of that part of the interval $(0,t]$ during which $J(u) = i$ for $i \in \mathbb{X}$ (mnemonically, the sojourn time in i during $(0,t]$), then

$$\xi(0,t] = \sum_{i \in \mathbb{X}} \nu_i \sigma_i(0,t] \quad (t \in \mathbb{R}_+), \tag{2.6}$$

and N_{Cox} consists of points evolving as a Poisson process at rate ν_i on the disjoint sets of support of σ_i for $i \in \mathbb{X}$. Equation (2.1) shows that in order to evaluate the variance of the Cox process, we must find

$$\operatorname{var} \xi(0,t] = \sum_{i,j \in \mathbb{X}} \nu_i \nu_j \operatorname{cov}\big(\sigma_i(0,t], \sigma_j(0,t]\big). \tag{2.7}$$

When $\mathbb{X}$ is a finite set, the finiteness conditions we impose are automatically satisfied, but for the sake of completeness, we allow the countably infinite case of $\mathbb{X}$ except where we know of proof only in the finite case (see (2.20) and Section 4). For N_{Cox} to be well defined, we want $\xi(0,t] < \infty$ a.s. for finite $t > 0$, which is the case when $\overline{\nu} \equiv \sum_{i \in \mathbb{X}} \nu_i \varpi_i < \infty$, where for stationary $J(\cdot)$, we set

$$\varpi_i = \Pr\{J(t) = i\} = \mathrm{E}[\sigma_i(0,1]] \quad (\text{all } t). \tag{2.8}$$

Then, $\mathrm{E}[\xi(0,t]] = \overline{\nu} t$ for all $t > 0$. Assuming (as we must for the conditions of stationarity to hold) that the chain $\{J_n\}$ is irreducible and has a stationary distribution $\{\check{p}_i\}$ (so $\check{p}_j = \sum_{i \in \mathbb{X}} \check{p}_i p_{ij}$), this is related to the distribution $\{\varpi_i\}$ via the mean holding times $\mu_i = \int_0^\infty \overline{H}_i(u)\mathrm{d}u = \mathrm{E}(X_n \mid J_n = i)$ as at (2.9). When $F_{ij}(t) = \Pr\{X_n \le t \mid J_n = i,\ J_{n+1} = j\}$, the process of termination of sojourns in state i is governed by the (in general) dishonest DFs $Q_{ij}(t) = p_{ij} F_{ij}(t)$ but such that the holding time DFs $H_i(t) = \sum_j Q_{ij}(t)$ are honest. We make the simplifying assumption that $p_{ii} = 0$ (all i).

Assume that the point process defined by such an MRP (i.e., the sequence of epochs $\{T_n\}$) can and does exist in a stationary state; in which case, its intensity λ is given by $\lambda^{-1} = \sum_{i \in \mathbb{X}} \check{p}_i \mu_i$, and

$$\varpi_i = \lambda \check{p}_i \mu_i = \Pr\{J(t) = i\} \quad (\text{all } t), \tag{2.9}$$

with the semi-Markov process $J(\cdot)$ here being stationary also. Since the rate of entry epochs into state i equals $\lambda \check{p}_i$, it follows that the mean time m_{ii} between successive entries into state i is given by

$$m_{ii} = \frac{1}{\lambda \check{p}_i} = \frac{\mu_i}{\varpi_i} \quad (i \in \mathbb{X}). \tag{2.10}$$

We assume that our MRP is irreducible (i.e., the Markov chain $\{J_n\}$ is irreducible), and therefore it can be studied via first passage distributions $G_{ij}(\cdot)$ (it is here that the assumption $p_{ii} = 0$ simplifies the discussion); define for every $i \in \mathbb{X}$ and $j \in \mathbb{X}$ except $j = i$

$$\begin{aligned} G_{ji}(t) &= \Pr\{\text{entry to } i \text{ occurs in } (0,t] \mid \text{state } j \neq i \text{ entered at } 0\}, \\ G_{ii}(t) &= \Pr\{\text{second entry to } i \text{ occurs in } (0,t] \mid \text{state } i \text{ entered at } 0\}. \end{aligned} \tag{2.11}$$

Then, for example,

$$G_{ii}(t) = \sum_{k \in \mathbb{X} \setminus \{i\}} p_{ik} \int_0^t F_{ik}(\mathrm{d}u) G_{ki}(t-u) = \sum_k (Q_{ik} * G_{ki})(t), \tag{2.12}$$

where our convention in writing the convolution $(A * B)(t)$ of a nonnegative function $B(\cdot)$ (like G_{ki}) with respect to a measure $A(\cdot)$ (like Q_{ik}) is that $(A * B)(t) = \int_0^t A(\mathrm{d}u) B(t-u)$, or in vector algebra notation when $\mathbf{A} = (A_{ij}(\cdot))$ and $\mathbf{B} = (B_{ij}(\cdot))$ are compatible,

$$\big((\mathbf{A} * \mathbf{B})(t)\big)_{ij} = \sum_{k \in \mathbb{X}} \int_0^t A_{ik}(\mathrm{d}u) B_{kj}(t-u). \tag{2.13}$$

When we consider only the point process N_{MRP} of epochs where entrances into states occur, for which we should count the number of entries N_i into state i and therefore have $N_{\mathrm{MRP}} = \sum_{i \in \mathbb{X}} N_i$, Sgibnev [7] has shown (under the condition of irreducibility) that there is a solidarity result; it implies that $m_{ii}^{-2} \operatorname{var} N_i(0,t] \sim m_{11}^{-2} \operatorname{var} N_1(0,t]$ as $t \to \infty$ when the number of visits to any one state has LRD behavior (and hence, that the point process of visits to any other given state is LRD also, and moreover the asymptotic behavior of the variance function $m_{ii}^{-2} \operatorname{var} N_i(0,t]$ is the same irrespective of the state i). Given this solidarity property, it is seemingly extraordinary that the variance of the amount of time spent in the various states need not have the same asymptotic behavior. The major aim now in considering a Cox process directed by a stationary MRP is to show that this asymptotic behavior is determined, as in the ARP case, by a linear mixture of integrals of certain functions that are crucial in Sgibnev [3, 7] (see also Appendix A), namely,

$$\mathcal{H}_i(t) \equiv \frac{1}{\mu_i} \int_0^\infty \min(u,t) \overline{H}_i(u) \mathrm{d}u. \tag{2.14}$$

We also write $\widetilde{H}_i(t) = (1/\mu_i) \int_0^t \overline{H}_i(u) \mathrm{d}u$; this equals $\mathcal{H}_i'(t)$. Write $\boldsymbol{\mathcal{H}}(t)$ and $\widetilde{\mathbf{H}}(t)$ for vectors with components $\mathcal{H}_i(t)$ and $\widetilde{H}_i(t)$, respectively.

Recall (2.7) writing alternatively

$$V(t) \equiv \operatorname{var}\xi(0,t] = \sum_{i,j\in\mathbb{X}} 2\nu_i\nu_j \int_0^t (t-u)[\varpi_{ij}(u) - \varpi_i\varpi_j]\,\mathrm{d}u, \tag{2.15}$$

where $\varpi_{ij}(u) = \Pr\{J(0+) = i,\, J(u) = j\}$ for $u > 0$ (stationarity of $J(\cdot)$ is assumed as around (2.9)). In terms of the distribution of $J(\cdot)$, only the uni- and bivariate distributions $\{\varpi_i\}$ and $\varpi_{ij}(u)$ are involved in (2.15), and LRD behavior is therefore associated with the integral of $\varpi_{ij}(u) - \varpi_i\varpi_j$ over large intervals. Since these bivariate probabilities are those of a semi-Markov process, each $\varpi_{ij}(\cdot)$ has a representation as a convolution involving DFs of lifetimes on the state space $\mathbb{X}$, and this leads to renewal function representations and use of asymptotics of renewal functions as we shall demonstrate.

Write $U_{ij}(t) = \mathrm{E}(N_j[0,t] \mid \text{state } i \text{ entered at } 0) = \delta_{ij} + \sum_{k\in\mathbb{X}} \int_0^t Q_{ik}(\mathrm{d}u)U_{kj}(t-u)$. Then, $U_i(x) = \sum_{j\in\mathbb{X}} U_{ij}(x)$ satisfies the same backwards equation with δ_{ij} and U_{kj} replaced by 1 and U_k. Writing $\mathbf{Q} = (Q_{ij})_{i,j\in\mathbb{X}}$, define

$$\begin{aligned}\mathbf{U} = \big(U_{ij}(t)\big)_{i,j\in\mathbb{X}} &= \mathbf{I} + \mathbf{Q} + \mathbf{Q}^{2*} + \cdots \\ &= \mathbf{I} + \mathbf{Q} * \mathbf{U} = \mathbf{I} + \mathbf{U} * \mathbf{Q}\end{aligned} \tag{2.16}$$

(note that $U_{ij}(t) = \mathrm{E}[N_j(0,t] \mid i \text{ entered at } 0]$ for $j \neq i$, while for $j = i$, since the N_j are orderly, $N_i[0,t] = 1 + N_i(0,t]$). Since $U_i = \mathbf{U}_i^T\mathbf{e}$, where $\mathbf{U}_i$ is the vector over $j \in \mathbb{X}$ of U_{ij} (all vectors are column vectors unless transposed as, e.g., $\mathbf{U}_i^T$),

$$U_{\mathrm{MRP}}(t) = \check{\mathbf{p}}^T\mathbf{U}\mathbf{e} = \mathrm{E}\big(N_{\mathrm{MRP}}[0,t] \mid \text{jump at 0 of stationary } J(\cdot)\big). \tag{2.17}$$

Now, substitute in the standard formula (e.g., Daley and Vere-Jones [4, page 62]) to give $\operatorname{var}N_{\mathrm{MRP}}(0,t]$ for the stationary point process generated by the jumps of a stationary MRP:

$$\operatorname{var}N_{\mathrm{MRP}}(0,t] = \lambda \int_0^t \big(2[U_{\mathrm{MRP}}(u) - \lambda u] - 1\big)\,\mathrm{d}u, \tag{2.18}$$

where in terms of the respective vectors $\check{\mathbf{p}}$ and $\boldsymbol{\mu}$ of the stationary jump distribution $\{\check{p}_i\}$ and mean sojourn times $\{\mu_i\}$ of the states $i \in \mathbb{X}$, $1/\lambda = \check{\mathbf{p}}^T\boldsymbol{\mu} = \sum_{i\in\mathbb{X}} \check{p}_i\mu_i$ as around (2.9). The integrand at (2.18) has uniformly bounded increments because $U_{\mathrm{MRP}}(t) \sim \lambda t$ $(t \to \infty)$ and it is subadditive (see Appendix B), like the renewal function (e.g., Daley and Vere-Jones [4, Exercise 4.4.5(b)]).

Let G_{kk} be the return time DF for some given state $k \in \mathbb{X}$. Sgibnev [7] showed that for $t \to \infty$ and all other $i, j \in \mathbb{X}$ for the stationary irreducible LRD MRP,

$$m_{jj}U_{ij}(t) - t \sim \mathcal{G}(t) = \frac{1}{m_{kk}} \int_0^\infty \min(t,u)\overline{G}_{kk}(u)\mathrm{d}u. \tag{2.19}$$

Then from (2.17), at least for a finite state space $\mathbb{X}$, it follows that

$$U_{\mathrm{MRP}}(t) - \lambda t = \sum_{i\in\mathbb{X}}\sum_{j\in\mathbb{X}} \check{p}_i(U_{ij}(t) - \lambda\check{p}_j t) \sim \sum_{i\in\mathbb{X}}\sum_{j\in\mathbb{X}} \check{p}_i \cdot \lambda\check{p}_j \cdot \mathcal{G}(t) = \lambda\mathcal{G}(t), \tag{2.20}$$

and hence that

$$\operatorname{var} N_{\mathrm{MRP}}(t) \sim 2\lambda^2 \int_0^t \mathcal{G}(u)\mathrm{d}u. \tag{2.21}$$

Whether (2.21) holds for countably infinite state space remains a question for another place; the discussion in Section 4 is relevant to the nature of the return time distribution G_{kk} in (2.19).

3. Recurrence relations for bivariate probabilities and asymptotics

In this section, we establish the result that extends the simpler conclusion of Daley [1] from an alternating renewal process to a Markov renewal process on a finite state space $\mathbb{X}$. So far, we do not know the nature of any extension to the case that $\mathbb{X}$ is countably infinite.

Theorem 3.1. *Let the Cox process N_{Cox} be driven by a long-range dependent random measure $\xi(A) = \int_A \nu_{J(u)}\mathrm{d}u$ determined by a stationary semi-Markov process $J(\cdot)$ on a finite state space $\mathbb{X}$. Then, N_{Cox} has the same Hurst index as the Markov renewal process N_{MRP} underlying $J(\cdot)$. Both $\operatorname{var} N_{\mathrm{Cox}}(0,x]$ and $\operatorname{var} N_{\mathrm{MRP}}(0,x]$ are asymptotically determined by the holding time DFs $\{H_i(\cdot) : i \in \mathbb{X}\}$ in the MRP, at least one of which must have infinite second moment. Under these conditions, for $t \to \infty$,*

$$\operatorname{var} N_{\mathrm{Cox}}(0,t] \sim 2\int_0^t \sum_{i\in\mathbb{X}} (\nu_i - \bar{\nu})^2 \varpi_i \mathcal{H}_i(u)\mathrm{d}u, \tag{3.1}$$

while $\operatorname{var} N_{\mathrm{MRP}}(0,t]$ is given by (2.21) in which

$$\mathcal{G}(u) \sim \sum_{i\in\mathbb{X}} \varpi_i \mathcal{H}_i(u) \quad (u \longrightarrow \infty), \tag{3.2}$$

where $\{\varpi_i\}$ is the stationary distribution for $J(\cdot)$ and the truncated second moment functions $\mathcal{H}_i(\cdot)$ and $\mathcal{G}(\cdot)$ are given by (2.14) and (2.19).

In general, $\operatorname{var} N_{\mathrm{MRP}}(0,t] \sim \lambda^2 \int_0^t \mathcal{G}(u)\mathrm{d}u \not\sim (\text{const.})\operatorname{var} N_{\mathrm{Cox}}(0,t]$, but if for some j, all the ratios $\mathcal{H}_i(t)/\mathcal{H}_j(t)$ $(i \in \mathbb{X} \setminus \{j\})$ converge as $t \to \infty$ to limits in $[0,\infty]$, then

$$\operatorname{var} N_{\mathrm{MRP}}(0,t] \sim (\text{const.})\operatorname{var} N_{\mathrm{Cox}}(0,t] \quad (t \longrightarrow \infty). \tag{3.3}$$

Proof. If all holding time DFs H_i have finite second moments, then because $\mathbb{X}$ is finite, so do all return time DFs G_{kk}, and the MRP cannot be LRD.

The last part of the theorem, given (3.1)–(3.2), is proved in the same way as the analogous statement for the alternating renewal case, so for the rest, we concentrate on demonstrating (3.1)–(3.2).

We develop expressions involving the bivariate probabilities $\varpi_{ij}(t)$ (see around (2.15)) for the stationary irreducible semi-Markov process $J(\cdot)$. The variance function $V(t) = \operatorname{var}\xi(0,t]$ at (2.15) describes the variance of the Cox process via (2.1). Equation (2.15) shows that $V(\cdot)$ is differentiable, with derivative

$$V'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\operatorname{var}\xi(0,t] = \sum_{i,j\in\mathbb{X}} 2\nu_i\nu_j \int_0^t [\varpi_{ij}(u) - \varpi_i\varpi_j]\mathrm{d}u, \tag{3.4}$$

which is already simpler to evaluate than (2.15) itself. In particular, when $\xi(\cdot)$ is LRD, $V(t)$ is larger than $O(t)$ for $t \to \infty$, so that when $V'(t) \sim g(t) \to \infty$ as $t \to \infty$ for some $g(\cdot)$ that is ultimately monotone, the asymptotic behavior of $V(t)$ for large t is the same as for $\int_0^t g(u)\mathrm{d}u$.

For a stationary irreducible semi-Markov process $J(\cdot)$ on $\mathbb{X}$ as we are considering, the joint distribution on $\mathbb{X} \times \mathbb{X} \times \mathbb{R}_+$ of the current state i, the state k next entered, and the forward recurrence time x for that next entry, is determined by the density function

$$\frac{\varpi_i \overline{Q}_{ik}(x)\mathrm{d}x}{\mu_i} \qquad (i \in \mathbb{X},\ k \in \mathbb{X} \setminus \{i\},\ 0 < x < \infty). \tag{3.5}$$

In (3.13), we use $\widetilde{\mathbf{Q}}(t)$ to denote the array with elements $(1/\mu_i)\int_0^t \overline{Q}_{ij}(u)\mathrm{d}u$. Note that the vector $\widetilde{\mathbf{H}}(t)$ as below (2.14) satisfies $\widetilde{\mathbf{H}}(t) = \widetilde{\mathbf{Q}}(t)\mathbf{e}$.

Define

$$\Pi_{j|i}(t) = \int_0^t \mathrm{E}\big(\delta_{j,J(u)} \mid J(0+) = i\big)\mathrm{d}u, \tag{3.6}$$

so that

$$\mathrm{E}\left[\delta_{i,J(0)} \int_0^t \delta_{j,J(u)}\mathrm{d}u\right] = \varpi_i \Pi_{j|i}(t) = \int_0^t \varpi_{ij}(u)\mathrm{d}u. \tag{3.7}$$

Setting $\mathbf{\Pi}(t) = (\Pi_{j|i}(t))_{i,j\in\mathbb{X}}$, it follows that (3.4) is expressible as

$$V'(t) = \sum_{i,j\in\mathbb{X}} 2\nu_i\nu_j\big[\varpi_i(\Pi_{j|i}(t) - \varpi_j t)\big] = 2\nu^T \operatorname{diag}(\boldsymbol{\varpi})\big(\mathbf{\Pi}(t) - \mathbf{e}\boldsymbol{\varpi}^T t\big)\nu. \tag{3.8}$$

We now develop expressions for $\Pi_{j|i}$ in terms of the truncated second moment functions at (2.14) and the related functions, discussed in Lemma 3.3,

$$M_{ij}(t) = \mathrm{E}\left[\int_0^t \delta_{j,J(u)}\mathrm{d}u \;\middle|\; \text{state } i \text{ entered at } 0\right]. \tag{3.9}$$

Lemma 3.2

$$\Pi_{j|i}(t) = \delta_{ji}\mathscr{H}_i(t) + \sum_{k\in\mathbb{X}} \int_0^t \frac{\overline{Q}_{ik}(v)}{\mu_i} M_{kj}(t-v)\mathrm{d}v, \tag{3.10a}$$

equivalently, with $\mathbf{M}(t) = (M_{ij}(t))_{i,j\in\mathbb{X}}$,

$$\mathbf{\Pi}(t) = \operatorname{diag}\big(\boldsymbol{\mathscr{H}}(t)\big) + (\widetilde{\mathbf{Q}} * \mathbf{M})(t). \tag{3.10b}$$

Proof. For $j \neq i$, we use the joint distribution at (3.5) and a backwards decomposition to write

$$\Pi_{j|i}(t) = \sum_{k\in\mathbb{X}} \int_0^t \frac{\overline{Q}_{ik}(x)}{\mu_i}\mathrm{d}x\, M_{kj}(t-x), \tag{3.11}$$

which is (3.10a) for $j \neq i$. For $j = i$,

$$\Pi_{i|i}(t) = \sum_{k\in\mathbb{X}} t\int_t^\infty \frac{\overline{Q}_{ik}(x)}{\mu_i}\mathrm{d}x + \sum_{k\in\mathbb{X}}\int_0^t \frac{\overline{Q}_{ik}(x)}{\mu_i}[x + M_{ki}(t-x)]\mathrm{d}x. \tag{3.12}$$

Grouping terms according to whether they involve any $M_{ki}(\cdot)$ or not leads to (3.10). □

Lemma 3.3 (Recurrence relations for $M_{ij}(\cdot)$)

$$M_{ij}(t) = \delta_{ji}\int_0^t \overline{H}_i(u)\mathrm{d}u + \sum_{k\in\mathbb{X}}\int_0^t Q_{ik}(\mathrm{d}u)M_{kj}(t-u) \quad (i,j\in\mathbb{X}), \tag{3.13}$$

hence $\mathbf{M}(t) = \mathrm{diag}(\tilde{\check{\mathbf{H}}}(t))\,\mathrm{diag}(\boldsymbol{\mu}) + (\mathbf{Q} * \mathbf{M})(t)$, *so*

$$M_{ij}(t) = \left(\mathbf{U} * \mathrm{diag}(\tilde{\check{\mathbf{H}}})\,\mathrm{diag}(\boldsymbol{\mu})(t)\right)_{ij} = \int_0^t \overline{H}_j(u)U_{ij}(t-u)\mathrm{d}u. \tag{3.14}$$

Proof. Equation (3.13) is established by a standard backwards decomposition. The equation is written more usefully in the form of a generalized Markov renewal equation as shown, from which the rest of the lemma follows. □

In the second term of (3.10b), substituting for $\mathbf{M}$ from (3.14) yields

$$(\tilde{\mathbf{Q}} * \mathbf{M})(t) = (\tilde{\mathbf{Q}} * \mathbf{U} * \mathrm{diag}(\tilde{\check{\mathbf{H}}}))(t). \tag{3.15}$$

Since $U_{ij}(t) \leq U_{jj}(t)$ for all $t > 0$ and all $i, j \in \mathbb{X}$, a dominated convergence argument involving $U_{ij}(t-u)/U_{jj}(t)$ in (3.14) implies that $\lim_{t\to\infty} M_{ij}(t)/U_{jj}(t) = \int_0^\infty \overline{H}_j(u)\ \mathrm{d}u = \mu_j$, and since $U_{jj}(t) \sim \lambda\check{p}_j t$ for $t \to \infty$, this implies, with (2.9), that

$$M_{ij}(t) \sim (\lambda\check{p}_j\mu_j)t = \varpi_j t \quad (\text{all } i). \tag{3.16}$$

The same arguments applied to (3.10a) show that $\Pi_{j|i}(t) \sim \varpi_j t$ for every i so that every element of $\boldsymbol{\Pi}(t) - \mathbf{e}\boldsymbol{\varpi}^T t$ in (3.8) is at most $o(t)$ for $t \to \infty$. We now find the exact asymptotics of these elements.

The components of $(\tilde{\mathbf{Q}} * \mathbf{M})(t)$ in (3.10b) can be written as

$$\begin{aligned}\left((\tilde{\mathbf{Q}} * \mathbf{M})(t)\right)_{ij} &= \sum_{k\in\mathbb{X}}\int_0^t \frac{\overline{Q}_{ik}(u)}{\mu_i}M_{kj}(t-u)\mathrm{d}u \\ &= \sum_{k\in\mathbb{X}}\int_0^t \frac{\overline{Q}_{ik}(u)}{\mu_i}\left(M_{kj}(t-u) - \varpi_j(t-u)\right)\mathrm{d}u + \sum_{k\in\mathbb{X}}\int_0^t \frac{\overline{Q}_{ik}(u)}{\mu_i}\varpi_j(t-u)\mathrm{d}u.\end{aligned} \tag{3.17}$$

The last term equals

$$\varpi_j\int_0^t \frac{\overline{H}_i(u)}{\mu_i}(t-u)\mathrm{d}u = \varpi_j\int_0^\infty \frac{\overline{H}_i(u)}{\mu_i}(t-u)_+\mathrm{d}u. \tag{3.18}$$

Consequently, from (3.10a), $\Pi_{j|i}(t) - \varpi_j t$ equals

$$\delta_{ji}\mathcal{H}_i(t) + \sum_{k\in\mathbb{X}} \int_0^t \frac{\overline{Q}_{ik}(u)}{\mu_i}\big(M_{kj}(t-u) - \varpi_j(t-u)\big)\mathrm{d}u - \varpi_j \int_0^\infty \frac{\overline{H}_i(u)}{\mu_i}\big(t-(t-u)_+\big)\mathrm{d}u, \tag{3.19}$$

and the last term equals $\varpi_j\mathcal{H}_i(t)$; so finally

$$\Pi_{j|i}(t) - \varpi_j t = (\delta_{ji} - \varpi_j)\mathcal{H}_i(t) + \sum_{k\in\mathbb{X}} \int_0^t \frac{\overline{Q}_{ik}(u)}{\mu_i}\big(M_{kj}(t-u) - \varpi_j(t-u)\big)\mathrm{d}u. \tag{3.20}$$

In vector algebra notation, writing $L(t) = t_+$, this reads

$$\mathbf{\Pi}(t) - \mathbf{e}\boldsymbol{\varpi}^T t = \operatorname{diag}(\boldsymbol{\mathcal{H}}(t)) - \boldsymbol{\mathcal{H}}(t)\boldsymbol{\varpi}^T + (\widetilde{\mathbf{Q}} * (\mathbf{M} - \mathbf{e}\boldsymbol{\varpi}^T L))(t). \tag{3.21}$$

This is not quite of the form we want; the first two terms on the right-hand side are expressed in terms of the truncated second moments of the sojourn time DFs H_i as at (2.14); it remains to consider the last term. Start by using the expression below (3.13) in writing

$$\begin{aligned}
&\big((\mathbf{M} - \mathbf{e}\boldsymbol{\varpi}^T L)(t)\big)_{ij} \\
&\quad = \big((\mathbf{U} * \operatorname{diag}(\widetilde{\mathbf{H}})\operatorname{diag}(\boldsymbol{\mu}))(t)\big)_{ij} - \varpi_j t \\
&\quad = \int_0^t U_{ij}(\mathrm{d}u)\int_0^{t-u} \overline{H}_j(v)\mathrm{d}v - \varpi_j t \\
&\quad = \int_0^t U_{ij}(u)\overline{H}_j(t-u)\mathrm{d}u - \varpi_j t \\
&\quad = \int_0^t \left[U_{ij}(u) - \frac{u}{m_{jj}}\right]\overline{H}_j(t-u)\mathrm{d}u + \frac{1}{m_{jj}}\int_0^t (t-v)\overline{H}_j(v)\mathrm{d}v - \varpi_j t \\
&\quad = \int_0^t [U_{ij}(u) - \lambda\check{p}_j u]\overline{H}_j(t-u)\mathrm{d}u + \frac{\varpi_j}{\mu_j}\int_0^\infty (t-v)_+\overline{H}_j(v)\mathrm{d}v - \varpi_j t\int_0^\infty \frac{\overline{H}_j(v)}{\mu_j}\mathrm{d}v \\
&\quad = \int_0^t [U_{ij}(u) - \lambda\check{p}_j u]\overline{H}_j(t-u)\mathrm{d}u - \varpi_j\mathcal{H}_j(t).
\end{aligned} \tag{3.22}$$

By (2.19), the integral here $\sim \mu_j\mathcal{G}(t)/m_{jj} = \varpi_j\mathcal{G}(t)$, so

$$(\mathbf{M} - \mathbf{e}\boldsymbol{\varpi}^T L)(t) \sim \mathbf{e}\boldsymbol{\varpi}^T\mathcal{G}(t) - \mathbf{e}\boldsymbol{\varpi}^T \operatorname{diag}(\boldsymbol{\mathcal{H}}(t)). \tag{3.23}$$

But in (3.21), $\widetilde{\mathbf{Q}}$ is a stochastic kernel, so the last term there has this same asymptotic behavior and

$$\mathbf{\Pi}(t) - \mathbf{e}\boldsymbol{\varpi}^T t \sim \operatorname{diag}(\boldsymbol{\mathcal{H}}(t)) - \boldsymbol{\mathcal{H}}(t)\boldsymbol{\varpi}^T + \mathbf{e}\boldsymbol{\varpi}^T(\mathbf{I}\mathcal{G}(t) - \operatorname{diag}(\boldsymbol{\mathcal{H}}(t))), \tag{3.24}$$

at least in the case of a finite state space $\mathbb{X}$. Finally then (cf. (3.8)),

$$\begin{aligned} V'(t) &\sim 2\sum_{i\in\mathbb{X}}\sum_{j\in\mathbb{X}} \nu_i\nu_j\varpi_i\Big(\delta_{ji}\mathcal{H}_i(t)+\varpi_j\big[-\mathcal{H}_i(t)+\mathcal{G}(t)-\mathcal{H}_j(t)\big]\Big) \\ &= 2\left(\sum_{i\in\mathbb{X}}\nu_i^2\varpi_i\mathcal{H}_i(t)-2\overline{\nu}\sum_i\nu_i\varpi_i\mathcal{H}_i(t)+(\overline{\nu})^2\mathcal{G}(t)\right) \\ &= 2\sum_{i\in\mathbb{X}}(\nu_i-\overline{\nu})^2\varpi_i\mathcal{H}_i(t)+2(\overline{\nu})^2\left(\mathcal{G}(t)-\sum_{i\in\mathbb{X}}\varpi_i\mathcal{H}_i(t)\right) \quad (t\to\infty). \end{aligned} \tag{3.25}$$

This establishes (3.1) except for showing that the coefficient of $(\overline{\nu})^2$ vanishes asymptotically, that is, (3.2) holds.

Recall (see above (2.16)) the function $U_i(x) = \mathrm{E}(N_{\mathrm{MRP}}[0,x] \mid \text{state } i \text{ entered at } 0)$. Just as the functions $M_{ij}(\cdot)$ satisfy generalized Markov renewal equations (see Lemma 3.3), so too do the functions $U_i(x)-\lambda x$. Using a backwards decomposition, we have

$$U_i(x) = 1+\sum_{k\in\mathbb{X}}\int_0^x Q_{ik}(\mathrm{d}u)U_k(x-u), \tag{3.26}$$

and therefore

$$\begin{aligned} U_i(x)-\lambda x &= 1-\lambda x+\sum_{k\in\mathbb{X}}\int_0^x Q_{ik}(\mathrm{d}u)\big[U_k(x-u)-\lambda(x-u)\big]+\lambda\int_0^x(x-u)H_i(\mathrm{d}u) \\ &= \sum_{k\in\mathbb{X}}\int_0^x Q_{ik}(\mathrm{d}u)\big(U_k(x-u)-\lambda(x-u)\big)+1-\lambda x+\lambda\int_0^x H_i(u)\mathrm{d}u \\ &= 1-\lambda\mu_i+\lambda\int_x^\infty\overline{H}_i(v)\mathrm{d}v+\sum_{k\in\mathbb{X}}\int_0^x Q_{ik}(\mathrm{d}u)\big(U_k(x-u)-\lambda(x-u)\big). \end{aligned} \tag{3.27}$$

Write $\mathbf{Z}(x)$ and $\mathbf{z}(x)$ for the vectors with respective components $U_i(x)-\lambda x$ and $1-\lambda\mu_i+\lambda\int_x^\infty\overline{H}_i(v)\mathrm{d}v$ ($i\in\mathbb{X}$). Then, $\mathbf{Z}=\mathbf{z}+\mathbf{Q}*\mathbf{Z}$ is a generalized Markov renewal equation, and therefore it has solution (under the condition that it is unique, which is the case when $\mathbb{X}$ is finite) $\mathbf{Z}(x)=(\mathbf{U}*\mathbf{z})(x)$. In terms of the components, this gives

$$\begin{aligned} U_i(x)-\lambda x &= \sum_{j\in\mathbb{X}}\int_0^x U_{ij}(\mathrm{d}u)\left[1-\lambda\mu_j+\lambda\int_{x-u}^\infty\overline{H}_j(v)\mathrm{d}v\right] \\ &= U_i(x)-\lambda x-\sum_{j\in\mathbb{X}}\lambda\mu_j\big[U_{ij}(x)-\lambda\check{p}_jx\big]+\lambda\sum_{j\in\mathbb{X}}\int_0^x U_{ij}(\mathrm{d}u)\int_{x-u}^\infty\overline{H}_j(v)\mathrm{d}v, \end{aligned} \tag{3.28}$$

that is,

$$\sum_{j\in\mathbb{X}}\frac{\lambda\mu_j}{m_{jj}}\big(m_{jj}U_{ij}(x)-x\big)=\lambda\sum_{j\in\mathbb{X}}\int_0^x U_{ij}(\mathrm{d}u)\int_{x-u}^\infty\overline{H}_j(v)\mathrm{d}v. \tag{3.29}$$

Now, our MRP is LRD, so by (2.10) and Sgibnev's [7] solidarity result quoted at (2.19), the left-hand side here $\sim\sum_j\lambda\varpi_j\mathcal{G}(x)=\lambda\mathcal{G}(x)$. For the right-hand side, we can apply

the asymptotic convergence lemma in Sgibnev [3] (see Appendix A), because $U_{ij}(x) \sim x/m_{jj} = \lambda \check{p}_j x$ $(x \to \infty)$, to deduce that the right-hand side of (3.29) $\sim \lambda^2 \sum_{j\in\mathbb{X}} \check{p}_j \mu_j \mathcal{H}_j(x)$ $= \lambda \sum_{j\in\mathbb{X}} \varpi_j \mathcal{H}_j(x)$; so (3.2) holds. □

In the setting in Daley [1] for the case of an alternating renewal process, we should have in our general notation above that $\nu_1 = 1$ for the ON state, 1, say, and $\nu_0 = 0$ for the OFF state, $\overline{\nu} = \varpi = \varpi_1 = \mathrm{E}(X_1)/\mathrm{E}(Y)$, where $Y = X_1 + X_0$ is a generic cycle time, $\check{p}_0 = \check{p}_1 = 1/2$, and $\varpi_0 = 1 - \varpi$. An ARP can be studied via cycle times (with generic duration Y), with return time distribution $G(x) = \Pr\{Y \le x\}$ for which $\mathcal{G}(\cdot)$ emerges naturally for (2.19) and (3.2). The right-hand side of (3.1) equals $(1-\varpi)^2 \varpi \mathcal{H}_1(t) + \varpi^2(1-\varpi)\mathcal{H}_0(t)$, so our theorem above is consistent with Daley [1].

4. Discussion

Our proof of the asymptotic relation at (2.21) for the behavior of $\mathrm{var}\, N_{\mathrm{MRP}}(0,t]$ when the MRP is LRD depends on Sgibnev's [7] solidarity result and, lacking any uniform convergence result over the state space $\mathbb{X}$, it is confined to the case that $\mathbb{X}$ is finite. Whether or not a relation like (2.21) persists in the countable case is not clear. We indicate one difficulty.

Consider a realization of our MRP. Let a "tour" consist of the successive states $\{j_n\}$ visited on a path starting from $j_0 = k$ until a first return to k, consisting of say, N_{tour} transitions in all, so $j_{N_{\mathrm{tour}}} = k$ and $j_n \ne k$ for $n = 1, \ldots, N_{\mathrm{tour}} - 1$; for such a path, represent the first return time Y_{kk}, with DF G_{kk}, and in self-evident notation, as

$$Y_{kk} = \sum_{\{j_n\}} X^{\mathrm{tour}}_{j_n, j_{n+1}} = \sum_{n=0}^{N_{\mathrm{tour}}-1} X^{\mathrm{tour}}_{j_n, j_{n+1}}. \tag{4.1}$$

Then, Y_{kk} has infinite second moment if and only if either (or both) of some X_{ij} and N_{tour} has infinite second moment. For a Markov chain in discrete time, only the latter is possible (because whenever $p_{ij} > 0, X_{ij} = 1$ a.s.). Trivially, a Markov chain in discrete time is also a Markov renewal process, and thus, in a LRD MRP with all holding times being equal to 1, say, a relation like (3.2) would be impossible because the left-hand side would be infinite but the right-hand side would be finite.

Appendices

A. An asymptotic convergence lemma

The result given below is the essence of Sgibnev [3, Theorem 4], used to establish the asymptotic behavior of the difference between a renewal function $U(t)$ and its asymptote λt when a generic lifetime r.v. has infinite second moment. Sgibnev's proof assumes that $U(\cdot)$ is a renewal function, but this is not needed in our proof below.

Lemma A.1. *Let the nonnegative function $z(x)$ $(x > 0)$ be monotonic decreasing and such that $L(t) \equiv \int_0^t z(u)\mathrm{d}u \to \infty$ for $t \to \infty$. Let the monotonic increasing nonnegative function $U(t)$ have uniformly bounded increments $U(x+1) - U(x) \le K < \infty$ (all $x > 0$) and let it be*

asymptotically linear, so that $U(t) \sim \lambda t$ $(t \to \infty)$ *for some finite positive* λ. *Then,*

$$L_1(t) \equiv (U * z)(t) \equiv \int_0^t z(t-u)U(\mathrm{d}u) \sim \lambda L(t) \quad (t \longrightarrow \infty). \tag{A.1}$$

Proof. Given $\epsilon > 0$, the asymptotic linearity of $U(\cdot)$ implies that there exists finite positive t_ϵ such that

$$\left| U(t) - \lambda t \right| \le \epsilon t \quad (\text{all } t \ge t_\epsilon). \tag{A.2}$$

Write

$$\begin{aligned} L_1(t) &= \int_0^t \left[z(t) + \int_{t-u}^t \left| \mathrm{d}z(v) \right| \right] U(\mathrm{d}u) \\ &= \left[U(t) - U(0+) \right] z(t) + \int_0^t \left[U(t) - U(t-v) \right] \left| \mathrm{d}z(v) \right| \\ &= \left[U(t) - U(0+) \right] z(t) + \left(\int_0^{t_\epsilon} + \int_{t_\epsilon}^t \right) \int_0^t \left[U(t) - U(t-v) \right] \left| \mathrm{d}z(v) \right| \\ &= \left[U(t) - U(0+) \right] z(t) + A_\epsilon(t) + \int_{t_\epsilon}^t \left[U(t) - U(t-v) \right] \left| \mathrm{d}z(v) \right|, \end{aligned} \tag{A.3}$$

where $0 < A_\epsilon(t) \le [z(0) - z(t_\epsilon)] K t_\epsilon$, uniformly in t. Then,

$$L_1(t) - A_\epsilon(t) = U(t)z(t_\epsilon) - U(0+)z(t) - \int_{t_\epsilon}^t U(t-v) \left| \mathrm{d}z(v) \right|. \tag{A.4}$$

For $t > 2t_\epsilon$, this integral equals $(\int_{t_\epsilon}^{t-t_\epsilon} + \int_{t-t_\epsilon}^t) U(t-v)|\mathrm{d}z(v)|$, in which the latter integral, $B_\epsilon(t)$, say, satisfies

$$0 \le B_\epsilon(t) \equiv \int_{t-t_\epsilon}^t U(t-v) \left| \mathrm{d}z(v) \right| \le U(t_\epsilon) z(t-t_\epsilon) \le (\lambda + \epsilon) t_\epsilon z(t-t_\epsilon), \tag{A.5}$$

which for a given ϵ is uniformly bounded, independently of t. The integral that remains equals $\int_{t_\epsilon}^{t-t_\epsilon} U(t-v)|\mathrm{d}z(v)|$ which by (A.1) is bounded above and below by

$$\begin{aligned} (\lambda \pm \epsilon) \int_{t_\epsilon}^{t-t_\epsilon} (t-v) \left| \mathrm{d}z(v) \right| &= (\lambda \pm \epsilon) \int_{t_\epsilon}^{t-t_\epsilon} \left| \mathrm{d}z(v) \right| \int_0^{t-v} \mathrm{d}w \\ &= (\lambda \pm \epsilon) \int_0^{t-t_\epsilon} \mathrm{d}w \int_{t_\epsilon}^{\min(t-w,\, t-t_\epsilon)} \left| \mathrm{d}z(v) \right| \\ &= (\lambda \pm \epsilon) \int_0^{t-t_\epsilon} \left[z(t_\epsilon) - z\left(\min\left(t-w, t-t_\epsilon \right) \right) \right] \mathrm{d}w \\ &= (\lambda \pm \epsilon) \int_{t_\epsilon}^{t} \left[z(t_\epsilon) - z\left(\min\left(w, t-t_\epsilon \right) \right) \right] \mathrm{d}w. \end{aligned} \tag{A.6}$$

Using the upper bound, we can therefore write

$$\begin{aligned} L_1(t) - A_\epsilon(t) - B_\epsilon(t) + U(0+)z(t) & \\ < [U(t) - (\lambda+\epsilon)(t-t_\epsilon)]z(t_\epsilon) + (\lambda+\epsilon)\int_{t_\epsilon}^{t} z(\min(w, t-t_\epsilon))\,\mathrm{d}w & \\ \le (\lambda+\epsilon)t_\epsilon z(t_\epsilon) + (\lambda+\epsilon)\int_{t_\epsilon}^{t} z(\min(w, t-t_\epsilon))\,\mathrm{d}w, & \end{aligned} \tag{A.7}$$

in which the second inequality comes from (A.1) because $t > t_\epsilon$. Divide each extreme of this inequality by $L(t)$, and observe that in the limit $t \to \infty$, the only term on the left-hand side that does not vanish is $L_1(t)/L(t)$, while the right-hand side (after division) converges to $\lambda + \epsilon$. It then follows that, because ϵ is arbitrary, $\limsup_{t\to\infty} L_1(t)/L(t) \le \lambda$.

Using the lower bound at (A.6) leads instead to

$$L_1(t) - A_\epsilon(t) - B_\epsilon(t) + U(0+)z(t) \ge (\lambda-\epsilon)t_\epsilon z(t_\epsilon) + (\lambda-\epsilon)\int_{t_\epsilon}^{t} z(\min(w, t-t_\epsilon))\,\mathrm{d}w, \tag{A.8}$$

and a similar argument as in using the upper bound gives $\liminf_{t\to\infty} L_1(t)/L(t) \ge \lambda$. □

B. Subadditivity of the renewal function $U_{\mathrm{MRP}}(\cdot)$

Lemma B.1. *The renewal function $U_{\mathrm{MRP}}(\cdot)$ defined on jump epochs of a stationary* MRP *is subadditive.*

Proof. For a stationary MRP on state space $\mathbb{X}$, recall the Palm expectations (see around (2.16) and (3.26) above)

$$U_i(x) = \mathrm{E}\big(N_{\mathrm{MRP}}[0,x] \mid \text{state } i \text{ entered at } 0\big) = \sum_j U_{ij}(x), \tag{B.1}$$

$$U_{\mathrm{MRP}}(x) = \sum_i \check{p}_i U_i(x). \tag{B.2}$$

For a stationary MRP, the stationary distribution $\{\check{p}_i\}$ for the embedded jump process $\{J_n\}$ satisfies both $\check{p}_j = \sum_{i\in\mathbb{X}} \check{p}_i p_{ij}$ and the equation for the state probability at the epoch of the first jump after any fixed time interval thereafter, that is, the semi-Markov process $J(t)$ satisfies $\check{p}_k = \Pr\{\text{first jump of } J(x+t) \text{ in } t>0 \text{ is to } k \mid J(\cdot) \text{ has jump at } 0\}$, namely,

$$\check{p}_k = \sum_i \check{p}_i \sum_j \int_0^x U_{ij}(\mathrm{d}u)\overline{H}_j(x-u)\int_0^\infty \frac{Q_{jk}(x-u+\mathrm{d}z)}{\overline{H}_j(x-u)} = \sum_i \check{p}_i \sum_j \int_0^x U_{ij}(\mathrm{d}u)\int_{x-u}^\infty Q_{jk}(\mathrm{d}z). \tag{B.3}$$

Now,

$$\begin{aligned} U_i(x+y) - U_i(x) &= \mathrm{E}\big(N(x,x+y] \mid \text{state } i \text{ entered at } 0\big) \\ &= \sum_j \sum_k \int_0^x U_{ij}(\mathrm{d}u) \int_{x-u}^{x+y-u} Q_{jk}(\mathrm{d}z) U_k(x+y-u-z) \\ &\le \sum_j \sum_k \int_0^x U_{ij}(\mathrm{d}u) \int_{x-u}^{x+y-u} Q_{jk}(\mathrm{d}z) U_k(y), \end{aligned} \tag{B.4}$$

because every $U_i(\cdot)$ is nondecreasing and every $Q_{jk}(\cdot)$ is a measure, and then, again because every $Q_{jk}(\cdot)$ is a measure and using (B.2), $U_{\mathrm{MRP}}(x+y) - U_{\mathrm{MRP}}(x)$ equals

$$\begin{aligned} \sum_i \check{p}_i \big[U_i(x+y) - U_i(x)\big] &\le \sum_i \check{p}_i \sum_j \sum_k \int_0^x U_{ij}(\mathrm{d}u) \int_{x-u}^{x+y-u} Q_{jk}(\mathrm{d}z) U_k(y) \\ &\le \sum_k U_k(y) \sum_i \check{p}_i \sum_j \int_0^x U_{ij}(\mathrm{d}u) \int_{x-u}^{\infty} Q_{jk}(\mathrm{d}z) \\ &= \sum_k U_k(y) \check{p}_k = U_{\mathrm{MRP}}(y). \end{aligned} \tag{B.5}$$

□

Acknowledgments

D. J. Daley's work was done using facilities as an Honorary Visiting Fellow in the Centre for Mathematics and its Applications at the ANU and as a holder of a Marie Curie Transfer of Knowledge Fellowship of the European Community's Sixth Framework Programme (Programme HANAP under Contract no. MTKD-CT-2004-13389 at the University of Wrocław). T. Rolski also acknowledges partial support from the Transfer of Knowledge Fellowship.

References

[1] D. J. Daley, "Long-range dependence in a Cox process directed by an alternating renewal process," submitted for publication.

[2] J. J. Hunter, "On the moments of Markov renewal processes," *Advances in Applied Probability*, vol. 1, pp. 188–210, 1969.

[3] M. S. Sgibnev, "On the renewal theorem in the case of infinite variance," *Sibirskiĭ Matematicheskiĭ Zhurnal*, vol. 22, no. 5, pp. 178–189, 224, 1981, Translation in *Siberian Mathematical Journal*, vol. 22, no. 5, pp. 787–796, 1981.

[4] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, Probability and Its Applications, Springer, New York, NY, USA, 2nd edition, 2003.

[5] E. Cinlar, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1975.

[6] V. G. Kulkarni, *Modeling and Analysis of Stochastic Systems*, Texts in Statistical Science Series, Chapman & Hall, London, UK, 1995.
[7] M. S. Sgibnev, "An infinite variance solidarity theorem for Markov renewal functions," *Journal of Applied Probability*, vol. 33, no. 2, pp. 434–438, 1996.

D. J. Daley: Centre for Mathematics and its Applications, The Australian National University, ACT 0200, Australia
Email address: daryl@maths.anu.edu.au

T. Rolski: Mathematical Institute, University of Wrocław, pl. Grunwaldzki 2/4, 50384 Wrocław, Poland
Email address: rolski@math.uni.wroc.pl

R. Vesilo: Department of Electronics, Macquarie University, North Ryde, NSW 2109, Australia
Email address: rein@ics.mq.edu.au

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 94515, 16 pages
doi:10.1155/2007/94515

Review Article

The Geometry of Statistical Efficiency and Matrix Statistics

K. Gustafson

Received 23 March 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

We will place certain parts of the theory of statistical efficiency into the author's operator trigonometry (1967), thereby providing new geometrical understanding of statistical efficiency. Important earlier results of Bloomfield and Watson, Durbin and Kendall, Rao and Rao, will be so interpreted. For example, worse case relative least squares efficiency corresponds to and is achieved by the maximal turning antieigenvectors of the covariance matrix. Some little-known historical perspectives will also be exposed. The overall view will be emphasized.

1. Introduction and Summary

Recently, Gustafson [1–3] was able to connect the theory of statistical efficiency to his operator trigonometry, which is a theory of antieigenvalues and antieigenvectors which he initiated in 1967 for a different purpose. The aim of this paper is to go beyond the [1–3] papers to provide a more overall view of these results and their implications. We will also use this opportunity to expose some historical perspectives that have been generally forgotten or which are otherwise little known.

The outline and summary of this paper are as follows. In Section 2, we obtain the statistical efficiency ratio of BLUE to OLSE covariance in terms of the geometry provided by the author's 1967 operator trigonometry. To fix ideas here, this result can be described as giving to the [4, 5] Bloomfield-Watson-Knott solution of the Durbin conjecture, that is, its geometrical meaning. In Section 3, we provide the reader with the basics of the operator trigonometry. This brief but adequate bibliographical citation is given from which further detail may be obtained. To augment the reader's intuition and appreciation for the operator trigonometry, and because we are writing here for an audience of statisticians,

in Section 4 we recall the origin of the operator trigonometry, that is, operator semigroups, with application to Markov processes. This problem essentially induced both of the key elements of the operator trigonometry. In Section 5, we return to the topic of statistical efficiency and provide some lesser-known historical background. This is augmented in Section 6 with a look at an interesting early paper of von Neumann. From the latter, we are able to make here an interesting new connection of statistical efficiency to partial differential equations. In Section 7, we develop the interesting and useful distinction between what we call inefficiency vectors versus antieigenvectors. Both satisfy related variational equations. Through this link, we may then relate in Section 8 certain considerations of canonical correlations as treated in [6] by Rao-Rao to the general mathematical setting of statistical efficiency and operator trigonometry—all three are now combined. Section 9 concludes the paper with some further discussion of the historical view of statistical efficiency as viewed through the context of this paper.

2. The geometry of statistical efficiency

What follows was shown in Gustafson [1–3]. Considering the general linear model, we follow Wang and Chow [7] for convenience:

$$y = X\beta + e, \tag{2.1}$$

where y is an n-vector composed of n random samplings of a random variable Y, X is an $n \times p$ matrix usually called the design or model matrix, β is a z-vector composed of p unknown nonrandom parameters to be estimated, and e is an n-vector of random errors incurred in observing y. The elements x_{ij} of X may have different statistical meanings depending on the application. We assume for simplicity that the error or noise e has expected value 0 and has covariance matrix $\sigma^2 V$, where V is a symmetric positive definite $n \times n$ matrix. Of course one can generalize to singular V and to unknown V and so on by using singular value decomposition and generalized inverses throughout to develop a more general theory, but we shall not do so here. We absorb the σ^2 or nonidentical row-dependent variances into V. A customary assumption on X is that $n \geqq 2p$, that is, one often thinks of X as having only a few (regressor) columns available. In fact, it is useful to often think of p as just 1 or 2. Generally, it seems to be usually assumed that the columns of X are linearly independent, and often it is assumed that those columns form an orthonormal set $X^*X = I_p$.

The relative statistical efficiency for comparing an ordinary least-squares estimator OLSE $\hat{\beta}$ and the best linear unbiased estimator BLUE β^* is defined as

$$\mathrm{RE}\,(\hat{\beta}) = \frac{|\,\mathrm{Cov}\,(\beta^*)\,|}{|\,\mathrm{Cov}\,(\hat{\beta})\,|} = \frac{1}{|\,X^*VX\,|\,|\,X^*V^{-1}X\,|}, \tag{2.2}$$

where $|\cdot|$ denotes determinant. A fundamental lower bound for statistical efficiency is

$$\mathrm{RE}\,(\hat{\beta}) \geqq \prod_{i=1}^{p} \frac{4\lambda_i\lambda_{n-i+1}}{(\lambda_i + \lambda_{n-i+1})^2}, \tag{2.3}$$

where $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n > 0$ are the eigenvalues of V. This lower bound is sometimes called the Bloomfield-Watson-Knott lower bound; see Section 5 for more historical particulars. In Gustafson [1], the following new and geometrical interpretation of the lower bound (2.3) was obtained. More specifics of the operator trigonometry, antieigenvalues, and antieigenvectors will be given in Section 3. The essential meaning of Theorem 2.1 is that the linear model's statistical efficiency is limited by the maximal turning angles of the covariance matrix V.

Theorem 2.1. *For the general linear model (2.1) with SPD covariance matrix $V > 0$, for $p = 1$, the geometrical meaning of the relative efficiency (2.2) of an OLSE estimator $\hat{\beta}$ against BLUE β^* is*

$$\mathrm{RE}\,(\hat{\beta}) \geqq \cos^2 \phi(V), \tag{2.4}$$

where $\phi(V)$ is the operator angle of V. For $p \leqq n/2$, the geometrical meaning is

$$\mathrm{RE}\,(\hat{\beta}) \geqq \prod_{i=1}^{p} \cos^2 \phi_i(V) = \prod_{i=1}^{p} \mu_i^2(V), \tag{2.5}$$

where the $\phi_i(V)$ are the successive decreasing critical turning angles of V, that is, corresponding to the higher antieigenvalues $\mu_i(V)$. The lower bound (2.3), as expressed geometrically in (2.4), is attained for $p = 1$ by either of the two first antieigenvectors of V:

$$x_{\pm} = \pm \left(\frac{\lambda_1}{\lambda_1 + \lambda_n} \right)^{1/2} x_n + \left(\frac{\lambda_n}{\lambda_1 + \lambda_n} \right)^{1/2} x_1. \tag{2.6}$$

For $p \leqq n/2$, the lower bound (2.3), as expressed geometrically in (2.5), is attained as

$$\prod_{i=1}^{p} \frac{\langle V x_{\pm}^i, x_{\pm}^i \rangle}{\| V x_{\pm}^i \| \, \| x_{\pm}^i \|}, \tag{2.7}$$

where $x_{\pm}^i$ denotes the ith higher antieigenvectors of V given by

$$x_{\pm}^i = \pm \left(\frac{\lambda_i}{\lambda_i + \lambda_{n-i+1}} \right)^{1/2} x_{n-i+1} + \left(\frac{\lambda_{n-i+1}}{\lambda_i + \lambda_{n-i+1}} \right)^{1/2} x_i. \tag{2.8}$$

In (2.6) and (2.8), x_i denotes the normalized ith eigenvector of V corresponding to the eigenvalue λ_i.

We remark that Theorem 2.1 follows rather immediately from (2.3) once one recognizes that the factors on the right-hand side of (2.3) are exactly the cosines of the critical turning angles of V. This connection was first pointed out in Gustafson [1]. In Gustafson [3], some related trace statistical efficiency bounds were also given an operator trigonometric interpretation.

3. The operator trigonometry: antieigenvalues and angles

For simplicity, let A be an $n \times n$ symmetric positive definite (SPD) matrix with eigenvalues $0 < \lambda_n \leqq \lambda_2 \leqq \cdots \leqq \lambda_1$. Then, the first antieigenvalue of A was defined to be

$$\mu_1 = \min_{x \neq 0} \frac{\langle Ax, x \rangle}{\|Ax\| \|x\|} \tag{3.1}$$

and a related entity

$$\nu_1 = \min_{\epsilon > 0} \|\epsilon A - I\| \tag{3.2}$$

also came naturally into the theory. How that came about will be described in Section 4. Because of the need for both μ_1 and ν_1, the author felt that ν_1 must also be trigonometric. Indeed it is. Gustafson [8] established the following key minmax result.

Theorem 3.1. *Given a strongly accretive operator B on a Hilbert space, then*

$$\sup_{\|x\| \leqq 1} \inf_{\epsilon} \|(\epsilon B - I)x\|^2 = \inf_{\epsilon > 0} \sup_{\|x\| \leqq 1} \|(\epsilon B - I)x\|^2. \tag{3.3}$$

In particular for an SPD matrix A, one has

$$\mu_1^2 + \nu_1^2 = 1. \tag{3.4}$$

Originally, the minimum (3.1) was called $\cos A$ for obvious reasons, and after Theorem 3.1 was realized, the minimum (3.2) could be called $\sin A$. This is an essential critical point to understand about the operator trigonometry. One must have both a $\sin A$ and a $\cos A$ if one wants some kind of trigonometry. Later, the better notations $\cos \phi(A)$ and $\sin \phi(A)$ were introduced so as to avoid any unwarranted confusion with cosine and sine functions in an operator's functional calculus. Moreover, it is clear that A does have a meaningful operator angle $\phi(A)$ defined equivalently by either (3.1) or (3.2). This operator maximal turning angle $\phi(A)$ is a real tangible angle in n-dimensional Euclidean space. It is attained by A's two (here normalized to norm 1) antieigenvectors:

$$x_{\pm} = \pm \left(\frac{\lambda_1}{\lambda_1 + \lambda_n} \right)^{1/2} x_n + \left(\frac{\lambda_n}{\lambda_1 + \lambda_n} \right)^{1/2} x_1, \tag{3.5}$$

where x_1 and x_n are any (normalized) eigenvectors from the eigenspaces corresponding to λ_1 and λ_n, respectively. The antieigenvectors are those that are turned to the maximal amount when operated on by A, and they thus attain the minimums in (3.1) and (3.2).

A more general theory has been developed, and for that and further history and other ramifications of the operator trigonometry and antieigenvalue-antieigenvector theory, we just refer to the books of Gustafson [9], Gustafson and Rao [10], and the surveys of Gustafson [11, 2]. One more basic ingredient which should be mentioned here is the Euler equation

$$2\|Ax\|^2 \|x\|^2 (\mathrm{Re}\, A)x - \|x\|^2 \,\mathrm{Re}\langle Ax, x \rangle A^* Ax - \|Ax\|^2 \,\mathrm{Re}\langle Ax, x \rangle x = 0 \tag{3.6}$$

which is satisfied by the antieigenvectors of A, for any strongly accretive matrix A. When A is Hermitian or normal, this Euler equation is satisfied not only by the first antieigenvectors $x_{\pm}$ of A, but also by all eigenvectors of A. Thus, the expression (3.1) generalizes the usual Rayleigh quotient theory for SPD matrices A to now include antieigenvectors $x_{\pm}$, which minimize it, and all eigenvectors, which maximize it.

Higher antieigenvalues $\mu_i(A)$ and their corresponding higher antieigenvectors were originally defined, Gustafson [12], in a way analogous to that for higher eigenvalues in the Rayleigh-Ritz theory. That is okay for some applications but later, Gustafson [13], the author formulated a better general combinatorially based theory in which the higher antieigenvectors are those stated in (2.8). To each such pair, we obtain via (3.1) a sequence of decreasing-in-size maximal interior operator turning angles $\phi_i(V)$ as indicated in (2.5) (see Gustafson [14] for more details).

It is interesting to note that antieigenvectors, including the higher ones, always occur in pairs. In retrospect, this is a hint that there are connections of that fact to the fact that the usual analyses of statistical efficiency also often end up at a point where one needs to consider certain pairs of vectors. We will return to this point in Section 7.

4. The origin of the operator trigonometry: Markov processes

The author's creation of the operator trigonometry in 1967 came out of an abstract operator theoretic question. Let X be a Banach space and let A be the densely defined infinitesimal generator of a contraction semigroup e^{tA} on X. In other words, consider the initial value problem

$$\begin{aligned} \frac{du}{dt} &= Au(t), \quad t > 0, \\ u(0) &= u_0 \quad \text{given} \end{aligned} \tag{4.1}$$

and its solution $u(t) = U_t u_0 \equiv e^{tA} u_0$ with the contraction property $\|U_t\| \leqq 1$. So one can think of the heat equation, the Schrödinger equation, or a linear Markov process. In fact, it was a question of introducing a stochastic time change into a Markov process e^{tA}, which led to the following question. When can one multiplicatively perturb A to BA and still retain the contraction semigroup infinitesimal generator property in BA? The result was as follows, Gustafson [15], stated here in now familiar terms.

Theorem 4.1. *Let A be the infinitesimal generator of a contraction semigroup on a Banach space X. Then, BA is still an infinitesimal generator of a contraction semigroup if B is a strongly accretive operator satisfying*

$$\sin\phi(B) \leqq \cos\phi(A). \tag{4.2}$$

But the proof of Theorem 4.1 in Gustafson [15] did not originally involve any entity $\sin\phi(B)$ because such entities did not exist yet. The proof instead needed $\|\epsilon B - I\| \leqq \mu_1(A)$ for some positive ϵ. By the minmax Theorem 3.1, this requirement becomes (4.2).

Therefore, to better understand these now trigonometric entities, the author quickly computed them for some operator classes. For the most definitive and most useful class A a SPD matrix with eigenvalues $0 < \lambda_n \leqq \lambda_{n-1} \leqq \cdots \leqq \lambda_1$, one has

$$\cos\phi(A) = \frac{2\sqrt{\lambda_1\lambda_n}}{\lambda_1+\lambda_n}, \qquad \sin\phi(A) = \frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}, \tag{4.3}$$

which are attained by the antieigenvector pair (3.5).

5. Some history of statistical efficiency

Although the theory of statistical efficiency is well documented in a number of books, and in the 1970's papers of Bloomfield-Watson [4], Knott [5], and others, in the writing of Gustafson [1] this author wanted to get some original feel of the history for himself. For one thing, it was wondered where the "Durbin conjecture" which led to the lower bound (2.3) was explicitly stated. This was not found. But some related historical perspectives were put into Gustafson [1, 3, Section 4]. There, for example, one finds a description of precursor work of Plackett [16], Aitken [17], and Durbin and Kendall [18]. The latter paper is quite explicitly geometrical, although, not operator theoretically. Plackett [16] takes the fundamental notions all the way back to Gauss.

A second more recent historical look has revealed some further interesting historical perspectives. In particular, the Watson [19] paper is probably the explicit source of the "Durbin conjecture." In fact, one finds it there, (3.5), with a footnote crediting it to J. Durbin. However, Watson [20] admits a flaw in his [19] argument and thus the verification of the Durbin conjecture remained an open problem until 1975.

Going back further to the two papers of Durbin-Watson [21, 22], one finds a more classical statistical analysis of (2.1) from the point of view of χ^2 distributions, which is of course of central importance to the theory of analysis of variance. In particular, the second paper is largely devoted to a study of the statistic

$$d = \frac{\sum(\Delta z)^2}{\sum z^2} \tag{5.1}$$

which is to be used for testing for serial correlation within error in terms of a regression model. We go back to the first paper (see [21, page 409]) and find that the principal issue is "the problem of testing the errors for independence forms the subject of this paper and its successor." Attribution is made to earlier papers by Anderson [23] and Anderson and Anderson [24], where possible serial correlations in least-squares residuals from Fourier regressions were tested. In Watson [20], which is a quite useful paper historically, study of the efficiency of least squares is said to follow that of Grenander [25] and Grenander and Rosenblatt [26]. In fact, we have traced efficiency explicitly back to Fisher [27]. See our further discussion in Section 9.

6. The von Neumann connection and a new connection to partial differential equations

In our historical search, tracing back through the two papers of Durbin and Watson [21, 22], one comes upon the interesting $n \times n$ matrix

$$A = \begin{bmatrix} 1 & -1 & 0 & \cdots & & 0 \\ -1 & 2 & -1 & \cdots & & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & & & \cdots & & 0 \\ & & & -1 & 2 & -1 \\ 0 & & & & -1 & 1 \end{bmatrix}. \tag{6.1}$$

It is stated there that this results from the statistic to be used to test for serial correlation

$$d = \frac{\sum(\Delta z)^2}{\sum z^2} = \frac{\langle Az, z\rangle}{\sum z^2}, \tag{6.2}$$

where z is the residual from linear regression. It was shown [22] that the mean and variance of the statistic d are given by

$$\begin{aligned} E(d) &= \frac{P}{n-k'-1}, \\ \operatorname{var}(d) &= \frac{2[Q-PE(d)]}{(n-k'-1)(n-k'+1)}, \end{aligned} \tag{6.3}$$

where

$$\begin{aligned} P &= \operatorname{tr} A - \operatorname{tr}\left(X'AX(X'X)^{-1}\right), \\ Q &= \operatorname{tr} A^2 - 2\operatorname{tr}\left(X'A^2X(X'X)^{-1}\right) + \operatorname{tr}\left((X'AX(X'X)^{-1}\right)^2, \end{aligned} \tag{6.4}$$

where k' is the number of columns of the matrix of observations of the independent variables

$$\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k'1} \\ \vdots & & & \\ x_{1n} & x_{2n} & \cdots & c_{k'n} \end{bmatrix}. \tag{6.5}$$

One wonders, or at least this author wonders, about how A came about. It turns out that this query became quite interesting as we now explain.

A more careful reading of Durbin and Watson [21] leads to a paper of von Neumann [28], and one cannot resist looking at it. As it is well known, von Neumann was a

polymath and this paper is not an exception. An in-depth study of the statistic

$$\eta = \frac{\delta^2}{s^2} \tag{6.6}$$

is carried out, where s^2 is the sample variance of a normally distributed random variable and $\delta^2 = \sum_{\mu=1}^{n-1}(x_{\mu+1} - x_\mu)^2/(n-1)$ is the mean square successive difference—the goal being to determine the independence or trend dependence of the observations $x_1,\ldots,x_n$. Thus, we find this paper to be an early and key precedent to all the work done by Durbin, Watson, and others in the period of 1950–1975.

Von Neumann's analysis is extensive and he obtains a number of theoretical results which, if we might paraphrase see Durbin and Watson [21, page 418], are more or less beyond use by conventional statisticians. However, both Durbin-Watson papers [21, 22] go ahead and use the matrix A to illustrate their theory. So one looks further into von Neumann's paper to better understand the origin of the matrix A of (6.1). One finds there (see [28, page 367]) the statement: "the reasons for the study of the distribution of the mean square successive difference δ^2, in itself as well as in its relationship to the variance s^2, have been set forth in a previous publication, to which the reader is referred." However, it is made clear that comparing observed values of the statistic η will be used to determine "whether the observations $x_1,\ldots,x_n$ are independent or whether a trend exists."

Since curiosity knows no bounds, we pushed the historical trace back to the previous publication of von Neumann, Kent, Bellison, and Hart [29]. The answer to our curiosity about why von Neumann became involved with this statistical regression problem is found there. To quote (see [29, page 154]), "the usefulness of the differences between successive observations only appears to be realized first by ballisticians, who faced the problem of minimizing effects due to wind variation, heat, and wear in measuring the dispersion of the distance traveled by shell." The 4-author paper originated from the Aberdeen Ballistic Research Laboratory, where von Neumann was consulting.

Returning to his analysis in von Neumann [28], we find that he begins with a now more or less classical multivariate analysis of normally distributed variables. By diagonalization, a quadratic form $\sum A_\mu x'_\mu$ is obtained where the A_μ, $\mu = 1,\ldots,n$, are the eigenvalues of the form $(n-1)\delta^2$. The smallest eigenvalue $A_n = 0$ is found, with eigenvector $x_0 = (1,\ldots,1)/\sqrt{n}$. A further analysis, using an interesting technique of assuming the $x'_1,\ldots,x'_{n-1}$ to be uniformly distributed over an $n-1$ unit sphere, shows that the statistic η of (6.5) is then distributed according to

$$\eta = \frac{n}{n-1}\sum_{\mu=1}^{n-1} A_\mu x_\mu^2. \tag{6.7}$$

Thus, the sought eigenvalues A_μ, $\mu = 1,\ldots,n$, are the eigenvalues of the quadratic form $(n-1)\delta^2$, which is then written as

$$(n-1)\delta^2 = x_1^2 + 2\sum_{\mu=2}^{n-1} x_\mu^2 + x_n^2 - 2\sum_{\mu=1}^{n-1} x_\mu x_{\mu+1}. \tag{6.8}$$

The matrix of this form is (6.1) and it is that matrix which is also borrowed and used in Durbin and Watson [21, 22]. Used as well are the eigenvalues

$$A_k = 4\sin^2\left(\frac{k\pi}{2n}\right), \quad k = 1,\ldots,n-1 \tag{6.9}$$

which von Neumann computes from the determinant of A.

Commentary. When we first saw the matrix A in Durbin and Watson [21, 22], our take was completely different. As this author is a specialist in partial differential equations, for example, see Gustafson [30], we immediately see the matrix A in (6.1) as the discretized Poisson-Neumann boundary value problem

$$\begin{aligned} -\frac{d^2u(x)}{dx^2} &= f(x), \quad 0 < x < 1, \\ \frac{du}{dx} &= 0 \quad \text{at } x = 0,1. \end{aligned} \tag{6.10}$$

In saying this, I am disregarding the exact interval and discrete Δx sizes.

This new connection between statistical efficiency and partial differential equations will be further explored elsewhere, especially as it will no doubt generalize to Dirichlet, Neumann, and Robin boundary value problems for the Laplacian operator $-\Delta = \sum \partial^2 u/\partial x^2$ in higher dimensions. The reverse implications for a more general context of statistical efficiency could also be interesting. Moreover, we have already worked out the complete operator trigonometry for the two-dimensional discretized Dirichlet problem in Gustafson [31].

We also comment in passing that a similar ballistic problem—that of control of rocket flight—was the motivating application in Japan during the Second World War that led Ito to develop his stochastic calculus, which is now so important in the theory of financial derivatives and elsewhere.

7. The inefficiency equation and the Euler equation

Following Wang and Chow [7], among others, one may apply a Lagrangian method to

$$\mathrm{RE}\left(\hat{\beta}\right)^{-1} = |XV^{-1}X|\,|X'VX| \tag{7.1}$$

with the general case having been reduced to that of $X'X = I_p$. By a differentiation of $F(x,\lambda) = \ln|X'V^{-1}X| + \ln|X'VX| - 2\,\mathrm{tr}(X'X\Lambda)$ and subsequent minimization, the relation

$$X'X(\Lambda + \Lambda') = \Lambda + \Lambda' = 2I_p \tag{7.2}$$

is obtained. Here, Λ is a $p \times p$ upper triangular matrix which is the Lagrange multiplier with respect to the constraint $X'X = I_p$. From this and further work including the simultaneous diagonalization of $X'V^2X$, $X'VX$, and $X'V^{-1}X$, one arrives at the result

$$\mathrm{RE}\left(\widehat{\beta}\right)^{-1} = \prod_{i=1}^{p} x_i' V x_i x_i' V^{-1} x_i, \tag{7.3}$$

where X is now the $n \times p$ column matrix $X = [(x_1) \cdots (x_p)]$ whose columns go into the expression (7.3). The Lagrange multiplier minimization leading to (7.3) has also now yielded the equation for the x_i:

$$\frac{V^2 x_i}{x_i' V x_i} + \frac{x_i}{x_i' V^{-1} x_i} = 2Vx_i, \quad i = 1, \ldots, p. \tag{7.4}$$

Clearly, the span $\{x_i, Vx_i\}$ is a two- (or one-)dimensional reducing subspace of V and it is spanned by two (or one) eigenvectors ψ_j and ψ_k of V. Writing each column $x_i = \sum_{j=1}^{n} \alpha_{ij} \psi_j$ in terms of the full eigenvector basis of V, (7.4) yields the quadratic equation

$$\frac{z^2}{x_i' V x_i} - 2z + \frac{1}{x_i' V^{-1} x_i} = 0 \tag{7.5}$$

for the two (or one) eigenvalues λ_j and λ_k associated to each x_i, $i = 1, \ldots, p$. Substituting those eigenvalues as found from (7.5) into (7.3) brings (7.3) to the statistical efficiency lower bound (2.3).

On the other hand, the Euler equation (3.6) from the operator trigonometry, for $n \times n$ SPD matrices A, becomes

$$\frac{A^2 x}{\langle A^2 x, x \rangle} - \frac{2Ax}{\langle Ax, x \rangle} + x = 0. \tag{7.6}$$

Comparison of (7.5), which we call the inefficiency equation, and the Euler equation (7.6) yield the following result.

Theorem 7.1. *For any $n \times n$ SPD covariance matrix V or more generally any $n \times n$ SPD matrix A, all eigenvectors x_j satisfy the inefficiency equation (7.4) and the Euler equation (7.6). The only other vectors satisfying the inefficiency equation (7.4) are the "inefficiency vectors"*

$$x_{\pm}^{j+k} = \pm \frac{1}{\sqrt{2}} x_j + \frac{1}{\sqrt{2}} x_k, \tag{7.7}$$

where x_j and x_k are any eigenvectors corresponding to any distinct eigenvalues $\lambda_j \neq \lambda_k$. The only other vectors satisfying the Euler equation (7.6) are the antieigenvectors

$$x_{\pm}^{jk} = \pm \left(\frac{\lambda_k}{\lambda_j + \lambda_k} \right)^{1/2} x_j + \left(\frac{\lambda_j}{\lambda_j + \lambda_k} \right)^{1/2} x_k. \tag{7.8}$$

For details of the proof of Theorem 7.1, see Gustafson [1, 3].

Commentary. The statistical interpretation of relative statistical inefficiency of an OLSE estimator $\widehat{\beta}$ in terms of (2.2) is that the design matrix X chosen for (2.1) unfortunately contains columns of the form (7.7). That is why we called those the inefficiency vectors of V. The most critical are of course those with $j = 1$ and $k = n$. On the other hand, the new geometrical interpretation of relative statistical inefficiency of an OLSE estimator $\widehat{\beta}$, now in terms of the bound (2.3) as seen trigonometrically according to Theorem 2.1, is now in the worst-case situation; the matrix X under consideration unfortunately contains columns of the form (7.8). These antieigenvectors represent the critical turning angles of the covariance matrix V. The worst case is when $j = 1$ and $k = n$.

8. Canonical correlations and Rayleigh quotients

The Euler equation for the antieigenvectors can be placed (at least in the case of A symmetric positive definite) within a context of stationary values of products of Rayleigh quotients. To do so, we refer to the paper of Rao and Rao [6], and references therein. If one considers the problem of obtaining the stationary values of an expression

$$\frac{x'Cx}{(x'Ax)^{1/2}(x'Bx)^{1/2}} \tag{8.1}$$

with A and B being symmetric positive definite and C being symmetric, then squaring (8.1) gives the product of two Rayleigh quotients

$$\frac{\langle Cx,x\rangle}{\langle Ax,x\rangle}\cdot\frac{\langle Cx,x\rangle}{\langle Bx,x\rangle}. \tag{8.2}$$

Taking the functional derivative of (8.1) with respect to x yields the equation

$$\frac{x'Cx}{x'Ax}Ax+\frac{x'Cx}{x'Bx}Bx=2Cx. \tag{8.3}$$

Note that if we let $C = T$, $A = T^2$, $B = 1$, then (8.1) becomes the antieigenvalue quotient (3.1). Similarly, (8.3), for the same operators and with x being normalized to $\|x\| = 1$, becomes the Euler equation (7.6). On the other hand, the full Euler equation (3.6) for any bounded accretive operator A on any Hilbert space is more general than (8.3) in the sense of operators treated. Moreover, one can easily put B and C operators into the coefficients by a similar derivation. Thus, a general theory encompassing statistical efficiency, operator trigonometry, and canonical correlations could be developed.

Commentary. In their analysis, Rao and Rao [6] arrive at two cases, the first case corresponds to stationary values equal to 1, and the second case corresponds to smaller stationary values. As regards the second case, they note that "there can be solutions of the form $x = ae_i + be_j$," where the e_i and e_j are eigenvectors. But we now know from the operator trigonometry that these are the two cases covered by our Euler equation (3.6), and that the solutions in the second case are the antieigenvectors.

9. Concluding discussion

Who first formulated that the definition $\mathrm{RE}(\widehat{\beta})$ of statistical efficiency was not clear to this author. Durbin and Kendall [18], certainly two great veterans in the field, specifically define E to be the efficiency of t' relative to t according to (see [18, page 151])

$$\rho(t,t') = \sqrt{\frac{\operatorname{var} t}{\operatorname{var} t'}} = \sqrt{E}. \tag{9.1}$$

Here $t = \sum_{j=1}^{n} \lambda_j x_j$ is a linear estimator of the mean. To be unbiased, the coefficients λ_j must satisfy $\sum \lambda_j = 1$. The variance of the estimator t is then $\sigma^2 \sum \lambda_j^2 = \sigma^2 (OP)^2$, where OP is the line segment from the origin to the $\sum \lambda_j = 1$ hyperplane in λ-space. Clearly, the smallest such variance arrives when one takes the point P to be the bottom of the line segment perpendicular to the hyperplane. Variance of t' is just $\sigma^2 (OP')^2$ for any other point P' in the hyperplane. So $E = \cos\phi$, where ϕ is the angle between the lines OP and OP'.

Durbin and Kendall [18] cite the book of Cramér [32] for statistical efficiency. There [32, Chapter 32, page 474], Cramér makes it clear that "in the sequel, we shall exclusively consider the measures of dispersion and concentration associated with the variance and its multidimensional generalizations." Then see [32, page 481], the efficiency $e(\alpha^*)$ is defined to be the ratio between the variance $D^2(\alpha^*)$ of an unbiased and regular estimate α^* and its smallest possible value

$$\frac{1}{n \displaystyle\int_{-\infty}^{\infty} \left(\frac{\partial \log f}{\partial \alpha} \right)^2 f \, dx}. \tag{9.2}$$

Here, $f(x,\alpha)$ is a continuous frequency function. The discrete case is also worked out in later pages. Cramér attributes the concept of efficient estimate to Fisher [27, 33]. Also mentioned (see [32, page 488]) are (later) papers by Neyman, Pearson, and Koopman. So the theory of statistical efficiency arises centrally out of the general theory of estimation of variance by maximum likelihood methods, and it seems, from the early days of that development.

In Freund's classic textbook (Miller and Miller [34]), one finds (see page 327) that the fact that $\operatorname{var}(\widehat{\theta}) \geqq$ the quantity in (9.2) is called the Cramér-Rao inequality. The denominator of (9.2) is interpreted as the information about the estimator θ which is supplied by the sample. Smaller variance is interpreted to mean greater information. Thus, as Cramér already made clear (see our quote above and Chapter 32 of his book), we are looking at central tendency as measured by second moments.

We decided to bite the bullet and go back to Fisher [27, 33]. Indeed, in his first paper see [27, page 309], he clearly defines efficiency of a statistic as "the ratio whose intrinsic accuracy bears to that of the most efficient statistic possible; it expresses the proportion of the total available relevant information of which that statistic makes use." He carefully attributes, designates, or, in any case, cites in connection with that definition a 1908 paper by Student and a 1763 paper by Bayes. Then, we find (on page 315) that "in 1908, Student broke new ground by calculating the distribution of the ratio which the deviation of

the mean from its population value bears to the standard deviation calculated from the sample." Of course, both papers [27, 33] also contain excellent discussions of the method of maximum likelihood and its pros and cons.

Here, this author must interject that in a classified naval intelligence task, in 1959, he first became aware of, and implemented, the χ^2 distribution for estimating goodness of fit for combinations of normally distributed random variables. The application was concerned with observations at several receiving sites of the bearings of received signal from a transmitting enemy submarine. For an unclassified account of this work, see the paper of Gustafson [35]. This author still remembers the genuine joy of operational naval personnel as they called out that "the χ^2 of the fit is…!" It is also perhaps an amusing irony that 45 years later this author, through the indirect and abstract path of his operator trigonometry, has arrived back at χ^2 testing.

A second point for discussion is that in this treatment, we have not gone into the more general theory of statistical efficiency utilizing generalized inverses. Certainly, it is natural and essential to do so for both theory and statistical applications. For example, when V is nonsingular, one has (e.g., see Puntanen and Styan [36]) in terms of generalized inverses

$$\begin{aligned} \text{BLUE}(X\beta) &= X\beta^* = X(X^*V^{-1}X)^{-}X^*V^{-1}y, \\ \text{OLSE}(X\beta) &= X\hat{\beta} = X(X^*X)^{-}X^*y. \end{aligned} \tag{9.3}$$

However, in this author's opinion, the essential points are first seen for $p = 1$, that is, in the case of X, a single regressor vector. In any case, the more general theory including generalized inverses is now so well worked out in the mathematical statistics literature that such a state of affairs should excuse the author from having to process it all. On the other hand, it is equally clear that the operator trigonometry of statistical efficiency should be extended to that setting including generalized inverses and, moreover, singular correlation matrices V. Possibly, we shall do that in the future, but such a comprehensive study is a task for another paper.

However, we here may "close the picture" from the other direction. From the usual assumption $X^*X = I_p$, where X is an $n \times p$ semi-unitary matrix, it is instructive to take its p orthonormal columns and conceptually add to them $n - p$ orthonormal columns. These may be thought of as "fictitious" additional regressors that one would like to have. How to do so is just the procedure in the proof of the classical Schur theorem. Call any one of these enlarged unitary regressor matrices X. Then, (9.3) is simplified to

$$\text{BLUE}\,(X\beta^*) = X^{-1}y, \qquad \text{OLSE}\,(X\hat{\beta}) = y. \tag{9.4}$$

Also, the efficiency (2.2) becomes 1, caused essentially by the unitarity of X. Although this exercise should not surprise anyone, still it seems to this author that the generalized inverse theory could be viewed as an "intermediate" theory dealing with how badly you have truncated and otherwise abused the fictitiously available large set of Schur unitaries. As a variation on this theme, for an arbitrary $n \times n$ matrix X written in its polar form $X = U|X|$, where U is the isometry from the range of the absolute value operator $|X|$ to the range of X, the operator trigonometry concerns itself only with the turning angles of the Hermitian polar factor $|X|$. See Gustafson [14] for more on this point. Thus, the essence

of the minimization of the Durbin lower bound (2.3) by its attainment by antieigenvector regression vectors as described in Theorem 2.1 has to do with the polar Hermitian factor of X, and not with its isometric factor U. So our thought experiment exercise leading to (9.4) says that the unitary factor of the design matrix X has no effect on its statistical efficiency.

To conclude, in this paper we have placed the theory of statistical efficiency into the geometrical setting of the author's operator trigonometry. There are many remaining aspects of both together with their further interconnection, with which we have not dealt.

Addendum

In the intervening two years since the IWMS 2005 conference, on which the work herein was first presented, I have written two further related papers that should be mentioned: Gustafson [37, 38].

In [37], what follows are rendered trigonometric: Khatri-Rao inequality, Khatri-Rao-Ando bound, Bartmann-Bloomfield bound, and Hotelling correlation coefficient. In [38], I provide a complete survey of the various applications of my operator trigonometry, from 1966 to the present.

Acknowledgment

Finally, I would like to add that following the IWMS 2005 conference, I was led to a great appreciation of all of Jeff Hunter's pioneering contributions to the linear algebra of Markov chains and, in particular, his extremely fine and important results for the perturbation theory using generalized inverses and finite rank updates.

References

[1] K. Gustafson, "On geometry of statistical efficiency," preprint, 1999.
[2] K. Gustafson, "An unconventional computational linear algebra: operator trigonometry," in *Unconventional Models of Computation, UMC'2K*, I. Antoniou, C. Calude, and M. Dinneen, Eds., pp. 48–67, Springer, London, UK, 2001.
[3] K. Gustafson, "Operator trigonometry of statistics and econometrics," *Linear Algebra and Its Applications*, vol. 354, no. 1–3, pp. 141–158, 2002.
[4] P. Bloomfield and G. S. Watson, "The inefficiency of least squares," *Biometrika*, vol. 62, no. 1, pp. 121–128, 1975.
[5] M. Knott, "On the minimum efficiency of least squares," *Biometrika*, vol. 62, no. 1, pp. 129–132, 1975.
[6] C. R. Rao and C. V. Rao, "Stationary values of the product of two Raleigh quotients: homologous canonical correlations," *Sankhyā B*, vol. 49, no. 2, pp. 113–125, 1987.
[7] S. G. Wang and S.-C. Chow, *Advanced Linear Models: Theory and Applications*, vol. 141 of *Statistics: Textbooks and Monographs*, Marcel Dekker, New York, NY, USA, 1994.
[8] K. Gustafson, "A min-max theorem," *Notices of the American Mathematical Society*, vol. 15, p. 799, 1968.
[9] K. Gustafson, *Lectures on Computational Fluid Dynamics, Mathematical Physics, and Linear Algebra*, World Scientific, River Edge, NJ, USA, 1997.
[10] K. Gustafson and D. K. M. Rao, *Numerical Range: The Field of Values of Linear Operators and Matrices*, Universitext, Springer, New York, NY, USA, 1997.

[11] K. Gustafson, "Commentary on topics in the analytic theory of matrices," in *Collected Works of Helmut Wielandt 2*, B. Huppert and H. Schneider, Eds., pp. 356–367, DeGruyters, Berlin, Germaney, 1996.

[12] K. Gustafson, "Antieigenvalue inequalities in operator theory," in *Proceedings of the 3rd Symposium on Inequalities*, O. Shisha, Ed., pp. 115–119, Academic Press, Los Angeles, Calif, USA, September 1972.

[13] K. Gustafson, "Antieigenvalues," *Linear Algebra and Its Applications*, vol. 208-209, pp. 437–454, 1994.

[14] K. Gustafson, "An extended operator trigonometry," *Linear Algebra and Its Applications*, vol. 319, no. 1–3, pp. 117–135, 2000.

[15] K. Gustafson, "A note on left multiplication of semigroup generators," *Pacific Journal of Mathematics*, vol. 24, no. 3, pp. 463–465, 1968.

[16] R. L. Plackett, "A historical note on the method of least squares," *Biometrika*, vol. 36, no. 3-4, pp. 458–460, 1949.

[17] A. C. Aitken, "On least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh Section A*, vol. 55, pp. 42–48, 1935.

[18] J. Durbin and M. G. Kendall, "The geometry of estimation," *Biometrika*, vol. 38, no. 1-2, pp. 150–158, 1951.

[19] G. S. Watson, "Serial correlation in regression analysis—I," *Biometrika*, vol. 42, no. 3-4, pp. 327–341, 1955.

[20] G. S. Watson, "Linear least squares regression," *Annals of Mathematical Statistics*, vol. 38, no. 6, pp. 1679–1699, 1967.

[21] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression—I," *Biometrika*, vol. 37, no. 3-4, pp. 409–428, 1950.

[22] J. Durbin and G. S. Watson, "Testing for serial correlation in least squares regression—II," *Biometrika*, vol. 38, no. 1-2, pp. 159–178, 1951.

[23] T. W. Anderson, "On the theory of testing serial correlation," *Skandinavisk Aktuarietidskrift*, vol. 31, pp. 88–116, 1948.

[24] R. L. Anderson and T. W. Anderson, "Distribution of the circular serial correlation coefficient for residuals from a fitted Fourier series," *Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 59–81, 1950.

[25] U. Grenander, "On the estimation of regression coefficients in the case of an autocorrelated disturbance," *Annals of Mathematical Statistics*, vol. 25, no. 2, pp. 252–272, 1954.

[26] U. Grenander and M. Rosenblatt, *Statistical Analysis of Stationary Time Series*, John Wiley & Sons, New York, NY, USA, 1957.

[27] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London A*, vol. 222, pp. 309–368, 1922.

[28] J. von Neumann, "Distribution of the ratio of the mean square successive difference to the variance," *Annals of Mathematical Statistics*, vol. 12, no. 4, pp. 367–395, 1941.

[29] J. von Neumann, R. H. Kent, H. R. Bellinson, and B. I. Hart, "The mean square successive difference," *Annals of Mathematical Statistics*, vol. 12, pp. 153–162, 1941.

[30] K. Gustafson, *Introduction to Partial Differential Equations and Hilbert Space Methods*, Dover, Mineola, NY, USA, 3rd edition, 1999.

[31] K. Gustafson, "Operator trigonometry of the model problem," *Numerical Linear Algebra with Applications*, vol. 5, no. 5, pp. 377–399, 1998.

[32] H. Cramér, *Mathematical Methods of Statistics*, vol. 9 of *Princeton Mathematical Series*, Princeton University Press, Princeton, NJ, USA, 1946.

[33] R. A. Fisher, "Theory of statistical estimation," *Proceedings of the Cambridge Philosophical Society*, vol. 22, pp. 700–725, 1925.

[34] I. Miller and M. Miller, *Freund's Mathematical Statistics*, Prentice-Hall, Upper Saddle River, NJ, USA, 6th edition, 1999.
[35] K. Gustafson, "Parallel computing forty years ago," *Mathematics and Computers in Simulation*, vol. 51, no. 1-2, pp. 47–62, 1999.
[36] S. Puntanen and G. P. H. Styan, "The equality of the ordinary least squares estimator and the best linear unbiased estimator," *The American Statistician*, vol. 43, no. 3, pp. 153–164, 1989.
[37] K. Gustafson, "The trigonometry of matrix statistics," *International Statistical Review*, vol. 74, no. 2, pp. 187–202, 2006.
[38] K. Gustafson, *Noncommutative trigonometry*, vol. 167 of *Operator Theory: Advances and Applications*, Birkhäuser Basel, Birkhäuser Verlag AG, Viaduktstraße 42, 4051 Basel, Switzerland, 2006.

K. Gustafson: Department of Mathematics, University of Colorado at Boulder, Boulder, Colo 80309, USA
Email address: gustafs@euclid.colorado.edu

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 90842, 11 pages
doi:10.1155/2007/90842

Research Article

Weibull Model Allowing Nearly Instantaneous Failures

C. D. Lai, Michael B. C. Khoo, K. Muralidharan, and M. Xie

Received 27 April 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

A generalized Weibull model that allows instantaneous or early failures is modified so that the model can be expressed as a mixture of the uniform distribution and the Weibull distribution. Properties of the resulting distribution are derived; in particular, the probability density function, survival function, and the hazard rate function are obtained. Some selected plots of these functions are also presented. An R script was written to fit the model parameters. An application of the modified model is illustrated.

1. Introduction

Many generalizations and extensions of the Weibull distribution have been proposed in reliability literature due to the lack of fits of the traditional Weibull distribution. A summary of these generalizations is given by Pham and Lai [1]. An extensive treatment on Weibull models is given by Murthy et al. [2].

A modified Weibull distribution that allows instantaneous or early failures is introduced by Muralidharan and Lathika [3]. It was pointed out that the occurrence of instantaneous or early (premature) failures in life testing experiment is a phenomenon observed in electronic parts as well as in clinical trials. These occurrences may be due to inferior quality, faulty construction, or nonresponse of the treatments.

It has been shown that the distribution may be represented as a mixture of a singular distribution at zero (or t_0) and a two-parameter Weibull distribution. Because of the singular nature of distribution, we have been unable to define the failure rate (hazard rate) function meaningfully. The aim of this paper is to provide a meaningful modification so that the resulting model can be expressed as a mixture of two continuous distributions. This modified form is more realistic as "true" instantaneous failures rarely occur. The

modification allows us to establish and study the failure rate function of this reliability model via mixture distributions. We also provide some graphical plots to illustrate some possible shapes for the survival functions as well as the failure rate functions.

The present paper focuses on the "nearly instantaneous" failure case as it has fewer parameters. This special case is more realistic than the "early failure" case since many products exhibit an "infant mortality" phenomenon due to initial defects.

2. Representation of the model

Let $F(t)$ and $R(t) = 1 - F(t)$ denote the cumulative distribution function and the survival function of the mixture, respectively. We assume that F is continuous and its density be given by $f(t) = F'(t)$. The component distribution functions and their survival functions are $F_i(t)$ and $R_i(t) = 1 - F_i(t)$, respectively, $i = 1,2$. The failure rate of a lifetime distribution is defined as $h(t) = f(t)/R(t)$ provided the density exists.

Instead of assuming an instant or an early failure to occur at a particular time point, as in the original model of Muralidharan and Lathika [3], we now represent this model as a mixture of a generalized Dirac delta function and the 2-parameter Weibull as opposed to a mixture of a singular distribution with a Weibull. Thus, the resulting modification gives rise to a density function:

$$f(t) = p\delta_d(t - t_0) + q\alpha\lambda^\alpha t^{\alpha-1}e^{-(\lambda t)^\alpha}, \quad p + q = 1,\ 0 < p < 1, \tag{2.1}$$

where

$$\delta_d(t - t_0) = \begin{cases} \dfrac{1}{d}, & t_0 \le t \le t_0 + d, \\ 0, & \text{otherwise}, \end{cases} \tag{2.2}$$

for *sufficiently small* d. Here $p > 0$ is the mixing proportion.

We note that

$$\delta(x - x_0) = \lim_{d\to 0} \delta_d(x - x_0), \tag{2.3}$$

where $\delta(\cdot)$ is the Dirac delta function. We may view the Dirac delta function as a normal distribution having a zero mean and standard deviation that tends to 0. For a fixed value of d, (2.2) denotes a uniform distribution over an interval $[t_0, t_0 + d]$ so the modified model is now effectively a mixture of a Weibull with a uniform distribution. Instead of including a possible instantaneous failure in the model, (2.2) allows for a possible "near instantaneous" failure to occur uniformly over a very small time interval.

Note that the case $t_0 = 0$ corresponds to instantaneous failures, whereas $t_0 \neq 0$ (but small) corresponds to the case with early failures.

Noting from (2.1) and (2.2), we see that the mixture density function is not continuous at t_0 and $t_0 + d$. However, both the distribution and survival functions are continuous.

Writing $f_1(t) = \delta_d(t - t_0)$ and $f_2(t) = \alpha\lambda^\alpha t^{\alpha-1}e^{-(\lambda t)^\alpha}$, $\alpha, \lambda > 0$; (2.1) can be written as

$$f(t) = pf_1(t) + qf_2(t), \quad p + q = 1,\ 0 < p < 1, \tag{2.4}$$

so

$$F(t) = pF_1(t) + qF_2(t), \tag{2.5}$$

$$R(t) = 1 - F(t) = p + q - \{pF_1(t) + qF_2(t)\} = pR_1(t) + qR_2(t). \tag{2.6}$$

Thus, the failure (hazard) rate function of the mixture distribution is

$$h(t) = \frac{pf_1(t) + qf_2(t)}{pR_1(t) + qR_2(t)}. \tag{2.7}$$

A mixture distribution involving two 2-parameter Weibull distributions has been thoroughly studied in Jiang and Murthy [4]. The mixture considered in this paper is more complex in the sense that one of the mixing distributions has a finite range which poses some challenges.

Via (2.4) simulated observations from this model are made by generating uniform variates and Weibull variates with proportions p and $q = 1 - p$, respectively.

3. Survival function and failure rate function of the model

Recently, failure rates of mixtures are discussed quite extensively. Lai and Xie [5, Section 2.8] provide a good summary. As demonstrated by Block et al. [6], various shapes of failure rate functions can arise with a mixture of two increasing failure rate distributions. Now, a Weibull has an increasing (decreasing) failure rate if its shape parameter α is greater (smaller) than 1. The uniform distribution also has an increasing failure rate if it is uniform over [0, a]. Thus we are interested to know what shapes would result from our model.

The reliability (survival) functions of the respective component distributions are given by

$$R_1(t) = \begin{cases} 1 & \text{if } 0 \le t < t_0, \\ \dfrac{d + t_0 - t}{d} & \text{if } t_0 \le t \le t_0 + d, \\ 0 & \text{if } t > t_0 + d, \end{cases} \tag{3.1}$$

$$R_2(t) = e^{-(\lambda t)^{\alpha}}, \quad t \ge 0,\ \alpha, \lambda > 0. \tag{3.2}$$

The failure rates are, respectively,

$$h_1(t) = \begin{cases} 0, & 0 \le t < t_0, \\ \dfrac{1}{d + t_0 - t}, & t_0 \le t \le t_0 + d, \\ \infty, & t > t_0 + d, \end{cases} \tag{3.3}$$

$$h_2(t) = \alpha\lambda(\lambda t)^{\alpha - 1}. \tag{3.4}$$

It can be shown from (2.4) and (2.6) that for any mixture of two continuous distributions, the failure rate function can be expressed as

$$h(t) = \frac{f(t)}{R(t)} = w(t)h_1(t) + [1 - w(t)]h_2(t), \tag{3.5}$$

where $w(t) = pR_1(t)/R(t)$ for all $t \geq 0$. In our case,

$$w(t) = \begin{cases} \dfrac{p}{R(t)} & \text{if } 0 \leq t < t_0, \\ \dfrac{pR_1(t)}{R(t)} & \text{if } 0 \leq t \leq t_0 + d, \\ 0 & \text{if } t > t_0 + d \end{cases} \tag{3.6}$$

with

$$w'(t) = w(t)[1 - w(t)]\{h_2(t) - h_1(t)\}. \tag{3.7}$$

(Note that equation (3.7) is valid for all cases).

Also a simple differentiation shows that

$$h'(t) = w'(t)h_1(t) + w(t)h_1'(t) - w'(t)h_2(t) + [1 - w(t)]h_2'(t). \tag{3.8}$$

Now, $w(t)h_1(t) = pR_1(t)/R(t) \times f_1(t)/R_1(t) = pf_1(t)/R(t)$, so (3.5) is well defined for all $t > 0$.

Summarized expression for $R(t)$ and $h(t)$ are, respectively, given as

$$R(t) = pR_1(t) + qR_2(t) = \begin{cases} p + qe^{-(\lambda t)^\alpha}, & 0 \leq t < t_0, \\ \dfrac{p(d + t_0 - t)}{d} + qe^{-(\lambda t)^\alpha}, & t_0 \leq t \leq t_0 + d, \\ qe^{-(\lambda t)^\alpha}, & t > t_0 + d, \end{cases} \tag{3.9}$$

$$h(t) = \begin{cases} \left(\dfrac{qe^{-(\lambda t)^\alpha}}{p + qe^{-(\lambda t)^\alpha}}\right)\alpha\lambda^\alpha t^{\alpha-1}, & 0 \leq t < t_0, \\ \dfrac{p + dqe^{-(\lambda t)^\alpha}\alpha\lambda^\alpha t^{\alpha-1}}{p(d + t_0 - t) + dqe^{-(\lambda t)^\alpha}}, & t_0 \leq t \leq t_0 + d, \\ \alpha\lambda^\alpha t^{\alpha-1}, & t > t_0 + d. \end{cases} \tag{3.10}$$

Recall that $h(t)$ is discontinuous at both $t = t_0$ and $t = t_0 + d$. Unlike the model of Muralidharan and Lathika [3], the survival function is continuous though not differentiable at $t = t_0$ and $t = t_0 + d$.

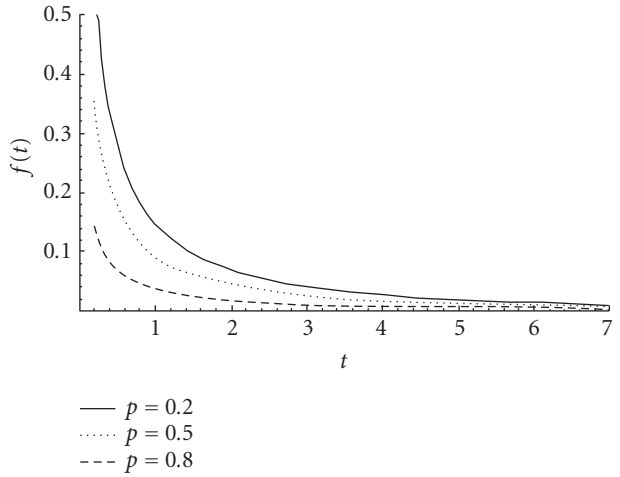


FIGURE 4.1. Plot of $f(t) : \lambda = 1, \alpha = 0.5, d = 0.2, t_0 = 0$.

4. Nearly instantaneous failure case ($t_0 = 0$)

Consider a special case of the model (2.1) whereby $t_0 = 0$. The model may be called the Weibull with "nearly instantaneous failure" model.

In this case, (3.3) is simplified giving the failure rate of the uniform distribution as

$$h_1(t) = \begin{cases} \dfrac{1}{d-t}, & 0 \le t \le d, \\ \infty, & t > d, \end{cases} \tag{4.1}$$

and (3.1) its survival function is given as

$$R_1(t) = \begin{cases} \dfrac{d-t}{d} & \text{if } 0 \le t \le d, \\ 0 & \text{if } t > d. \end{cases} \tag{4.2}$$

The Weibull model with "nearly instantaneous failure" occurring uniformly over $[0, d]$ has

$$R(t) = \begin{cases} \dfrac{p(d-t)}{d} + qe^{-(\lambda t)^{\alpha}}, & 0 \le t \le d, \\ qe^{-(\lambda t)^{\alpha}}, & t > d, \end{cases} \tag{4.3}$$

$$h(t) = \begin{cases} \dfrac{p + dqe^{-(\lambda t)^{\alpha}} \alpha \lambda^{\alpha} t^{\alpha-1}}{p(d-t) + dqe^{-(\lambda t)^{\alpha}}}, & 0 \le t \le d, \\ \alpha \lambda^{\alpha} t^{\alpha-1}, & t > d. \end{cases} \tag{4.4}$$

Graphs of Survival, Density, and Failure Rate Functions. Graphical plots are important for ageing distributions. It is not the aim of this paper to present complete characterizations

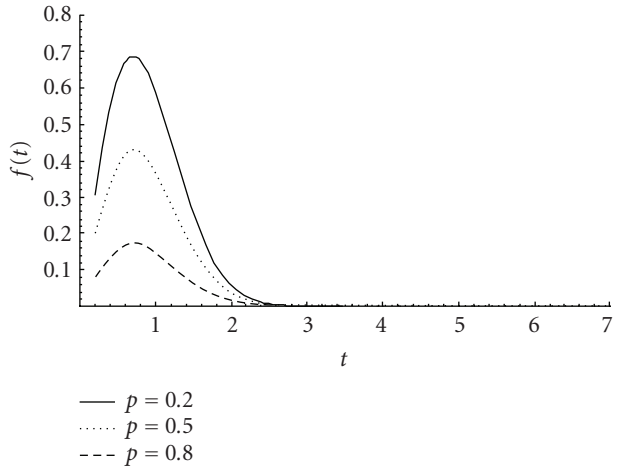


FIGURE 4.2. Plot of $f(t) : \lambda = 1, \alpha = 2, d = 0.2, t_0 = 0$.

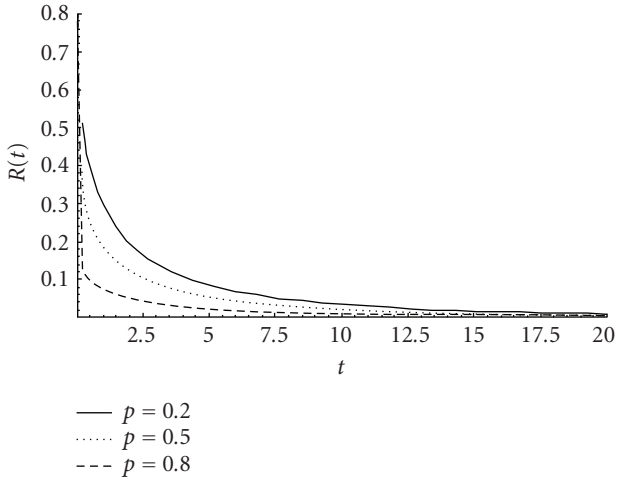


FIGURE 4.3. Plot of $R(t) : \lambda = 1, \alpha = 0.5, d = 0.2, t_0 = 0$.

for the survival, density, and the failure rate functions. Instead, snapshots are taken of some possible shapes from this model, as it is important to identify whether the model is useful for specific datasets for which empirical plots are available.

Consider in detail the special case when $t_0 = 0$, that is, the Weibull with "nearly" instantaneous failure model.

Density functions. In both Figures 4.1 and 4.2, three density functions with $p = 0.2$, 0.5, and 0.8 are plotted. In all figures, the smallest mixing proportion p is given by the solid line.

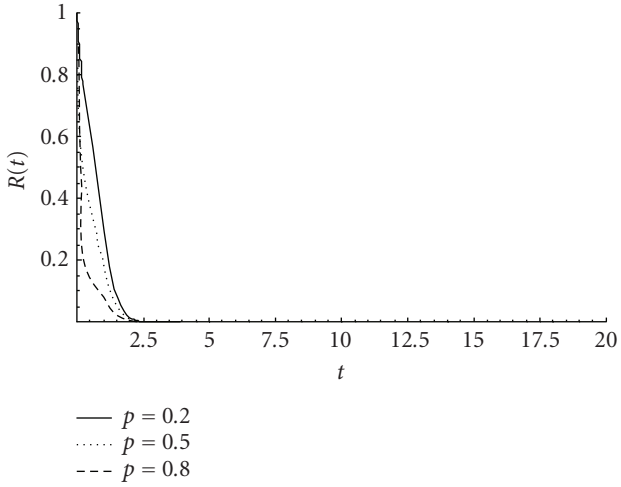


FIGURE 4.4. Plot of $R(t) : \lambda = 1, \alpha = 2, d = 0.2, t_0 = 0$.

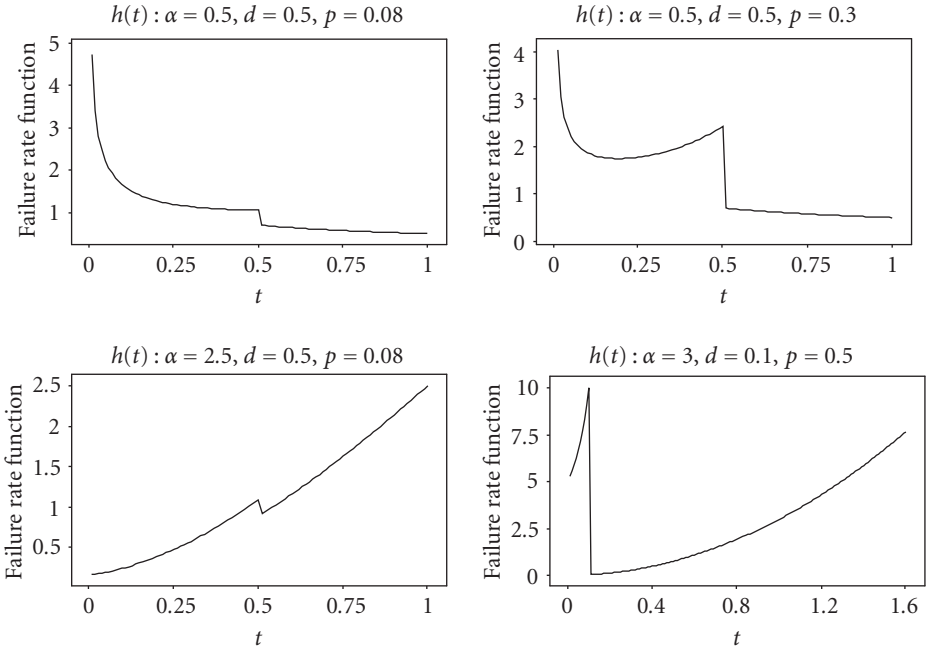


FIGURE 4.5. Plots of failure rate functions.

Survival functions. The survival functions are given by Figures 4.3 and 4.4 which correspond to the density functions in Figures 4.1 and 4.2, respectively.

Failure rate functions. The failure rate function $h(t)$ is given as in (4.4). Clearly, its shape is the same as the Weibull failure rate after d. Thus we focus on the segment from 0 to d.

Table 5.1. Instantaneous failures at $t_i = 0$, $i = 1,2,\ldots,28$.

	$\hat{p}$	$\hat{\alpha}$	$1/\hat{\lambda}$
Estimates	0.6999992	1.1929632	0.9315360
Standard error	0.07244318	0.28878116	0.23615878

Since the scale parameter λ does not alter the shape, it is set to one. Figure 4.5 shows that $h(t)$ can be increasing, decreasing, or bathtub shaped for $0 \le t \le d$.

From the plots, it can be seen that the failure rate function of the model gives rise to several different shapes and bumps; this is expected as mixing distributions with a component distribution that has a finite range often cause some problems. Although the second part can be either increasing or decreasing, the first segment can achieve various shapes. This finding is consistent with Block et al. [6].

5. Data fitting and an application

Mixture distributions are used widely in the statistics literature. Bebbington et al. [7] have used a mixture of two generalized Weibull distributions to fit human mortality data. A mixture distribution may give rise to different failure rate thus it can provide pseudo-demarcation of various phases of a lifespan. Bebbington et al. [7] also use their mixture distribution to identify the differences between (sub)populations.

While software for fitting mixture distributions is available, such as the MIX package for the R environment (Macdonald [8]), such packages do not handle uniform distributions, or mixtures of unlike distributions.

To fit the model to a dataset, an R script was written. One can write a code in R to fit all 4 parameters (p,λ,α,d) and another to fit 3 parameters (p,λ,α) with d given and held fixed. The second case always works, and works very well, but the first never gives good results when the edge of the uniform (parameter d) is inside the peak of the Weibull (personal communication with Professor Macdonald). This is because there is not enough information in the data to fit the 4th parameter in this situation. In practice, the value of d can be manually estimated quite accurately from the dataset.

5.1. Application. A sample of 40 boards of woods were checked for their dryness on a particular area of a board. The actual observations were degrees of dryness measured as a percentage. This dataset was analysed in Muralidharan and Lathika [3] with $t_i = 0$, $i = 1,2,\ldots,28$ and the other positive observations are as follows: 0.0463741, 0.0894855, 0.4, 0.42517, 0.623441, 0.6491, 0.73346, 1.35851, 1.77112, 1.86047, 2.12125, 2.12389.

Treating the degree of dryness as the "failure time," we apply the Weibull model with "nearly instantaneous" failures model to this data (see Table 5.1).

It is reasonable to spread the zeros uniformly over an interval $[0,d)$. For illustration, we select $d = 0.135$ so that $t_1 = 0$, $t_2 = 0.0005$, $t_3 = 0.001,\ldots,t_{28} = 0.0135$. Applying the MLE method in R, we found the following see Table 5.2.

It seems to us the sizes of standard errors of the three estimates are reasonable in view of a small sample size.

Table 5.2. Uniform spread of "nearly" instantaneous failure times.

	$\hat{p}$	$\hat{\alpha}$	$1/\hat{\lambda}$
Estimates	0.6981846	1.1656527	0.9431262
Standard error	0.07292797	0.30176818	0.24822803

Table 5.3. Exponential with "nearly" instantaneous failures: $d = 0.0135$.

	$\hat{p}$	$1/\hat{\lambda}$
v1–estimates (se)	0.6959071 (0.0734413)	1.0031716 (0.2896755)
v2–estimates (se)	0.6959356 (0.07343414)	1.0033558 (0.28970177)

Since the shape parameter $\hat{\alpha} = 1.16$ and the fact that three parameters are being estimated from few data points, it would be more realistic to specify $\alpha = 1$ *a priori* (i.e., an exponential) as a special case of the Weibull model. This would reduce the number of parameters and therefore the uncertainty in the parameter estimates. Table 5.3 summarizes the parameter estimates of the exponential mixing with the uniform model.

Note. v1 consists of of measurements from the original dataset. The 28 zeros in v1 are then calibrated to spread over uniformly over $[0, d]$. v2 is formed by replacing the first 28 cells of v1 by the calibrated values.

We note from the preceding table that the precision for the second parameter estimate actually deteriorates in comparison with the Weibull case. Perhaps the Akaike information criteria (AIC) or BIC would be a more objective way to evaluate this comparison.

5.2. Sensitivity analysis. In the above model fitting, we have chosen d manually and the value $d = 0.0135$ is chosen because it is sufficiently apart from the first nonzero observed value 0.0463741. We have assessed the sensitivity of the selection of the parameter d. For the "instantaneous" failures case, we let d vary between 0.01 and 0.04; the resulting estimates and their standard errors are virtually unchanged. However, it becomes sensitive when d is too close to $t = 0$ or to the first Weibull failure time $t = 0.0463741$. If d is set to 0.135, the parameter estimates are then given by $\hat{p} = 0.7497827$, $\hat{\alpha} = 1.8827550$, and $1/\hat{\lambda} = 0.7327841$ indicating that these estimates change noticeably as d encroaches on the Weibull "territory." It also shows that for this value of d, the model with the uniform mixing with the exponential can longer be an alternative because the estimate for the shape parameter α is now close to 2.

6. Conclusion

The Weibull distribution has been widely used as a life model in reliability applications. However, one often finds that it does not fit well in the early part of a lifespan for various reasons. In particular, in the cases where initial defects are present causing early failures, the Weibull distribution is found inadequate to model such phenomenon. The proposed model of a modified Weibull mixing with a uniform distribution to model the first phase of a lifespan should provide a useful alternative.

Appendix

R code for fitting the model

```
 v1<-rep(0,40)
v1[29:40]<-c(0.0463741,0.0894855,0.4,0.42517,0.623441,0.6491,
            0.73346,1.35851,1.77112,1.86047,2.12125,2.12389)
v2 <- c(0, 0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003, 0.0035,
0.004, 0.0045, 0.005, 0.0055, 0.006, 0.0065, 0.007, 0.0075, 0.008,
0.0085, 0.009, 0.0095, 0.01, 0.0105, 0.011, 0.0115, 0.012, 0.0125,
0.013, 0.0135, 0.046374, 0.089486, 0.4, 0.42517, 0.623441, 0.6491,
0.73346, 1.35851, 1.77112, 1.86047, 2.12125, 2.12389)

uniweib <-cbind(v1, v2)

neglluniweib<-function(p,x,d){
  pp<-p[1]
  shape<-p[2]
  rate<-p[3]
  if(pp>0 & shape>0 & rate>0 & pp<1) {
    -sum(log(pp*dunif(x,0,d)+(1- pp)*dweibull(x,shape,1/rate)))
  }else{
    1e+200
  }
}

uniweibmle <- function(d,prop,shape,rate,x) {
  p<-c(prop,shape,rate)
  fit<-nlm(neglluniweib,p=p,x=x,d=d,hessian=T)
  fit$se<-sqrt(diag(solve(fit$hessian)))
 fit[c(1,2,7,3,4,5,6)]

}

uniweibmle(0.0135,0.1,2,1,v1) uniweibmle(0.0135,0.1,2,1,v2)
```

Acknowledgments

The first author is privileged to have been a former student and a colleague of Professor Jeff Hunter, who has been an inspiration to many. The authors wish him well in his retirement. The authors are very grateful to Professor Peter Macdonald who helped them to write an R script for the model and to the referees for their helpful comments. The authors also wish to acknowledge Dr. K. Govindaraju for his help to improve the quality of the figures.

References

[1] H. Pham and C. D. Lai, "On recent generalizations of the Weibull distribution," *IEEE Transactions on Reliability*, vol. 54, no. 3, pp. 454–458, 2007.
[2] D. N. P. Murthy, M. Xie, and R. Jiang, *Weibull Models*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, NJ, USA, 2004.
[3] K. Muralidharan and P. Lathika, "Analysis of instantaneous and early failures in Weibull distribution," *Metrika*, vol. 64, no. 3, pp. 305–316, 2006.
[4] R. Jiang and D. N. P. Murthy, "Mixture of Weibull distributions—parametric characterization of failure rate function," *Applied Stochastic Models and Data Analysis*, vol. 14, no. 1, pp. 47–65, 1998.
[5] C. D. Lai and M. Xie, *Stochastic Ageing and Dependence for Reliability*, Springer, New York, NY, USA, 2006.
[6] H. W. Block, Y. Li, and T. H. Savits, "Initial and final behaviour of failure rate functions for mixtures and systems," *Journal of Applied Probability*, vol. 40, no. 3, pp. 721–740, 2003.
[7] M. Bebbington, C. D. Lai, and R. Zitikis, "Modeling human mortality using mixtures of bathtub shaped failure distributions," *Journal of Theoretical Biology*, vol. 245, no. 3, pp. 528–538, 2007.
[8] P. D. M. Macdonald, "MIX for the R Environment," distributed by ICHTHUS DATA SYSTEMS, http://www.math.mcmaster.ca/peter/mix/mix.html.

C. D. Lai: Institute of Information Sciences and Technology, Massey University,
Palmerston North, New Zealand
Email address: c.lai@massey.ac.nz

Michael B. C. Khoo: School of Mathematical Sciences, Universiti Sains Malaysia,
11800 Pulau Pinang, Malaysia
Email address: mkbc@usm.my

K. Muralidharan: Department of Statistics, The Maharaja Sayajirao University of Baroda,
Vadodara 390 002, India
Email address: lmv_murali@yahoo.com

M. Xie: Department of Industrial and Systems Engineering, National University of Singapore,
Singapore 119077
Email address: mxie@nus.edu.sg

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 86180, 23 pages
doi:10.1155/2007/86180

Research Article

On the Semiparametric Efficiency of the Scott-Wild Estimator under Choice-Based and Two-Phase Sampling

Alan Lee

Received 30 April 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

Using a projection approach, we obtain an asymptotic information bound for estimates of parameters in general regression models under choice-based and two-phase outcome-dependent sampling. The asymptotic variances of the semiparametric estimates of Scott and Wild (1997, 2001) are compared to these bounds and the estimates are found to be fully efficient.

1. Introduction

Suppose that for each number of subjects, we measure a response y and a vector of covariates x, in order to estimate the parameters β of a regression model which describes the conditional distribution of y given x. If we have sampled directly from the conditional distribution, or even the joint distribution, we can estimate β without knowledge of the distribution of the covariates.

In the case of a discrete response, which takes one of J values $y_1,\ldots,y_J$, say, we often estimate β using a case-control sample, where we sample from the conditional distribution of X given $Y = y_j$. This is particularly advantageous if some of the values y_j occur with low probability. In case-control sampling, the likelihood involves the distribution of the covariates, which may be quite complex, and direct parametric modelling of this distribution may be too difficult. To get around this problem, the covariate distribution can be treated nonparametrically. In a series of papers (Scott and Wild [1, 2] Wild [3]) Scott and Wild have developed an estimation technique which yields a semiparametric estimate of β. They dealt with the unknown distribution of the covariates by profiling it out of the likelihood, and derived a set of estimating equations whose solution is the semiparametric estimator of β.

This technique also works well for more general sampling schemes, for example, for two-phase outcome-dependent stratified sampling. Here, the sample space is partitioned into S disjoint strata which are defined completely by the values of the response and possibly some of the covariates. In the first phase of sampling, a prospective sample of size N is taken from the joint distribution of x and y, but only the stratum to which the individual belongs is observed. In the second phase, for $s = 1,\ldots,S$, a sample of size $n_1^{(s)}$ is selected from the $n_0^{(s)}$ individuals in stratum s which were selected in the first phase, and the rest of the covariates are measured. Such a sampling scheme can reduce the cost of studies by confining the measurement of expensive variables to the most informative subjects. It is also an efficient design for elucidating the relationship between a rare disease and a rare exposure, in the presence of confounders.

Another generalized scheme that falls within the Scott-Wild framework is that of case-augmented sampling, where a prospective sample is augmented by a further sample of controls. In the prospective sample, we may observe both disease state and covariates, or covariates alone. Such schemes are discussed in Lee et al. [4].

In this paper, we introduce a general method for demonstrating that the Scott-Wild procedures are fully efficient. We use a (slightly extended) version of the theory of semiparametric efficiency due to Bickel et al. [5] to derive an "information bound" for the asymptotic variance of the estimates. We then compute the asymptotic variances of the Scott-Wild estimators, and demonstrate their efficiency by showing that the asymptotic variance coincides with the information bound in each case.

The efficiency of these estimators has been studied by several authors, who have also addressed this question using semiparametric efficiency theory. This theory assumes an i.i.d. sample, and so various ingenious devices have been used to apply it to the case of choice-based sampling. For example, Breslow et al. [6] consider case-control sampling, that the data are generated by Bernoulli sampling, where either a case or a control is selected by a randomisation device with known selection probabilities, and the covariates of the resulting case or control are measured. The randomisation at the first stage means that the i.i.d. theory can be applied.

The efficiency of regression models under an approximation to the two-phase sampling scheme has been considered by Breslow et al. [7] using missing value theory. In this approach, a single prospective sample is taken. For some individuals, the response and the covariates are both observed. For the rest, only the response is measured and the covariates are regarded as missing values. The efficiency bound is obtained using the missing value theory of Robins et al. [8].

In this paper, we adopt a more direct approach. First, we sketch an extension of Bickel-Klaassen-Ritov-Wellner theory to cover the case of sampling from several populations, which we require in the rest of the paper. Such extensions have also been studied by McNeney and Wellner [9], and Bickel and Kwon [10]. Then information bounds for the regression parameters are derived assuming that separate prospective samples are taken from the case and control populations.

The minor modifications to the standard theory required for the multisample efficiency bounds are sketched in Section 2. This theory is then applied to case-control sampling and an information bound derived in Section 3. We also derive the asymptotic

variance of the Scott-Wild estimator and show that it coincides with the information bound.

In Section 4, we deal with the two-phase sampling scheme. We argue that a sampling scheme, equivalent to the two-phase scheme described above is to regard the data as arising from separate independent sampling from $S+1$ populations. This allows the application of the theory sketched in Section 2. We derive a bound and again show that the asymptotic variance of the Scott-Wild estimator coincides with the bound. Finally, mathematical details are given in Section 5.

In the context of data that are independently and identically distributed, Newey [11] characterises the information bound in terms of a population version of a profile likelihood, rather than a projection. A parallel approach to calculating the information bound for the case-control and two-phase problems, using Newey's "profile" characterisation, is contained in Lee and Hirose [12].

2. Multisamples, information bounds, and semiparametric efficiency

In this section, we give a brief account of the theory of semiparametric efficiency when the data are not independently and identically distributed, but rather consist of separate independent samples from different populations.

Suppose we have J populations. From each population, we independently select separate i.i.d. samples so that for $j = 1,\dots,J$, we have a sample $\{x_{ij},\ i = 1,\dots,n_j\}$ from a distribution with density p_j, say. We call the combined sample a multisample. We will consider asymptotics where $n_j/n \to w_j$, and $n = n_1 + \cdots + n_J$.

Suppose that p_j is a member of the family of densities

$$\mathcal{P} = \{p_j(x,\beta,\eta),\ \beta \in \mathcal{B},\ \eta \in \mathcal{N}\}, \tag{2.1}$$

where $\mathcal{B}$ is a subset of $\mathcal{R}_k$ and $\mathcal{N}$ is an infinite-dimensional set. We denote the true values of β and η by β_0 and η_0, and $p_j(x,\beta_0,\eta_0)$ by p_{j0}. Consider *asymptotically linear* estimates of β of the form

$$\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \phi_j(x_{ij}) + o_p(1), \tag{2.2}$$

where $E_j\phi_j(X) = 0$, E_j denoting expectation with respect to p_{j0}. The functions ϕ_j are called the *influence functions* of the estimate and its asymptotic variance is

$$\sum_{j=1}^{J} w_j E_j[\phi_j \phi_j^T]. \tag{2.3}$$

The *semiparametric information bound* is a matrix $\mathbf{B}$ that is a lower bound for the asymptotic variance of all asymptotically linear estimates of β. We have

$$\operatorname{Avar} \hat{\beta} = \sum_j w_j E_j\left[\phi_j \phi_j^T\right] \geq \mathbf{B}, \tag{2.4}$$

where the ϕ_j are the influence functions of $\hat{\beta}$.

The efficiency bound is found as follows. Let T be a subset of $\mathcal{R}_p$ so that $\mathcal{P}_T = \{p_j(x,\beta, \eta(t)),\ \beta \in \mathcal{B},\ t \in T\}$ is a p-dimensional submodel of $\mathcal{P}$. We also suppose that if η_0 is the true value of η, then $\eta(t_0) = \eta_0$ for some $t_0 \in T$. Thus, the submodel includes the true model, having $\beta = \beta_0$ and $\eta = \eta_0$.

Consider the vector-valued score functions

$$\dot{l}_{j,\eta} = \frac{\partial \log p_j(x,\beta,\eta(t))}{\partial t}, \tag{2.5}$$

whose elements are assumed to be members of $L_2(P_{j0})$, where P_{j0} is the measure corresponding to $p_j(x,\beta_0,\eta_0)$. Consider also the space $L_{2k}(P_{j0})$, the space of all $\mathcal{R}_k$-valued square-integrable functions with respect to P_{j0}, and the Cartesian product $\mathcal{H}$ of these spaces, equipped with the norm defined by

$$||(f_1,\ldots,f_J)||^2_{\mathcal{H}} = \sum_{j=1}^{J} w_j \int ||f_j||^2 dP_{j0}. \tag{2.6}$$

The subspace of $\mathcal{H}$ generated by the score functions $(\dot{l}_{1,\eta},\ldots,\dot{l}_{J,\eta})$ is the set of all vector-valued functions of the form $(\mathbf{A}\dot{l}_{1,\eta},\ldots,\mathbf{A}\dot{l}_{J,\eta})$, where $\mathbf{A}$ ranges over all k by p matrices. Thus, to each finite-dimensional sub-family of $\mathcal{P}$, there correspond a score function and subspace of $\mathcal{H}$ generated by the score function. The closure in $\mathcal{H}$ of the span(over all such subfamilies) of all these subspaces is called the *nuisance tangent space* and denoted by $\mathcal{T}_\eta$.

Consider also the score functions

$$\dot{l}_{\beta,j} = \frac{\partial \log p_j(x,\beta,\eta)}{\partial \beta}. \tag{2.7}$$

The projection $\dot{l}^*$ in $\mathcal{H}$ of $\dot{l}_\beta = (\dot{l}_{\beta,1},\ldots,\dot{l}_{\beta,J})$ onto the orthogonal complement of $\mathcal{T}_\eta$ is called the *efficient score*, and its elements (which are members of $L_{2,k}(P_{j0})$) are denoted by $\dot{l}^*_j$. The matrix $\mathbf{B}$ (the efficiency bound) is given by

$$\mathbf{B}^{-1} = \sum_{j=1}^{J} w_j E_j\left[\dot{l}^*_j \dot{l}^{*T}_j\right]. \tag{2.8}$$

The functions $\mathbf{B}\dot{l}^*_j$ are called the *efficient influence functions*, and any multisample asymptotically linear estimate of β having these influence functions is asymptotically efficient.

3. The efficiency of the Scott-Wild estimator in case-control studies

In this section, we apply the theory sketched in Section 2 to regression models, where the data are obtained by case-control sampling. Suppose that we have a response Y (assumed as discrete with possible values $y_1,\ldots,y_J$) and a vector X of covariates, and we want to model the conditional distribution of Y given X using a regression function

$$f_j(x,\beta) = P(Y = y_j \mid X = x), \tag{3.1}$$

say, where β is a k-vector of parameters. If the distribution of the covariates X is specified by a density g, then the joint distribution of X and Y is

$$f_j(x,\beta)g(x) \tag{3.2}$$

and the conditional distribution of x given $Y = y_j$ is

$$p_j(x,\beta,\eta) = \frac{f_j(x,\beta)g(x)}{\pi_j}, \tag{3.3}$$

where

$$\pi_j = \int f_j(x,\beta)g(x)dx. \tag{3.4}$$

In case-control sampling, the data are not sampled from the joint distribution, but rather from the conditional distributions of X given $Y = y_j$. We are thus in the situation of Section 2 with g playing the role of η and

$$p_j(x,\beta,g) = \frac{f_j(x,\beta)g(x)}{\pi_j}. \tag{3.5}$$

3.1. The information bound in case-control studies. To apply the theory of Section 2, we must identify the nuisance tangent space $\mathcal{T}_\eta$ and calculate the projection of $\dot{l}_\beta$ on this space. Direct calculation shows that

$$\dot{l}_{\beta,j} = \frac{\partial \log f_j(x,\beta)}{\partial\beta} - \mathcal{E}_j\left[\frac{\partial \log f_j(x,\beta)}{\partial\beta}\right], \tag{3.6}$$

where $\mathcal{E}_j$ denotes expectation with respect to the true density p_{j0}, given by $p_{j0}(x) = p_j(x,\beta_0,g_0)$, where β_0 and g_0 are the true values of β and g. Here, and in what follows, all derivatives are evaluated at the true values of parameters.

Also, for any finite-dimensional family $\{g(x,t)\}$ of densities with $g(x,t_0) = g_0(x)$, we have

$$\dot{l}_{\eta,j} = \frac{\partial \log g(x,t)}{\partial t} - \mathcal{E}_j\left[\frac{\partial \log g(x,t)}{\partial t}\right]. \tag{3.7}$$

It follows by the arguments of Bickel et al. [5, page 52] that the nuisance tangent space is of the form

$$\mathcal{T}_\eta = \{(h - \mathcal{E}_1[h],\ldots,h - \mathcal{E}_J[h]) : h \in L_{2,k}(G_0)\}, \tag{3.8}$$

where $dG_0 = g_0 dx$, and $L_{2,k}(G_0)$ is the space of all k-dimensional functions f satisfying the condition $\int \|f\|^2 dG_0(x) < \infty$.

The efficient score, the projection of $\dot{l}_\beta$ on the orthogonal complement of $\mathcal{T}_\eta$, is described in our first theorem. In the theorem, we use the notations $\pi_{j0} = \int f_j(x,\beta_0)\,dG_0(x)$,

$$\begin{aligned} f^*(x) &= \sum_{j=1}^{J} \frac{w_j}{\pi_j} f_j(x), \\ \dot{l}_{\beta,j} &= \left(\dot{l}_{\beta,j1},\ldots,\dot{l}_{\beta,jk}\right)^T, \\ \phi_l(x) &= \sum_{j=1}^{J} \frac{w_j}{\pi_{j0}} \dot{l}_{\beta,jl} f_j(x,\beta_0). \end{aligned} \tag{3.9}$$

Then we have the following result.

Theorem 3.1. *Let A be the operator $L_2(G_0) \to L_2(G_0)$ defined by*

$$(Ah)(x) = f^*(x)h(x) - \sum_{j=1}^{J} \frac{w_j}{\pi_j} f_j(x) \left(\frac{f_j}{\pi_j}, h\right)_2, \tag{3.10}$$

where $(\cdot,\cdot)_2$ is the inner product in $L_2(G_0)$. Then the efficient score has j, l element

$$\dot{l}_{\beta,jl} - h_l^* + E_j[h_l^*], \tag{3.11}$$

where h_l^* is any solution in $L_2(G_0)$ of the operator equation

$$Ah_l^* = \phi_l. \tag{3.12}$$

A proof is given in Section 5.1.

It remains to identify a solution to (3.12). Define $P_j(x) = (w_j/\pi_{j0}) f_j(x,\beta_0)/f^*(x)$ and $v_{jj'} = \int P_j P_{j'} f^*\,dG_0$. Let $\mathbf{V} = (v_{jj'})$, $\mathbf{W} = \operatorname{diag}(w_1,\ldots,w_J)$, and $\mathbf{M} = \mathbf{W} - \mathbf{V}$. Note that the row and column sums of $\mathbf{M}$ are zero since

$$w_j - \sum_{j'=1}^{J} \int P_j P_{j'} f^*\,dG_0 = w_j - \frac{w_j}{\pi_j} \int f_j\,dG_0 = 0. \tag{3.13}$$

Using these definitions and (3.10), we get

$$Ah_l = h_l f^* - \sum_{j=1}^{J} \left(h_l, \frac{f_j}{\pi_j}\right)_2 P_j f^* \tag{3.14}$$

so that $Ah_l = \phi_l$ if and only if

$$h_l = \frac{\phi_l}{f^*} + \sum_{j=1}^{J} \left(h_l, \frac{f_j}{\pi_j}\right)_2 P_j. \tag{3.15}$$

This suggests that h_l^* will be of the form

$$h_l^* = \frac{\phi_l}{f^*} + \sum_{j=1}^{J} c_j P_j \tag{3.16}$$

for some constants $c_1,\ldots,c_J$. In order that h_l^* satisfy (3.12), we must have

$$c_j - \sum_{j'=1}^{J} c_{j'} \left(P_{j'}, \frac{f_j}{\pi_j} \right)_2 - w_j^{-1} (\phi_l, P_j)_2 = 0, \quad j = 1,\ldots,J. \tag{3.17}$$

Now,

$$\left(P_{j'}, \frac{f_j}{\pi_j} \right)_2 = \pi_j^{-1} \int P_{j'} f_j \, dG_0 = w_j^{-1} \int P_{j'} P_j f^* \, dG_0 = (\mathbf{W}^{-1}\mathbf{V})_{jj'} \tag{3.18}$$

so that (3.17) will be satisfied if the vector $c = (c_1,\ldots,c_J)^T$ satisfies

$$\mathbf{M}c = d_{(l)}, \tag{3.19}$$

where $d_l = (d_{1l},\ldots,d_{Jl})^T$ with $d_{jl} = (\phi_l, P_j)_2$. Thus, we require that $c = \mathbf{M}^- d_{(l)}$, where $\mathbf{M}^-$ is a generalised inverse of $\mathbf{M}$.

Our next result gives the information bound.

Theorem 3.2. *Let* $\mathbf{D} = (d_1,\ldots,d_k)$ *and* $\phi = (\phi_1,\ldots,\phi_k)^T$. *The inverse of the information bound* $\mathbf{B}$ *is given by*

$$\mathbf{B}^{-1} = \sum_{j=1}^{J} w_j \mathscr{E}_j [\dot{l}_{\beta,j} \dot{l}_{\beta,j}^T] - \int \frac{\phi\phi^T}{f^*} \, dG_0 - \mathbf{D}^T \mathbf{M}^- \mathbf{D}. \tag{3.20}$$

See Section 5.2 for a proof.

3.2. Efficiency of the Scott-Wild estimator in case-control studies. Suppose that we have J disease states (typically $J = 2$, with disease-state case and control), and we choose n_j individuals at random from disease population j, $j = 1,\ldots,J$, observing covariates $x_{1,j},\ldots,x_{n_j,j}$ for the individuals sampled from population j. Also suppose that we have a regression function $f_j(x,\beta)$, $j = 1,\ldots,J$, giving the conditional probability that an individual with covariates x has disease state j. The unconditional density g of the covariates is unspecified. The true values of β and g are denoted by β_0 and g_0, and the true probability of being in disease state j is $\pi_{j0} = \int f(x,\beta_0) g_0(x) \, dx$.

Under the case-control sampling scheme, the log-likelihood (Scott and Wild [2]) is

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \log f_j(x_{ij},\beta) + \sum_{j=1}^{J} \sum_{i=1}^{n_j} \log g(x_{ij}) - \sum_{j=1}^{J} n_j \log \pi_j. \tag{3.21}$$

Scott and Wild show that the nonparametric MLE of β is the "beta" part of the solution of the estimating equation

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} \frac{\partial \log P_j^*(x_{ij},\beta,\rho)}{\partial\theta} = 0, \tag{3.22}$$

where $\theta = (\beta, \rho)$, $\rho = (\rho_1, \ldots, \rho_{J-1})$,

$$P_j^*(x,\beta,\rho) = \frac{e^{\rho_j} f_j(x,\beta)}{\sum_{l=1}^{J-1} e^{\rho_l} f_l(x,\beta) + f_J(x,\beta)}, \quad j = 1,\ldots,J-1,$$
$$P_J^*(x,\beta,\rho) = \frac{f_J(x,\beta)}{\sum_{l=1}^{J-1} e^{\rho_l} f_l(x,\beta) + f_J(x,\beta)}. \tag{3.23}$$

A Taylor series argument shows that the solution of (3.22) is an asyptotically linear estimate.

Thus, to estimate β, we are treating the function $l^*(\theta) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \log P_j^*(x_{ij},\beta,\rho)$ as though it were a log-likelihood. Moreover, Scott and Wild indicate that we can obtain a consistent estimate of the standard error by using the second derivative $-\partial^2 l^*(\theta)/\partial\theta\partial\theta^T$, which they call the "pseudo-information matrix."

Now let $n = n_1 + \cdots + n_J$, let the n_j's converge to infinity with $n_j/n \to w_j$, $j = 1,\ldots,J$, and let $\rho_0 = (\rho_{01},\ldots,\rho_{0,J-1})^T$, where $\exp(\rho_{0j}) = (w_j/\pi_{0j})/(w_J/\pi_{0J})$. It follows from the law of large numbers and the results of Scott and Wild that the asymptotic variance of the estimate of β is the $\beta\beta$ block of the inverse of the matrix

$$\mathbf{I}^* = -\sum_{j=1}^{J} w_j \mathscr{E}_j\left[\frac{\partial^2 \log P_j^*(x_{ij},\beta,\rho)}{\partial\theta\partial\theta^T}\right], \tag{3.24}$$

where all derivatives are evaluated at (β_0, ρ_0). Using the partitioned matrix inverse formula, the $\beta\beta$ block of $(\mathbf{I}^*)^{-1}$ is

$$\left(\mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho}\left(\mathbf{I}^*_{\rho\rho}\right)^{-1}\mathbf{I}^*_{\rho\beta}\right)^{-1}, \tag{3.25}$$

where $\mathbf{I}^*$ is partitioned as

$$\mathbf{I}^* = \begin{bmatrix} \mathbf{I}^*_{\beta\beta} & \mathbf{I}^*_{\beta\rho} \\ \mathbf{I}^*_{\rho\beta} & \mathbf{I}^*_{\rho\rho} \end{bmatrix}. \tag{3.26}$$

To prove the efficiency of the estimator, we show that the information bound (3.20) coincides with the asymptotic variance (3.25). To prove this, the following representation of the matrix $\mathbf{I}^*$ will be useful. Let $\mathbf{S}$ be the $J \times k$ matrix with j,l element $S_{jl} = (\partial \log f_j(x,\beta)/\partial\beta_l)\,|_{\beta=\beta_0}$ and jth row S_j, and let $\mathbf{E}$ be the $J \times k$ matrix with j,l element $\mathscr{E}_j[S_{jl}]$. Also note that $P_j(x) = P_j^*(x,\beta_0,\rho_0)$ and write $P = (P_1,\ldots,P_S)^T$. Then we have the following theorem.

Theorem 3.3. (1) $\mathbf{I}^*_{\beta\beta} = \sum_{j=1}^{J} w_j \mathscr{E}_j[S_j S_j^T] - \int \mathbf{S}^T P P^T \mathbf{S} f^*\, dG_0$.

(2) *Let* $\mathbf{U} = \mathbf{W}\mathbf{E} - \int PP^T\mathbf{S}f^*\,dG_0$. *Then* $\mathbf{I}^*_{\rho\beta}$ *consists of the first* $J-1$ *rows of* $\mathbf{U}$.

(3) $\mathbf{I}^*_{\rho\rho}$ *consists of the first* $J-1$ *rows and columns of* $\mathbf{M} = \mathbf{W} - \mathbf{V}$.

A proof is given in Section 5.3.

Now we show that the information bound coincides with the asymptotic variance. Using the definition $\phi_l(x) = \sum_{j=1}^{J}(w_j/\pi_{j0})\dot{l}_{\beta,jl} f_j(x,\beta_0)$, we can write $\phi = (\mathbf{S} - \mathbf{E})^T P f^*$,

and substituting this and the relationship $\dot{l}_\beta = \mathbf{S} - \mathbf{E}$ into (3.20), we get

$$\mathbf{B}^{-1} = \sum_{j=1}^{J} w_j E_j \left[S_j S_j^T \right] - \mathbf{E}^T \mathbf{W} \mathbf{E} - \int (\mathbf{S} - \mathbf{E})^T P P^T (\mathbf{S} - \mathbf{E}) f^* \, dG_0(x) - \mathbf{D}^T \mathbf{M}^- \mathbf{D}. \tag{3.27}$$

Moreover,

$$\mathbf{D} = \int P \phi^T \, dG_0(x) = \int P P^T (\mathbf{S} - \mathbf{E}) f^* \, dG_0(x) = \mathbf{W}\mathbf{E} - \mathbf{U} - \mathbf{V}\mathbf{E} = \mathbf{M}\mathbf{E} - \mathbf{U}. \tag{3.28}$$

Substituting this into (3.27) and using the relationships described in Theorem 3.3, we get

$$\mathbf{B}^{-1} = \mathbf{I}^*_{\beta\beta} - \mathbf{U}^T \mathbf{M}^- \mathbf{U} - \mathbf{E}^T (\mathbf{I} - \mathbf{M}\mathbf{M}^-) \mathbf{U} - \mathbf{U}^T (\mathbf{I} - \mathbf{M}^- \mathbf{M}) \mathbf{E}. \tag{3.29}$$

By Theorem 3.3, the matrix

$$\begin{bmatrix} \mathbf{I}^{*\,-1}_{\rho\rho} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{0} \end{bmatrix} \tag{3.30}$$

is a generalised inverse of $\mathbf{M}$, so $\mathbf{U}^T \mathbf{M}^- \mathbf{U} = \mathbf{I}^*_{\beta\rho} \mathbf{I}^{*\,-1}_{\rho\rho} \mathbf{I}^*_{\rho\beta}$. Also,

$$(\mathbf{I} - \mathbf{M}\mathbf{M}^-)\mathbf{U} = (\mathbf{I} - \mathbf{M}\mathbf{M}^-)(\mathbf{M}\mathbf{E} - \mathbf{D}) = (\mathbf{I} - \mathbf{M}\mathbf{M}^-)\mathbf{M}\mathbf{E} - (\mathbf{I} - \mathbf{M}\mathbf{M}^-)\mathbf{M}\mathbf{C} = \mathbf{0} \tag{3.31}$$

by the properties of a generalised inverse. Thus, $\mathbf{B}^{-1} = \mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho} \mathbf{I}^{*\,-1}_{\rho\rho} \mathbf{I}^*_{\rho\beta}$ and the Scott-Wild estimate is efficient.

4. Efficiency of the Scott-Wild estimator under two-stage sampling

In this section, we use the same techniques to show that the Scott-Wild nonparametric MLE is also efficient under two-stage sampling.

4.1. Two stage sampling. In this sampling scheme, the population is divided into S strata, where stratum membership is completely determined by an individual's response y and possibly some of the covariates x—typically those that are cheap to measure. In the first sampling stage, a random sample of size n_0 is taken from the population, and the stratum to which the sampled individuals belong is recorded. For the ith individual, let $Z_{is} = 1$ if the individual is in stratum s, and zero otherwise. Then $n_0^{(s)} = \sum_{i=1}^{n_1} Z_{is}$ is the number of individuals in stratum s. In the second sampling stage, for each stratum s, a simple random sample of size $n_1^{(s)}$ is taken from the $n_0^{(s)}$ individuals in the stratum. Let x_{is}, $i = 1,\ldots,n_1^{(s)}$ and y_{is}, $i = 1,\ldots,n_1^{(s)}$ be the covariates and responses for those individuals. Note that $n_1^{(s)}$ depends on $n_0^{(s)}$ and must be regarded as random since $n_0^{(s)} \geq n_1^{(s)}$ for $s = 1,\ldots,S$. We assume that the distribution of $n_1^{(s)}$ depends only on $n_0^{(s)}$, and that, conditional on the $n_0^{(s)}$'s, the $n_1^{(s)}$'s are independent.

As in Section 3, let $f(y \mid x,\beta)$ be the conditional density of y given x, which depends on a finite number of parameters β, which are the parameters of interest. Let g denote the density of the covariates. We will regard g as an infinite-dimensional nuisance parameter.

The conditional density of (x, y), conditional on being in stratum s, using Bayes theorem, is

$$\frac{I_s(x,y)f(y \mid x,\beta)g(x)}{\iint I_s(x,y)f(y \mid x,\beta)g(x)\,dx\,dy}, \tag{4.1}$$

where $I_s(x, y)$ is the stratum indicator, having value 1 if an individual having covariates x and response y is in stratum s, and zero otherwise. The unconditional probability of being in stratum s in the first phase is

$$Q_s = \iint I_s(x,y)f(y \mid x,\beta)g(x)\,dx\,dy. \tag{4.2}$$

Introduce the function $Q_s(x,\beta) = \int I_s(x,y)f(y \mid x,\beta)\,dy$. Then,

$$Q_s = \int Q_s(x,\beta)g(x)\,dx. \tag{4.3}$$

Under two-phase sampling, the log-likelihood (Wild [3], Scott and Wild [2]) is

$$\sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}} \log f(y_{is} \mid x_{is},\beta) + \sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}} \log g(x_{is}) + \sum_{s=1}^{S} m_s \log Q_s, \tag{4.4}$$

where $m_s = n_0^{(s)} - n_1^{(s)}$. Scott and Wild show that the semiparametric MLE $\hat{\beta}$ (i.e., the "β" part of the maximiser $(\hat{\beta},\hat{g})$ of (4.4)) is equal to the "β" part of the solution of the estimating equations

$$\frac{\partial \ell^*}{\partial \beta} = 0, \qquad \frac{\partial \ell^*}{\partial \rho} = 0. \tag{4.5}$$

The function ℓ^* is given by

$$\ell^*(\beta,\rho) = \sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}} \log f(y_{is} \mid x_{is},\beta) - \sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}} \log\left[\sum_r \mu_r(\rho)Q_r(x_{is},\beta)\right] + \sum_{s=1}^{S} m_s \log Q_s(\rho), \tag{4.6}$$

where $Q_1(\rho),\ldots,Q_S(\rho)$ are probabilities defined by $\sum_{s=1}^{S} Q_s(\rho) = 1$ and $\log Q_s/Q_S = \rho_s$, $s = 1,\ldots,S$, and $\mu_s(\rho) = c(n_0 - m_s/Q_s(\rho))$. The μ_s's depend on the quantity c and the m_s's, and for fixed values of these quantities, they are completely determined by the $S-1$ quantities ρ_s. Note that the estimating equations (4.5) are invariant under choice of c. It will be convenient to take c as N^{-1}, where $N = n_0 + n_1$, where $n_1 = \sum_{s=1}^{S} n_1^{(s)}$.

In order to apply the theory of Section 2 to two-phase sampling, we will prove that the asymptotics under two-phase sampling are the same as those under the following multi-sample sampling scheme.

(1) As in the first scheme, take a random sample of n_0 individuals and record the stratum in which they fall. This amounts to taking an i.i.d. sample $\{(Z_{i1},\ldots,Z_{iS}), i = 1,\ldots,n_0\}$ of size n_0 from MULT$(1,Q_1,\ldots,Q_S)$.

(2) For $s = 1,\ldots,S$, take independent i.i.d. samples $\{(x_{is},y_{is}),\ i = 1,\ldots,n_1^{(s)}\}$ of size $n_1^{(s)}$ from the conditional distribution of (x,y) given s, having density $p_s(x,y,\beta,g) = I_s(x,y)f(y \mid x,\beta)g(x)/Q_s$.

We note that the likelihood under this modified sampling scheme is the same as before, and we show in Theorem 4.1 that the asymptotic distribution of the parameter estimates is also the same. It follows that if an estimate is efficient under the multisampling scheme, it must also be efficient under two-phase sampling.

THEOREM 4.1. *Let $N = n_0 + n_1$, where $n_1 = \sum_{s=1}^{S} n_1^{(s)}$, and suppose that $\sqrt{N}(n_0/N - w_0) \xrightarrow{p} 0$ and $\sqrt{N}(n_1^{(s)}/N - w_s) \xrightarrow{p} 0$, $s = 1,\ldots,S$.*

Let $\hat{\theta}$ be the solution of the estimating equation (4.5), and let θ_0 be the solution to the equation

$$w_0 \mathscr{E}[\psi_0(Z_1,\ldots,Z_s,\theta)] + \sum_{s=1}^{S} \mathscr{E}_s[\psi_s(x,y,\theta)] = 0, \tag{4.7}$$

where $\mathscr{E}_s$ denotes expectation with respect to p_s,

$$\begin{aligned} \psi_0(Z_1,\ldots,Z_s,\theta) &= \frac{\partial}{\partial\theta}\sum_{s=1}^{S} Z_s \log Q_s, \\ \psi_s(x,y,\theta) &= \frac{\partial}{\partial\theta}\left\{\log f(y \mid x,\beta) - \log\left[\sum_s \mu_s Q_s(x,\beta)\right] - \log Q_s\right\}, \quad s = 1,\ldots,S. \end{aligned} \tag{4.8}$$

Then $\sqrt{N}(\hat{\theta} - \theta_0)$ is asymptotically $N(0,(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1})$ under both sampling schemes, where $\mathbf{V} = \sum_{s=0}^{S} w_s E_s[(\psi_s - E_s[\psi_s])(\psi_s - E_s[\psi_s])^T]$ and $\mathbf{I}^* = -\sum_{s=0}^{S} w_s E_s[\partial\psi_s/\partial\theta]$.

A proof is given in Section 5.4.

4.2. The information bound. Now we derive the information bound for two-stage sampling. By the arguments of Section 4.1, the information bound for two-phase sampling is the same as that for the case of independent sampling from the $S+1$ densities $p_s(x,y,\beta,g)$,

where

$$p_s(x,y,\beta,g) = \frac{I_s(x,y)f(y \mid x,\beta)g(x)}{Q_s}, \quad s = 1,\ldots,S,$$
$$p_0(x,y,\beta,g) = Q_1^{Z_1} \cdots Q_J^{Z_J}, \tag{4.9}$$

where $Z_s = I_s(x,y)$ is the sth stratum indicator.

First, we identify the form of the nuisance tangent space (NTS) for this problem. As in Section 3, we see that the score functions for this problem are

$$\dot{l}_0 = \frac{\partial \log p_0(x,y,\beta,g)}{\partial \beta} = \sum_{s=1}^{S} Z_s \mathscr{E}_s[\mathscr{S}],$$
$$\dot{l}_s = \frac{\partial \log p_s(x,y,\beta,g)}{\partial \beta} = \mathscr{S} - \mathscr{E}_s[\mathscr{S}], \quad s = 1,\ldots,S, \tag{4.10}$$

where $\mathscr{S} = \partial \log f(y \mid x,\beta)/\partial\beta$ and $\mathscr{E}_s$ denotes expectation with respect to the true density $p_s(x,y,\beta_0,g_0)$. Similarly, if $g(x,t)$ is a finite-dimensional subfamily of densities, then $\partial \log p_s(x,y,\beta,g(x,t))/\partial t = h - \mathscr{E}_s[h]$, $s = 1,\ldots,S$, and

$$\frac{\partial \log p_0(x,y,\beta,g(x,t))}{\partial t} = \sum_{s=1}^{S} Z_s \mathscr{E}_s[h], \tag{4.11}$$

where $h = \partial \log g(x,t)/\partial t$. Arguing as in Section 3, we see that the NTS consists of all elements of the form

$$T(h) = \left(\sum_{s=1}^{S} Z_s(\mathscr{E}_s[h] - \mathscr{E}[h]), h - \mathscr{E}_1[h], \ldots, h - \mathscr{E}_s[h] \right), \tag{4.12}$$

where $\mathscr{E}$ denotes expectation with respect to G_0.

As before, the efficient score is $\dot{l}^* = \dot{l} - T(h^*)$, where h^* is the element of $L_{2k}(G_0)$ which minimises $\|\dot{l} - T(h)\|^2_{\mathscr{H}}$. An explicit expression for this squared distance is

$$\sum_{j=1}^{k} \left\{ w_0 \sum_{s=1}^{S} \mathscr{E}\left[Z_s \{\mathscr{E}_s[\mathscr{S}_j] - \mathscr{E}_s[h_j] + \mathscr{E}[h_j]\}^2 \right] + \sum_{s=1}^{S} w_s \mathscr{E}_s\left[\{\mathscr{S}_j - \mathscr{E}_s[\mathscr{S}_j] - h_j + \mathscr{E}_s[h_j]\}^2 \right] \right\}, \tag{4.13}$$

where h_j and $\mathscr{S}_j$ are the jth elements of h and $\mathscr{S}$, respectively. To obtain the projection, we must choose h_j to minimise the term in the braces in (4.13). Some algebra shows that this term may be written as

$$(h_j, Ah_j)_2 - 2(h_j, \phi_j)_2 + \sum_{s=1}^{S} (w_0 Q_{s0} - w_s)\mathscr{E}_s[\mathscr{S}_j]^2 + \sum_{s=1}^{S} w_s \mathscr{E}_s[\mathscr{S}_j^2], \tag{4.14}$$

where $Q_{s0} = \int Q(x,\beta_0)g_0(x)\,dx$ is the true value of Q_s, $(\cdot,\cdot)_2$ is the inner product on $L_2(G_0)$, and A is a selfadjoint nonnegative definite operator on $L_2(G_0)$ defined by

$$Ah = Q^*\left\{h - \sum_{r=1}^{S}\sum_{s=1}^{S}(\delta_{rs}-\gamma_{rs})\frac{\int h(x)Q_r(x,\beta_0)g_0(x)\,dx}{Q_{r0}}P_s\right\},$$

$$Q^*(x) = \sum_{s=1}^{S}\frac{w_s}{Q_{s0}}Q_s(x,\beta_0),$$

$$P_s(x) = \frac{(w_s/Q_{s0})Q_j(x,\beta_0)}{Q^*(x)},$$

$$\gamma_{rs} = \begin{cases} \dfrac{w_0Q_r(1-Q_r)}{w_r}, & r = s, \\ -\dfrac{w_0Q_rQ_s}{w_r}, & r \neq s, \end{cases}$$

$$\phi_j(x) = \sum_{s=1}^{S}\frac{w_s}{Q_{s0}}Q_s(x,\beta_0)\frac{\partial \log Q_s(x,\beta)}{\partial\beta_j}\,|_{\beta=\beta_0} - \sum_{s=1}^{S}\sum_{r=1}^{S}Q^*(x)P_r(x)(\delta_{rs}-\gamma_{rs})\mathscr{E}_s(\mathscr{S}_j). \tag{4.15}$$

As in Section 3, (4.14) is minimised when $h_j = h_j^*$, where h_j is a solution of $Ah_j = \phi_j$, which must be of the form

$$h_j^* = \frac{\phi_j}{f^*} + \sum_{r=1}^{S}c_{rj}P_r \tag{4.16}$$

for constants c_{rj} which satisfy the equation

$$c_{rj} - \sum_{s=1}^{S}\sum_{t=1}^{S}\frac{(\delta_{rs}-\gamma_{rs})}{w_s}v_{st}c_{tj} = \sum_{s=1}^{S}\frac{(\delta_{rs}-\gamma_{rs})}{w_s}d_{sj}, \tag{4.17}$$

where $v_{rs} = \int P_rP_sQ^*\,dG_0$ and $d_{sj} = (P_s,\phi_j)_2$. Writing $\mathbf{\Gamma} = (\gamma_{rs})$, $\mathbf{C} = (c_{rj})$, $\mathbf{D} = (d_{rj})$, $\mathbf{W} = \operatorname{diag}(w_1,\ldots,w_S)$, and $\mathbf{V} = (v_{rs})$, (4.17) can be expressed in matrix terms as

$$\mathbf{MC} = \mathbf{D}, \tag{4.18}$$

where $\mathbf{M} = \mathbf{W}(\mathbf{I}-\mathbf{\Gamma})^{-1} - \mathbf{V}$. These results allow us to find the efficient score and hence the information bound, which is described in the following theorem.

Theorem 4.2. *The information bound* $\mathbf{B}$ *is given by*

$$\mathbf{B}^{-1} = \sum_{s=1}^{S}w_s\mathscr{E}_s[\mathscr{S}\mathscr{S}^T] + \sum_{s=1}^{S}(w_0Q_{s0}-w_s)\mathscr{E}_s[\mathscr{S}]\mathscr{E}_s[\mathscr{S}]^T - \int\frac{\phi\phi^T}{Q^*}\,dG_0(x) - \mathbf{D}^T\mathbf{M}^{-}\mathbf{D}. \tag{4.19}$$

The proof is similar to that of Theorem 3.2 and hence omitted.

4.3. Efficiency of the Scott-Wild estimator. Let $\widehat{\theta} = (\widehat{\beta}, \widehat{\rho})$ be the solutions of the estimating equations (4.5). By Theorem 4.1, under suitable regularity conditions, $\widehat{\theta}$ is asymptotically normal with asymptotic variance

$$\mathbf{I}^{*-1}\mathbf{V}\mathbf{I}^{*-1}, \tag{4.20}$$

where $\mathbf{I}^*$ and $\mathbf{V}$ are as in Theorem 4.1. It turns out that the matrix $\mathbf{V}$ is of the form

$$\mathbf{V} = \mathbf{I}^* - \mathbf{I}^* \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{A} \end{pmatrix} \mathbf{I}^* \tag{4.21}$$

for some matrix $\mathbf{A}$. Thus, the asymptotic variance of $\widehat{\theta}$ is

$$\mathbf{I}^{*-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{A} \end{pmatrix}, \tag{4.22}$$

and it follows from the partitioned matrix inverse formula that the asymptotic variance matrix of $\widehat{\beta}$ is the inverse of

$$\mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho}(\mathbf{I}^*_{\rho\rho})^{-1}\mathbf{I}^*_{\rho\beta}, \tag{4.23}$$

where $\mathbf{I}^*$ is partitioned as

$$\mathbf{I}^* = \begin{bmatrix} \mathbf{I}^*_{\beta\beta} & \mathbf{I}^*_{\beta\rho} \\ \mathbf{I}^*_{\rho\beta} & \mathbf{I}^*_{\rho\rho} \end{bmatrix}. \tag{4.24}$$

To demonstrate the efficiency of $\widehat{\beta}$, we must show that (4.23) and (4.19) coincide. To do this, we need a more explicit formula for $\mathbf{I}^*$. Let $\mathbf{S}$ be the $S \times k$ matrix with s, j element $(\partial \log Q_s(x,\beta_j)/\partial\beta)\,|_{\beta=\beta_0}$, and let $\mathbf{E}$ be the $S \times k$ matrix with lth row $E_s = \mathscr{E}_s[\mathcal{S}]$, where $\mathcal{S} = (\partial \log f(y \mid x,\beta)/\partial\beta)\,|_{\beta=\beta_0}$. Also define

$$P^*_s(x,\beta,\rho) = \frac{\mu_s(\rho)Q_s(x,\beta)}{\sum_{r=1}^{S}\mu_r(\rho)Q_r(x,\beta)} \tag{4.25}$$

and note that $P_s(x) = P^*_s(x,\beta_0,\rho_0)$, where ρ_0 satisfies $Q_s(\rho_0) = Q_{s0}$, $s = 1,\ldots,S$. Finally, write $P = (P_1,\ldots,P_S)^T$. Then we have the following theorem.

Theorem 4.3.

(1) $\mathbf{I}^*_{\beta\beta} = \sum_{s=1}^{S} w_s \mathscr{E}_s[\mathcal{S}\mathcal{S}^T] - \int \mathbf{S}^T PP^T \mathbf{S} Q^*\, dG_0(x)$.

(2) *Let* $\mathbf{U} = \mathbf{W}\mathbf{E} - \int PP^T\mathbf{S}Q^*\, dG_0(x)$. *Then* $\mathbf{I}^*_{\rho\beta} = \mathbf{A}^T\mathbf{U}_0$, *where* $\mathbf{U}_0$ *consists of the first* $S-1$ *rows of* $\mathbf{U}$ *and* $\mathbf{A}$ *is a nonsingular* $(S-1)\times(S-1)$ *matrix.*

(3) $\mathbf{I}^*_{\rho\rho} = \mathbf{A}^T\mathbf{M}_0\mathbf{A}$, *where* $\mathbf{M}_0$ *consists of the first* $S-1$ *rows and columns of* $\mathbf{M} = \mathbf{W}(\mathbf{I} - \mathbf{\Gamma})^{-1} - \mathbf{V}$.

The proof is given in Section 5.5.

We now use Theorems 4.1 and 4.2 to show that the efficiency bound (4.19) equals the asymptotic variance (4.23). Arguing as in Section 3, we get

$$\mathbf{B}^{-1} = \mathbf{I}^*_{\beta\beta} - \mathbf{I}^*_{\beta\rho}\mathbf{I}^{*\,-1}_{\rho\rho}\mathbf{I}^*_{\rho\beta} + \left\{ \sum_{s=1}^{S} (w_0 Q_{s0} - w_s) E_s E_s^T + \mathbf{E}^T \mathbf{W}(\mathbf{I} - \boldsymbol{\Gamma})\mathbf{E} \right\}. \tag{4.26}$$

We complete the argument by showing that the term in the braces in (4.26) is zero. We have

$$\begin{aligned} \mathbf{E}^T \mathbf{W}(\mathbf{I} - \boldsymbol{\Gamma})\mathbf{E}^T &= \sum_{s=1}^{S} (w_s - w_0 Q_{s0}) E_s E_s^T + w_0 \left(\sum_{s=1}^{S} Q_{s0} E_s \right) \left(\sum_{s=1}^{S} Q_{s0} E_s \right)^T \\ &= \sum_{s=1}^{S} (w_s - w_0 Q_{s0}) E_s E_s^T \end{aligned} \tag{4.27}$$

since $\sum_{s=1}^{S} Q_{s0} E_s = 0$. Since the term in the braces in (4.26) is zero, the asymptotic variance coincides with the information bound and so the Scott-Wild estimator has full semiparametric efficiency.

5. Proofs

5.1. Proof of Theorem 3.1. The efficient score is the projection of $\dot{l}_\beta$ onto $\mathcal{T}_\eta^\perp$, and so it is of the form $\dot{l}_\beta - g$, where g is the unique minimiser of $\|\dot{l}_\beta - g\|^2_{\mathcal{H}}$ in $\mathcal{T}_\eta$. By (3.8), this is $\dot{l}_\beta - T(h^*)$, where h^* is the (unique) minimiser of $\|\dot{l}_\beta - T(h)\|^2_{\mathcal{H}}$ in $L_{2,k}(G_0)$. Write $h^* = (h_1^*, \ldots, h_k^*)$. Then,

$$\|\dot{l}_\beta - T(h^*)\|^2_{\mathcal{H}} = \sum_{l=1}^{k} \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int \left(\dot{l}_{\beta,jl} - h_l^* - E_j[h_l^*]\right)^2 f_j \, dG_0 \tag{5.1}$$

so that we must choose h_l^* to minimise

$$\sum_{j=1}^{J} \frac{w_j}{\pi_j} \int \left(\dot{l}_{\beta,jl} - h_l^* - E_j[h_l^*]\right)^2 f_j \, dG_0 = \sum_{j=1}^{J} w_j E_j\left[\dot{l}^2_{\beta,jl}\right] + (Ah_l^*, h_l^*)_2 - 2(\phi_l, h_l^*)_2. \tag{5.2}$$

Now let h_l^* be any solution in $L_2(G_0)$ to (3.12). Then for any h in $L_2(G_0)$, using the fact that A is selfadjoint and positive-definite, we get

$$\begin{aligned} \sum_{j=1}^{J} w_j E_j\left[\dot{l}^2_{\beta,jl}\right] + (Ah, h)_2 - 2(\phi_l, h)_2 &= \sum_{j=1}^{J} w_j E_j\left[\dot{l}^2_{\beta,jl}\right] - (Ah_l^*, h_l^*)_2 + (h - h_l^*, A(h - h_l^*))_2 \\ &\geq \sum_{j=1}^{J} w_j E_j(\dot{l}^2_{\beta,jl}) - (Ah_l, h_l^*)_2 \end{aligned} \tag{5.3}$$

with equality if $h = h_l^*$ so that the efficient score has j, l element $S_{\beta,jl} - h_l^* + E_j[h_l^*]$ as asserted.

5.2. Proof of Theorem 3.2. The l, l' element of $\mathbf{B}^{-1}$ is

$$\begin{aligned}
\sum_{j=1}^{J} w_j E_j\left[\dot{l}^*_{\beta,jl}\dot{l}^*_{\beta,jl'}\right] &= \sum_{j=1}^{J}\frac{w_j}{\pi_j}\int\left(\dot{l}_{\beta,jl}-h^*_l-E_j(h^*_l)\right)\left(\dot{l}_{\beta,jl'}-h^*_{l'}-E_j(h^*_{l'})\right)f_j\,dG_0 \\
&= \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}\dot{l}_{\beta,jl'}] + (Ah^*_l,h^*_{l'})_2 - (\phi_l,h^*_{l'})_2 - (\phi_{l'},h^*_l)_2 \\
&= \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}\dot{l}_{\beta,jl'}] - (\phi_l,h^*_{l'})_2 \\
&= \sum_{j=1}^{J} w_j E_j[\dot{l}_{\beta,jl}\dot{l}_{\beta,jl'}] - \int\frac{\phi_l\phi_{l'}}{f^*}\,dG_0 - d^T_{(l)}\mathbf{M}^{-}d_{(l')}.
\end{aligned} \tag{5.4}$$

5.3. Proof of Theorem 3.3. First, we note the formula

$$\frac{\partial^2\log P^*_j}{\partial\theta\partial\theta^T} = \frac{\partial^2 P^*_j}{\partial\theta\partial\theta^T}\frac{1}{P^*_j} - \frac{\partial\log P^*_j}{\partial\theta}\frac{\partial\log P^*_j}{\partial\theta^T} \tag{5.5}$$

and the fact that

$$\begin{aligned}
\sum_{j=1}^{J} w_j E_j\left[\frac{\partial^2 P^*_j}{\partial\theta\partial\theta^T}\frac{1}{P^*_j}\right] &= \sum_{j=1}^{J}\frac{w_j}{\pi_j}\int\frac{\partial^2 P^*_j}{\partial\theta\partial\theta^T}\frac{1}{P^*_j}f_j\,dG_0(x) \\
&= \sum_{j=1}^{J}\int\frac{\partial^2 P^*_j}{\partial\theta\partial\theta^T}f^*\,dG_0(x) \\
&= \frac{\partial^2}{\partial\theta\partial\theta^T}\int f^*\,dG_0(x) \\
&= 0
\end{aligned} \tag{5.6}$$

since $\sum_{j=1}^{J} P^*_j = 1$. Hence

$$\mathbf{I}^* = -\sum_{j=1}^{J} w_j E_j\left[\frac{\partial^2\log P^*_j}{\partial\theta\partial\theta^T}\right] = \sum_{j=1}^{J} w_j E_j\left[\frac{\partial\log P^*_j}{\partial\theta}\frac{\partial\log P^*_j}{\partial\theta^T}\right]. \tag{5.7}$$

Next, we note the derivatives

$$\begin{aligned}
\frac{\partial\log P^*_j(x,\beta,\rho)}{\partial\beta} &= S_j - \sum_{s=1}^{J} S_s P_s, \\
\frac{\partial\log P^*_j(x,\beta,\rho)}{\partial\rho_r} &= \delta_{j,r} - P_r,
\end{aligned} \tag{5.8}$$

when the derivatives are evaluated at (β_0, ρ_0). Thus

$$\begin{aligned}
\mathbf{I}^*_{\beta\beta} &= \sum_{j=1}^{J} w_j E_j \left[\frac{\partial \log P_j^*}{\partial \beta} \frac{\partial \log P_j^*}{\partial \beta^T} \right] \\
&= \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int \left(S_j - \sum_{s=1}^{J} S_s P_s \right) \left(S_j - \sum_{s=1}^{J} S_s P_s \right)^T f_j(x)\, dG_0(x) \\
&= \sum_{j=1}^{J} w_j E_j \left[S_j S_j^T \right] - \int \left(\sum_{s=1}^{J} S_s P_s \right) \left(\sum_{s=1}^{J} S_s P_s \right)^T f^*(x)\, dG_0(x) \\
&= \sum_{j=1}^{J} w_j E_j \left[S_j S_j^T \right] - \int \mathbf{S}^T P P^T \mathbf{S} f^*\, dG_0(x),
\end{aligned} \tag{5.9}$$

which proves part 1. Also

$$\begin{aligned}
\mathbf{I}^*_{\rho\beta,r} &= \sum_{j=1}^{J} w_j E_j \left[\frac{\partial \log P_j^*}{\partial \rho_r} \frac{\partial \log P_j^*}{\partial \beta} \right] \\
&= \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int (\delta_{r,s} - P_r) \left(S_j - \sum_{s=1}^{J} S_s P_s \right) f_j(x)\, dG_0(x) \\
&= w_r E_r [S_{rl}] - \int \left(\sum_{j=1}^{J} S_j P_j \right) P_r f^*(x)\, dG_0(x),
\end{aligned} \tag{5.10}$$

which proves part 2. Finally,

$$\begin{aligned}
\mathbf{I}^*_{\rho\rho,rs} &= \sum_{j=1}^{J} w_j E_j \left[\frac{\partial \log P_j^*}{\partial \rho_r} \frac{\partial \log P_j^*}{\partial \rho_s} \right] \\
&= \sum_{j=1}^{J} \frac{w_j}{\pi_j} \int (\delta_{jr} - P_r)(\delta_{js} - P_s) f_j(x)\, dG_0(x) \\
&= \int (\delta_{rs} - P_s) P_r f^*(x)\, dG_0(x) \\
&= \delta_{rs} w_r - v_{rs} \\
&= M_{rs}.
\end{aligned} \tag{5.11}$$

5.4. Proof of Theorem 4.1. Under the two-stage sampling scheme, the joint distribution of $\{n_0^{(s)}\}$, $\{n_1^{(s)}\}$ and $\{(x_{is}, y_{is}),\ i = 1,\ldots,n_1^{(s)},\ s = 1,\ldots,S\}$ (Wild [3]) is

$$\begin{aligned}
&\prod_{s=1}^{S} P\left[n_1^{(s)} \mid n_0^{(s)} \right] \times \frac{n_0!}{n_0^{(1)}! \cdots n_0^{(S)}!} Q_1^{n_0^{(1)}} \cdots Q_S^{n_0^{(S)}} \\
&\qquad \times \prod_{s=1}^{S} \frac{\left\{ \prod_{i=1}^{n_1^{(s)}} I_s(x_{is}, y_{is}) f(y_{is} \mid x_{is}, \beta) g(x_{is}) \right\}}{Q_s^{n_1^{(s)}}}.
\end{aligned} \tag{5.12}$$

Thus, conditional on the $\{n_0^{(s)}\}$ and $\{n_1^{(s)}\}$, the random variables $\{(x_{is}, y_{is}),\ i = 1,\ldots,n_1^{(s)},\ s = 1,\ldots,S\}$ are independent, with $\{(x_{is}, y_{is}),\ i = 1,\ldots,n_1^{(s)}\}$ being an i.i.d. sample from the conditional distribution of (x, y), conditional on being in stratum s, having density

$$p_s(x,y,\beta,g) = \frac{I_s(x,y)f(y \mid x,\beta)g(x)}{Q_s}. \tag{5.13}$$

Define

$$\begin{aligned}\psi_s^{(N)}(x,y,\theta) &= \frac{\partial}{\partial\theta}\left\{\log f(y \mid x,\beta) - \log\left[\sum_s \mu_s Q_s(x,\beta)\right] - \log Q_s\right\}, \quad s = 1,\ldots,S,\\ \psi_0^{(N)}(Z_1,\ldots,Z_s,\theta) &= \frac{\partial}{\partial\theta}\sum_{s=1}^{S} Z_s \log Q_s.\end{aligned} \tag{5.14}$$

Then the estimating equations (4.5) can be written in the form

$$\sum_{i=1}^{n_0}\psi_0^{(n_0)}(Z_{i1},\ldots,Z_{is},\theta) + \sum_{s=1}^{S}\sum_{i=1}^{n_0^{(s)}}\psi_s^{(n_0)}(x_{is},y_{is},\theta) = 0. \tag{5.15}$$

Note that the functions $\psi_s^{(N)}$ depend on N, the $n_1^{(s)}$'s and the $n_0^{(s)}$'s through the μ_s's, and the Q_s's. As $N \to \infty$, the functions converge to

$$\begin{aligned}\psi_s(x,y,\theta) &= \frac{\partial}{\partial\theta}\left\{\log f(y \mid x,\beta) - \log\left[\sum_s \mu_s Q_s(x,\beta)\right] - \log Q_s\right\}, \quad s = 1,\ldots,S,\\ \psi_0(x,y,\theta) &= \frac{\partial}{\partial\theta}\sum_{s=1}^{S} Z_s \log Q_s,\end{aligned} \tag{5.16}$$

where $\mu_s = w_0 - (w_0 Q_{s0} - w_s)/Q_s$.

Put

$$S_N(\theta) = \sum_{i=1}^{n_0}\psi_0^{(N)}(Z_{i1},\ldots,Z_{iS},\theta) + \sum_{s=1}^{S}\sum_{i=1}^{n_1^{(s)}}\psi_s^{(N)}(x_{is},y_{is},\theta). \tag{5.17}$$

A standard Taylor expansion argument gives

$$\sqrt{N}(\hat\theta - \theta_0) = \left(-\frac{1}{N}\frac{\partial S_N}{\partial\theta}\bigg|_{\theta=\theta_0}\right)^{-1}\frac{1}{\sqrt{N}}S(\theta_0) + \frac{1}{\sqrt{N}}\left(-\frac{1}{N}\frac{\partial S_N}{\partial\theta}\bigg|_{\theta=\theta_0}\right)^{-1}R, \tag{5.18}$$

where the jth element of R is

$$R_j = \frac{1}{2}(\hat\theta - \theta_0)^T\frac{\partial^2 N_j}{\partial\theta\partial\theta^T}\bigg|_{\theta=\bar\theta}(\hat\theta - \theta_0), \tag{5.19}$$

S_{Nj} is the jth element of S_N and $\|\bar\theta - \theta_0\| \le \|\hat\theta - \theta_0\|$.

Consider first $S_N(\theta_0)/\sqrt{N}$. We have

$$\begin{aligned}\frac{S_N(\theta_0)}{\sqrt{N}} &= \sqrt{\frac{n_0}{N}}\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\left\{\psi_0^{(N)}(Z_{i1},\ldots,Z_{iS},\theta_0)-\mathscr{E}[\psi_0]\right\}\\ &\quad+\sum_{s=1}^{S}\sqrt{\frac{n_1^{(s)}}{N}}\frac{1}{\sqrt{n_1^{(s)}}}\sum_{i=1}^{n_1^{(s)}}\left\{\psi_s^{(N)}(x_{is},y_{is},\theta)-\mathscr{E}_s[\psi_s]\right\}\\ &\quad+\sqrt{N}\sum_{s=1}^{S}\left(\frac{n_0}{N}-w_0\right)\mathscr{E}[\psi_0]+\sqrt{N}\sum_{s=1}^{S}\left(\frac{n_1^{(s)}}{N}-w_s\right)\mathscr{E}_s[\psi_s].\end{aligned}\tag{5.20}$$

Since $\sqrt{N}(n_0/(N)-w_0)$ and $\sqrt{N}(n_1^{(s)}/(N)-w_s)$ converge to zero in probability, we see that

$$\begin{aligned}\frac{S(\theta_0)}{\sqrt{N}} &= \sqrt{w_0}\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\sum_{s=1}^{S}\left\{\psi_0^{(N)}(Z_{is},\theta_0)-\mathscr{E}[\psi_0]\right\}\\ &\quad+\sum_{s=1}^{S}\sqrt{w_s}\frac{1}{\sqrt{n_0^{(s)}}}\sum_{i=1}^{n_0^{(s)}}\left\{\psi_s^{(N)}(x_{is},y_{is},\theta)-\mathscr{E}_s[\psi_s]\right\}+o_p(1).\end{aligned}\tag{5.21}$$

So it suffices to consider $S_N = S_N^{(1)} + S_N^{(2)}$, where $S_N^{(1)}$ and $S_N^{(2)}$ are the first and second terms above.

Under the alternative multisampling scheme, $S_N^{(1)}$ and $S_N^{(1)}$ are independent, as are the S summands of $S_N^{(2)}$. Thus, by the CLT, provided $\psi_s^{(N)}$ converges to ψ_s sufficiently quickly, we see that S_N is asymptotically normal with zero mean and asymptotic variance $\mathbf{V} = \sum_{s=0}^{S} w_s \operatorname{Var}\psi_s$.

Conversely, under two-phase sampling, the characteristic function of S_N is

$$E\left[e^{itS_N}\right] = \sum_{(0)} E\left[e^{itS_N} \mid \left\{n_0^{(s)}\right\},\left\{n_1^{(s)}\right\}\right]P\left[\left\{n_0^{(s)}\right\},\left\{n_1^{(s)}\right\}\right],\tag{5.22}$$

where $\sum_{(0)}$ denotes summation over all possible values of the $\{n_0^{(s)}\}$ and $\{n_1^{(s)}\}$. Since $S_N^{(2)}$ depends on $\{n_0^{(s)}\}$ only through $\{n_1^{(s)}\}$, (5.22) equals

$$E\left[e^{itS_N}\right] = \sum_{(0)} E\left[e^{itS_N^{(1)}}E\left[e^{itS_N^{(2)}} \mid \left\{n_1^{(s)}\right\}\right]\right]P\left[\left\{n_0^{(s)}\right\},\left\{n_1^{(s)}\right\}\right].\tag{5.23}$$

Let $\mathbf{V}_2 = \sum_{s=1}^{S} w_s \operatorname{Var}[\psi_s]$. Assuming that the $\psi_s^{(N)}$ converge sufficiently quickly to the ψ_s, it follows that $E[e^{itS_N^{(2)}} \mid \{n_1^{(s)}\}] \to \exp\{-(1/2)t^T\mathbf{V}_2 t\}$ since the distribution of $S_N^{(2)}$, conditional on $\{n_0^{(s)}\}$ and $\{n_1^{(s)}\}$, is the same as that (unconditionally) under multisampling.

Now let ϵ be arbitrary and let N_0 be such that

$$\left|E\left[e^{itS_N^{(2)}} \mid \left\{n_1^{(s)}\right\}\right] - \exp\left\{-\frac{1}{2}t^T\mathbf{V}_2 t\right\}\right| < \frac{\epsilon}{2},\tag{5.24}$$

whenever $n_1^{(s)} \geq N_0$ for $s = 1,\ldots,S$. Also, assume that the (random) sample sizes ultimately get large, in the sense that there exists N^* such that

$$P\left[n_1^{(1)} \geq N_0,\ldots,n_S^{(1)} \geq N_0\right] \geq 1 - \frac{\epsilon}{4}, \tag{5.25}$$

whenever $N > N^*$. Denote by $\sum_{(1)}$ summation over all values of $\{n_0^{(s)}\}$ and $\{n_1^{(s)}\}$ for which $n_1^{(s)} \geq N_0$ for $s = 1,\ldots,S$, and let $\sum_{(2)}$ denote summation over all remaining values. Then,

$$\begin{aligned} E\left[e^{itS_N}\right] = E\left[e^{itS_N^{(1)}}\right]\exp\left\{-\frac{1}{2}t^T\mathbf{V}_2t\right\} &+ \sum_{(1)}E\left[e^{itS_N^{(1)}}\left(E\left[e^{itS_N^{(2)}} \mid \left\{n_1^{(s)}\right\}\right] - \exp\left\{-\frac{1}{2}t^T\mathbf{V}_2t\right\}\right)\right] \\ &+ \sum_{(2)}E\left[e^{itS_N^{(1)}}\left(E\left[e^{itS_N^{(2)}} \mid \left\{n_1^{(s)}\right\}\right] - \exp\left\{-\frac{1}{2}t^T\mathbf{V}_2t\right\}\right)\right]. \end{aligned} \tag{5.26}$$

If $n_0 > N^*$, the sum of the second two terms is less than ϵ in absolute value. So

$$E\left[e^{itS_N}\right] = E\left[e^{itS_N^{(1)}}\right]\exp\left\{-\frac{1}{2}t^T\mathbf{V}_2t\right\} + o(1). \tag{5.27}$$

Again by the same arguments as above, $[e^{itS_N^{(1)}}]$ converges to $\exp\{-(1/2)t^T\mathbf{V}_1t\}$, where $\mathbf{V}_1$ is $w_0\,\mathrm{Var}[\psi_0(Z_1,\ldots,Z_S,\theta_0)]$ so that $E[e^{itS_N}]$ converges to $\exp\{-(1/2)t^T\mathbf{V}t\}$, and hence S_N converges in distribution to a multivariate normal with variance $\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2$.

Assuming that $\hat{\theta}$ is $\sqrt{N}$-consistent, similar arguments show that $-(1/N)(\partial S/\partial\theta)|_{\theta=\theta_0}$ converges in probability to $\mathbf{I}^*$ under both sampling schemes, and that $R/\sqrt{N}$ is $o_p(1)$. Thus, as asserted, in both cases, $\sqrt{N}(\hat{\theta} - \theta_0)$ converges to a multivariate normal with variance $(\mathbf{I}^*)^{-1}\mathbf{V}(\mathbf{I}^*)^{-1}$.

5.5. Proof of Theorem 4.3. Let

$$P_s^\dagger(x,y,\beta,\rho) = \frac{\mu_s(\rho)I_s(x,y)f(y \mid x,\beta)}{\sum_r \mu_r(\rho)Q_r(x,\beta)}. \tag{5.28}$$

From the definition of $\mathbf{I}^*$ in Theorem 4.1 and the law of large numbers, we get

$$\begin{aligned} \mathbf{I}^* &= -w_0\mathscr{E}\left[\sum_{s=1}^S Z_s\frac{\partial^2\log Q_s}{\partial\theta\partial\theta^T}\right] - \sum_{s=1}^S w_s\mathscr{E}_s\left[\frac{\partial^2\log P_s^\dagger}{\partial\theta\partial\theta^T} - \frac{\partial^2\log Q_s\mu_s}{\partial\theta\partial\theta^T}\right] \\ &= \sum_{s=1}^S w_s\mathscr{E}_s\left[\frac{\partial\log P_s^\dagger}{\partial\theta}\frac{\partial\log P_s^\dagger}{\partial\theta^T}\right] - \sum_{s=1}^S w_s\mathscr{E}_s\left[\frac{1}{P_s^\dagger}\frac{\partial^2 P_s^\dagger}{\partial\theta\partial\theta^T}\right] \\ &\quad + \sum_{s=1}^S w_s\frac{\partial^2\log Q_s\mu_s}{\partial\theta\partial\theta^T} - \sum_{s=1}^S w_0Q_{s0}\frac{\partial^2\log Q_s}{\partial\theta\partial\theta^T}. \end{aligned} \tag{5.29}$$

The second term of this expression is zero since

$$\begin{aligned}\sum_{s=1}^{S} w_s \mathscr{E}_s\left[\frac{1}{P_s^\dagger}\frac{\partial^2 P_s^\dagger}{\partial\theta\partial\theta^T}\right] &= \sum_{s=1}^{S}\int \frac{\partial^2}{\partial\theta\partial\theta^T}\int P_s^\dagger\, dy Q^*\, dG_0(x)\\ &= \sum_{s=1}^{S}\frac{\partial^2}{\partial\theta\partial\theta^T}\int P_s Q^*\, dG_0(x)\\ &= \frac{\partial^2}{\partial\theta\partial\theta^T}\int Q^*\, dG_0(x)\\ &= 0.\end{aligned} \tag{5.30}$$

Now, we evaluate $\mathbf{I}^*_{\beta\beta}$. For the $\beta\beta$ submatrix, the third and fourth terms of (5.29) are zero. Thus, using the derivative

$$\frac{\partial P_s^\dagger}{\partial\beta} = \mathcal{S} - \mathbf{S}^T P, \tag{5.31}$$

we get

$$\begin{aligned}\mathbf{I}^*_{\beta\beta} &= \sum_{s=1}^{S} w_s \mathscr{E}_s\left[\frac{\partial \log P_s^\dagger}{\partial\beta}\frac{\partial \log P_s^\dagger}{\partial\beta^T}\right]\\ &= \sum_{s=1}^{S}\frac{w_s}{Q_{s0}}\iint (\mathcal{S}-\mathbf{S}^T P)(\mathcal{S}-\mathbf{S}^T P)^T I_s(x,y) f(y\mid x,\beta_0)\, dy\, dG_0(x)\\ &= \sum_{s=1}^{S}\frac{w_s}{Q_{s0}}\iint \mathcal{S}\mathcal{S}^T I_s(x,y) f(y\mid x,\beta_0)\, dy\, dG_0(x) - \int \mathbf{S}^T P(\mathbf{S}^T P)^T Q^*(x)\, dG_0(x)\\ &= \sum_{s=1}^{S} w_s \mathscr{E}_s[\mathcal{S}\mathcal{S}^T] - \int \mathbf{S}^T P P^T \mathbf{S} Q^*\, dG_0(x),\end{aligned} \tag{5.32}$$

which proves part 1.

Now, consider $\mathbf{I}^*_{\rho\beta,rj}$. Again, the third and fourth terms of (5.29) are zero. Introduce the parameters $\lambda_1,\ldots,\lambda_{S-1}$ defined by

$$\lambda_r = \log\left(\frac{\mu_r(\rho)}{\mu_S(\rho)}\right), \quad r = 1,\ldots,S-1. \tag{5.33}$$

Then,

$$\frac{\partial P_s^\dagger}{\partial\rho_r} = \sum_{p=1}^{S-1}\frac{\partial\lambda_p}{\partial\rho_r}\frac{\partial P_s^\dagger}{\partial\lambda_p} = \sum_{p=1}^{S-1}\frac{\partial\lambda_p}{\partial\rho_r}(\delta_{sp} - P_p). \tag{5.34}$$

Thus,

$$\begin{aligned}\mathbf{I}^*_{\rho\beta,rj} &= \sum_{s=1}^{S} w_s \mathscr{E}_s\left[\frac{\partial \log P_s^\dagger}{\partial \rho_r}\frac{\partial \log P_s^\dagger}{\partial \beta_j}\right] \\ &= \sum_{s=1}^{S}\frac{w_s}{Q_{s0}}\iint\left[\sum_{p=1}^{S-1}\frac{\partial \lambda_p}{\partial \rho_r}(\delta_{sp}-P_p)\right](\mathcal{S}-\mathbf{S}P)_j I_s(x,y) f(y\mid x,\beta_0)\,dy\,dG_0(x) \\ &= \sum_{p=1}^{S-1}\frac{\partial \lambda_p}{\partial \rho_r}u_{pj},\end{aligned} \tag{5.35}$$

where

$$u_{pj} = \sum_{s=1}^{S}\frac{w_s}{Q_{s0}}\iint(\delta_{ps}-P_p)(\mathcal{S}-\mathbf{S}P)_j I_s(x,y) f(y\mid x,\beta_0)\,dy\,dG_0(x). \tag{5.36}$$

Then, as in Theorem 3.3, we see that u_{pj} is the p, j element of $\mathbf{U}$, and so part 2 of the theorem is true with $\mathbf{A}_{pr} = \partial\lambda_p/\partial\rho_r$.

The $\rho\rho$ submatrix is

$$\begin{aligned}\mathbf{I}^*_{\rho\rho} &= \sum_{s=1}^{S} w_s E_s\left[\frac{\partial \log P_s^\dagger}{\partial \rho}\frac{\partial \log P_s^\dagger}{\partial \rho^T}\right] - \sum_{s=1}^{S} w_0 Q_{s0}\frac{\partial^2 \log Q_s}{\partial \rho \partial \rho^T} + \sum_{s=1}^{S} w_s \frac{\partial^2 \log Q_s \mu_s}{\partial \rho \partial \rho^T} \\ &= \sum_{s=1}^{S} w_s E_s\left[\frac{\partial \log P_s^\dagger}{\partial \rho}\frac{\partial \log P_s^\dagger}{\partial \rho^T}\right] - w_0\sum_{s=1}^{S}\frac{1}{\kappa_s}\frac{\partial Q_s}{\partial \rho}\frac{\partial Q_s}{\partial \rho^T},\end{aligned} \tag{5.37}$$

where $\kappa_s = Q_{s0}w_s/c_s$. It follows from (5.34) that $\mathbf{I}^*_{\rho\rho} = \mathbf{A}^T\mathbf{M}_0\mathbf{A}$, where $\mathbf{M}_0$ has p, q element

$$\sum_{s=1}^{S} w_s E_s\left[\frac{\partial \log P_s^\dagger}{\partial \lambda_p}\frac{\partial \log P_s^\dagger}{\partial \lambda_q}\right] - w_0\sum_{s=1}^{S}\frac{1}{\kappa_s}\frac{\partial Q_s}{\partial \lambda_p}\frac{\partial Q_s}{\partial \lambda_q}. \tag{5.38}$$

As in Section 5.3, the first term of this expression is $\delta_{pq}w_p - \nu_{pq}$. Routine calculations using the relationships $\lambda_p = \log(\mu_p/\mu_S)$ and $\mu_p = w_0 - c_p/Q_p$ give

$$\frac{\partial Q_p}{\partial \lambda_q} = \delta_{pq}\kappa_p - \frac{\kappa_p\kappa_q}{\kappa^*}, \tag{5.39}$$

where $\kappa^* = \sum_{p=1}^{S}\kappa_p$. This representation implies that

$$\sum_{s=1}^{s}\frac{1}{\kappa_s}\frac{\partial Q_s}{\partial \lambda_p}\frac{\partial Q_s}{\partial \lambda_q} = \frac{\partial Q_p}{\partial \lambda_q} \tag{5.40}$$

so that the p, q element of $\mathbf{M}_0$ is $\delta_{pq}w_p - \nu_{pq} - w_0(\partial Q_p/\partial\lambda_q)$.

By the Sherman-Morrison formula, the p, q element of the matrix $\mathbf{W}(\mathbf{I}-\boldsymbol{\Gamma})^{-1} - \mathbf{W}$ is $-w_0(\partial Q_p/\partial\lambda_q)$. So the matrix $\mathbf{M}_0$ consists of the first $S-1$ rows and columns of $\mathbf{W} - \mathbf{V} + \mathbf{W}(\mathbf{I}-\boldsymbol{\Gamma})^{-1} - \mathbf{W} = \mathbf{W}(\mathbf{I}-\boldsymbol{\Gamma})^{-1} - \mathbf{V} = \mathbf{M}$.

References

[1] A. J. Scott and C. J. Wild, "Fitting regression models to case-control data by maximum likelihood," *Biometrika*, vol. 84, no. 1, pp. 57–71, 1997.

[2] A. J. Scott and C. J. Wild, "Maximum likelihood for generalised case-control studies," *Journal of Statistical Planning and Inference*, vol. 96, no. 1, pp. 3–27, 2001.

[3] C. J. Wild, "Fitting prospective regression models to case-control data," *Biometrika*, vol. 78, no. 4, pp. 705–717, 1991.

[4] A. Lee, A. J. Scott, and C. J. Wild, "Fitting binary regression models with case-augmented samples," *Biometrika*, vol. 93, no. 2, pp. 385–397, 2006.

[5] P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins Series in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, Md, USA, 1993.

[6] N. E. Breslow, J. M. Robins, and J. A. Wellner, "On the semi-parametric efficiency of logistic regression under case-control sampling," *Bernoulli*, vol. 6, no. 3, pp. 447–455, 2000.

[7] N. E. Breslow, B. McNeney, and J. A. Wellner, "Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling," *The Annals of Statistics*, vol. 31, no. 4, pp. 1110–1139, 2003.

[8] J. M. Robins, F. Hsieh, and W. Newey, "Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates," *Journal of the Royal Statistical Society*, vol. 57, no. 2, pp. 409–424, 1995.

[9] B. McNeney and J. A. Wellner, "Application of convolution theorems in semiparametric models with non-i.i.d. data," *Journal of Statistical Planning and Inference*, vol. 91, no. 2, pp. 441–480, 2000.

[10] P. J. Bickel and J. Kwon, "Inference for semiparametric models: some questions and an answer," *Statistica Sinica*, vol. 11, no. 4, pp. 863–960, 2001.

[11] W. K. Newey, "The asymptotic variance of semiparametric estimators," *Econometrica*, vol. 62, no. 6, pp. 1349–1382, 1994.

[12] A. Lee and Y. Hirose, "Semi-parametric efficiency bounds for regression models under case-control sampling: the profile likelihood approach," submitted to *Annals of the Institute of Statistical Mathematics*.

Alan Lee: Department of Statistics, University of Auckland, Auckland 1142, New Zealand
Email address: aj.lee@auckland.ac.nz

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 51801, 13 pages
doi:10.1155/2007/51801

Research Article

Correlations in Output and Overflow Traffic Processes in Simple Queues

Don McNickle

Received 11 April 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

We consider some simple Markov and Erlang queues with limited storage space. Although the departure processes from some such systems are known to be Poisson, they actually consist of the superposition of two complex correlated processes, the overflow process and the output process. We measure the cross-correlation between the counting processes for these two processes. It turns out that this can be positive, negative, or even zero (without implying independence). The models suggest some general principles on how big these correlations are, and when they are important. This may suggest when renewal or moment approximations to similar processes will be successful, and when they will not.

1. Introduction

We consider a finite capacity queueing system, usually of the M/G/C/L class. Customers who arrive when there are already $L + C$ customers in the system overflow. We can identify three traffic processes in the customers leaving the system: the *overflow* process formed by customers who find the system full, the *output* process formed by customers who complete service, and these two processes, superposed together, from the *departure* process.

It is well known that in a number of cases in this class of queues, the departure process, when considered as an isolated process, is Poisson. For example:

(i) in the limit, the departure process from an M/M/C/L queue is Poisson (Boes [1]);
(ii) interdeparture times from an M/G/C/0 queue are exponentially distributed and are independent in the limit (Shanbag and Tambouratzis [2]).

Yet the complex behaviour present in these processes cannot be determined by examining the departure process in isolation. If we take the simplest case, of outputs and overflows

from an M/M/1/0 queue, then although both the output and the overflow process are renewal, neither is Poisson. The overflow process has a hyperexponential distribution and the output process is of phase type with two phases. Yet when superposed, they produce a Poisson process. Now the superposition of two independent renewal processes is Poisson if and only if both are Poisson processes. So the conclusion must be that the overflow and output processes cannot be independent. In this paper, we investigate the degree of this dependence by considering the cross-correlation functions of the two processes. We show that these cross-correlations can be positive, negative, or even uniformly zero, depending on the parameters of the process. Only in the simplest case do they have a clear form. However the models do suggest some general principles for the signs and sizes of these correlations, and when they are important. There have been a number of methods suggested for moment- or renewal-type approximations to these processes or similar ones, for example, the equivalent random method from classical telephone theory (Cooper [3]). (See also Whitt [4], Albin [5], Albin and Kai [6], and Johnson [7]). Having some qualitative knowledge about correlations should enhance the reliability and the appropriate use of these methods. There are a large number of papers on queues with similar correlated arrivals; see, for example, Adan and Kulkarni [8], and Heindl [9] (and other papers in the special issue of *Performance Evaluation*).

Individually, the characteristics of the overflow and output traffic processes are usually quite easy to determine. The output process is of Markov renewal type, (see Disney and Kiessler [10, Theorem 3.3, page 172]). This collapses to a renewal output process if and only if either (i) the service times are all zero with probability 1, (ii) $L = 0$, (iii) $L = 1$ with the service times being constant, or (iv) $L = \infty$ with the service times being exponential (see Disney and Kiessler [10, Theorem 3.5] for the proof of this).

Where the departure process has a Markov renewal representation, the overflow process is renewal, since it consists of the times between entries to a particular state in the Markov renewal process. Hence the distribution of times between overflows can be established by the usual filtering arguments.

2. Methodology

2.1. Cross-correlations between the overflow and the output processes. The measures that we will use for the dependence between the overflow and output processes are the *crosscovariance* and *cross-correlation* of the processes that count the number of outputs, $N^o(t)$, and overflows, $N^{ov}(t)$, in $(0,t]$. Thus we define

$$\begin{aligned} c\operatorname{cov}\left(N^o(t)N^{ov}(t)\right) &= E\left[N^o(t)N^{ov}(t)\right] - E\left[N^o(t)\right]E\left[N^{ov}(t)\right], \\ c\operatorname{cor}\left(N^o(t)N^{ov}(t)\right) &= \frac{c\operatorname{cov}\left(N^o(t)N^{ov}(t)\right)}{\sqrt{\operatorname{Var}\left(N^o(t)\right)\operatorname{Var}\left(N^{ov}(t)\right)}}. \end{aligned} \tag{2.1}$$

Since we need the joint distribution of $N^o(t)$ and $N^{ov}(t)$, we consider systems where the entire departure process can be represented by an n-state Markov renewal process (X,T) with semi-Markov kernel $Q^d(t)$. Usually the state will be the number of customers left behind by a departure. Entries to one subclass of states (often the nth state) represent overflows, while entries to the remaining states are outputs.

The Laplace-Stieltjes transform of the Markov renewal kernel is

$$R_s = \left(I - Q_s^d\right)^{-1}, \tag{2.2}$$

where Q_s^d is the Laplace-Stieltjes transform of $Q^d(t)$. Çinlar [11, page 165] gives a general expression for the expectation of the product of the numbers of visits by time t, $N_j(t)$ and $N_k(t)$, to states j and k, for a general finite-state Markov renewal process. That is, conditional on the process starting in state i:

$$E_i\left[N_j(t)N_k(t)\right] = \delta_{jk}R'_{ij}(t) + R'_{ij} * R'_{jk}(t) + R_{ik} * R'_{kj}(t). \tag{2.3}$$

The expected product of the numbers of visits to states j and k in steady state is

$$\sum_{i=1}^{n} \alpha_i E_i\left[N_j(t)N_k(t)\right]. \tag{2.4}$$

Here $*$ stands for the convolution operation, $R'_{ij}(t) = R_{ij}(t) - \delta_{ij}$, to ensure that the state occupied at time zero is not counted in the expected number of entries; n is the number of states, and $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_n]$ is the steady state vector for the Markov chain imbedded in the Markov renewal process.

2.2. The calculation method and verification steps. The calculations were carried out using the symbolic algebra package Maple to do the matrix operations and invert the resulting Laplace-Stieltjes transforms. Since these programmes are reasonably complex (up to 100 lines of Maple), a number of checks to verify the calculation were carried out. For example, in each of the cases where the marginal departure process is Poisson with rate λ, it was verified that $\boldsymbol{\alpha}$ and Q_s^d satisfy ($\mathbf{e}$ is a column vector of ones)

$$\alpha Q_s^d e = \frac{\lambda}{\lambda + s}. \tag{2.5}$$

The matrix of expected products of the numbers of visits to states j and k in steady state was verified to be symmetric in j and k (this is a particularly good check since it is produced at the end of a long sequence of distinctly asymmetric calculations), and finally the simulation study (see Section 4) was also run on a number of the cases that could be solved analytically. It produced entirely consistent results.

3. Results

3.1. M/M/1/0. We start with the simplest case of no storage with all distributions being exponential. While this is a very simple model, it is also the simplest case of the Erlang-B formula from classical telephone theory. We let the state of the system be the number of customers left behind by a departure. Thus each entry to state 0 is an output and entries to state 1 are overflows. With λ being the rate of the arrival Poisson process, and μ being the service rate, the Laplace-Stieltjes transform of the semi-Markov kernel of the departure

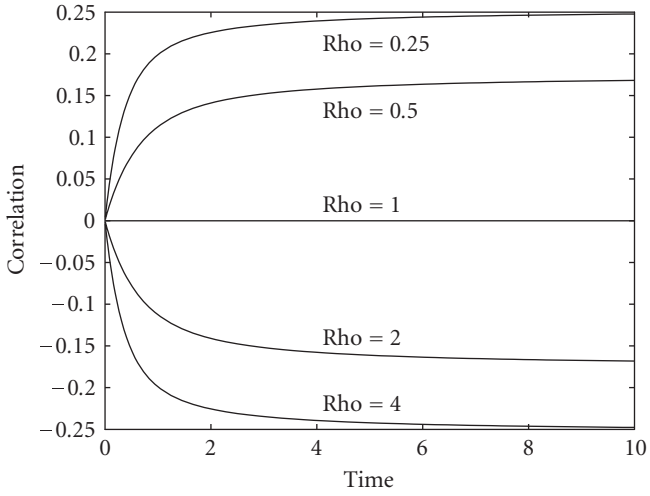


Figure 3.1. M/M/1/0 output and overflow cross-correlations.

process (see also Disney and Kiessler [10, page 84]) is

$$Q_s^d = \begin{pmatrix} \dfrac{\lambda\mu}{(\lambda+s)(\lambda+\mu+s)} & \dfrac{\lambda^2}{(\lambda+s)(\lambda+\mu+s)} \\ \dfrac{\mu}{(\lambda+\mu+s)} & \dfrac{\lambda}{(\lambda+\mu+s)} \end{pmatrix}. \tag{3.1}$$

The Laplace-Stieltjes transform of the Markov renewal kernel is

$$R_s = \left(I - Q_s^d\right)^{-1} = \begin{pmatrix} \dfrac{(\lambda+s)(\mu+s)}{s(\lambda+\mu+s)} & \dfrac{\lambda^2}{s(\lambda+\mu+s)} \\ \dfrac{\mu(\lambda+s)}{s(\lambda+\mu+s)} & \dfrac{\lambda^2+2\lambda s+\mu s+s^2}{s(\lambda+\mu+s)} \end{pmatrix}. \tag{3.2}$$

Applying (2.3) and (2.4) to appropriate terms from (3.2) leads (with the assistance of Maple) to the following expression for the cross-covariance of the number of outputs and the number of overflows by time t (starting from steady state):

$$c\,\mathrm{cov}\left(N^o(t)N^{ov}(t)\right) = \frac{\mu\lambda^2\left(\lambda-\mu-\left(\lambda^2-\mu^2\right)t-(\lambda-\mu)e^{-(\lambda+\mu)t}\right)}{(\lambda+\mu)^4}. \tag{3.3}$$

Thus we immediately see that when $\lambda = \mu$, the cross-covariance is zero for all times. Of course by Disney and Kiessler [10, Theorem 3.5], this cannot imply that the overflow and output processes are independent. So this is one of those peculiar situations where zero correlations do not imply independence. The variances of the numbers of outputs and of

overflows are

$$\begin{aligned}\operatorname{Var}\left(N^{o}(t)\right) &= \frac{\mu\lambda\left(\left(\lambda^3+\mu^3+\mu^2\lambda+\mu\lambda^2\right)t+2\lambda\mu\left(1-e^{-(\lambda+\mu)t}\right)\right)}{(\lambda+\mu)^4},\\ \operatorname{Var}\left(N^{ov}(t)\right) &= \frac{\lambda^2\left(\left(\lambda^3+\mu^3+5\mu^2\lambda+5\mu\lambda^2\right)t-2\lambda\mu\left(1-e^{-(\lambda+\mu)t}\right)\right)}{(\lambda+\mu)^4}.\end{aligned} \tag{3.4}$$

The cross-correlation of the number of outputs and the number of overflows by time t can then be calculated from (3.3) and (3.4). The cross-correlation can be seen to be identical, except for a change of sign, when the values of λ and μ are interchanged. Thus if $\lambda < \mu$, the number of outputs and the number of overflows are positively correlated over any time interval, while if $\lambda > \mu$ they are negatively correlated. To demonstrate this symmetry, the cross-correlations are plotted in Figure 3.1 for values of $(\lambda,\mu) = \{(1,4),(1,2),(1,1),(2,1),(4,1)\}$ (traffic intensities of Rho $= 0.25, 0.5, 1, 2, 4$).

So the cross-correlations reduce monotonically with increasing traffic intensity, and increase monotonically in absolute value with time. Taking the limit of the cross-correlation expression gives

$$\lim_{t\to\infty} c\operatorname{cor}\left(N^{o}(t)N^{ov}(t)\right) = \frac{\lambda\mu(\mu-\lambda)}{\sqrt{\left(\lambda^2+4\lambda\mu+\mu^2\right)\left(\lambda^3\mu+\lambda\mu^3\right)}}. \tag{3.5}$$

This limiting expression was also found by a simple direct method in Disney and Kiessler.

3.2. Systems with storage, M/M/1/1, M/M/1/2. We now add one or two units of storage to the system. Again an appropriate state for the departure Markov renewal process is the number of customers left behind by a departure. We give only the results for M/M/1/2. Those for M/M/1/1 are similar, but less pronounced. Labelling the states as $\{0,1,2,3\}$, each entry to state 0, 1, or 2 is an output and entries to state 3 are overflows. The Laplace-Stieltjes transform of the semi-Markov kernel of the departure process is

$$Q_s^d = \begin{pmatrix} \dfrac{\lambda\mu}{(\lambda+s)(\lambda+\mu+s)} & \dfrac{\lambda^2\mu}{(\lambda+s)(\lambda+\mu+s)^2} & \dfrac{\lambda^3\mu}{(\lambda+s)(\lambda+\mu+s)^3} & \dfrac{\lambda^4}{(\lambda+s)(\lambda+\mu+s)^3} \\ \dfrac{\mu}{\lambda+\mu+s} & \dfrac{\lambda\mu}{(\lambda+\mu+s)^2} & \dfrac{\lambda^2\mu}{(\lambda+\mu+s)^3} & \dfrac{\lambda^3}{(\lambda+\mu+s)^3} \\ 0 & \dfrac{\mu}{\lambda+\mu+s} & \dfrac{\lambda\mu}{(\lambda+\mu+s)^2} & \dfrac{\lambda^2}{(\lambda+\mu+s)^2} \\ 0 & 0 & \dfrac{\mu}{(\lambda+\mu+s)} & \dfrac{\lambda}{(\lambda+\mu+s)} \end{pmatrix}. \tag{3.6}$$

The Markov renewal matrix does not now have an informative form, but the method is as before. Since the results are no longer symmetric in λ and μ, we take, as we will from now on, the mean service time to be 1. Plots of the cross-correlations, for traffic intensities of 0.5, 0.8, 1, and 2 are given in Figure 3.2. The smallest cross-correlations occur at a traffic intensity of about 0.8. So the symmetry and monotone nature of the M/M/1/0 results have gone (possibly due to the fact that the output process is no longer renewal), but

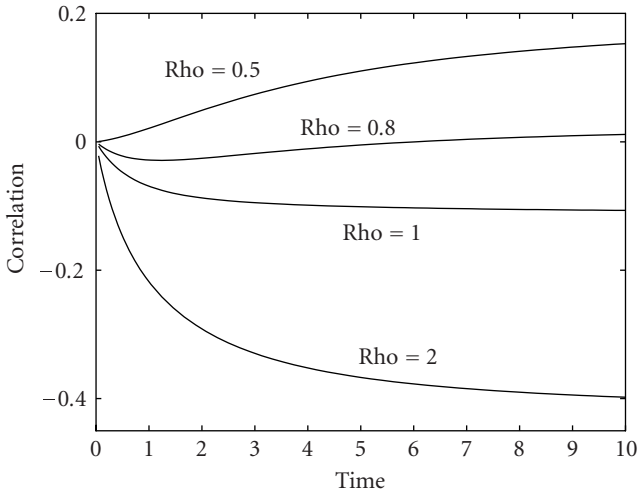


FIGURE 3.2. M/M/1/2 output and overflow cross-correlations.

the general principles of positive cross-correlations at low traffic intensities and negative cross-correlations at high traffic intensities, which will be discussed further in Section 3.5, are starting to emerge.

3.3. M/M/2/0. With two servers and no storage we assume that when both servers are idle, an arrival selects a server by tossing a coin (the results when the servers are tested in a fixed order are very similar). We take the state of the process to be $\{i,j,k\}$, where $i = 1$ or 2 is the server from which the last output occurred, and $\{j,k\} \in \{0,1\}\times\{0,1\}$ is the number of customers left behind at servers 1 and 2. Since all the service and interarrival distributions are negative exponential, this process is Markov renewal. The five possible states are $\{1,0,0\}$, $\{2,0,0\}$, $\{2,1,0\}$, $\{1,0,1\}$, and the overflow state, which does not require the index of the last output, $\{*,1,1\}$. With the states in that order, and using the notation $\lambda+s=a$, $\lambda+\mu+s=b$, $\lambda+2\mu+s=c$, the semi-Markov kernel is then

$$Q_s^d = \begin{pmatrix} \frac{\lambda\mu}{2ab} & \frac{\lambda\mu}{2ab} & \frac{\lambda^2\mu}{abc} & \frac{\lambda^2\mu}{abc} & \frac{\lambda^3}{abc} \\ \frac{\lambda\mu}{2ab} & \frac{\lambda\mu}{2ab} & \frac{\lambda^2\mu}{abc} & \frac{\lambda^2\mu}{abc}s & \frac{\lambda^3}{abc} \\ \frac{\mu}{b}s & 0 & \frac{\lambda\mu}{abc} & \frac{\lambda\mu}{abc} & \frac{\lambda^2}{bc} \\ 0 & \frac{\mu}{b} & \frac{\lambda\mu}{abc} & \frac{\lambda\mu}{abc} & \frac{\lambda^2}{bc} \\ 0 & 0 & \frac{\mu}{b} & \frac{\mu}{b} & \frac{\lambda}{c} \end{pmatrix}. \tag{3.7}$$

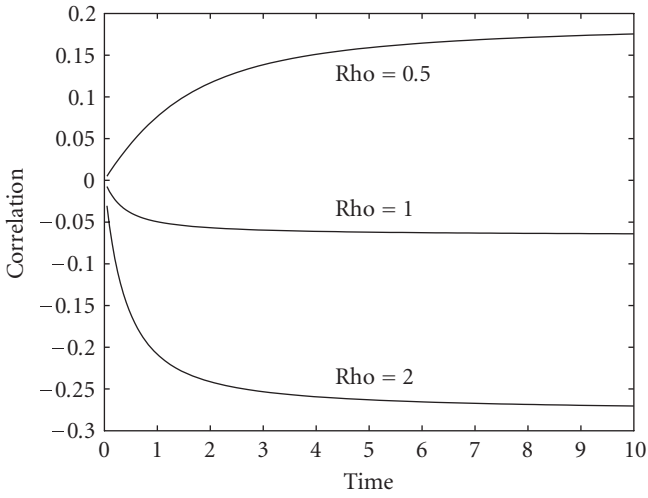


FIGURE 3.3. M/M/2/0 total output and overflow correlations.

Now that we have two departure streams, there are a number of cross-correlations that could be considered. We give only those between the total output process (i.e., the superposition of the outputs from the two servers) and the overflow process, in Figure 3.3. Thus the results are qualitatively similar to the first system, M/M/1/0.

3.4. A remark on the equivalent random method. Since this is a simple case of the classic Erlang-B situation, it is also interesting to look at the cross-correlation between the two output streams. Even in the case where the servers are tested in fixed order, these turn out to be very small (<0.05 in absolute value.) This may suggest another reason why mean-variance methods like the equivalent random method (see Cooper [3, page 165]) have been found to work so well in telephone networks. If all the output streams, including those from a single set of lines, are nearly uncorrelated, characterising the carried traffic at a subsequent link by only its first two moments is more likely to work.

3.5. Some comments on the results so far. From the results so far, we can form explanations which give some insight into the processes involved. We note that when the traffic intensity is low, the cross-correlations are positive; when the traffic intensity is about 1, the cross-correlations are very small; and that when the traffic intensity is much greater than 1, the cross-correlations are strongly negative. Our explanation for this goes as follows. When the traffic intensity is low, the dependence in the departure process is basically being driven by fluctuations in the arrival process. Hence the output and overflow processes tend to move together; when there are an abnormally large number of outputs there are also an abnormally large number of overflows. For high traffic intensity, on the other hand, the server is almost always busy so the output and overflow processes are complements of each other—abnormal excess in one is associated with abnormal paucity in the other, and hence the cross-correlations are negative.

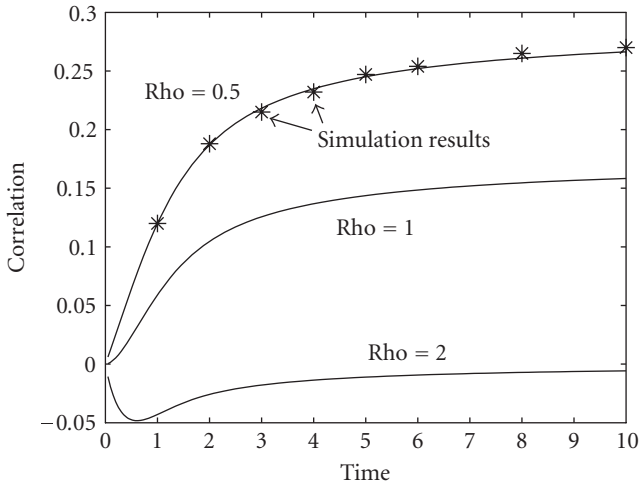


FIGURE 3.4. $M/E_2/1/0$ output and overflow cross-correlations.

3.6. Nonexponential service time distributions. In general, cross-correlations for these systems are difficult to measure analytically, since the departure process does not have a compact representation as a Markov renewal process. One tractable case is $M/E_2/1/0$. The state is the phase of the customer (if any) left behind in service by a departure. Thus state 0 corresponds to an output, while state 1 is an overflow that occurred while the customer in service was in the first phase of service, and state 2 is an overflow that occurred during the second phase of service. The semi-Markov kernel is

$$Q_s^d = \begin{pmatrix} \dfrac{\lambda\mu^2}{(\lambda+s)(\lambda+\mu+s)^2} & \dfrac{\lambda^2}{(\lambda+s)(\lambda+\mu+s)} & \dfrac{\lambda^2\mu}{(\lambda+s)(\lambda+\mu+s)^2} \\ \dfrac{\mu^2}{(\lambda+\mu+s)^2} & \dfrac{\lambda}{\lambda+\mu+s} & \dfrac{\lambda\mu}{(\lambda+\mu+s)^2} \\ \dfrac{\mu}{\lambda+\mu+s} & 0 & \dfrac{\lambda}{\lambda+\mu+s} \end{pmatrix}. \tag{3.8}$$

We again make the mean service time 1, so each stage of the service time distribution has the rate 2, and plot the output and overflow cross-correlations in Figure 3.4, along with some simulation results for the same system (see Section 4).

So the trend of positive cross-correlations at low traffic intensities continues with Erlang service distributions, but the negative cross-correlations at high traffic intensities, although present, are less pronounced, perhaps due to the lower rate of overflows.

4. Some simulation results

For Erlang service time distributions with $L > 0$ (and for M/M/1/L with $L > 2$), the Markov renewal representation of the departure process is either too complex or has too

many states to use the analytic approach detailed in Section 2. Simulating such systems is quite easy however. A simulation programme in GPSS/H was written to see if the effects noted previously carry on to systems with more storage. The simulation programme was first verified against the analytic results for $M/E_2/1/0$ from the previous section. For each time value 100 000 pairs of the number of outputs and the number of overflows were recorded, along with the sample cross-correlation calculated from these. The half width of an approximate 95% confidence interval for these cross-correlations would be $1.96/\sqrt{100\,000} = 0.006$. All of the simulation results (some are plotted on Figure 3.4 in the previous section) are within this distance of the analytic values.

$M/E_2/1/2$. We keep the mean service time to be 1 (so each stage of the Erlang distribution has the rate 2).

Traffic intensity	Time	0	2	4	6	8	10
$\rho = 0.5$		0	0.089	0.155	0.190	0.211	0.221
$\rho = 1$		0	−0.014	0.025	0.050	0.063	0.069
$\rho = 2$		0	−0.222	−0.257	−0.274	−0.284	−0.288

Now with some storage, the results have returned to a pattern compatible with our explanation of the effect of traffic intensity, positive correlations at low traffic intensity as both the overflow and output processes move together, and negative correlations at high traffic intensity, where they tend to be the complements of each other.

$D/M/1/2$. The system change we are after here is the reduced variability in the arrival process. If our explanation of the qualitative effect of traffic intensity on the cross-correlations is correct, we might expect that the low traffic intensity effect of positive cross-correlations would be reduced or eliminated, while the high traffic intensity effect of negative cross-correlations should be at least preserved or possibly enhanced. That is precisely what happens:

Traffic intensity	Time	0	2	4	6	8	10
$\rho = 0.5$		0	−0.007	−0.032	−0.061	−0.082	−0.106
$\rho = 1$		0	−0.329	−0.485	−0.591	−0.665	−0.717
$\rho = 2$		0	−0.705	−0.833	−0.885	−0.911	−0.928

Two other examples, $E_2/M/1/0$ and $E_2/E_2/1/0$, support the hypothesis that it is the variability of the arrival process that is responsible for the positive cross-correlations at low traffic intensities.

$E_2/M/1/0$.

Traffic intensity	Time	0	2	4	6	8	10
$\rho = 0.5$		0	−0.018	−0.054	−0.071	−0.080	−0.089
$\rho = 1$		0	−0.019	−0.249	−0.268	−0.279	−0.283
$\rho = 2$		0	−0.349	−0.391	−0.405	−0.411	−0.415

$E_2/E_2/1/0$.

Traffic intensity	Time	0	2	4	6	8	10
$\rho = 0.5$		0	0.071	0.064	0.051	0.045	0.043
$\rho = 1$		0	−0.071	−0.091	−0.107	−0.109	−0.114
$\rho = 2$		0	−0.215	−0.238	−0.245	−0.250	−0.254

5. The effects of these correlations

5.1. How big are the effects of these correlations? Can we ever get away with treating the overflow and output streams as being independent? For the simplest model, M/M/1/0, it is possible to answer this exactly by considering, as an alternative, a model consisting of two independent M/M/1/0 systems and combining the outputs from the first system and the overflows from the second system. So this alternative combined process is what would result if we were to treat the overflow and output streams as being independent. As a reference model to measure the effects of this assumption, we compare the blocking probability at a subsequent server with no storage, called server 3, both for the alternative model and the correct (Poisson input) model.

We first need to determine the departure process from the two-independent-systems model. We take the state of the process to be $\{i, j, k\}$, where $i = 1$ or 2 is the system from which the last output occurred, and $\{j, k\} \in \{0, 1\} \times \{0.1\}$ is the number of customers left behind at servers 1 and 2. Since all the service and interarrival distributions are negatively exponential, this process is Markov renewal. The eight possible states are outputs from node 1: $\{1,0,0\}$, $\{1,0,1\}$, overflows from node 1: $\{1,1,0\}$, $\{1,1,1\}$, outputs from node 2: $\{2,0,0\}$, $\{2,0,1\}$, and overflows from node 2: $\{2,1,0\}$, $\{2,1,1\}$. The semi-Markov kernel is then

$$Q = \begin{pmatrix} \frac{\lambda_1\mu_1}{ab} & \lambda_1\lambda_2\mu_1 e & \frac{\lambda_1^2}{ab} & \lambda_1^2\lambda_2 e & \frac{\lambda_2\mu_2}{ac} & \frac{\lambda_2^2}{ac} & \lambda_1\lambda_2\mu_1 e & \lambda_1\lambda_2^2 e \\ 0 & \frac{\lambda_1\mu_1}{bd} & 0 & \frac{\lambda_1^2}{cd} & \frac{\mu_2}{c} & \frac{\lambda_2}{c} & \frac{\lambda_1\mu_2}{cd} & \frac{\lambda_1\lambda_2}{cd} \\ \frac{\mu_1}{b} & \frac{\lambda_2\mu_1}{bd} & \frac{\lambda_1}{b} & \frac{\lambda_2\lambda_1}{bd} & 0 & 0 & \frac{\mu_2}{d} & \frac{\lambda_2}{d} \\ 0 & \frac{\mu_1}{d} & 0 & \frac{\lambda_1}{d} & 0 & 0 & \frac{\mu_2}{d} & \frac{\lambda_2}{d} \\ \frac{\lambda_1\mu_1}{ab} & \lambda_1\lambda_2\mu_1 e & \frac{\lambda_1^2}{ab} & \lambda_1^2\lambda_2 e & \frac{\lambda_2\mu_2}{ac} & \frac{\lambda_2^2}{ac} & \lambda_1\lambda_2\mu_1 e & \lambda_1\lambda_2^2 e \\ 0 & \frac{\lambda_1\mu_1}{bd} & 0 & \frac{\lambda_1^2}{cd} & \frac{\mu_2}{c} & \frac{\lambda_2}{c} & \frac{\lambda_1\mu_2}{cd} & \frac{\lambda_1\lambda_2}{cd} \\ \frac{\mu_1}{b} & \frac{\lambda_2\mu_1}{bd} & \frac{\lambda_1}{b} & \frac{\lambda_2\lambda_1}{bd} & 0 & 0 & \frac{\mu_2}{d} & \frac{\lambda_2}{d} \\ 0 & \frac{\mu_1}{d} & 0 & \frac{\lambda_1}{d} & 0 & 0 & \frac{\mu_2}{d} & \frac{\lambda_2}{d} \end{pmatrix}. \tag{5.1}$$

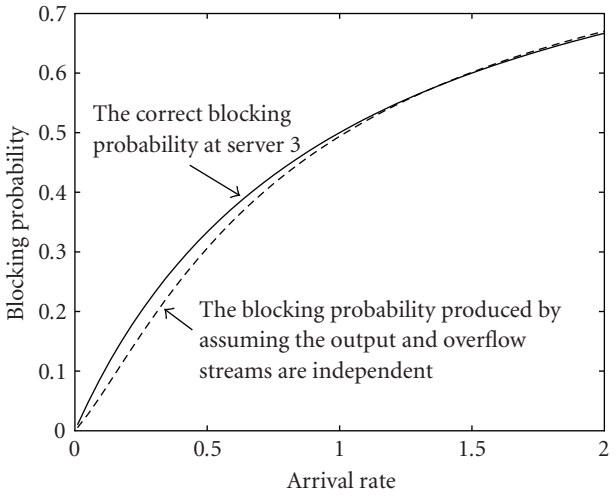


FIGURE 5.1. The effect of treating output and overflow streams as being independent.

Here $a = \lambda_1 + \lambda_2 + s$, $b = \lambda_1 + \lambda_2 + \mu_1 + s$, $c = \lambda_1 + \lambda_2 + \mu_2 + s$, $d = \lambda_1 + \lambda_2 + \mu_1 + \mu_2 + s$, and $e = 1/abd + 1/acd$. The input process that we require constitutes entries to states 1, 2, 7, and 8 so this is obtained from Q by the usual filtering arguments. That is, if $S = \{1,2,7,8\}$ and $T = \{3,4,5,6\}$, the Laplace-Stieltjes transform of the semi-Markov kernel of the input process is found from: $Q_i = Q_{SS} + Q_{ST}(I - Q_{TT})^{-1}Q_{TS}$, where Q_{SS}, Q_{ST}, Q_{TT}, and Q_{TS} are the corresponding submatrices of Q.

Finally when used as input to a subsequent node (server 3) with no storage and service rate μ_3, the state distribution as seen by an arriving customer has a semi-Markov kernel given by

$$Q_3(s) = \begin{pmatrix} Q_i(s) - Q_i(s+\mu_3) & Q_i(s+\mu_3) \\ Q_i(s) - Q_i(s+\mu_3) & Q_i(s+\mu_3) \end{pmatrix}. \tag{5.2}$$

Since states 5, 6, 7, 8 of this process are the states in which an overflow at server 3 occurs, the probability of overflow at node 3 is $\pi_5 + \pi_6 + \pi_7 + \pi_8$, taken from the steady state distribution of the imbedded Markov chain $Q_3(0)$.

If we take $\lambda_1 = \lambda_2$, and all of the service rates to be one, the fraction of customers who overflow from server 3 is simplified to

$$\pi_5 + \pi_6 + \pi_7 + \pi_8 = \frac{\lambda\left(4\lambda^5 + 25\lambda^4 + 67\lambda^3 + 88\lambda^2 + 49\lambda + 60\right)}{\left(4\lambda^4 + 13\lambda^3 + 16\lambda^2 + 9\lambda + 1\right)\left(\lambda^2 + 4\lambda + 6\right)}. \tag{5.3}$$

We can compare this with the overflow fraction which would occur using the correctly correlated overflow and output processes. This is easy to calculate as server 3 is actually

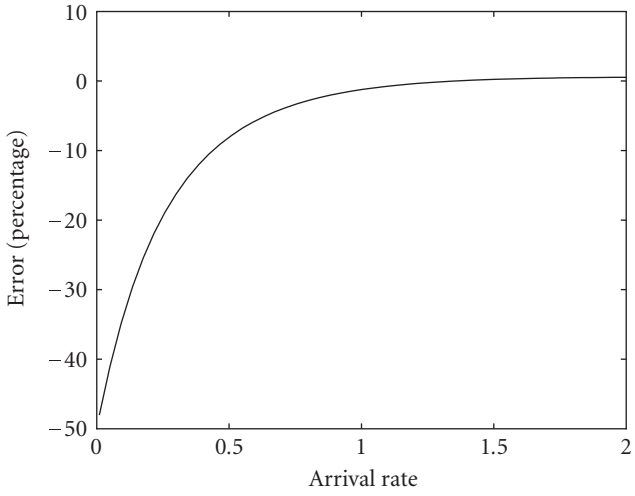


Figure 5.2. Percentage error due to assuming independence.

an M/M/1/0 system. Plotting the error against traffic intensity (see Figure 5.1) shows that while for high traffic intensities the error is small, for intensities less than 1, treating the overflow and output processes as being independent leads to an underestimation of the blocking probability by up to 50% (see Figure 5.2).

5.2. How much extra information can we get from the correlations? In the same case, M/M/1/0, we can show that there is enough information in the cross-correlations to perfectly reconstruct the departure process. We assume that we know the marginal distributions of the time between outputs and of the time between overflows. However, in addition, we assume we know, or have measured, the cross-covariances (like equation (3.3), but conditioned on the starting state). It is easy to show that we now have enough information in the four equations to solve for, for example, the Markov renewal kernel (3.2). So at least in the class of two-state Markov renewal processes, we would have exactly determined the departure process. Thus at least in this simple case, the addition of this information on cross-correlations is enough to reduce the error to zero.

6. Conclusions

We have shown that considerable dependencies, as measured by cross-correlations, can arise in the output and overflow processes from simple queueing models. These can be large, and either positive or negative. Positive cross-correlations are associated with lower traffic intensities, which is also the situation in which ignoring these correlations may produce the largest percentage errors. The variability of the arrival process is an important factor in this. The smallest correlations usually occur at a traffic intensity close to 1. The qualitative insight these models provide may prove useful in determining a circumstance under which moment or renewal approximations will work well.

References

[1] D. C. Boes, "Note on the output of a queuing system," *Journal of Applied Probability*, vol. 6, no. 2, pp. 459–461, 1969.
[2] D. N. Shanbhag and D. G. Tambouratzis, "Erlang's formula and some results on the departure process for a loss system," *Journal of Applied Probability*, vol. 10, no. 1, pp. 233–240, 1973.
[3] R. B. Cooper, *Introduction to Queueing Theory*, Edward Arnold, London, UK, 2nd edition, 1981.
[4] W. Whitt, "Approximating a point process by a renewal process—I: two basic methods," *Operations Research*, vol. 30, no. 1, pp. 125–147, 1982.
[5] S. L. Albin, "Approximating a point process by a renewal process—II: superposition arrival processes to queues," *Operations Research*, vol. 32, no. 5, pp. 1133–1162, 1984.
[6] S. L. Albin and S.-R. Kai, "Approximation for the departure process of a queue in a network," *Naval Research Logistics Quarterly*, vol. 33, no. 1, pp. 129–143, 1986.
[7] M. A. Johnson, "Markov MECO: a simple Markovian model for approximating nonrenewal arrival processes," *Communications in Statistics. Stochastic Models*, vol. 14, no. 1-2, pp. 419–442, 1998.
[8] I. J. B. F. Adan and V. G. Kulkarni, "Single-server queue with Markov-dependent inter-arrival and service times," *Queueing Systems*, vol. 45, no. 2, pp. 113–134, 2003.
[9] A. Heindl, "Decomposition of general queueing networks with MMPP inputs and customer losses," *Performance Evaluation*, vol. 51, no. 2–4, pp. 117–136, 2003, special issue on queueing networks with blocking, Kouvatsos and Balsamo, eds.
[10] R. L. Disney and P. C. Kiessler, *Traffic Processes in Queueing Networks: A Markov Renewal Approach*, vol. 4 of *Johns Hopkins Series in the Mathematical Sciences*, Johns Hopkins University Press, Baltimore, Md, USA, 1987.
[11] E. Çinlar, "Markov renewal theory," *Advances in Applied Probability*, vol. 1, pp. 123–187, 1969.

Don McNickle: Department of Management, University of Canterbury, Christchurch 8140, New Zealand
Email address: don.mcnickle@canterbury.ac.nz

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 37475, 15 pages
doi:10.1155/2007/37475

Research Article

An M-Estimation-Based Procedure for Determining the Number of Regression Models in Regression Clustering

C. R. Rao, Y. Wu, and Q. Shao

Received 16 June 2007; Accepted 16 July 2007

Recommended by Paul Cowpertwait

In this paper, a procedure based on M-estimation to determine the number of regression models for the problem of regression clustering is proposed. We have shown that the true classification is attained when n increases to infinity under certain mild conditions, for instance, without assuming normality of the distribution of the random errors in each regression model.

1. Introduction

Cluster analysis is a statistical tool to classify a set of objects into groups so that objects within a group are "similar" and objects in different groups are "dissimilar." The purpose of clustering is to discover "natural" structure hidden in a data set. Regression clustering technique is often used to classify the data and recover the underlying structure, when the data set is believed to be a random sample from a population comprising of a fixed, but unknown, number of subpopulations, each of which is characterized by a distinct regression model. Regression clustering is one of the most commonly used model-based clustering techniques. It has been studied by Bock [1], Quandt and Ramsey [2], and Späth [3] among others, and has applications in a variety of disciplines, for example, in market segmentation by DeSarbo and Cron [4] and quality control systems by Lou et al. [5].

A fundamental problem, as well as a preliminary step in regression clustering, is to determine the underlying "true" number of regression models in a data set. Shao and Wu [6] proposed an information-based criterion (named criterion "LS-C" in the sequel) to tackle this problem. The limiting behavior of LS-C is given in their paper.

However, it is well known that the least squares (LS) method is very sensitive to outliers and violation of the normality assumption of the data. This instability also exists in the

LS-based procedures for both selecting the number of regression models and classifying the data in the context of regression clustering.

During the past three decades, numerous efforts have been made for developing robust statistical procedures for statistical inferences. Among them, procedures based on M-estimators, which are maximum likelihood-type estimators (Hampel et al. [7] and Huber [8]), play an important role. The M-estimation-based model selection criteria are considered by Konishi and Kitagawa [9], Machado [10], and Wu and Zen [11] among others.

To overcome the instability of the LS-based procedures in regression clustering, we propose an M-estimation-based procedure for determining the number of regression models, which is an extension of M-estimation-based information criterion for linear model selection developed by Wu and Zen [11]. Its asymptotic behavior will be investigated.

The structure of this paper is arranged as follows. In Section 2, we build a probabilistic framework for our problem and introduce some notations used in this paper. Section 3 lists all the assumptions needed for our study. In Section 4, we study the limiting behavior of the proposed criterion. Some ancillary results required for our proofs are presented in the appendix.

2. Notation and preliminaries

We consider the clustering problem for n objects $\mathbb{O}^{(n)} = \{1,\ldots,n\}$, where for each object j, $(\mathbf{x}_j, y_j)$ has been recorded, where $\mathbf{x}_j = (x_j^{(1)},\ldots,x_j^{(p)})' \in \mathbb{R}^p$ is a nonrandom explanatory p-vector and $y_j \in \mathbb{R}$ is a random response variable. The set of these n objects is a random sample from a structured population as specified below.

Suppose that there exists an underlying partition $\Pi_{k_0}^{(n)} = \{\mathbb{O}_1^{(n)},\ldots,\mathbb{O}_{k_0}^{(n)}\}$ for these n objects, and each component $\mathbb{O}_i^{(n)} \triangleq \{i_1,\ldots,i_{n_i}\} \subseteq \mathbb{O}^{(n)}$ is characterized by a linear regression model:

$$y_{j,\mathbb{O}_i} = \mathbf{x}'_{j,\mathbb{O}_i}\boldsymbol{\beta}_{0i} + e_{j,\mathbb{O}_i}, \quad j = i_1,\ldots,i_{n_i}, \tag{2.1}$$

where $n_i = |\mathbb{O}_i|$ is the number of observations in the ith component $\mathbb{O}_i$, $i = 1,\ldots,k_0$, and $\sum_{i=1}^{k_0} n_i = n$. Note that $\mathbb{O}_i$ and $\mathbb{O}_i^{(n)}$ are used interchangeably to denote the ith component of the underlying partition $\Pi_{k_0}^{(n)}$. $(\mathbf{x}_{j,\mathbb{O}_i}, y_{j,\mathbb{O}_i})$ $(j = i_1,\ldots,i_{n_i},\ i = 1,\ldots,k_0)$ is a relabeled observation $(\mathbf{x}_j, y_j)$ $(j = 1,\ldots,n)$ to represent the jth object in the ith component $\mathbb{O}_i$ of the true partition $\Pi_{k_0}^{(n)}$. We will use this double-index notation for any object $(\mathbf{x}_j, y_j)$ throughout this paper to identify the component to which it belongs. $\boldsymbol{\beta}_{0i} \in \mathbb{R}^p$ are k_0 pairwise distinct p-vectors of unknown regression parameters, and $e_{j,\mathbb{O}_i}$, $j = i_1,\ldots,i_{n_i}$, are independently and identically distributed random errors for $i = 1,\ldots,k_0$.

However, this underlying structure (2.1) is not observable. What we observe is just a random sample of n objects with the data values $(\mathbf{x}_j, y_j)$ for each of the $p+1$ variables associated with each object. Our task is then to reconstruct the hidden structure (2.1) from the observed data by first estimating the number of regression models k_0 and then classifying the data and estimating the regression parameters in each regression model accordingly.

Consider any possible classification of these n objects: $\Pi_k^{(n)} = \{\mathscr{C}_1^{(n)},\ldots,\mathscr{C}_k^{(n)}\}$, where $k \le K$ is a positive integer. For this partitioning, we fit k M-estimator-based linear regression models and obtain kM-estimates $\widehat{\boldsymbol{\beta}}_s$, $s = 1,\ldots,k$, separately. Then the M-estimator-based criterion for estimating the number of regression models is given as follows: let $q(k)$ be a strictly increasing function of k and let A_n be a sequence of constants. We define

$$R_n\left(\Pi_k^{(n)}\right) = \sum_{s=1}^{k} \sum_{j\in\mathscr{C}_s} \rho\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s}\widehat{\boldsymbol{\beta}}_s\right) + q(k)A_n, \tag{2.2}$$

where ρ is a convex discrepancy function. As an example, ρ can be chosen as Huber's discrepancy function

$$\rho_c(t) = \begin{cases} \dfrac{1}{2}t^2, & |t| < c, \\ c|t| - \dfrac{1}{2}c^2, & |t| \ge c. \end{cases} \tag{2.3}$$

Also, in (2.2), $\sum_{j\in\mathscr{C}_s}$ stands for the summation made over all the observations in the class $\mathscr{C}_s$ and $\widehat{\boldsymbol{\beta}}_s$ is the M-estimator in the sth class such that

$$\sum_{j\in\mathscr{C}_s} \rho\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s}\widehat{\boldsymbol{\beta}}_s\right) = \min_{\boldsymbol{\beta}_s} \sum_{j\in\mathscr{C}_s} \rho\left(y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s}\boldsymbol{\beta}_s\right). \tag{2.4}$$

Again, $\mathscr{C}_s$ and $\mathscr{C}_s^{(n)}$ are used interchangeably in the above equations to denote the sth class in the partition $\Pi_k^{(n)}$. We will continue this convenient usage without further explanation in the sequel. It can be seen that in (2.2), the first term is a generalization of a minimum negative log-likelihood function and the second term is the penalty for over-fitting.

Then the estimate of the underlying number of regression models k_0, $\widehat{k}_n$ is obtained by minimizing the criterion (2.2), that is,

$$\widehat{k}_n = \arg\min_{1\le k\le K} \min_{\Pi_k^{(n)}} R_n\left(\Pi_k^{(n)}\right). \tag{2.5}$$

We will call this criterion MR-C, which stands for the M-estimator-based regression clustering. Moreover, criterion MR-C in (2.5) shows that it actually determines the optimal number of regression models and the associated partitioning simultaneously.

3. Assumptions

Let $\mathbb{O}_l = \{l_1,\ldots,l_{n_l}\}$ be any component or a subset of a component associated with the underlying true partition $\Pi_{k_0}^{(n)}$ of $\mathbb{O}^{(n)}$, and $n_l = |\mathbb{O}_l|$. If we let $X_{n_l} = (\mathbf{x}_{l_1,\mathbb{O}_l},\ldots,\mathbf{x}_{l_{n_l},\mathbb{O}_l})'$ be the design matrix in $\mathbb{O}_l$, then $W_{n_l} = X'_{n_l}X_{n_l}$, $d_{n_l}^2 = \max_{1\le j\le n_l} \mathbf{x}'_{j,\mathbb{O}_l} W_{n_l}^{-1}\mathbf{x}_{j,\mathbb{O}_l}$.

To facilitate the study on the limiting behavior of the criterion MR-C, we need the following assumptions.

(A) For the true partition $\Pi_{k_0}^{(n)} = \{\mathbb{O}_1,\ldots,\mathbb{O}_{k_0}\}$ and $n_i = |\mathbb{O}_i|$, there exists a fixed constant $a_0 > 0$ such that

$$a_0 n \le n_i \le n \quad \forall\, i = 1,\ldots,k_0. \tag{3.1}$$

Remark 3.1. This assumption is equivalent to the explicit assumption that the population comprises k_0 subpopulations with proportions $\pi_1,\ldots,\pi_{k_0}$ where $0<\pi_i<1$, $i=1,\ldots,k_0$, $\sum_{i=1}^{k_0}\pi_i=1$. Then $a_0=\min_{1\le i\le k_0}\pi_i$ would satisfy (3.1).

(B1) $\rho(\cdot)$ is a convex function on $\mathbb{R}^1$.

(B2) $\mathrm{E}[\rho(e_{j,\mathbb{O}_i})]$ is finite for all $j\in\mathbb{O}_i$ and $i=1,\ldots,k_0$.

(B3) For any $\boldsymbol{\beta}$ and observations in $\mathbb{O}_l$,

$$\liminf_{n_l\to\infty}\frac{1}{n_l}\sum_{j\in\mathbb{O}_l}\mathrm{E}\left[\rho\left(e_{j,\mathbb{O}_l}-\mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}\right)-\rho\left(e_{j,\mathbb{O}_l}\right)\right]\ge g(\boldsymbol{\beta}),\tag{3.2}$$

where $g(\cdot)$ is a nonnegative convex function and is strictly convex in a neighborhood of $\mathbf{0}$.

If ρ has a first-order derivative, in order to find M-estimator of $\boldsymbol{\beta}_s$ in the sth-class, one may first find all first-order partial derivatives of $\sum_{j\in\mathbb{O}_s}\rho(y_{j,\mathbb{O}_s}-\mathbf{x}'_{j,\mathbb{O}_s}\boldsymbol{\beta}_s)$ and then set them to be equal to zeros. The simultaneous solutions of these equations give the M-estimator of $\boldsymbol{\beta}_s$. However in some cases, ρ does not have a first-order derivative. Note that for any convex function, it always has subgradients, which are just partial derivatives if they do exist (see Rockafellar [12]). Let $\psi(\cdot)$ be any choice of the subgradient of $\rho(\cdot)$ and denote by $\mathcal{U}$ the set of discontinuity points of ψ, which is the same for all choices of ψ.

(C1) The common distribution function F of $e_{j,\mathbb{O}_i}$, $j\in\mathbb{O}_i$, is unimodal and satisfies $F(\mathcal{U})=0$. $\mathrm{E}[\psi(e_{j,\mathbb{O}_i})]=0$, $\mathrm{E}[\psi^2(e_{j,\mathbb{O}_i})]=\sigma_i^2<\infty$ for any $i=1,\ldots,k_0$, and

$$\mathrm{E}\left[\psi\left(e_{j,\mathbb{O}_i}+u\right)\right]=a_iu+o(|u|),\quad \text{as } u\longrightarrow 0,\tag{3.3}$$

where a_i, $i=1,\ldots,k_0$, are finite positive constants.

(C2) There exist positive constants ζ and h_0 such that for any $h\in[0,h_0]$ and any u,

$$\psi(u+h)-\psi(u)\le\zeta.\tag{3.4}$$

(C3) The moment generating function $M_i(t)=\mathrm{E}[\exp\{t\psi(e_{j,\mathbb{O}_i})\}]$ exists for $|t|\le\Delta$, where $i=1,\ldots,k_0$.

(C4) $\mathrm{E}[|\psi(e_{j,\mathbb{O}_i})|^3]<\infty$, $j\in\mathbb{O}_i$, $i=1,\ldots,k_0$.

Denote the eigenvalues of a symmetric matrix B of order p by $\lambda_1(B)\ge\cdots\ge\lambda_p(B)$.

(X) There are constants a_1 and a_2 such that

$$0<a_1n_l\le\lambda_p\left(W_{n_l}\right)\le\lambda_1\left(W_{n_l}\right)\le a_2n_l\quad\text{for large enough } n_l.\tag{3.5}$$

The following three assumptions are on d_{n_l}. Recall that $d_{n_l}^2=\max_{1\le j\le n_l}\mathbf{x}'_{j,\mathbb{O}_l}W_{n_l}^{-1}\mathbf{x}_{j,\mathbb{O}_l}$.

(X1) $d_{n_l}(\log\log n_l)^{1/2}\to 0$ as $n_l\to\infty$.

(X2) $d_{n_l}(\log n_l)^{1+\iota}=O(1)$, where $\iota>0$ is a constant.

(X3) When n_l is large enough, there exists a constant $\omega>0$ such that $d_{n_l}\le\omega n_l^{-1/2}$.

Remark 3.2. Assumptions (X) and (X1)–(X3) describe essentially the behavior of the explanatory variables. Assumptions (X1)–(X3) are imposed so that d_{n_l} converges to 0 at certain rates. It can be seen that Assumption (X) is satisfied almost surely if $\mathbf{x}_i$, $i=1,2,\ldots$, are independently and identically distributed observations of a random vector $\mathbf{X}$ with

strictly positive definite covariance matrix. If we further assume that $|\mathbf{X}|$ is finite, then (X1)–(X3) are met almost surely.

(Z) The sequence $\{A_n\}$ satisfies

$$\frac{A_n}{n} \longrightarrow 0, \qquad \frac{A_n}{\log\log n} \longrightarrow \infty. \tag{3.6}$$

Excluding Assumption (A), all other assumptions are ordinarily used in the study of limiting behavior of an M-estimator. The only difference is that we now require them to hold in any sth-class, $1 \le s \le k$.

4. Limiting behavior of the criterion MR-C

Suppose that (B1)–(B3), (C1)–(C3), (X), (X1), and (Z) hold.

Let $\Pi_{k_0}^{(n)}$ be the true underlying partition of the n objects with the model structure (2.1). Observe that the true partition $\Pi_{k_0}^{(n)}$ is a sequence of naturally nested classifications as n increases, that is,

$$\mathbb{O}_i^{(n)} \subseteq \mathbb{O}_i^{(n+1)}, \quad i = 1,\ldots,k_0, \text{ for large } n. \tag{4.1}$$

Consider a given sequence of classifications with k clusters $\Pi_k^{(n)} = \{\mathcal{C}_1^{(n)},\ldots,\mathcal{C}_k^{(n)}\}$ of $\mathbb{O}^{(n)}$ such that

$$\mathcal{C}_s^{(n)} \subseteq \mathcal{C}_s^{(n+1)}, \quad s = 1,\ldots,k, \text{ for large } n, \tag{4.2}$$

when n increases. For simplicity, when no confusion appears, n will be suppressed in $\Pi_{k_0}^{(n)}$, $\Pi_k^{(n)}$, $\mathbb{O}_i^{(n)}$, $1 \le i \le k_0$, and $\mathcal{C}_s^{(n)}$, $1 \le s \le k$.

Consider the following two cases.

Case 1. $k_0 < k < K$, where $K < \infty$ is a fixed constant:

$$\begin{aligned} &R_n(\Pi_k^{(n)}) - R_n(\Pi_{k_0}^{(n)}) \\ &\quad = \sum_{s=1}^{k} \sum_{j\in\mathcal{C}_s} \rho(y_{j,\mathcal{C}_s} - \mathbf{x}'_{j,\mathcal{C}_s}\widehat{\boldsymbol{\beta}}_s) - \sum_{i=1}^{k_0} \sum_{j\in\mathbb{O}_i} \rho(y_{j,\mathbb{O}_i} - \mathbf{x}'_{j,\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}) + (q(k) - q(k_0))A_n, \end{aligned} \tag{4.3}$$

where

$$\widehat{\boldsymbol{\beta}}_s = \arg\min_{\boldsymbol{\beta}} \sum_{j\in\mathcal{C}_s} \rho(y_{j,\mathcal{C}_s} - \mathbf{x}'_{j,\mathcal{C}_s}\boldsymbol{\beta}), \tag{4.4}$$

$$\widehat{\boldsymbol{\beta}}_{0i} = \arg\min_{\boldsymbol{\beta}} \sum_{j\in\mathbb{O}_i} \rho(y_{j,\mathbb{O}_i} - \mathbf{x}'_{j,\mathbb{O}_i}\boldsymbol{\beta}). \tag{4.5}$$

Since we have $k_0 < k < K < \infty$, the number of possible intersection sets $\mathcal{C}_s \cap \mathbb{O}_i$ is finite, and

$$\mathbb{O}^{(n)} = \cup_{i=1}^{k_0}\mathbb{O}_i = \cup_{s=1}^{k}\mathcal{C}_s = \cup_{s=1}^{k}\cup_{i=1}^{k_0}(\mathcal{C}_s \cap \mathbb{O}_i). \tag{4.6}$$

Hence

$$\begin{aligned}
&R_n(\Pi_k^{(n)}) - R_n(\Pi_{k_0}^{(n)}) \\
&= \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] + (q(k) - q(k_0))A_n \\
&= \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)\right] \\
&\quad + \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\quad + (q(k) - q(k_0))A_n,
\end{aligned} \tag{4.7}$$

where $\widehat{\boldsymbol{\beta}}_{0si}$ is the M-estimator defined by

$$\widehat{\boldsymbol{\beta}}_{0si} = \arg\min_{\boldsymbol{\beta}} \sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\boldsymbol{\beta}\right). \tag{4.8}$$

By (4.4) and (4.8), we have

$$\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)\right] \geq 0. \tag{4.9}$$

By (A.3) of Lemma A.2, (4.5), (4.8), and the fact that $\mathscr{C}_s \cap \mathbb{O}_i$ is a subset of the true class $\mathbb{O}_i$, we have that

$$\begin{aligned}
\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\right)\right] &= O(\log\log n), \\
\sum_{i=1}^{k_0}\sum_{j\in\mathbb{O}_i}\left[\rho\left(y_{j,\mathbb{O}_i} - \mathbf{x}'_{j,\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right) - \rho\left(e_{j,\mathbb{O}_i}\right)\right] &= O(\log\log n).
\end{aligned} \tag{4.10}$$

Note that $\bigcup_{s=1}^{k}\bigcup_{i\in k_0}\bigcup_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\{e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\}$ is the same as $\bigcup_{i=1}^{k_0}\bigcup_{j\in\mathbb{O}_i}\{e_{j,\mathbb{O}_i}\}$. Hence we have that

$$\sum_{i=1}^{k_0}\sum_{j\in\mathbb{O}_i}\rho\left(e_{j,\mathbb{O}_i}\right) \equiv \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\rho\left(e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\right). \tag{4.11}$$

We further have

$$\begin{aligned}
&\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i}-\mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)-\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i}-\mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right]\\
&\quad=\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i}\left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i}-\mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)-\rho\left(e_{j,\mathscr{C}_s\cap\mathbb{O}_i}\right)\right]\\
&\quad\quad-\sum_{i=1}^{k_0}\sum_{j\in\mathbb{O}_i}\left[\rho\left(y_{j,\mathbb{O}_i}-\mathbf{x}'_{j,\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)-\rho\left(e_{j,\mathbb{O}_i}\right)\right]=O(\log\log n).
\end{aligned}\tag{4.12}$$

Therefore, by (4.9), (4.12), Assumption (Z), and the fact that $q(k)-q(k_0)>0$, we obtain that

$$R_n\left(\Pi_k^{(n)}\right)-R_n\left(\Pi_{k_0}^{(n)}\right)>0,\quad \text{a.s.}\tag{4.13}$$

for n large enough.

Case 2. $k<k_0$.

By [6, Lemma 3.1] for any partition $\Pi_k^{(n)}=\{\mathscr{C}_1,\ldots,\mathscr{C}_k\}$ of $\mathbb{O}^{(n)}$, there exist one class in $\Pi_k^{(n)}$ and two distinct components in the true partition $\Pi_{k_0}^{(n)}=\{\mathbb{O}_1,\ldots,\mathbb{O}_{k_0}\}$, say $\mathscr{C}_1\in\Pi_k^{(n)}$ and $\mathbb{O}_1,\mathbb{O}_2\in\Pi_{k_0}^{(n)}$ such that

$$b_0 n<\left|\mathscr{C}_1\cap\mathbb{O}_1\right|<n,\qquad b_0 n<\left|\mathscr{C}_1\cap\mathbb{O}_2\right|<n,\tag{4.14}$$

where $b_0=a_0/k_0>0$ is a constant.

Let $d_0=\min_{1\le i\ne l\le k_0}|\boldsymbol{\beta}_{0i}-\boldsymbol{\beta}_{0l}|$. Then $d_0>0$ is a fixed constant. Consider

$$\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right),\qquad \sum_{j\in\mathscr{C}_1\cap\mathbb{O}_2}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_2}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_2}\widehat{\boldsymbol{\beta}}_1\right),\tag{4.15}$$

where $\widehat{\boldsymbol{\beta}}_1$ is the M-estimator in $\mathscr{C}_1$ defined in (4.4) with $s=1$. Then in view of the convexity of $\rho(\cdot)$, by (4.4), (4.14), and the fact that $\boldsymbol{\beta}_{01}$, $\boldsymbol{\beta}_{02}$ are two distinct underlying true parameter vectors in the model structure (2.1), at least one of the following two inequalities must hold:

$$\begin{aligned}
&\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right)\\
&\quad>\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widetilde{\boldsymbol{\beta}}\right)\quad\forall\widetilde{\boldsymbol{\beta}}:\left|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_{01}\right|\le\frac{d_0}{4},
\end{aligned}\tag{4.16}$$

$$\begin{aligned}
&\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_2}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_2}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_2}\widehat{\boldsymbol{\beta}}_1\right)\\
&\quad>\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_2}\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_2}-\mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_2}\widetilde{\boldsymbol{\beta}}\right),\quad\forall\widetilde{\boldsymbol{\beta}}:\left|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_{02}\right|\le\frac{d_0}{4}.
\end{aligned}\tag{4.17}$$

Without loss of generality, we assume that (4.16) holds. Now let us focus our discussion on the set $\mathscr{C}_1 \cap \mathbb{C}_1$. Let $n_{11} = |\mathscr{C}_1 \cap \mathbb{C}_1|$ be the number of objects in the set $\mathscr{C}_1 \cap \mathbb{C}_1$. We intend to find the order of

$$\sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_1) - \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011})] \stackrel{\text{def}}{=} T \tag{4.18}$$

in terms of n as n increases to infinity. In the above expression for T, $\widehat{\boldsymbol{\beta}}_{01}$ is the M-estimator in $\mathbb{C}_1$ defined in (4.5) with $i = 1$ and $\widehat{\boldsymbol{\beta}}_{011}$ is the M-estimator in $\mathscr{C}_1 \cap \mathbb{C}_1$ defined as follows:

$$\widehat{\boldsymbol{\beta}}_{011} = \arg\min_{\boldsymbol{\beta}} \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \boldsymbol{\beta}). \tag{4.19}$$

$\mathscr{C}_1 \cap \mathbb{C}_1$ is a subset of $\mathbb{C}_1 \in \Pi_{k_0}^{(n)}$ which is the underlying true classification of $\mathbb{C}^{(n)}$. By (A.4), Lemma A.2, with probability one, $|\widehat{\boldsymbol{\beta}}_{011} - \boldsymbol{\beta}_{01}| < d_0/4$ for n_{11} large enough, where $\widehat{\boldsymbol{\beta}}_{011}$ is defined in (4.19). Let $\overline{D} \stackrel{\text{def}}{=} \{\widetilde{\boldsymbol{\beta}} : |\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_{01}| = d_0/4\}$. Then by (4.16), it is certain that there exists a point $\widetilde{\boldsymbol{\beta}}_{\overline{D}} \in \overline{D}$ such that

$$\sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_1) > \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widetilde{\boldsymbol{\beta}}_{\overline{D}}). \tag{4.20}$$

Hence

$$\begin{aligned} T &> \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widetilde{\boldsymbol{\beta}}_{\overline{D}}) - \rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011})] \\ &= \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widetilde{\boldsymbol{\beta}}_{\overline{D}}) - \mathrm{E}(\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}))] \\ &\quad - \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011}) - \mathrm{E}(\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}))] \stackrel{\text{def}}{=} T_1 + T_2. \end{aligned} \tag{4.21}$$

By (A.6), Lemma A.3, there exists a constant $\delta > 0$ such that

$$T_1 \geq \delta n_{11} + o(n_{11}), \quad \text{a.s.} \tag{4.22}$$

Write $T_2 = T_{21} + T_{22}$ with

$$\begin{aligned} T_{21} &= \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(y_{j,\mathscr{C}_1 \cap \mathbb{C}_1} - \mathbf{x}'_{j,\mathscr{C}_1 \cap \mathbb{C}_1} \widehat{\boldsymbol{\beta}}_{011}) - \rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1})], \\ T_{22} &= \sum_{j \in \mathscr{C}_1 \cap \mathbb{C}_1} [\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}) - \mathrm{E}(\rho(e_{j,\mathscr{C}_1 \cap \mathbb{C}_1}))]. \end{aligned} \tag{4.23}$$

By (A.3), Lemma A.2, and (4.2), we have

$$T_{21} = O(\log\log n_{11}), \quad \text{a.s.} \tag{4.24}$$

By (B2), (4.2), and the strong law of large numbers, we obtain

$$T_{22} = o(n_{11}), \quad \text{a.s.} \tag{4.25}$$

Hence, by (4.24) and (4.25), we have

$$T_2 = o(n_{11}), \quad \text{a.s.} \tag{4.26}$$

In view of (4.21), (4.22), and (4.26), it follows that

$$\sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1} \left[\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right) - \rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_{011}\right)\right] > \delta n_{11} + o(n_{11}), \quad \text{a.s.} \tag{4.27}$$

By (4.9), we can express our object function as follows:

$$\begin{aligned}
&R_n\left(\Pi_k^{(n)}\right) - R_n\left(\Pi_{k_0}^{(n)}\right) \\
&\quad = \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_s\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right)\right] \\
&\qquad + \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\qquad + \left(q(k) - q(k_0)\right)A_n \\
&\quad \geq \sum_{j\in\mathscr{C}_1\cap\mathbb{O}_1} \left[\rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_1\right) - \rho\left(y_{j,\mathscr{C}_1\cap\mathbb{O}_1} - \mathbf{x}'_{j,\mathscr{C}_1\cap\mathbb{O}_1}\widehat{\boldsymbol{\beta}}_{011}\right)\right] \\
&\qquad + \sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\qquad + \left(q(k) - q(k_0)\right)A_n,
\end{aligned} \tag{4.28}$$

where $\widehat{\boldsymbol{\beta}}_s$, $1 \leq s \leq k$, and $\widehat{\boldsymbol{\beta}}_{0i}$, $1 \leq i \leq k_0$, are defined in (4.4) and (4.5). By the same argument as used in Case 1, we have

$$\begin{aligned}
&\sum_{s=1}^{k}\sum_{i=1}^{k_0}\sum_{j\in\mathscr{C}_s\cap\mathbb{O}_i} \left[\rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0si}\right) - \rho\left(y_{j,\mathscr{C}_s\cap\mathbb{O}_i} - \mathbf{x}'_{j,\mathscr{C}_s\cap\mathbb{O}_i}\widehat{\boldsymbol{\beta}}_{0i}\right)\right] \\
&\qquad = O(\log\log n) = o(n).
\end{aligned} \tag{4.29}$$

Therefore, by (3.6), (4.14), (4.27), and (4.29), we have

$$\begin{aligned}
&R_n\left(\Pi_k^{(n)}\right) - R_n\left(\Pi_{k_0}^{(n)}\right) \\
&\qquad > \delta n_{11} + o(n_{11}) + o(n) + \left[q(k) - q(k_0)\right]A_n \\
&\qquad \geq \delta b_0 n + o(n) + \left[q(k) - q(k_0)\right]A_n > 0, \quad \text{a.s.}
\end{aligned} \tag{4.30}$$

for n large enough.

Table 5.1. Parameter values used in the simulation study of regression clustering.

Case	k_0	Regression coefficients	No. of obs.
1	1	$\boldsymbol{\beta}_0 = \binom{1}{6}$	$n = 120$
2	2	$\boldsymbol{\beta}_{01} = \binom{20}{9}, \boldsymbol{\beta}_{02} = \binom{1}{6}$	$n_1 = 70$ $n_2 = 50$
3	3	$\boldsymbol{\beta}_{01} = \binom{30}{9}, \boldsymbol{\beta}_{02} = \binom{12}{8}, \boldsymbol{\beta}_{03} = \binom{-2}{9}$	$n_1 = 35$ $n_2 = 35$ $n_3 = 50$

Therefore, combining the results from (4.13) in Case 1 and (4.30) in Case 2, we have showed that the true classification is attained when n increases to infinity.

Remark 4.1. In the above discussion, the set of the conditions (B1)–(B3), (C1)–(C3), (X), (X1), and (Z) can be replaced by any set of the following conditions:

(a) (A), (B1)–(B3), (C1)-(C2), (C4), (X), (X2), and (Z);

(b) (A), (B1)-(B2), (C1)-(C2), (C4), (X), (X3), and (Z);

(c) (A), (B1)-(B2), (C1)–(C3), (X), (X3), and (Z).

Remark 4.2. Hannan and Quinn [13] show that $A_n = c\log\log n$ is sufficient for strong consistency in a classical estimation procedure for the order of an autoregression. By computing the upper bound in our proofs carefully, we can show that $A_n = c\log\log n$ also works here.

Remark 4.3. The above study is not feasible when all possible classifications are considered simultaneously. For simplicity, we consider the quadratic ρ function, that is, $\rho(t) = t^2$. Let $D_n = \{\text{all nonempty subsets of } \mathbb{O}\}$, then for any $l \in \{1, 2, \ldots, p\}$,

$$\max_{d\in D_n}\left|\sum_{j\in d} x_j^{(l)} e_j\right| \geq \max\left(\sum_{j:x_j^{(l)}e_j>0} x_j^{(l)} e_j, \sum_{j:x_j^{(l)}e_j<0} (-x_j^{(l)} e_j)\right) \geq \frac{1}{2}\sum_{j=1}^{n}\left|x_j^{(l)} e_j\right|. \tag{4.31}$$

Note that in general, $\sum_{j=1}^{n} |x_j^{(l)} e_j| = O(n)$ for any $l \in \{1, 2, \ldots, p\}$. Hence the key equation (A.2), Lemma A.2 does not hold uniformly for all possible subsets of $\mathbb{O} = \{1, 2, \ldots, n\}$.

5. A simulation study

In this section, we present a simulation study for the finite sample performance of the criterion MR-C. In this simulation, $q(k) = 3k(p + 3)$, where p is the known number of regression coefficients in the model structure (2.1) and k is the number of clusters we consider. Since $\lim_{t\to 0}[(\log n)^t - 1]/t = \log\log n$ holds, by Remark 4.2 in Section 4, we let $A_n^{(i)} = (1/\lambda_i)((\log n)^{\lambda_i}) - 1$, with $\lambda_1 = 1.6$, $\lambda_2 = 1.8$, $\lambda_3 = 2.0$, and $\lambda_4 = 2.2$ employed in the simulation.

We consider one cluster, two cluster and three cluster cases, respectively. In all cases, the covariate is generated from $N(0, 1)$. The parameter values used for each case are given in Table 5.1. $N(0, 1)$ and Cauchy(0, 1) random error terms are used to generate the data

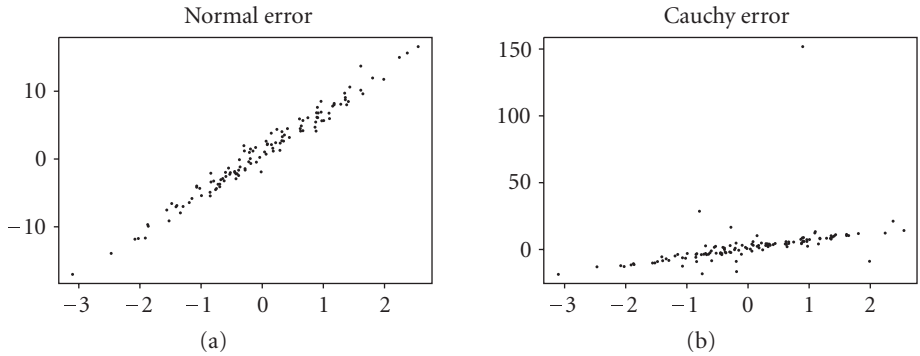


FIGURE 5.1. Plots of simulated data for one homogeneous cluster.

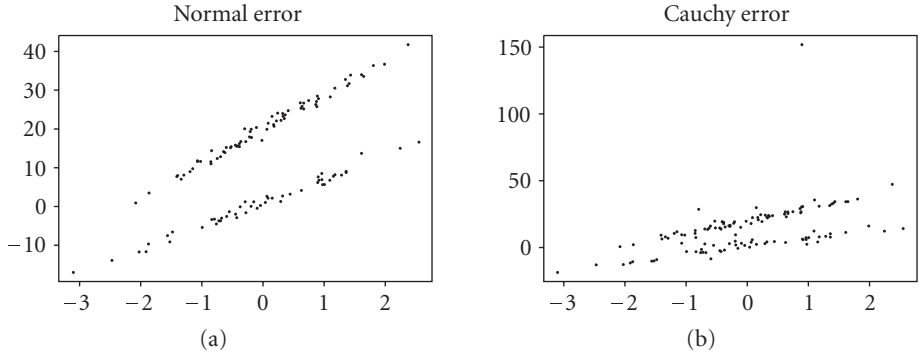


FIGURE 5.2. Plots of simulated data for two separated linear patterns.

for each of the above three cases, respectively. Therefore, in all, we actually consider six models. We use a shorthand notation to identify them:

(i) N1C1 (C1C1) Case 1, one single line, normal (Cauchy) errors;

(ii) N1C2 (C1C2) Case 2, two separated lines, normal (Cauchy) errors;

(iii) N1C3 (C1C3) Case 3, three separated lines, normal (Cauchy) errors.

The ρ functions we employed for M-estimator are (1) $\rho_1(u) = u^2$ (LS); (2) $\rho_2(u) = 0.5u^2$ if $|u| \leq 1.345$ and $\rho_2(u) = 1.345|u| - 0.5 \times 1.345^2$ otherwise (Huber ρ). When ρ is the quadratic discrepancy function, MR-C coincides with LS-C. In the following, MR-C stands for the M-estimator-based regression clustering procedure with Huber's ρ. In order to keep the same scale between LS-C and MR-C, the actual LS-C implemented in this simulation study is to minimize $\sum_{s=1}^{k} \sum_{j \in \mathscr{C}_s} (y_{j,\mathscr{C}_s} - \mathbf{x}'_{j,\mathscr{C}_s} \widehat{\boldsymbol{\beta}}_s)^2/2 + q(k)A_n$ over all possible partitions. It is clear that this slight modification does not affect the asymptotic property of LS-C.

Figures 5.1, 5.2, and 5.3 give us an intuitive idea of what the data look like for Cases 1, 2, and 3 with $N(0,1)$ and Cauchy(0, 1) errors, respectively. These figures show that the groupings of linear patterns are quite apparent and clear in each case for the normal error

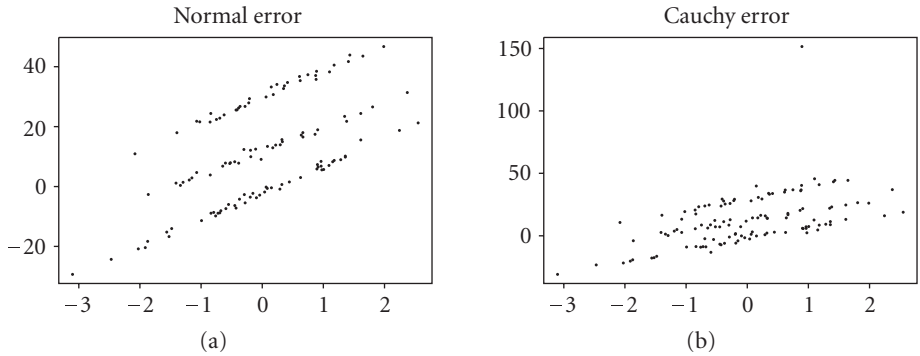


Figure 5.3. Plots of simulated data for three separated linear patterns.

Table 5.2. Relative frequencies of selecting k based on 500 simulations (Case 1).

	$e_j \sim N(0,1)$, N1C1							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k=2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=3$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$e_j \sim$ Cauchy$(0,1)$, C1C1							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.588	0.772	0.860	0.926
$k=2$	0.160	0.170	0.186	0.216	0.150	0.076	0.042	0.010
$k=3$	0.368	0.392	0.412	0.432	0.130	0.078	0.054	0.042
$k=4$	0.472	0.438	0.402	0.352	0.132	0.074	0.044	0.022

† The true number of clusters $k_0 = 1$.

models while there are some outliers far away from the whole pattern for each case with Cauchy errors.

For each of the aforementioned six models, we generate the data by the model structure (2.1), we then use LS-C and MR-C to select the number of clusters and classify the data. This process is then repeated 500 times separately. To reduce the computation burden, we only fit models with possible numbers of clusters as 1, 2, 3, 4 when the true number of clusters k_0 is 1 or 2; and we only consider possible cluster size of 1, 2, 3, 4, and 5, when k_0 is 3.

In the simulation study, LS-C and MR-C are used to select the best k, respectively. Tables 5.2, 5.3, and 5.4 display the relative frequencies of selecting k for each of the six models using LS-C and MR-C separately. It is apparent that both Huber ρ and LS functions perform extremely well for these models with normal errors. However, as shown in

Table 5.3. Relative frequencies of selecting k based on 500 simulations (Case 2).

	$e_j \sim N(0,1)$, N1C2							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k=3$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$e_j \sim$ Cauchy$(0,1)$, C1C2							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	0.086	0.088	0.098	0.114	0.898	0.924	0.956	0.970
$k=3$	0.278	0.294	0.318	0.346	0.054	0.036	0.022	0.018
$k=4$	0.636	0.618	0.584	0.540	0.048	0.040	0.022	0.012

† The true number of clusters $k_0 = 2$.

Table 5.4. Relative frequencies of selecting k based on 500 simulations (Case 3).

	$e_j \sim N(0,1)$, N1C3							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=3$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$k=4$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=5$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	$e_j \sim$ Cauchy$(0,1)$, C1C3							
	LS-C				MR-C			
Model	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$	$A_n^{(1)}$	$A_n^{(2)}$	$A_n^{(3)}$	$A_n^{(4)}$
$k=1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k=2$	0.052	0.052	0.052	0.052	0.000	0.000	0.000	0.000
$k=3$	0.148	0.174	0.196	0.214	0.932	0.950	0.958	0.972
$k=4$	0.370	0.382	0.396	0.394	0.056	0.040	0.034	0.022
$k=5$	0.430	0.392	0.356	0.340	0.012	0.010	0.008	0.006

† The true number of clusters $k_0 = 3$.

these tables, in contrast to the nearly perfect performance of both criteria in the normal error models, when the random errors used to generate the data in each case are from Cauchy(0, 1), MR-C with Huber ρ function still selects the underlying true numbers of clusters with promising high proportions of correctness while LS-C loses the power of detecting the underlying numbers of clusters significantly.

Appendix

Let $\mathbb{O}_l$ be any component or a subset of a component of the underlying true partition $\Pi_{k_0}^{(n)} = \{\mathbb{O}_1^{(n)}, \ldots, \mathbb{O}_{k_0}^{(n)}\}$ of the n objects $\mathbb{O}^{(n)}$. Let $n_l = |\mathbb{O}_l|$. The following lemmas hold in $\mathbb{O}_l$ and can be proved similarly as by Wu and Zen [11].

Lemma A.1. *Suppose that (B1), (C1)-(C2), (X), and (X1) hold. Then,*

$$\frac{1}{n_l} \sum_{j \in \mathbb{O}_l} [\gamma_j - \mathrm{E}(\gamma_j)] \longrightarrow 0, \quad a.s., \tag{A.1}$$

where $\gamma_j = \rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}) - \rho(e_{j,\mathbb{O}_l}) + \mathbf{x}'_{j,\mathbb{O}_l}(\boldsymbol{\beta} - \boldsymbol{\beta}_{0l})\psi(e_{j,\mathbb{O}_l})$ *if* $|\boldsymbol{\beta} - \boldsymbol{\beta}_{0l}| > 0$.

Lemma A.2. *Suppose that the Assumptions (B1)–(B3), (C1)–(C3), (X), and (X1) hold. Then*

$$\sum_{j \in \mathbb{O}_l} \mathbf{x}_{j,\mathbb{O}_l} \psi(e_{j,\mathbb{O}_l}) = O\left((n_l \log\log n_l)^{1/2}\right), \quad a.s. \tag{A.2}$$

$$\sum_{j \in \mathbb{O}_l} [\rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\hat{\boldsymbol{\beta}}_{0l}) - \rho(e_{j,\mathbb{O}_l})] = O(\log\log n_l), \quad a.s. \tag{A.3}$$

$$\hat{\boldsymbol{\beta}}_{0l} = \boldsymbol{\beta}_{0l} + O\left((\log\log n_l / n_l)^{1/2}\right), \quad a.s., \tag{A.4}$$

where

$$\hat{\boldsymbol{\beta}}_{0l} = \arg\min_{\boldsymbol{\beta}} \sum_{j \in \mathbb{O}_l} \rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}). \tag{A.5}$$

Lemma A.3. *Suppose that the Assumptions (B1), (B2), (B3), (C2), (X), and (X1) hold. Then there exists a constant* $\delta > 0$ *such that*

$$\sum_{j \in \mathbb{O}_l} [\rho(y_{j,\mathbb{O}_l} - \mathbf{x}'_{j,\mathbb{O}_l}\boldsymbol{\beta}^*) - \mathrm{E}(\rho(e_{j,\mathbb{O}_l}))] \geq \delta n_l + o(n_l), \quad a.s. \tag{A.6}$$

holds for all $\boldsymbol{\beta}^* \in \overline{\mathcal{D}}$ *and* n_l *large enough, where* $\overline{\mathcal{D}}$ *is defined in the preceding lemma.*

Acknowledgments

The authors would like to thank the referees for comments and suggestions that improved the presentation of this paper. The research was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] H.-H. Bock, "Probability models and hypotheses testing in partitioning cluster analysis," in *Clustering and Classification*, P. Arabie, L. J. Hubert, and G. De Soete, Eds., pp. 377–453, World Scientific, River Edge, NJ, USA, 1996.

[2] R. E. Quandt and J. B. Ramsey, "Estimating mixtures of normal distributions and switching regressions," *Journal of the American Statistical Association*, vol. 73, no. 364, pp. 730–752, 1978.

[3] H. Späth, "A fast algorithm for clusterwise linear regression," *Computing*, vol. 29, no. 2, pp. 175–181, 1982.

[4] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression," *Journal of Classification*, vol. 5, no. 2, pp. 249–282, 1988.

[5] S. Lou, J. Jiang, and K. Keng, "Clustering objects generated by linear regression models," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1356–1362, 1993.

[6] Q. Shao and Y. Wu, "A consistent procedure for determining the number of clusters in regression clustering," *Journal of Statistical Planning and Inference*, vol. 135, no. 2, pp. 461–476, 2005.

[7] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, New York, NY, USA, 1986.

[8] P. J. Huber, "Robust regression: asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, vol. 1, pp. 799–821, 1973.

[9] S. Konishi and G. Kitagawa, "Generalised information criteria in model selection," *Biometrika*, vol. 83, no. 4, pp. 875–890, 1996.

[10] J. A. F. Machado, "Robust model selection and M-estimation," *Econometric Theory*, vol. 9, no. 3, pp. 478–493, 1993.

[11] Y. Wu and M. M. Zen, "A strongly consistent information criterion for linear model selection based on M-estimation," *Probability Theory and Related Fields*, vol. 113, no. 4, pp. 599–625, 1999.

[12] R. T. Rockafellar, *Convex Analysis*, Princeton Mathematical Series, no. 28, Princeton University Press, Princeton, NJ, USA, 1970.

[13] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society. Series B*, vol. 41, no. 2, pp. 190–195, 1979.

C. R. Rao: Department of Statistics, Penn State University, University Park, PA 16802, USA
Email address: crr1@psu.edu

Y. Wu: Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3
Email address: wuyh@mathstat.yorku.ca

Q. Shao: Novartis Pharmaceuticals Corporation, East Hanover, NJ 07936, USA
Email address: qing.shao@novartis.com

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 56372, 12 pages
doi:10.1155/2007/56372

Research Article

Methods for Stratified Cluster Sampling with Informative Stratification

Alastair Scott and Chris Wild

Received 24 April 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

We look at fitting regression models using data from stratified cluster samples when the strata may depend in some way on the observed responses within clusters. One important subclass of examples is that of family studies in genetic epidemiology, where the probability of selecting a family into the study depends on the incidence of disease within the family. We develop the survey-weighted estimating equation approach for this problem, with particular emphasis on the estimation of superpopulation parameters. Full maximum likelihood for this class of problems involves modelling the population distribution of the covariates which is simply not feasible when there are a large number of potential covariates. We discuss efficient semiparametric maximum likelihood methods in which the covariate distribution is left completely unspecified. We further discuss the relative efficiencies of these two approaches.

1. Introduction

In this paper, we look at the problem of fitting models to data from stratified cluster samples. We are particularly interested in situations where the probability that a cluster is in a particular stratum depends on the value of its response. Sometimes this dependence is explicit and obvious. An important special case of this situation is the case-control family study, which is widely used in genetic epidemiology (see Neuhaus et al. [1, 2]). In a simple case-control study, we stratify the population into cases (people with a disease under study, say) and controls (people without the disease), choose independent random samples from each stratum, and record the values of potential covariates for each person selected in the study. In a case-control family study, we record the same information

and, in addition, we identify a set of family members for each person selected in the case-control study and record the disease status and the values of the covariates of each of these family members. For example, Whittemore [3] considers a case-control family study of ovarian cancer. Here, the clusters consist of mother-daughter pairs. The case stratum contains all pairs in which the daughter has been diagnosed with ovarian cancer, and the control stratum contains all the other pairs. Other examples of similar retrospective family studies are given in Zhao et al. [4].

Another example where the strata are determined explicitly by the response is given by Neuhaus and Jewell [5]. They consider data from a stratified cluster sample of individuals enrolled in the Federal Employees Health Benefit Plan in which the response variable indicates whether or not someone used outpatient mental health services during the previous year for each of the years 1979–1981. Here, the clusters consist of the three observations for a single person, and four strata were defined by the total number of times (0, 1, 2, or 3) that the person used the service in the three years of the study.

In all these examples, stratum membership is determined exactly by the value of the (multivariate) response. In most surveys, however, the relationship between the response and stratum membership is less clearcut, with the strata determined by such things as administrative convenience or the availability of a suitable list. This is true even for some case-control family studies. For example, in the study that motivated this work, Wrensch et al. [6] conducted a population-based case-control study of glioma, the most common type of malignant brain tumour, in the San Francisco Bay Area. The investigators gathered all cases of glioma diagnosed in a specified time interval and a population-based sample of comparable controls through random digit dialling. They also gathered the brain tumour status and covariate information from family members of the original case-control sample participants. In this case, the case stratum contains all families with at least one member diagnosed with glioma in the specified time interval. The chance of a family being included in this stratum depends on the number of family members with glioma, but is not completely determined by this.

To cover these more general cases, we consider situations in which we may have to fit a parametric model, $P_h(\mathbf{y}, \mathbf{X}; \boldsymbol{\gamma})$, for the conditional probability of a cluster being included in the hth stratum given values of the response vector, $\mathbf{y}$, and the matrix of covariates, $\mathbf{X}$. Note that there are no problems if this stratum inclusion model only involves $\mathbf{X}$. However, if the model depends on the response, $\mathbf{y}$, as well or, more generally depends on a design variable that is associated with $\mathbf{y}$ but is not included in the model, then the sampling is not ignorable (cf. Rubin [7]) and will affect the likelihood.

A possible strategy that is sometimes suggested for coping with informative stratification is to include the stratum indicator as a covariate in the model. This strategy avoids the technical problems but clearly makes no sense when the stratification depends solely on the response as in many of the examples above. Even in situations where the stratification does not depend directly on the response, it may distort the relationship between $\mathbf{y}$ and $\mathbf{X}$, which is the quantity of interest. For example, Lee et al. [8] consider a secondary analysis of data from a case-control study of Sudden Infant Death Syndrome (SIDS). The response in this analysis was an indicator of immunization, and clearly there would be

little sense in including the stratifying variable (SIDS) as a predictor in the model. Similarly, in our motivating brain cancer example, the researchers did not want to include the date of diagnosis in their predictive model.

The standard survey approach is through weighted estimating equations, with weights inversely proportional to the selection probabilities, as in Binder [9] or Rao et al. [10]. This works well when the weights are reasonably homogeneous but can be inefficient when the weights vary widely as they tend to do in retrospective studies. De Mets and Halperin [11] and Smith [10] looked at a more efficient approach which involved modelling the joint distribution of response, covariates, and design variables used for the stratification. This is efficient but becomes very difficult to implement when there are a large number of potential explanatory variables. In this paper, we look at an efficient intermediate approach based on semiparametric maximum likelihood in which the marginal distribution of the covariates is left unspecified. The general setup is described in Section 2. In Section 3, we examine the survey-weighted approach in some detail and the semi-parametric theory is developed in Section 4. We conclude with a brief discussion.

2. Basic setup

As in the introduction, we let $\mathbf{y}$ denote the vector of responses for the units in a cluster and we let $\mathbf{X}$ be the corresponding matrix of covariate values. In addition, we define a stratum indicator variable Z which takes the value $Z = h$ if the cluster is assigned to the hth stratum ($h = 1,\ldots,L$). We assume that the values in our finite population of N clusters are generated by random sampling from the joint distribution of $(\mathbf{y},\mathbf{X},Z)$. The clusters are then sorted into L strata, $\mathcal{S}_1,\ldots,\mathcal{S}_L$, according to the values of Z, resulting in N_h clusters in $\mathcal{S}_h$ ($\sum_1^L N_h = N$). Finally, we draw independent simple random samples, D_h, of n_h clusters from the N_h clusters in $\mathcal{S}_h$ ($h = 1,\ldots,L$) and observe the corresponding $(\mathbf{y},\mathbf{X})$ values. Let $(\mathbf{y}_{hj},\mathbf{X}_{hj})$ represent observed values for the jth cluster in the hth stratum ($h = 1,\ldots,L$; $j = 1,\ldots,n_h$). Our data are thus of the form $\{(\mathbf{y}_{hj},\mathbf{X}_{hj}, j \in D_h),\ N_h;\ h = 1,\ldots,L\}$. Note that the observed stratum sizes, $N_1,\ldots,N_L$, are random variables in this scenario and contain valuable information.

We are interested in modelling $f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta})$, the conditional distribution of the vector of cluster responses $\mathbf{y}$ given $\mathbf{X}$, the matrix of cluster covariates, and, in cases where it is needed, the conditional probabilities that the cluster falls into stratum $\mathcal{S}_h$, $h = 1,\ldots,L$ given $\mathbf{y}$ and $\mathbf{X}$:

$$\operatorname{pr}(\text{cluster} \in \mathcal{S}_h \mid \mathbf{y},\mathbf{X}) = \operatorname{pr}(Z = h \mid \mathbf{y},\mathbf{X}) = P_h(\mathbf{y},\mathbf{X};\boldsymbol{\gamma}). \tag{2.1}$$

Using an argument similar to that given in Scott and Wild [12, Appendix B], the likelihood function can be shown to be given by

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\gamma},g) &= \prod_{h=1}^{L}\left\{\prod_{j\in D_h}\operatorname{pr}(\mathbf{y}_{hj},\mathbf{X}_{hj} \mid \text{cluster} \in \mathcal{S}_h)\right\}Q_h^{N_h} \\ &= \prod_h\left(\prod_{D_h}\{f(\mathbf{y}_{hj} \mid \mathbf{X}_{hj};\boldsymbol{\theta})g(\mathbf{X}_{hj})\}Q_h^{N_h-n_h}\right), \end{aligned} \tag{2.2}$$

where $g(\mathbf{X})$ denotes the marginal density of X in the population and

$$Q_h = Q_h(\boldsymbol{\theta}, \boldsymbol{\gamma}, g) = \mathrm{pr}(Z = h) = \iint P_h(\mathbf{y}, \mathbf{X}; \boldsymbol{\gamma}) f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\theta}) g(\mathbf{X}) d\mathbf{y}\, d\mathbf{X} \tag{2.3}$$

denotes the marginal probability that a cluster is in stratum $\mathcal{S}_h$. Strictly speaking, we need to describe how we choose the sample sizes, $n_1, \ldots, n_L$. However, the kernel of the likelihood is as above for any scheme satisfying the condition that $\{n_1, \ldots, n_L\}$ depends only on $\{N_1, \ldots, N_L\}$ and not on the realized values of $(\mathbf{y}, \mathbf{X})$ (see Wild [13] for details).

If we had drawn a simple random sample from the whole population, or if the stratification depended only on $\mathbf{X}$, then the likelihood would factor into two terms, one involving $\boldsymbol{\theta}$ alone and the other involving $\boldsymbol{\gamma}$ and $g(\mathbf{X})$. This means that we could make inferences about $\boldsymbol{\theta}$ conditional on the observed values of $\mathbf{X}$ and not have to bother about terms involving $g(\mathbf{X})$. Unfortunately, we cannot ignore $g(\mathbf{X})$ when $P_h(\mathbf{y}, \mathbf{X}; \boldsymbol{\gamma})$ involves $\mathbf{y}$; just as in a case-control study, we cannot separate $\boldsymbol{\theta}$ from $g(\mathbf{x})$ because both are involved in Q_h. The most common way of coping with this is through weighted estimating equations with weights inversely proportional to the selection probabilities as in Binder [9]. We examine this approach in more detail in Section 3. It is relatively simple to implement and works well in many situations. However, it tends to be very inefficient if the selection probabilities vary widely as they often do in retrospective studies such as the case-control family studies described in the introduction. A more efficient alternative is to build a full parametric model for $g(\mathbf{x})$ and use ordinary maximum likelihood. A good description of this approach is given in Smith and Nathan [14]. It does indeed produce very efficient estimators but, unfortunately, it rapidly becomes impractical when the number of potential covariates increases. This limits its application when we have a large number of potential covariates with a mixture of continuous, categorical, and count variables, as is the case in many surveys.

Ideally, we would like a method that combines the simplicity of the weighted approach with the efficiency of maximum likelihood. In Section 4, we look at a semiparametric approach in which the marginal distribution of $\mathbf{X}$ is treated nonparametrically. In this approach, $g(\mathbf{X})$ becomes a (potentially infinite dimensional) nuisance parameter in the likelihood. The resulting estimators turn out to be very efficient while, perhaps more surprisingly, still reasonably simple to obtain.

3. Weighted estimators

If we had observed the values of $\{\mathbf{y}, \mathbf{X}, Z\}$ for every cluster in the finite population, we would estimate $\boldsymbol{\theta}$ by solving the "census" likelihood equation

$$\mathbf{S}(\boldsymbol{\theta}) = \sum_{h=1}^{L} \sum_{j=1}^{N_h} \mathbf{U}_{hj}(\boldsymbol{\theta}) = \mathbf{0}, \tag{3.1}$$

where $\mathbf{U}_{hj}(\boldsymbol{\theta}) = \partial \log f(\mathbf{y}_{hj} \mid \mathbf{X}_{hj}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}$. We will assume that the standard regularity conditions for likelihood (see, e.g., Lehmann [15, Section 7.3]) are satisfied so that

$$E\{\mathbf{S}(\boldsymbol{\theta})\} = \underline{0}, \qquad \operatorname{Cov}\{\mathbf{S}(\boldsymbol{\theta})\} = -E\left\{\frac{\partial \mathbf{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right\} = N\mathcal{J}(\boldsymbol{\theta}), \tag{3.2}$$

at the true value, $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Of course, we do not observe the whole population. However, for any fixed value of $\boldsymbol{\theta}$, $\mathbf{S}(\boldsymbol{\theta})$ is simply a vector of population totals and thus can be estimated from the sample observations by the weighted sample score,

$$\mathbf{S}_W(\boldsymbol{\theta}) = \sum_h \sum_{D_h} \frac{N_h}{n_h} \mathbf{U}_{hj}(\boldsymbol{\theta}). \tag{3.3}$$

The weighted estimator, $\widehat{\boldsymbol{\theta}}_W$, is then defined as the solution to the weighted pseudo-likelihood equation, $\mathbf{S}_W(\boldsymbol{\theta}) = \mathbf{0}$.

Under suitable regularity conditions on $\{\mathbf{U}_{hj}\}$, $\widehat{\boldsymbol{\theta}}_W$ is a consistent estimator of the finite population (or census) regression parameter, $\boldsymbol{\theta}_C$, defined as the solution to (3.1) (see, e.g., Binder [9] or Rao et al. [10]). Our interest here, however, is not in descriptive inferences about a particular finite population, but rather about modelling the underlying processes that lead different units to have different responses y. Thus, in survey sampling terminology, we are interested in estimating the superpopulation parameters. We have to take some care in deriving the properties of $\widehat{\boldsymbol{\theta}}_W$ in this framework since $N_1, \ldots, N_L$ are now random variables rather than fixed constants as in the standard finite population setup.

In sampling terminology, we can think of our situation as being equivalent to two-phase sampling for stratification. In the first phase, the finite population is generated as a random sample of size N from an (infinite) super population and the stratum to which each cluster (i.e., the value of Z) belongs is recorded. At the second phase, we draw a simple random sample of size n_h from the N_h clusters in stratum $\mathcal{S}_h$, with the values of $\{n_1, \ldots, n_L\}$ depending only on $\{N_1, \ldots, N_L\}$, and observe $\{\mathbf{y}_{hj}, \mathbf{X}_{hj}, j \in D_h\}$ for $h = 1, \ldots, L$.

We establish the results by first conditioning on $\mathbf{Z}_\mathbf{N}$, the vector of stratum indicators for the realized finite population and then averaging over the distribution of $\mathbf{Z}_\mathbf{N}$. Given $\mathbf{Z}_\mathbf{N}$, $\{N_1, \ldots, N_L\}$, and hence $\{n_1, \ldots, n_L\}$, are fixed constants and $\mathbf{U}_{hj}(\boldsymbol{\theta})$, $j \in D_h$, are i.i.d. observations from the conditional distribution of $\mathbf{U}(\boldsymbol{\theta}) = (\partial \log f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\theta}))/\partial\boldsymbol{\theta}$ given $Z = h$. Let $\boldsymbol{\mu}_h(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_h(\boldsymbol{\theta})$ denote the mean vector and covariance matrix of this conditional distribution, and let $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ denote the corresponding quantities for the unconditional distribution of $\mathbf{U}(\boldsymbol{\theta})$. Recall that $\boldsymbol{\mu}(\boldsymbol{\theta}_0) = \underline{0}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = \mathcal{J}(\boldsymbol{\theta}_0)$ under our regularity conditions. The unconditional distribution of $\{N_1, \ldots, N_L\}$ is multinomial $(N; Q_1, \ldots, Q_L)$ where, as before, Q_h is the marginal probability that $Z = h$ for $h = 1, \ldots, L$.

Note that

$$E\left\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\right\} = N\left(\sum_h \boldsymbol{\mu}_h(\boldsymbol{\theta}) Q_h\right) = N\boldsymbol{\mu}(\boldsymbol{\theta}). \tag{3.4}$$

It follows that

$$E\{\mathbf{S}_W(\boldsymbol{\theta})\} = E\left\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\right\} = N\boldsymbol{\mu}(\boldsymbol{\theta}). \tag{3.5}$$

Thus $E\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\} = \underline{0}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. In addition, using the standard results for conditional variances,

$$\operatorname{Cov}\{\mathbf{S}_W(\boldsymbol{\theta})\} = \operatorname{Cov}\left\{\sum_h N_h \boldsymbol{\mu}_h(\boldsymbol{\theta})\right\} + E\left\{\sum_h N_h^2 \frac{\boldsymbol{\Sigma}_h(\boldsymbol{\theta})}{n_h}\right\}. \tag{3.6}$$

To proceed further, we need to specify how the n_hs are chosen. We will assume that the sampling fractions n_h/N_h are fixed constants f_h with $0 < f_h \leq 1$ for $h = 1, \ldots, L$. (Of course we cannot always achieve this exactly in small samples but if $n_h = [f_h N_h]$ the difference is negligible asymptotically.) Then

$$\begin{aligned}
\operatorname{Cov}\{\mathbf{S}_W(\boldsymbol{\theta})\} &\simeq N\left\{\sum \boldsymbol{\mu}_h \boldsymbol{\mu}_h^T Q_h(1 - Q_h) - 2\sum\sum_{\ell<h} \boldsymbol{\mu}_h \boldsymbol{\mu}_\ell^T Q_h Q_\ell + \sum Q_h \frac{\boldsymbol{\Sigma}_h}{f_h}\right\} \\
&= N\left\{\sum_h Q_h\left(\boldsymbol{\mu}_h \boldsymbol{\mu}_h^T + \frac{\boldsymbol{\Sigma}_h}{f_h}\right) - \left(\sum_h Q_h \boldsymbol{\mu}_h\right)\left(\sum_h Q_h \boldsymbol{\mu}_h\right)^T\right\} \\
&= N\left\{\sum_h Q_h\left(\boldsymbol{\mu}_h \boldsymbol{\mu}_h^T + \frac{\boldsymbol{\Sigma}_h}{f_h}\right) - \boldsymbol{\mu}\boldsymbol{\mu}^T\right\} \\
&= N\sum_h Q_h\left\{\frac{\boldsymbol{\Sigma}_h}{f_h} + (\boldsymbol{\mu}_h - \boldsymbol{\mu})(\boldsymbol{\mu}_h - \boldsymbol{\mu})^T\right\}.
\end{aligned} \tag{3.7}$$

The first term is the covariance matrix that we would get if the weights were known and the second term represents the penalty we pay for incomplete knowledge about the weights. Using the relation

$$\operatorname{Cov}\{\mathbf{U}(\boldsymbol{\theta})\} = E\{\operatorname{Cov}\{\mathbf{U} \mid Z = h\}\} + \operatorname{Cov}\{E\{\mathbf{U} \mid Z = h\}\}, \tag{3.8}$$

we can also rewrite this variance in the form

$$\operatorname{Cov}\{\mathbf{S}_W(\boldsymbol{\theta})\} = N\left\{\Sigma(\boldsymbol{\theta}) + \sum_h Q_h\left(\frac{1}{f_h} - 1\right)\boldsymbol{\Sigma}_h\right\}. \tag{3.9}$$

Now the first term is the covariance matrix that we would have obtained by sampling all clusters in the finite population so that, in this representation, the second term represents the penalty that we pay for incomplete enumeration at the second phase.

Finally, it follows from the results of Chen and Rao [16] that $\mathbf{S}_W(\boldsymbol{\theta})$ is asymptotically multivariate normal as $N \to \infty$ with $n_h/N_h \to f_h$ for $h = 1,\ldots,L$ fixed. Having established the properties of $S_W(\boldsymbol{\theta})$, we use standard results for unbiased estimating equations (see, e.g., Amari and Kawanabe [17]) to invert the equation $\mathbf{S}_W(\hat{\boldsymbol{\theta}}_W) = \mathbf{0}$ and infer results for $\hat{\boldsymbol{\theta}}_W$. In particular, it follows that $\sqrt{N}(\hat{\boldsymbol{\theta}}_W - \boldsymbol{\theta})$ converges in distribution to a multivariate normal random variable with mean vector $\mathbf{0}$ and covariance matrix $N\mathbf{V}(\hat{\boldsymbol{\theta}})$, where

$$N\mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathcal{J}^{-1}(\boldsymbol{\theta}_0)\left(\sum_h Q_h\left[\frac{\boldsymbol{\Sigma}_h}{f_h} + (\boldsymbol{\mu}_h - \boldsymbol{\mu})(\boldsymbol{\mu}_h - \boldsymbol{\mu})^T\right]\right)\mathcal{J}^{-1}(\boldsymbol{\theta}_0), \tag{3.10}$$

with

$$\mathcal{J}(\boldsymbol{\theta}) = E\left\{-\frac{\partial^2 \log\left(f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right\}, \tag{3.11}$$

as $N \to \infty$. Recall that $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This means that we can rewrite $\mathbf{V}(\hat{\boldsymbol{\theta}})$ as

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = \frac{1}{N}\left[\mathcal{J}^{-1}(\boldsymbol{\theta}_0) + \mathcal{J}^{-1}(\boldsymbol{\theta}_0)\left\{\sum_h Q_h\left(\frac{1}{f_h} - 1\right)\boldsymbol{\Sigma}_h\right\}\mathcal{J}^{-1}(\boldsymbol{\theta}_0)\right]. \tag{3.12}$$

(Here we have used (3.9) to represent $\mathrm{Cov}(\mathbf{S}_W)$). The first term is what we would get if we sampled the whole population and the second term again represents the cost of incomplete enumeration.

We can estimate $\mathbf{V}(\hat{\boldsymbol{\theta}})$ by substituting $\hat{\mathbf{J}} = -(1/N)\partial\mathbf{S}_W(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}^T$ for $\mathcal{J}(\boldsymbol{\theta}_0)$, $\hat{\boldsymbol{\mu}}_h = \sum_j \mathbf{U}_{hj}/n_h$ for $\boldsymbol{\mu}_h$, $W_h = N_h/N$ for Q_h, and the ordinary within-stratum sample variance for $\boldsymbol{\Sigma}_h$. This, in conjunction with (3.10), leads to the estimator

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1}\left(\sum_h W_h^2\frac{\hat{\boldsymbol{\Sigma}}_h}{n_h}\right)\hat{\mathbf{J}}^{-1} + \frac{1}{N}\hat{\mathbf{J}}^{-1}\left\{\sum_h W_h\hat{\boldsymbol{\mu}}_h\hat{\boldsymbol{\mu}}_h^T\right\}\hat{\mathbf{J}}^{-1}. \tag{3.13}$$

The first term of (3.13), which is $O(1/n)$, is the variance estimate we would use if we assumed that the N_hs were fixed and the second term, which is $O(1/N)$, measures the effect of not knowing the N_hs in advance. This second term will be negligible in many applications.

The weighted method is relatively straightforward and most large statistical packages now include procedures for implementing it for linear and logistic regression models, although all will assume that the $\{N_h\}$s are fixed constants and thus will produce a slight underestimate of the standard errors. A big advantage over more efficient procedures is that it does not require any modeling of stratum inclusion probabilities. One important consequence of this is that the same procedure can be used for stratified two-stage sampling, where simple random subsamples are chosen from each selected cluster. More complex subsampling schemes can be handled simply by adjusting the weights in the pseudo likelihood (3.1).

In general, weighting works pretty well for standard sampling situations where the sampling fractions do not vary too much among strata. It does not work so well in situations where the sampling fractions vary widely, as they tend to do in retrospective designs like the case-control family studies discussed in the introduction. For example, Lawless et al. [18] report efficiencies of less than 15% (compared to the semiparametric estimators discussed in Section 4) for some unclustered case-control designs and Scott and Wild [19] report similar values for some special clustered designs. An appealing feature of the weighted method is its robustness to departures from the model. When the assumed regression model $f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta})$ is not valid, the fitted model produced by the weighted method can still be interpreted as estimating the best fitting model for the whole population; see Scott and Wild [20] for further discussion of this point.

4. Semiparametric estimators

We now return to the likelihood function $L(\boldsymbol{\theta},\boldsymbol{\gamma},g)$ given in (2.2). The semi-parametric maximum likelihood estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are obtained by maximizing $\ell(\boldsymbol{\theta},\boldsymbol{\gamma},g) = \log L(\boldsymbol{\theta},\boldsymbol{\gamma},g)$ over $\boldsymbol{\theta}$, $\boldsymbol{\gamma}$, and g leaving the density function $g(\cdot)$ completely undefined. Essentially, we treat $g(\cdot)$ as a (potentially infinite-dimensional) nuisance parameter. Although it might seem at first glance that this would be formidable task, it turns out that the semi-parametric MLE of $\boldsymbol{\theta}$ (and $\boldsymbol{\gamma}$) can be calculated relatively easily.

We start by reducing the problem to the simpler case in which stratum membership is determined directly by the cluster response. First, we augment the response vector $\mathbf{y}$ with the stratum indicator Z to give modified response variable $\tilde{\mathbf{y}} = \binom{\mathbf{y}}{Z}$. Next, we set $\tilde{\boldsymbol{\theta}} = \binom{\boldsymbol{\theta}}{\boldsymbol{\gamma}}$. Our problem is then reduced to that of fitting the model $f(\tilde{\mathbf{y}} \mid \mathbf{X};\tilde{\boldsymbol{\theta}})$, where

$$f(\tilde{\mathbf{y}} \mid \mathbf{X};\tilde{\boldsymbol{\theta}}) = f(z \mid \mathbf{y},\mathbf{X};\boldsymbol{\gamma})\, f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta}) = P_z(\mathbf{y},\mathbf{X};\boldsymbol{\gamma})\, f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta}), \tag{4.1}$$

to data from a stratified sample where the strata, $\mathcal{S}_h$ $(h = 1,\ldots,L)$, are determined completely by the response, $\tilde{\mathbf{y}}$. The estimating equations for the semi-parametric maximum likelihood equations in this reduced case are derived in Scott and Wild [12, 19], following earlier work by Cosslett [21]. In a companion paper in this issue, Lee [22] establishes the asymptotic efficiency of this estimator and shows that $\mathcal{J}^*(\hat{\boldsymbol{\phi}})^{-1}$ provides a consistent estimator of the variance. Similar results are obtained in Lee and Hirose [23] using a different approach based on the profile likelihood methods of Newey [24]. In the remainder of this section, we summarize the results of translating these results for the reduced case back into our original notation.

First, we define a pseudo-log-likelihood function

$$\begin{aligned}\ell^*(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi}) = &\sum_h \sum_{D_h} \log f_h^*(\mathbf{y}_{hj} \mid \mathbf{X}_{hj};\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi}) \\ &- \sum_h \{(N_h - n_h)\log(1-\pi_h) + n_h \log \pi_h\},\end{aligned} \tag{4.2}$$

where

$$f_h^*(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi}) \propto \pi_h P_h(\mathbf{y},\mathbf{X};\boldsymbol{\gamma})\, f(\mathbf{y} \mid \mathbf{X};\boldsymbol{\theta}) \tag{4.3}$$

and $\boldsymbol{\pi}$ is an L-dimensional vector of nuisance parameters. Then the semi-parametric maximum likelihood estimators, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$, of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are the appropriate components of $\hat{\boldsymbol{\phi}}$, the solution of the pseudo score equation

$$\mathbf{S}^*(\boldsymbol{\phi}) = \frac{\partial \ell^*(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \mathbf{0}, \tag{4.4}$$

where $\boldsymbol{\phi} = (\boldsymbol{\theta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\pi}^T)^T$. This means that, for the purposes of calculating the MLE of $\binom{\boldsymbol{\theta}}{\boldsymbol{\gamma}}$, we can act as if $\ell^*(\boldsymbol{\phi})$ is the log-likelihood; in essence, we have replaced a problem involving an infinite dimensional nuisance parameter, $g(\cdot)$, with one involving an L-dimensional nuisance parameter, $\boldsymbol{\pi}$.

The pseudoscore, $\mathbf{S}^*(\boldsymbol{\phi})$, has many of the properties of a standard score function. In the first place, with appropriate standardization, $\mathbf{S}^*$ is asymptotically normal as $N \to \infty$ provided $n_h/N_h \to f_h$ with $0 < f_h \leq 1$ for $h = 1, \ldots, L$. Secondly, $E\{\mathbf{S}^*(\boldsymbol{\phi})\} = \mathbf{0}$ at the true value, even though the individual terms in $\mathbf{S}^*(\boldsymbol{\phi})$ are neither identically distributed nor have expected value zero under the stratified sampling design. Finally, if we let $\mathcal{J}^*$ denote the observed (pseudo-) information matrix,

$$\mathcal{J}^*(\boldsymbol{\phi}) = -\frac{\partial \mathbf{S}^*(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}^T} = -\frac{\partial^2 \ell^*}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T}, \tag{4.5}$$

and let $\mathcal{J}^*$ denote its expected value, then $\mathbf{S}^*(\boldsymbol{\phi})$ is asymptotically normal with asymptotic covariance matrix equal to

$$\operatorname{Cov}\{\mathbf{S}^*(\boldsymbol{\phi})\} = \mathcal{J}^*(\boldsymbol{\phi}_0) - \mathcal{J}^*(\boldsymbol{\phi}_0) \begin{pmatrix} \mathbf{0} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{K} \end{pmatrix} \mathcal{J}^*(\boldsymbol{\phi}_0), \tag{4.6}$$

where $\mathbf{K}$ is some $L \times L$ symmetric matrix. Properties of $\hat{\boldsymbol{\phi}}$ then follow from standard results for estimating equations (see, e.g., Amari and Kawanabe [17]). In particular, $\hat{\boldsymbol{\phi}}$ is asymptotically normal with mean $\boldsymbol{\phi}_0$ and covariance matrix

$$\mathcal{J}^*(\boldsymbol{\phi}_0)^{-1} - \begin{pmatrix} \mathbf{0} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{K} \end{pmatrix}. \tag{4.7}$$

We are only interested in the block corresponding to the components of $\binom{\boldsymbol{\theta}}{\boldsymbol{\gamma}}$ and this does not involve $\mathbf{K}$. All this means that, for the purpose of estimating $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, we can operate as if $\mathbf{S}^*(\boldsymbol{\phi})$ is a genuine score function. The semiparametric MLE, $\binom{\hat{\boldsymbol{\theta}}}{\hat{\boldsymbol{\gamma}}}$, is obtained by setting $\mathbf{S}^*(\boldsymbol{\phi}) = \mathbf{0}$ and its covariance matrix can be estimated using the appropriate components of the inverse observed information matrix, $\mathcal{J}^*(\hat{\boldsymbol{\phi}})^{-1}$. (Note that some care has to be taken with solving the pseudoscore equations numerically as $\hat{\boldsymbol{\phi}}$ often corresponds to a saddle point of ℓ^* rather than a maximum.)

In principle, we can extend the results to stratified two-stage sampling where subsamples are drawn from the chosen clusters (or primary sampling units). However, to apply (4.3), we need the conditional probability of stratum membership given the observed $(\mathbf{y}, \mathbf{X})$, which requires integration over the values for the unsampled units in the cluster.

This is a nontrivial task in general so that the extension of the semiparametric approach to two-stage sampling is much less straightforward in practice than the weighted approach.

5. Discussion

We have sketched in Section 4 the development of semi-parametric methods for fitting regression models to data from stratified samples as an alternative to the weighted methods of Section 3, which are simple to implement but can be inefficient in retrospective studies, or full maximum likelihood, which is efficient but difficult to implement because it requires fitting a model for the joint distribution of all covariates. The methods are relatively simple to implement and simulations so far (see, e.g., Lawless et al. [18]) suggest that they are much more efficient than weighted methods in situations where the latter perform badly. In the very limited examples that we have looked at so far, they seem to be almost as efficient as full maximum likelihood but much more work needs to be done here.

A number of alternative general approaches to inference from complex surveys have been suggested in the literature and all of these can be specialized to informative stratified sampling. Nathan and Holt [25] and Smith and Nathan [14] suggest alternatives to full maximum likelihood that do not require the fitting of a complete model for the joint distribution of the covariates and design variables. Krieger and Pfeffermann [26] and Pfeffermann and Sverchkov [27] explore methods based on the induced distribution of $\mathbf{y}$ given $\mathbf{X}$ in the sample after the (possible informative) selection mechanism has been taken into account. The general semi-parametric methods for missing data problems that have been developed by Robins and his collaborators (see, e.g., Robins et al.[28, 29]) may also be applicable to our setup here. All of these methods seem to have connections to the methods that we have developed here and we are in the process of exploring these connections.

Acknowledgment

Jeff Hunter has been a colleague and friend for more than 35 years. We would like to take this opportunity to thank him for his advice and moral support throughout that time and to wish him a happy and productive retirement.

References

[1] J. M. Neuhaus, A. Scott, and C. Wild, "The analysis of retrospective family studies," *Biometrika*, vol. 89, no. 1, pp. 23–37, 2002.

[2] J. M. Neuhaus, A. Scott, and C. Wild, "Family-specific approaches to the analysis of case-control family data," *Biometrics*, vol. 62, no. 2, pp. 488–494, 2006.

[3] A. S. Whittemore, "Logistic regression of family data from case-control studies," *Biometrika*, vol. 82, no. 1, pp. 57–67, 1995.

[4] L. P. Zhao, L. Hsu, S. Holte, Y. Chen, F. Quiaoit, and R. L. Prentice, "Combined association and aggregation analysis of data from case-control family studies," *Biometrika*, vol. 85, no. 2, pp. 299–315, 1998.

[5] J. M. Neuhaus and N. P. Jewell, "The effect of retrospective sampling on binary regression models for clustered data," *Biometrics*, vol. 46, no. 4, pp. 977–990, 1990.

[6] M. Wrensch, M. Lee, R. Miike, et al., "Familial and personal medical history of cancer and nervous system conditions among adults with glioma and controls," *American Journal of Epidemiology*, vol. 145, no. 7, pp. 581–593, 1997.

[7] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[8] A. J. Lee, L. McMurchy, and A. Scott, "Re-using data from case-control studies," *Statistics in Medicine*, vol. 16, no. 12, pp. 1377–1389, 1997.

[9] D. A. Binder, "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review*, vol. 51, no. 3, pp. 279–292, 1983.

[10] J. N. K. Rao, A. Scott, and C. J. Skinner, "Quasi-score tests with survey data," *Statistica Sinica*, vol. 8, no. 4, pp. 1059–1070, 1998.

[11] D. DeMets and M. Halperin, "Estimation of a simple regression coefficient in samples arising from a sub sampling procedure," *Biometrics*, vol. 33, no. 1, pp. 47–56, 1977.

[12] A. Scott and C. Wild, "Maximum likelihood for generalised case-control studies," *Journal of Statistical Planning and Inference*, vol. 96, no. 1, pp. 3–27, 2001.

[13] C. Wild, "Fitting prospective regression models to case-control data," *Biometrika*, vol. 78, no. 4, pp. 705–717, 1991.

[14] T. M. F. Smith and G. Nathan, "The effect of selection on regression analysis," in *Current Analysis of Complex Surveys*, C. J. Skinner, D. Holt, and T. M. F. Smith, Eds., pp. 149–163, Wiley, New York, NY, USA, 1989.

[15] E. L. Lehmann, *Asymptotic Theory*, John Wiley & Sons, New York, NY, USA, 1999.

[16] J. Chen and J. N. K. Rao, "Asymptotic normality under two-phase sampling designs," *Statistica Sinica*, vol. 17, no. 2, pp. 1047–1064, 2007.

[17] S. Amari and M. Kawanabe, "Estimating functions in semiparametric statistical models," in *Selected Proceedings of the Symposium on Estimating Functions (Athens, Ga, 1996)*, I. V. Basawa, V. P. Godambe, and R. L. Taylor, Eds., vol. 32 of *IMS Lecture Notes Monograph Series*, pp. 65–81, Institute of Mathematical Statistics, Hayward, Calif, USA, 1997.

[18] J. F. Lawless, J. D. Kalbfleisch, and C. Wild, "Semiparametric methods for response-selective and missing data problems in regression," *Journal of the Royal Statistical Society. Series B*, vol. 61, no. 2, pp. 413–438, 1999.

[19] A. Scott and C. Wild, "The analysis of clustered case-control studies," *Journal of the Royal Statistical Society C*, vol. 50, pp. 389–401, 2001.

[20] A. Scott and C. Wild, "On the robustness of weighted methods for fitting models to case-control data," *Journal of the Royal Statistical Society. Series B*, vol. 64, no. 2, pp. 207–219, 2002.

[21] S. Cosslett, "Efficient estimation of discrete-choice models," in *Structural Analysis of Discrete Data with Econometric Applications*, C. F. Manski and D. McFadden, Eds., pp. 51–111, Wiley, New York, NY, USA, 1981.

[22] A. J. Lee, "On the semi-parametric efficiency of the Scott-Wild estimator under choice-based and two-phase sampling," to appear in *Journal of Applied Mathematics and Decision Sciences*.

[23] A. J. Lee and Y. Hirose, "Semi-parametric efficiency bounds for regression models under case-control sampling: the profile likelihood approach," to appear in *Annals of the Institute of Statistical Mathematics*.

[24] W. K. Newey, "The asymptotic variance of semiparametric estimators," *Econometrica*, vol. 62, no. 6, pp. 1349–1382, 1994.

[25] G. Nathan and D. Holt, "The effect of survey design on regression analysis," *Journal of the Royal Statistical Society. Series B*, vol. 42, no. 3, pp. 377–386, 1980.

[26] A. M. Krieger and D. Pfeffermann, "Maximum likelihood estimation from complex sample surveys," *Survey Methodology*, vol. 18, pp. 225–239, 1992.

[27] D. Pfeffermann and M. Sverchkov, "Parametric and semi-parametric estimation of regression models fitted to survey data," *Sankhya B*, vol. 61, no. 1, pp. 166–186, 1999.

[28] J. M. Robins, A. Rotnitzky, and L. P. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 846–866, 1994.

[29] J. M. Robins, F. S. Hsieh, and W. Newey, "Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates," *Journal of the Royal Statistical Society. Series B*, vol. 57, no. 2, pp. 409–424, 1995.

Alastair Scott: Department of Statistics, University of Auckland, Private Bag,
Auckland 92019, New Zealand
Email address: scott@stat.auckland.ac.nz

Chris Wild: Department of Statistics, University of Auckland, Private Bag,
Auckland 92019, New Zealand
Email address: wild@stat.auckland.ac.nz

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 13749, 10 pages
doi:10.1155/2007/13749

Review Article

A Philatelic Excursion with Jeff Hunter in Probability and Matrix Theory

George P. H. Styan and Götz Trenkler

Received 15 May 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

We present an excursion with Jeff Hunter, visiting some of his research topics. Specifically, we will present some facts about certain people whose work seems to have influenced Jeff in his scientific career; we illustrate our presentation with postage stamps that have been issued in honour of these people. Our main guide is Hunter's two-volume book entitled *Mathematical Techniques of Applied Probability* (Academic Press, 1983).

1. Introduction

In the following we are going on an excursion with Jeff Hunter, visiting some of his research topics. Specifically, we will present some facts about certain people whose work seems to have influenced Jeff in his scientific career. We illustrate our presentation with postage stamps that have been issued in honour of these people; we know that Jeff collects stamps. Our main guide will be the two-volume book [1, 2] by Hunter. (All but one of the postage stamps depicted in this article are identified with their *Scott* catalog numbers as found in the *Scott 2007 Standard Postage Stamp Catalogue* [3].)

2. Bernoulli, Seki, Leibniz and Chebyshev

One of the first entries of this two-volume treatment of applied probability is "Bernoulli trial" [1, page 10]: "… a sequence of independent trials is called a sequence of Bernoulli trials." A Bernoulli trial is an experiment whose outcome is random and can be either "success" and "failure" [4].

Figure 2.1. (Left panel) Jacob Bernoulli I: Switzerland 1994 (Scott 939); (right panel) Takakazu Seki Kôwa: Japan 1992 (Scott 2147).

The eldest of four brothers, Jacob Bernoulli I (also known as James Bernoulli; first name in German: Jakob, in French: Jacques) lived from 1654 to 1705, mainly in Switzerland. He is one of eight prominent mathematicians in the Bernoulli family [4] and is famous for his *Ars Conjectandi* [5] (published posthumously in 1713), a ground-breaking work on probability theory, and for his research concerning the law of large numbers, see Figure 2.1 (left panel).

The Bernoulli numbers B_k $(k = 0, 1, 2, \ldots)$ are a sequence of signed rational numbers that are implicitly defined by the identity

$$\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} \frac{B_k x^k}{k!}. \tag{2.1}$$

Actually, the Japanese mathematician Takakazu Seki Kôwa (1642–1708) discovered these numbers before Jacob Bernoulli I. Seki also discovered determinants before Leibniz, see, for example, [6–8].

Gottfried Wilhelm von Leibniz (the surname "Leibniz" is sometimes written as "Leibnitz" [6]) (1646–1716) was a German polymath "of remarkable breadth of knowledge, who made original contributions to optics, mechanics, statistics, logic, and probability theory. He conceived the idea of calculating machines and of a universal language. He wrote on history, law, and political theory" [9]; see also [6].

The postage stamp from St. Vincent (in Figure 2.2) was issued for the "750th Anniversary of Hannover: Gottfried Wilhelm Leibniz. Head librarian for the electors of Hannover & co-inventor of calculus."

The notion of "Markov chain" is already mentioned by Jeff Hunter in [1], but the main theory of Markov chains in discrete time is developed only in [2]. The early research of Andrey Andreyevich Markov (1856–1922) was much influenced by Pafnuty Lvovich Chebyshev (his surname is transliterated in various other ways, for example, Chebychev, Chebyshov, Tchebycheff, or Tschebyscheff [10, 11]) (1821–1894), see Figure 2.2 (right panel). Chebyshev is probably best known for his inequality, also known as

FIGURE 2.2. (Left panel) Gottfried W. von Leibniz: St. Vincent 1991 (Scott 1557); (right panel) Pafnuty Lvovich Chebyshev: USSR 1946 (Scott 1051).

the Bienaymé-Chebyshev inequality:

$$\text{Prob}\left(|X-\mu| \geq k\sigma\right) \leq \frac{1}{k^2}, \tag{2.2}$$

where X has expected value μ and variance σ^2, and k is any positive real number. We have not found a postage stamp that honours either Markov or the statistician Irenée–Jules Bienaymé (1796–1878), who proved (2.2) some years before Chebyshev, see, for example, Heyde and Seneta [12, page 132].

3. Wiener, Brownian Motion, Gauss, Laplace, Abel

Andrey Andreyevich Markov introduced the concept of a Markov chain as a working model for the study of dependent random variables. Brownian motion is a very good example of a continuous Markov process and is often called a Wiener process; Norbert Wiener (1894–1964) "was an American theoretical and applied mathematician, who is perhaps best known as the founder of cybernetics" [4].

Brownian motion, which is named after the botanist Robert Brown (1773–1858), "is either the random movement of particles suspended in a fluid or the mathematical model used to describe such random movements" [4].

A graph of Brownian motion is shown on a postage stamp (in Figure 3.1 (right panel) right-hand side, center) from Czechoslovakia. This stamp was issued in celebration of the 125th anniversary of the Union of Czechoslovakian Mathematicians and Physicists. Shown on the stamp is "geographical measurement from A. M. Malletta's book (we have not been able to further identify this book), 1672, earth fold and Brownian motion diagrams" [3, volume 2, page 714].

We observe, however, that Jeff Hunter's research interests focus mainly on discrete models. But even in discrete models, continuous distributions play a major role (see, e.g., volume 1, where the distribution of the number of recurrent events is shown to be

FIGURE 3.1. (Left panel) Norbert Wiener: Moldova 2000 (Scott 348); (right panel) Brownian motion: Czechoslovakia 1987 (Scott 2665).

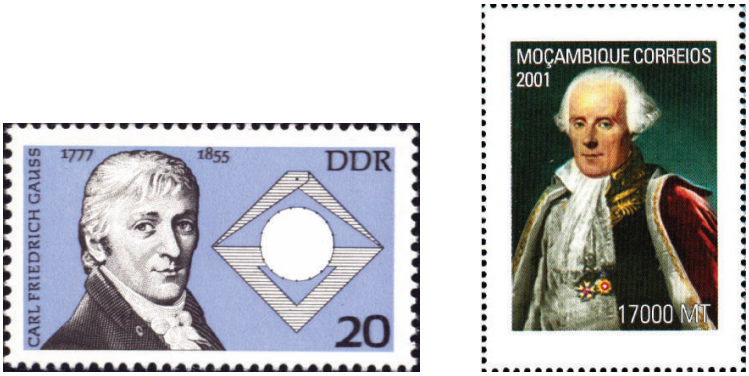


FIGURE 3.2. (Left panel) Carl Friedrich Gauss: German Democratic Republic 1977 (Scott 1811); (right panel) Pierre-Simon Laplace: Mozambique 2001 (Scott seemingly unlisted (we could not find this stamp listed in the *Scott 2007 Standard Postage Stamp Catalogue* [3])).

FIGURE 3.3. Niels Henrik Abel: Norway 2002 (Scott 1333).

approximately normal). The normal distribution is attributed to Carl Friedrich Gauss (the surname "Gauss" is sometimes written as "Gauß") (1777–1855).

Beyond mathematics, Gauss did research in astronomy, physics, and geodesy. His principal contributions to statistics are in the theory of least squares estimation, where major work was also done by the mathematician Pierre-Simon, Marquis de Laplace (1749–1827).

In connection with the behaviour of probability-generating functions and the application of discrete models, convergence problems of infinite series arise. In both volumes of his book, Jeff has to apply Abel's convergence theorem to solve these problems. Niels Henrik Abel (1802–1829) was a Norwegian mathematician, who is best known for proving the impossibility of finding, in closed form, the roots of a polynomial of degree greater than 4. Beyond group theory ("Abelian group"), Abel was especially interested in the theory of functions. In 2001 the government of Norway announced that the bicentennial of Abel's birth (1802) would mark the commencement of a new prize for mathematicians, named after Abel. The Abel Prize is similar to the Nobel Prize which is awarded in Sweden and Norway, but excludes mathematics. Sathamangalam R. Srinivasa Varadhan (born 1940) received the 2007 Abel Prize for his fundamental contributions to probability theory [4].

4. Euler, Banachiewicz, Penrose triangle, and Hawking

Matrix theory is a powerful tool which has been widely used by Jeff Hunter in his research papers to further the understanding of the main features of Markov chains. Jeff also devoted one chapter of his book to these techniques. Although Leonhard Euler (1707–1783) did not explicitly use the notion of a matrix (explicitly created only in the mid-19th century), in 1776 Euler created several magic squares, including a 4×4 magic square [13, page 8], [14, 15] which may be represented by the matrix

$$M = \begin{pmatrix} 8 & 10 & 15 & 1 \\ 11 & 5 & 4 & 14 \\ 2 & 16 & 9 & 7 \\ 13 & 3 & 6 & 12 \end{pmatrix}. \tag{4.1}$$

Dividing each element of M by 34 yields a doubly stochastic matrix [1, page 127].

When inverting large-scale matrices, Jeff applies results from the theory of partitioned matrices. The latter notion can be traced back to Tadeusz Banachiewicz (1882–1954), who was a Polish astronomer, mathematician, and geodesist, see, for example, [8, 16].

The Banachiewicz inversion formula can be stated as

$$\begin{pmatrix} P & Q \\ R & S \end{pmatrix}^{-1} = \begin{pmatrix} P^{-1} + P^{-1}QT^{-1}RP^{-1} & -P^{-1}QT^{-1} \\ -T^{-1}RP^{-1} & T^{-1} \end{pmatrix}, \tag{4.2}$$

where $T = S - RP^{-1}Q$. The matrix T is known as a Schur complement, a concept that has become increasingly important in statistics and applied mathematics in recent years, see, for example, Zhang [17]. In 1925, Banachiewicz invented a theory of "cracovians" (Cracovian calculus), a special kind of matrix algebra, which brought him international recognition. This theory solved several astronomical, geodesic, mechanical, and mathematical problems [4].

FIGURE 4.1. (Left panel) Leonhard Euler: USSR 1957 (Scott 1932); (right panel) Tadeusz Banachiewicz: Poland 1983 (Scott 2565).

FIGURE 4.2. (Left panel) Penrose triangle: Sweden 1982 (Scott 1396); (right panel) Stephen Hawking: Zambia 2000 (Scott 856n).

Jeff Hunter has participated in numerous conferences and workshops dedicated to matrix theory and probability, and related fields. In some of these meetings he met the statistician Calyampudi Radhakrishna Rao (born 1920), where both were invited speakers. Rao's tool of a generalized inverse of a matrix was one of Jeff's favourites (see [2] and his research papers, e.g., [18, 19]). (Calyampudi Radhakrishna Rao was the Ph.D. thesis adviser for the 2007 Abel Prize laureate Sathamangalam R. Srinivasa Varadhan, see Section 3.) As is well known, every matrix has a generalized inverse which, however, need not be unique. The Moore-Penrose inverse of a complex matrix A is, however, the unique matrix G satisfying the four equations

$$AGA = A, \quad GAG = G, \quad AG = (AG)^*, \quad GA = (GA)^*, \tag{4.3}$$

where the superscript $*$ denotes conjugate transpose. See, for example, [20, page 40].

FIGURE 5.1. Jules Henri Poincaré, Kurt Gödel, and Andrey Nikolaevich Kolmogorov: Portugal 2000 (Scott 2345b).

The Moore-Penrose inverse is named after the mathematician Eliakim Hastings Moore (1862–1932) and the renowned mathematical physicist Sir Roger Penrose (born 1931). It should be mentioned that Penrose, after his research on generalized inverses of matrices in the 1950s, apparently did not pursue his matrix interests any further, but started to work on singularities in the framework of general relativity, together with the famous theoretical physicist Stephen William Hawking (born 1942). The Penrose triangle (Figure 4.2, left panel) is an impossible object first created by the Swedish artist Oscar Reutersvärd (1915–2002) in 1934. Sir Roger Penrose independently devised and popularised it in the 1950s, describing it as "impossibility in its purest form."

5. Poincaré, Gödel, Kolmogorov, Carroll, and Hua

Jules Henri Poincaré (1854–1912) "was one of France's greatest mathematicians and theoretical physicists, and a philosopher of science" [4] and in 1886 he was appointed to the Chair of Mathematical Physics and Probability at the Sorbonne. Poincaré is famous for geometric probability. In matrix theory there is the Poincaré separation theorem for eigenvalues, see, for example, [21]:

$$\mathrm{ch}_{n-p+i}(A) \le \mathrm{ch}_i(F'AF) \le \mathrm{ch}_i(A), \quad i = 1,\dots,p \le n, \tag{5.1}$$

where ch_i denotes the ith largest eigenvalue. Here the matrix A is a real symmetric $n \times n$ matrix and F is a real $n \times p$ matrix such that $F'F = FF' = I_p$, the $p \times p$ identity matrix and F' is the transpose of F.

The postage stamp shown in Figure 5.1 was issued by Portugal as part of a large set in celebration of the 20th century, and includes Kurt Gödel (1906–1978) in addition to Poincaré and Andrey Nikolaevich Kolmogorov (1903–1987).

Kurt Gödel was a mathematician and philosopher. "One of the most significant logicians of all time, Gödel's work has had an immense impact upon scientific and philosophical thinking in the 20th century" [4].

Andrey Nikolaevich Kolmogorov, who is widely considered one of the prominent mathematicians of the 20th century, made major advances in probability theory (as well

FIGURE 5.2. (left panel) Lewis Carroll: Mali 1982 (Scott C443); (right panel) Loo-Keng Hua: China 1988 (Scott 2148).

as topology, intuitionistic logic, turbulence, classical mechanics, and computational complexity). Kolmogorov observed that "The theory of probability as a mathematical discipline can and should be developed from axioms in exactly the same way as geometry and algebra" [4].

In connection with matrices, the notion of determinant is indispensable. We have already mentioned that Seki invented determinants a few years before Leibniz. Jeff's book contains a nice collection of rules for determinants.

The first books solely dedicated to determinants, see, for example, [22], were written by the mathematician and physicist William Spottiswoode (1825–1883) and by the author, mathematician, Anglican clergyman, and photographer Rev. Charles Lutwidge Dodgson (1832–1908), who is better known as Lewis Carroll [23]; his books [24, 25] entitled *Condensation of Determinants* [24] and *An Elementary Treatise on Determinants* [25] are far less well known than his famous *Alice's Adventures in Wonderland* (1865) and its sequel *Through the Looking-Glass, and What Alice Found There* (1871), recently reissued (with an introduction by Martin Gardner) as a single paperback [26]. (The 6-page booklet *Condensation of Determinants* [24] is just a reprint of an article published in the *Proceedings of the Royal Society of London* but *An Elementary Treatise on Determinants* [25] is a 143-page book published by Macmillan, London, 1867.)

Loo-Keng Hua (1910–1985) was a great mathematician and a Chinese legendary hero. (The name "Loo-Keng Hua" is also written as "Hua Loo-keng" [27].) He had little formal education, but made enormous contributions to number theory, algebra, complex analysis, matrix geometry, and applied mathematics" [28, 29]. In particular the inequality

$$\det(I - A^*A) \cdot \det(I - B^*B) \leq \left| \det(I - A^*B) \right|^2 \tag{5.2}$$

is known as the "Hua determinantal inequality" [18, 30, 21]. Here the matrices A and B are contractive, that is, the singular values all lie in the half-open unit interval $[0,1)$; on the right-hand side of (5.2), the symbol $|\cdot|$ denotes absolute value.

We conclude this little excursion into the world of stamps and mathematics by expressing again our appreciation for Jeff's outstanding research activities, especially in matrix theory. We are sure that his retirement will be the starting point for new adventures in this attractive field.

Go ahead, Jeff!

Acknowledgments

The authors are most grateful to three anonymous referees for their helpful comments. Many of the images of stamps have been taken, with permission, from the excellent website of Jeff Miller [31]; see also the Reinhardt website [32] and the beautiful book by Robin Wilson [33]. For biographical material the authors consulted Abbott [34], Crystal [9], Heyde and Seneta [12], Johnson and Kotz [35], and Meyer [35], as well as the excellent open-access websites of *MacTutor* [11] and *Wikipedia* [4]. For closely related material, please see [8]. Many thanks to Evelyn Matheson Styan for her help; she also joins the authors in wishing Jeff a very happy retirement. This research has been supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

[1] J. J. Hunter, *Mathematical Techniques of Applied Probability, Volume 1, Discrete Time Models: Basic Theory*, Operations Research and Industrial Engineering, Academic Press, New York, NY, USA, 1983.

[2] J. J. Hunter, *Mathematical Techniques of Applied Probability, Volume 2, Discrete Time Models: Techniques and Applications*, Operations Research and Industrial Engineering, Academic Press, New York, NY, USA, 1983.

[3] J. J. Hunter, "Characterizations of generalized inverses associated with Markovian kernels," *Linear Algebra and Its Applications*, vol. 102, pp. 121–142, 1988.

[4] G. P. H. Styan, Ed., "IMAGE philatelic corner [Takakazu Seki Kôwa]," *IMAGE: The Bulletin of the International Linear Algebra Society*, no. 23, p. 28, 1999.

[5] J. Bernoulli, *Ars Conjectandi*, Thurnisiorum, Basel, Switzerland, 1713.

[6] R. W. Farebrother, G. P. H. Styan, and G. J. Tee, "Gottfried Wilhelm von Leibniz: 1646–1716," *IMAGE: The Bulletin of the International Linear Algebra Society*, no. 30, pp. 13–16, 2003.

[7] J. Reinhardt, "Physics-related stamps," June 2007, http://th.physik.uni-frankfurt.de/~jr/physstamps.html.

[8] A. J. Scott and G. P. H. Styan, "On a separation theorem for generalized eigenvalues and a problem in the analysis of sample surveys," *Linear Algebra and Its Applications*, vol. 70, pp. 209–224, 1985.

[9] D. Crystal, *The Cambridge Biographical Encyclopedia*, Cambridge University Press, Cambridge, UK, 1994.

[10] P. J. Davis, *The Thread: A Mathematical Yarn*, Harcourt, Boston, Mass, USA, 2nd edition, 1989.

[11] C. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, Pa, USA, 2000.

[12] C. C. Heyde and E. Seneta, Eds., *Statisticians of the Centuries*, Springer, New York, NY, USA, 2001.

[13] L. Euler, "On magic squares," http://arxiv.org/pdf/math/0408230 (English) .

[14] L. Euler, "De quadratis magicis [On magic squares]," *Commentationes Arithmeticae*, vol. 2, pp. 593–602, 1849, (Latin).

[15] L. Euler, *Opera Omnia, Series 1: Opera Mathematica*, vol. 7, Birkhäuser, Boston, Mass, USA, 1992.

[16] J. Grala, A. Markiewicz, and G. P. H. Styan, "Tadeusz Banachiewicz: 1882–1954," *IMAGE: The Bulletin of the International Linear Algebra Society*, no. 25, p. 24, 2000.

[17] Wikipedia, "The Free Encyclopedia," http://en.wikipedia.org/.

[18] L.-K. Hua, "Inequalities involving determinants," *Acta Mathematica Sinica*, vol. 5, pp. 463–470, 1955 (Chinese).

[19] J. J. Hunter, "Generalized inverses and their application to applied probability problems," *Linear Algebra and Its Applications*, vol. 45, pp. 157–198, 1982.

[20] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, vol. 15 of *CMS Books in Mathematics*, Springer, New York, NY, USA, 2nd edition, 2003.
[21] C. C. Paige, G. P. H. Styan, B.-Y. Wang, and F. Zhang, "Hua's matrix inequality and Schur complements," *International Journal of Information and System Sciences*, vol. 4, no. 1, 2008.
[22] R. W. Farebrother, S. T. Jensen, and G. P. H. Styan, "Sir Thomas Muir and nineteenth-century books on determinants," *IMAGE: The Bulletin of the International Linear Algebra Society*, no. 28, pp. 6–15, 2002.
[23] R. W. Farebrother, S. T. Jensen, and G. P. H. Styan, "Charles Lutwidge Dodgson: a biographical and philatelic note," *IMAGE: The Bulletin of the International Linear Algebra Society*, no. 25, pp. 22–23, 2000.
[24] C. L. Dodgson, "Condensation of determinants: being a new and brief method for computing their arithmetical values," *Proceedings of the Royal Society of London*, vol. 15, no. 84, pp. 150–155, 1866.
[25] C. L. Dodgson, *An Elementary Treatise on Determinants, with Their Application to Simultaneous Linear Equations and Algebraical Geometry*, Macmillan, London, UK, 1867.
[26] L. Carroll, *Alice's Adventures in Wonderland & through the Looking-Glass. With an Introduction by Martin Gardner*, Signet Classics, New York, NY, USA, Reissue edition, 2000.
[27] H. Halberstam, "Loo-Keng Hua: obituary," *Acta Arithmetica*, vol. 51, no. 2, pp. 99–117, 1988.
[28] H. Halberstam, Ed., *Loo-Keng Hua Selected Papers*, Springer, New York, NY, USA, 1983.
[29] J. Miller, "Images of Mathematicians on Postage Stamps," June 2007, http://jeff560.tripod.com/.
[30] L.-K. Hua, "Inequalities involving determinants," *Translations of the American Mathematical Society, Series II*, vol. 32, pp. 265–272, 1963 (English).
[31] J. E. Kloetzel, Ed., *Scott 2007 Standard Postage Stamp Catalogue*, Scott, Sidney, Ohio, USA, 163rd edition, 2006, on paper in 6 volumes and on 12 CD-ROMs with text-searchable pdf files.
[32] J. J. O'Connor and E. F. Robertson, "The MacTutor History of Mathematics Archive," April 2007, http://www-history.mcs.st-andrews.ac.uk/history/index.html.
[33] G. P. H. Styan and G. Trenkler, "IMAGE philatelic corner reprise," *IMAGE: The Bulletin of the International Linear Algebra Society*, vol. 38, pp. 9–12, 2007.
[34] D. Abbott, Ed., *The Biographical Dictionary of Scientists: Mathematicians*, Blond Educational, London, UK, 1985.
[35] N. L. Johnson and S. Kotz, Eds., *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1997.

George P. H. Styan: Department of Mathematics and Statistics, McGill University,
805 Ouest, Rue Sherbrooke, Montréal, Québec, Canada H3A 2K6
Email address: styan@math.mcgill.ca

Götz Trenkler: Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany
Email address: trenkler@statistik.uni-dortmund.de

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2007, Article ID 68280, 10 pages
doi:10.1155/2007/68280

Research Article

A Paradox in a Queueing Network with State-Dependent Routing and Loss

Ilze Ziedins

Received 25 May 2007; Accepted 8 August 2007

Recommended by Paul Cowpertwait

Consider a network of parallel finite tandem queues with two stages, where each arrival attempts to minimize its own cost due to loss. It is known that the user optimal and asymptotic system optimal policies may differ—we give examples showing that they may differ for finite systems and that as the service rate is increased at the second stage the user optimal policy may change in such a way that the total expected cost due to loss increases.

1. Introduction and background

Queueing networks often exhibit seemingly paradoxical behaviour, where adding capacity either in the form of extra capacity on links or at nodes, or even extra links or routes, does not always lead to an improvement in performance, and may even lead to a severe degradation in performance. The classic and very well-known example of this is Braess's paradox, which has been much studied, both in the traffic literature, and in the queueing theory literature (see [1, 2] for the original paper and [3] for a comprehensive list of references maintained by Braess). This was one of the examples mentioned by Professor Jeff Hunter in his inaugural lecture. It is therefore a great pleasure to write on a different kind of paradox in this festschrift for Jeff.

In addition to Braess's paradox, there are several other well-known paradoxes (see, e.g., Arnott and Small [4] and the references therein). The paradoxes can be ascribed to the difference between system optimal and perceived user optimal behaviour—that is, if individuals behave selfishly in a way that minimizes their own measure of cost (e.g., the delay in transit through a network), then the system as a whole can suffer, and flows can alter in such a way that all individuals see worse performance. This is most clearly

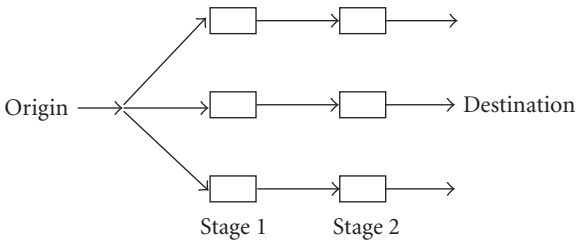


FIGURE 1.1. Three parallel two-stage tandem queues.

seen in traffic and transportation networks, and many of the early papers were directed to this application (see, e.g., [1, 5–8]). More recently, the phenomenon of selfish routing has become increasingly important in the context of telecommunication and computer networks (e.g., [9–13]).

The model considered here consists of a system of K parallel finite tandem queues with two stages, with a stream of arrivals who can be sent to any one of the queues. Figure 1.1 illustrates such a system with 3 parallel tandem queues. The objective is to minimize the cost due to loss, rather than minimize delay, the usual performance measure. Loss occurs when an individual attempts to enter a finite queue, but is unable to do so because it is full—that individual is then lost to the system. This assumption, while not at all realistic for traffic networks, is realistic for computer networks, where packets attempting to join a buffer that is full are lost and, if necessary, resent from the origin. This model was earlier studied in [14] (which obtained some properties of user optimal routing policies) and [15] (which obtained the asymptotic as $K \to \infty$ system optimal routing policies); and these papers should be referred to for a more complete list of earlier, relevant references. In [14] it was shown that user optimal policies may be paradoxical in the sense that arrivals may choose a queue with greater occupancy to minimize their probability of loss. The main contribution of this paper is to give an example showing that this paradoxical behaviour may then have further consequences—under user optimal routing it is possible for the expected cost, both to the system as a whole, and to the individual user, to *increase* when the service rate at the second queue is increased.

The model discussed here differs in some respects from those commonly studied in the literature. The most obvious difference is that it is a queueing system with loss. Paradoxes in loss networks have been studied [9, 10], but networks with finite queues and loss have been seen more rarely. The usual performance measure of interest is delay, and in most papers it is then also assumed that queues are infinite when studying paradoxes. Another difference is that here we concentrate on state-dependent routing, whereas routing paradoxes have most commonly been studied with probabilistic routing (see, e.g., [11, 16, 17] for some examples of other studies with state-dependent routing). For the model in this paper, with state-dependent routing, it has already been shown in [14] that when individuals are able to choose a route to minimize their own loss, they may choose a route

where the occupancy is greater, particularly at the first stage of the tandem. Here, we illustrate by example the difference in the expected cost between the user optimal policy and the system optimal policy when the number of queues, K, is finite, and we compare both with the asymptotically optimal cost. We then give an example showing that when the service rate at the second stage of the tandem is increased, the expected cost under the user optimal policy may increase, rather than decrease as might be expected.

In Section 2, we give a detailed description of the parallel tandem queue model, and we give some of the results that we will be using from [14, 15]. In Section 3, we give examples comparing the performance of state-dependent user optimal and system optimal routing schemes, looking both at asymptotic results taken from [15] and results for a finite system. We conclude with a short discussion in Section 4.

2. Definitions and preliminaries

This section gives a detailed description of the network and outlines some results from previous papers, in particular [14, 15].

Consider a network with K parallel tandem queues. Each tandem queue has two single server finite queues in sequence, which we refer to as stages, the first stage having service rate μ_1 and capacity C_1 and the second service rate μ_2 and capacity C_2. Arrivals to the system are as a Poisson process with rate $K\lambda$ so that the arrival rate scales with the size of the system. An arrival can be routed to any one of the K queues, but once it has joined a tandem queue, it cannot change to a different queue, so there is no interaction between the queues beyond that induced by any controls over the routing of arrivals. We assume that all interarrival times and service times are exponentially distributed and independent of one another, so the system can be modelled as a Markov process.

Since the queues are finite, not all arrivals will necessarily be accepted into the system. Moreover, since the second queue in the tandem is finite, it is possible for an individual to finish service at the first queue, but find that the second queue is full, so that they are unable to join it—we assume that in that case, the individual leaves the system, and is lost, without completing service at both queues.

We assume that the objective is to minimize the cost due to losing or blocking individuals. In a system such as this, it may also be possible to consider minimizing the delay, conditional on not being lost, but we do not do so here. Previous analyses of paradoxes in queueing networks (not loss networks) have often focussed on routing or other controls that minimize or equalize delay rather than loss, because the common assumption is that the queues have infinite capacity—however, in practice, queues are often finite, so that minimizing loss in networks with finite queues also needs to be considered.

Let d_1 be the cost of losing an individual on entry to queue 1, and d_2 the cost of losing an individual on entry to queue 2. We will examine both system optimal and user optimal policies, as well as a range of intermediate policies.

Under probabilistic routing, each arrival chooses queue k with probability p_k, $1 \le k \le K$, where $\sum_k p_k = 1$ independently of all other routing decisions, service times, and arrival times. In that case it is sufficient to consider the tandem queues separately, with arrival rate $K\lambda p_k$ at the kth tandem queue. The state for a single queue is then given by (i, j), where i is the occupancy at the first queue, and j the occupancy at the second. This single

queue has state space $S_1 = \{(i,j) : 0 \le i \le C_1,\ 0 \le j \le C_2\}$ and transition rates

$$(i,j) \longrightarrow \begin{cases} (i+1,j) & K\lambda p_k \quad \text{if } i < C_1, \\ (i-1,j+1) & \mu_1 \quad \text{if } i > 0,\ j < C_2, \\ (i-1,j) & \mu_1 \quad \text{if } i > 0,\ j = C_2, \\ (i,j-1) & \mu_2 \quad \text{if } j > 0. \end{cases} \tag{2.1}$$

Let the equilibrium distribution for a single tandem queue with these transition probabilities be denoted by $\widetilde{\pi}(n)$, $n \in S_1$. The system optimal policy is found by minimizing a weighted sum of the cost of loss for the individual queues. In contrast, the user optimal policy is one where the costs of loss at all queues that are in use are equalized. For this model, under probabilistic routing, the system optimal and user optimal policies coincide with $p_k = 1/K$, $1 \le k \le K$ and the flow of arrivals is divided equally between the tandem queues, so that each tandem queue has arrival rate λ.

Under state-dependent routing the analysis is considerably more complicated and it is necessary to consider the state of the whole network simultaneously. Now, let n_{ij} be the number of tandem queues with occupancy i at the first queue, and j at the second queue. Then $n = (n_{ij} : 0 \le i \le C_1,\ 0 \le j \le C_2)$ is a Markov process with state space $S = \{n : \sum_{ij} n_{ij} = K,\ n_{ij} \in \{0,1,2,\ldots,K\},\ 0 \le i \le C_1,\ 0 \le j \le C_2\}$ and transition rates partly depending on the routing rule. Denote by $r_n(i,j)$ the probability an arrival is sent to a tandem queue in state (i,j) if the network is in state n, and let $r_n(b)$ be the probability that an arrival is lost in state n, where $r_n(b) + \sum_{ij} r_n(i,j) = 1$ for all $n \in S$. Let $R = \{r_n(i,j), r_n(b);\ n \in S,\ 0 \le i \le C_1,\ 0 \le j \le C_2,\ 0 \le r_n(i,j), r_n(b) \le 1\}$ denote a particular state-dependent admission and routing policy. Note that for a finite system, since this is a Markov decision process, the system optimal policy will have $r_n(i,j), r_n(b) \in \{0,1\}$. Given some linear ordering of the states (i,j), let e_{ij} denote the $(C_1+1)\times(C_2+1)$ unit vector with the ijth entry equal to 1, and the remaining entries equal to 0. Then the transition rates under policy R are given by

$$n \longrightarrow \begin{cases} n - e_{ij} + e_{i+1,j} & K\lambda r_n(i,j) \quad \text{for } (i,j) \in S_1, \\ n - e_{ij} + e_{i-1,j+1} & n_{ij}\mu_1 \quad \text{if } i > 0,\ j < C_2, \\ n - e_{ij} + e_{i-1,j} & n_{ij}\mu_1 \quad \text{if } i > 0,\ j = C_2, \\ n - e_{ij} + e_{i,j-1} & n_{ij}\mu_2 \quad \text{if } j > 0. \end{cases} \tag{2.2}$$

We denote by $\pi_R(n)$, $n \in S$ the equilibrium distribution under a given policy R.

The state space grows rapidly as the capacities C_1, C_2, and the number of queues increase. The system optimal policy can be found using the theory of Markov decision processes, but apart from some special cases (e.g., when $C_1 = C_2 = 1$), the exactly optimal policy will, in general, not only require considerable computational effort to calculate, but also, just as importantly, substantial effort to implement. In Section 3, we therefore limit ourselves to state-dependent policies that are relatively easy both to analyse and to implement (although note that the asymptotic result given below gives optimality over all state-dependent policies). The system optimal policy minimizes over all policies R the

expected cost per queue per unit time, which is given by

$$\lambda \sum_{n:n\in S} \pi_R(n) r_n(b) + \mu_1 \sum_{n:n\in S} \pi_R(n) \sum_i n_{i,C_2}/K. \tag{2.3}$$

The user optimal policy, however, is one that chooses the route that will give the lowest expected cost due to loss for an arrival. This can be calculated explicitly (see [14]), and the details are not given here, although the calculations are done for the examples in Section 3.

In the numerical examples below, in addition to giving exact results for the system with a small number of queues, found by calculating the equilibrium distribution numerically, we also give the asymptotic costs and policy, as the number of queues becomes large. The following results from [15], which are obtained using the methods of [18, 19], give the basis for obtaining the asymptotic results given below. In the following, instead of considering n as the state, we instead consider $\mathbf{x}^K = n/K$. Here, x_{ij}^K is the proportion of tandem queues in state (i, j).

Consider the sequence of networks indexed by K, the Kth network operating under any admissible acceptance and routing policy (an admissible policy must be nonanticipating). Let $Kw_{ij}^K(t)$ be the number of arrivals that have been accepted at a tandem queue in state (i, j) by time t, $0 \le i \le C_1$, $0 \le j \le C_2$ and let $Kw_b^K(t)$ be the number of arrivals that have been lost at entry (i.e., not accepted into the system) by time t. Then $\{(\mathbf{x}^K, \mathbf{w}^K)(\cdot)\}$ is relatively compact and the limit of any convergent sequence has the following properties.

(1) $\sum_{ij} x_{ij}(t) = 1$ for all $t \ge 0$.

(2) $x_{ij}(t) \ge 0$ for all $t \ge 0$, $0 \le i \le C_1$, $0 \le j \le C_2$.

(3) There exists $\mathbf{z}(\cdot)$ such that, almost surely, $z_{ij}(t), z_b(t) \ge 0$, and

$$\begin{aligned} x_{ij}(t) = x_{ij}(0) + \int_0^t \big(& z_{i-1,j}(s) I_{\{i\neq 0\}} - z_{ij}(s) I_{\{i\neq C_1\}} \\ & + \mu_1 x_{i+1,j-1}(s) I_{\{i\neq C_1, j\neq 0\}} + \mu_1 x_{i+1,j}(s) I_{\{i\neq C_1, j=C_2\}} \\ & + \mu_2 x_{i,j+1}(s) I_{\{j\neq C_2\}} - x_{ij}(s)\big(\mu_1 I_{\{i\neq 0\}} + \mu_2 I_{\{j\neq 0\}}\big)\big) ds, \end{aligned} \tag{2.4}$$

$$\lambda = \sum_{ij} z_{ij}(t) + z_b(t),$$

for all $t \ge 0$, $0 \le i \le C_1$, $0 \le j \le C_2$.

The equilibrium distribution is then a solution to the following system of equations:

$$\begin{aligned} & z_{ij} + x_{ij}\big(\mu_1 I_{\{i\neq 0\}} + \mu_2 I_{\{j\neq 0\}}\big) \\ & \quad = z_{i-1,j} I_{\{i\neq 0\}} + \mu_1 x_{i+1,j-1} I_{\{i\neq C_1, j\neq 0\}} \\ & \qquad + \mu_1 x_{i+1,j} I_{\{i\neq C_1, j=C_2\}} + \mu_2 x_{i,j+1} I_{\{j\neq C_2\}}, \quad 0 \le i \le C_1,\ 0 \le j \le C_2, \\ & \lambda = \sum_{ij} z_{ij} + z_b, \quad \sum_{ij} x_{ij} = 1, \quad x_{ij}, z_{ij}, z_b \ge 0. \end{aligned} \tag{2.5}$$

These equations are balance equations for the asymptotic system. The z_{ij} here give the rate at which arrivals are entering queues in state (i, j) under the given policy, while z_b gives the rate at which arrivals are blocked.

In [15] these equations are constraints for the linear optimisation problem

$$\text{minimize } F(\mathbf{x}, z_b) = d_1 z_b + d_2 \mu_1 \sum_{i=1}^{C_1} x_{i,C_2}. \tag{2.6}$$

This can be solved to find the asymptotically optimal value of the objective function, and hence derive the asymptotically optimal control. However, the balance equations above can more generally be used to find the asymptotic costs for any routing policy of interest. In particular, we will give the asymptotic costs for the two main policies of interest, which are to accept all arrivals if possible, and to accept arrivals only if they can be routed to a tandem queue that has total occupancy less than C_2 (i.e., for the designated tandem queue, $n_1 + n_2 < C_2$).

3. Examples

Consider a system of parallel tandem queues with $C_1 = C_2 = 2$. If arrivals are accepted into the system, they will be routed to a queue in one of the states $(0,0), (1,0), (0,1), (1,1), (0,2), (1,2)$. In the state-dependent case, under user optimal routing, arrivals choose the queue that will minimize their own cost. Let $p_d(\mathbf{n})$ be the probability that an arrival joining a tandem queue in state $\mathbf{n} = (n_1, n_2)$ will reach its destination (the success probability). When $C_1 = C_2 = 2$,

$$\begin{aligned} p_d(1,1) &= 1 - \left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^2, \\ p_d(1,2) &= 1 - \left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^2 \left(1 + \frac{\mu_2}{\mu_1 + \mu_2}\right), \\ p_d(0,2) &= 1 - \frac{\mu_1}{\mu_1 + \mu_2} \end{aligned} \tag{3.1}$$

with, trivially, $p_d(0,0) = p_d(1,0) = p_d(0,1) = 1$ and $p_d(2,0) = p_d(2,1) = p_d(2,2) = 0$ (see [14] for details). We see immediately that $p_d(1,1) > p_d(1,2) > p_d(0,2)$. When $\mu_1 = \mu_2 = 1$, for instance, $p_d(1,1) = 3/4$, $p_d(1,2) = 5/8$ and $p_d(0,2) = 1/2$. Thus, somewhat paradoxically, an arrival wishing to minimize their own blocking probability at the second stage would prefer to join a queue in state $(1,2)$, rather than one in state $(0,2)$, even though the number of individuals in the former is greater. A queue in state $(1,1)$ is preferred to a queue in state $(0,2)$. In both cases the increased delay for the new arrival allows additional time for individuals ahead of them in the tandem to leave, thus reducing the blocking probability for the new arrival. Under user optimal routing, in general, arrivals may join queues that are in state (i,j) provided $d_1 > d_2(1 - p_d(i,j))$. Thus when $\mu_1 = \mu_2 = 1$, for instance, they may join queues in state $(1,1)$ if $d_2 < 4d_1$, in state $(1,2)$ if $d_2 < d_1 8/3$, and in state $(0,2)$ if $d_2 < 2d_1$.

The policy under user optimal routing is in strong contrast to the asymptotically system optimal policy, which is to accept arrivals if possible when $d_2 < d_1$, and otherwise to only accept arrivals into queues in one of the states $(0,0)$, $(1,0)$, or $(0,1)$, that is, a queue in such a state that the probability the arrival is lost at the second stage is 0 (see [15] for

details). The results in that paper also yield the asymptotic average costs (to first order). Let

$$\lambda^* = \mu_1 \frac{1 + (\mu_1/\mu_2)}{1 + (\mu_1/\mu_2) + (\mu_1/\mu_2)^2}. \tag{3.2}$$

Then the asymptotic average costs of the two policies are as follows.

(1) If $\lambda < \lambda^*$, then all arrivals (to first order) can be routed to queues where there is no blocking, and the average cost is 0.
(2) For the policy that accepts all arrivals if possible the average cost is $d_2(\lambda - \lambda^*)$ when $\lambda^* < \lambda < \mu_1$, and $d_1(\lambda - \mu_1) + d_2(\mu_1 - \lambda^*)$ when $\lambda > \mu_1$.
(3) For the policy that only accepts arrivals into queues with occupancy less than $C_1 + C_2$, the average cost is $d_1(\lambda - \lambda^*)$ for $\lambda > \lambda^*$.

For a finite number of queues, as already observed, state-dependent optimal policies can be found using the theory of Markov decision processes, but are complex. Instead we consider a number of policies intermediate between the two asymptotically optimal ones. The costs of these are calculated numerically, and although closed form expressions can be given, we do not do so here, since they are tedious and not at all illuminating. In the examples below, where $C_1 = C_2 = 2$, we consider the following policies. A policy here consists of a list of possible states for queues into which an arrival can be accepted, listed in order of preference with the most preferred first. The policies compared below are

(1) $(0,0),(0,1),(1,0)$,
(2) $(0,0),(0,1),(1,0),(1,1)$,
(3) $(0,0),(0,1),(1,0),(0,2)$,
(4) $(0,0),(0,1),(1,0),(1,1),(1,2)$,
(5) $(0,0),(0,1),(1,0),(1,1),(0,2)$,
(6) $(0,0),(0,1),(1,0),(1,1),(1,2),(0,2)$.

For instance, policy (2) is to send an arrival to a queue in state $(0,0)$ if possible, otherwise to a queue in state $(0,1)$, otherwise to a queue in state $(1,0)$, and finally, if there is no queue in any of these three states, to a queue in state $(1,1)$. If there are no queues in any of these four states, then the arrival is lost. Some candidate policies have been omitted from the list. The policy $(0,0),(0,1),(1,0),(1,2)$ has the same average cost as policy (1), since the state $(1,2)$ for a single queue is transient under this policy (to see this, observe that to reach the state $(1,2)$ from any of the other three states included in this policy, it needs to pass through a state with $n_1 + n_2 = 2$, but no such state is included in this policy). Also, in policies (4), (5), and (6) we have assumed that queues in state $(1,1)$ are preferred to queues in state $(0,2)$.

The first example has $C_1 = C_2 = 2$, with $\mu_1 = \mu_2 = 1$. In Figure 3.1 we plot the expected cost per unit time for each of the six policies when $d_1 = 1$ and $d_2 = 3$ for a system of four queues. We note that policy (3) is the user optimal policy in this case, although the system optimal policy is to only accept arrivals into tandem queues that have occupancy no more than 1. For comparison purposes, the asymptotic cost as the number of queues, $K \to \infty$, under the asymptotically optimal policy (policy (1)) is also given. The expected cost is lowest for policy (1), and highest for policy (6), with policy (2) having lower cost than policy (4), which has lower cost again than policies (3) and (5). Thus, in accordance with

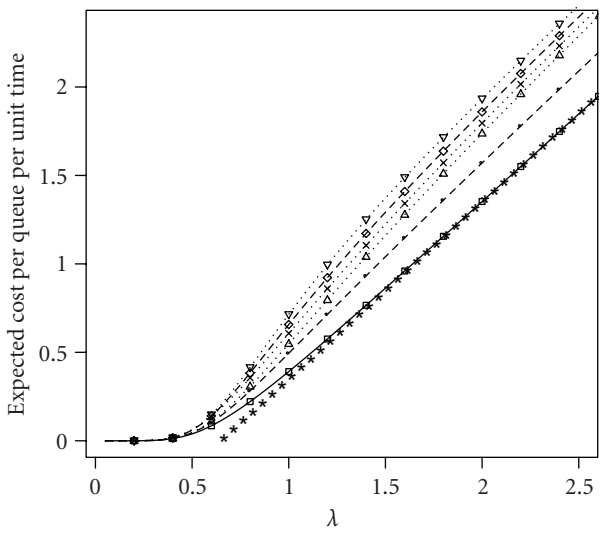


FIGURE 3.1. Expected cost per queue per unit time, $C_1 = C_2 = 2$, $\mu_1 = \mu_2 = 1$, $d_1 = 1$, $d_2 = 3$. Plotting symbols for each policy are 1 □, 2 ∘, 3 ×, 4 △, 5 ◇ 6 ▽, asymptotically optimal policy ⋆. Four tandem queues.

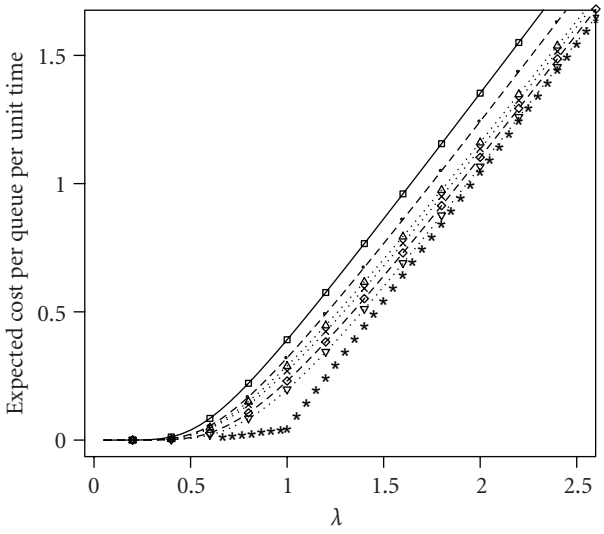


FIGURE 3.2. Expected cost per queue per unit time, $C_1 = C_2 = 2$, $\mu_1 = \mu_2 = 1$, $d_1 = 1$, $d_2 = 0.1$. Plotting symbols for each policy are 1 □, 2 ∘, 3 ×, 4 △, 5 ◇ 6 ▽, asymptotically optimal policy ⋆. Four tandem queues.

the user optimal policy, sending arrivals to queues in state $(0,2)$ also gives the highest average cost. However, when $d_2 < d_1$, this is largely reversed. Figure 3.2 gives a similar plot, but now with $d_2 = 0.1$, and we see that the policies are reversed, with policy (6) having the lowest cost, and policy (1) the highest.

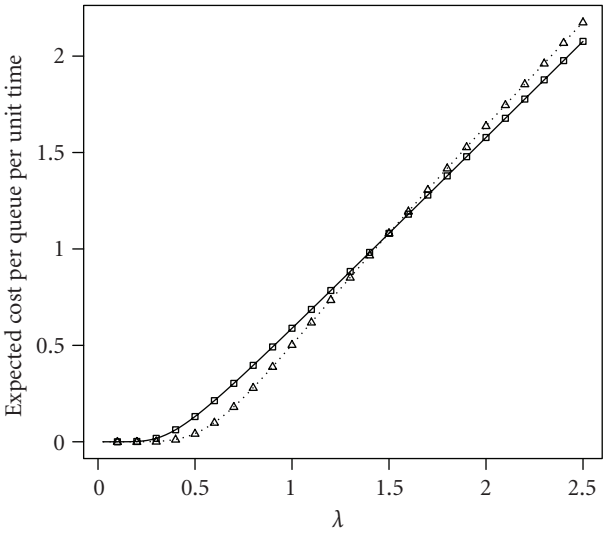


FIGURE 3.3. Expected cost per queue per unit time, $C_1 = C_2 = 2$, $\mu_1 = 1$, $d_1 = 1$, $d_2 = 2.5$ under user optimal policies when $\mu_2 = 0.5, 1.0$. Plotting symbols: $\mu_2 = 0.5$ □, $\mu_2 = 1.0$ △. Four queues.

Finally, Figure 3.3 plots the expected cost under the user optimal policy when $d_1 = 1$ and $d_2 = 2.5$ for $C_1 = C_2 = 2$, $\mu_1 = 1$, and $\mu_2 = 0.5, 1.0$. For both values of μ_2, the system optimal policy is policy (1). When $\mu_2 = 0.5$, the user optimal policy coincides with the system optimal policy, however, when $\mu_2 = 1.0$, policy (4) is user optimal since the expected cost of using queues in states $(1,1)$ and $(1,2)$ is less than d_1 in this case. We see from the plot that if arrivals follow user optimal policies, increasing the service rate at stage 2 of the tandem queues gives a higher expected cost overall for high values of λ.

These examples have all had unrealistically small capacities. This has been because the state space grows rapidly with C_1, C_2, and K. In all cases above, the equilibrium distribution, when the number of queues is finite, has been calculated explicitly to obtain the expected costs (rather than estimating from simulation). However, we conjecture that the finding in this case will carry over to larger capacities, that is, for sufficiently high arrival rates, as μ_2 increases, the expected cost may also increase, when d_2 is greater than d_1 (note that $d_2 > d_1$ is a reasonable scenario for a system where there may be a greater cost attached to losing an individual on whom some service has already been expended).

4. Conclusions

The numerical examples of the previous section have shown that permitting otherwise indistinguishable arrivals to use queues in certain states may lead to greater expected costs, when arrivals attempt to minimize their own costs due to loss. The difference here between the expected cost under user optimal and system optimal policies can be substantial. Furthermore, increasing the service rate, as in the classical paradoxes, may lead to worse overall performance, if user optimal policies are permitted. We have given a numerical example where increasing the service rate at the second stage of the tandem leads to increased expected cost.

Acknowledgment

This paper was completed while the author was visiting Lund University on a Solander Fellowship.

References

[1] D. Braess, "Über ein Paradoxon aus der Verkehrsplanung," *Unternehmensforschung*, vol. 12, no. 1, pp. 258–268, 1968.
[2] D. Braess, A. Nagurney, and T. Wakolbinger, "On a paradox of traffic planning," *Transportation Science*, vol. 39, no. 4, pp. 446–450, 2005, English language translation of [1].
[3] http://homepage.ruhr-uni-bochum.de/Dietrich.Braess/#paradox.
[4] R. Arnott and K. Small, "The economics of traffic congestion," *American Scientist*, vol. 82, no. 5, pp. 446–455, 1994.
[5] A. Downs, "The law of peak-hour expressway congestion," *Traffic Quarterly*, vol. 16, no. 3, pp. 393–409, 1962.
[6] R. Steinberg and W. I. Zangwill, "The prevalence of Braess' paradox," *Transportation Science*, vol. 17, no. 3, pp. 301–318, 1983.
[7] J. M. Thomson, *Great Cities and Their Traffic*, Gollancz, London, UK, Peregrine edition, 1977.
[8] J. G. Wardrop, "Some theoretical aspects of road traffic research," *ICE Proceedings, Engineering Divisions*, vol. 1, no. 3, pp. 325–362, 1952.
[9] E. Altman, R. El Azouzi, and V. Abramov, "Non-cooperative routing in loss networks," *Performance Evaluation*, vol. 49, no. 1–4, pp. 257–272, 2002.
[10] N. G. Bean, F. P. Kelly, and P. G. Taylor, "Braess's paradox in a loss network," *Journal of Applied Probability*, vol. 34, no. 1, pp. 155–159, 1997.
[11] F. P. Kelly, "Network routing," *Philosophical Transactions of the Royal Society of London. Series A*, vol. 337, no. 1647, pp. 343–367, 1991.
[12] A. Orda, R. Rom, and N. Shimkin, "Competitive routing in multiuser communication networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 5, pp. 510–521, 1993.
[13] T. Roughgarden and É. Tardos, "How bad is selfish routing?" *Journal of the ACM*, vol. 49, no. 2, pp. 236–259, 2002.
[14] S. Spicer and I. Ziedins, "User-optimal state-dependent routeing in parallel tandem queues with loss," *Journal of Applied Probability*, vol. 43, no. 1, pp. 274–281, 2006.
[15] R. Sheu and I. Ziedins, "Asymptotically optimal control of parallel tandem queues with loss," The University of Auckland, preprint, 2007.
[16] B. Calvert, "The Downs-Thomson effect in a Markov process," *Probability in the Engineering and Informational Sciences*, vol. 11, no. 3, pp. 327–340, 1997.
[17] B. Calvert, W. Solomon, and I. Ziedins, "Braess's paradox in a queueing network with state-dependent routing," *Journal of Applied Probability*, vol. 34, no. 1, pp. 134–154, 1997.
[18] P. J. Hunt and C. N. Laws, "Asymptotically optimal loss network control," *Mathematics of Operations Research*, vol. 18, no. 4, pp. 880–900, 1993.
[19] C. N. Laws and Y. C. Teh, "Alternative routeing in fully connected queueing networks," *Advances in Applied Probability*, vol. 32, no. 4, pp. 962–982, 2000.

Ilze Ziedins: Department of Statistics, The University of Auckland, Private Bag 92019,
Auckland 1142, New Zealand
Email address: ilze@stat.auckland.ac.nz