Matrix Analytic Methods in Applied Probability with a View towards Engineering Applications

Nielsen, Bo Friis

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Nielsen, B. F. (2013). *Matrix Analytic Methods in Applied Probability with a View towards Engineering Applications*. Technical University of Denmark.

Matrix Analytic Methods in Applied Probability with a View towards Engineering Applications

Bo Friis Nielsen

Department of Applied Mathematics
and Computer Science
Technical University of Denmark

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, building 303B,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Author: Bo Friis Nielsen (Copyright ©2013 by Bo Friis Nielsen)
Title: Matrix Analysis Methods in Applied Probability with a View towards Engineering Applications

Doctoral Thesis: ISBN 978-87-643-1291-1

Denne afhandling er af Danmarks Tekniske Universitet antaget til forsvar for den tekniske doktorgrad. Antagelsen er sket efter bedømmelse af den foreliggende afhandling.

Kgs. Lyngby, den 17. september 2013

Anders O. Bjarklev
Rektor

/Henrik C. Wegener
Prorektor

This thesis has been accepted by the Technical University of Denmark for public defence in fulfilment of the requirements for the degree of Doctor Technices. The acceptance is based on an evaluation of the present dissertation.

Kgs. Lyngby, 17 September 2013

Anders O. Bjarklev
President

/Henrik C. Wegener
Provost

Contents

Acknowledgements

First and foremost I would like to express my thanks and gratitude towards my coauthors, students, good colleagues, and friends. These include Allan Andersen, Nigel Bean, Jan Beyer, Mogens Bladt, Thomas Kaare Christensen, Luz Judith Rodriquez Esparza, Villy Bæk Iversen, Marcel Neuts, Fredrik Nilsson, Flemming Nielson, Hanne Riis Nielson, Vaidyanathan Ramaswami, and Uffe Thygesen.

Several colleagues have supported me in the task of compiling the material into the final presentation as a thesis. Explicitly I would like to mention Knut Conradsen who encouraged me to write it and Kaj Madsen who, in addition to his encouragement, also gave constructive remarks before the thesis was finalised.

Last but not least, I would like to thank my family, in particular my two children Emil and Laura, and my wife Anne. My two children Laura and Emil have given comments on several versions of the manuscripts. Their comments, I am certain, have helped me improve the presentation of the thesis. My wife Anne has had to endure with my endless list of broken deadlines, nevertheless, she never ceased to give her support throughout and never stopped encouraging me to continue and finalise the work.

Bistrup 22nd November 2012

Bo Friis Nielsen

Dansk resumé

Kødannelse opstår i forbindelse med de fleste kommercielle aktiviteter. Mennesker står i kø i butikker, i banker og i restauranter. Biler, skibe og fly venter på veje, i havne og på landingsbaner, medens elektroniske meddelelser venter ved transmissionslinier og på at blive behandlet i servere. Den matematiske modellering af køfænomener har været stigende siden køteoriens fødsel, der sædvanligvis tilskrives AK Erlang, i begyndelsen af det 20ende århundrede. Erlang er nok den fremmeste repræsentant blandt adskillelige skandinaviske, heriblandt flere danske, forskere som har bidraget væsentligt til køteorien og den nært beslægtede risikoteori med forsikringsvidenskab som sit største anvendelsesområde. Ved modellering af køsystemer er det nødvendigt med en model til beskrivelse af den tilfældige ankomststrøm af efterspørgsel efter betjening, i dennes forskellige former som mennesker, biler, skibe eller elektroniske meddelelser. Teorien for punktprocesser inden for anvendt sandsynlighed er netop udviklet til at imødekomme dette behov.

Den Markovske ankomst Process (MAP) er en af de vigtigste konkrete manifestationer af teorien for punkt processer. MAPpen er således en vigtig byggesten i matrix analytiske metoder en disciplin i køteorien, der er udviklet af Neuts og medforfattere. Teorien for matrix analytiske metoder er bekvem ud fra et praktisk synspunkt, idet mange systemer kan evalueres analytisk og numerisk ved brug af denne teoridannelse. I denne afhandling præsenteres bidrag til den teoretiske udvikling af feltet, herunder en generalisering til flerdimensionale fordelinger. Yderligere demonstreres anvendeligheden af teorien gennem eksempler fra telekommunikation og datalogi.

Afhandlingen er baseret på en række originale bidrag med en indledende sammenfatning. Strukturen af sammenfatningen er som følger.

Klassen af MAPper og den relaterede klasse af fasetype (PH) fordelinger tilhører de lidt større klasser af processer og fordelinger betegnet som henholdsvis Rationelle ankomst Processer (RAP) og matrixeksponentielle (ME) fordelinger. I kapitel 2 præsenteres de grundlæggende konstruktioner af fasetype og matrixeksponentielle fordelinger sammen med Markovske og rationelle ankomst processer. Den grundlæggende teori samt klassiske egenskaber

beskrives relativt kort, medens egne bidrag til teorien er beskrevet mere detaljeret. Kapitel 3 er afsat til diskussion af parameterestimation i de modeller, der er beskrevet i kapitel 2. Der gives en kort gennemgang af nuværende estimeringsmetoder, igen med et fokus på egne bidrag.

I kapitel 4 og 5 beskrives forskellige anvendelsesaspekter. Markovske ankomst Processer er et alsidigt redskab til følsomhedsanalyser af stokastiske systemer, eftersom de fleste punktproces deskriptorer for en MAP kan beregnes relativt enkelt. Følsomhedsanalyser baseret på MAPper er beskrevet i kapitel 4. Dette kapitel er af noget mere generisk natur end kapitel 5, hvor nogle konkrete eksempler på tekniske anvendelser præsenteres.

I kapitel 6 beskrives to forskellige bevisteknikker for, hvordan matrix analytiske resultater relateret til de klassiske modeller for fasetype fordelinger og Markovske ankomst processer udvides til at omfatte matrixeksponentielle fordelinger og rationelle ankomst processer.

I kapitel 7 introduces klasserne af multivariate matrixeksponentielle og bilaterale multivariate matrixeksponentielle fordelinger. Kapitlet starter med en lille gennemgang af tidligere arbejder om multivariate fasetype fordelinger, mens resten af kapitlet indeholder de seneste resultater af egen forskning.

Afhandlingens vigtigste bidrag er beskrevet i kapitel 4, 6 og 7. Kapitel 4 er vigtigt ud fra en teknisk synsvinkel. Fremgangsmåden beskrevet i [8] var på daværende tidspunkt noget kontroversiel. Målinger i pakkebaserede kommunikationsnetværk viste, at trafikken udviste betragtelig variabilitet med variation over flere tidsskalaer. Disse målinger fik adskillige forskere til at mene, at Markovkæde baserede modeller ville blive af mindre betydning fremover, da disse ikke skulle kunne tage højde for variabilitet over flere tidsskalaer. På basis af dette mente disse forskere, at der var behov for et paradigmeskift i køteorien. Imidlertid viste resultaterne fra [8], at den Markovske ankomst proces kunne forblive et nyttigt værktøj til modellering af moderne kommunikationssystemer. Artiklen og en foreløbig udgave [7] har været meget citeret. Også [4] er vigtig, da denne artikel demonstrerer, hvordan følsomhedsanalyser af køsystemer udført ved brug af Markovske ankomst Processer ofte kan føre til konklusioner af almen gyldighed.

De to sidste kapitler, 6 og 7, indeholder betydelige teoretiske bidrag. Betydningen af kapitel 6 ligger på nuværende tidspunkt primært i dets matematiske indhold. Det har været tilfredsstillende endeligt at kunne fastslå, at den almindelige forventning om, at resultater for PH fordelinger og MAPper kan overføres direkte til de mere generelle tilfælde med ME fordelinger og RAPper, er korrekt. De sædvanlige matrix analytiske ræsonnementer baserer sig på probabilistisk argumentation ud fra den tidsmæssige udvikling af den underliggende Markovkæde. Disse argumenter kan ikke umiddelbart udvides til ME og RAP tilfældet, idet man her ikke har en underliggende Markovkæde. To forskellige bevis teknikker blev anvendt. I [17] er analysen baseret på et “last entrance time” argument, medens analysen i [18] er baseret på en indlejret Markov kæde med et generelt tilstandsrum.

Endeligt beskrives bidrag fra [22, 24, 25, 28] i kapitel 7 indeholdende definitionen af en vigtig klasse af bilaterale multivariate matrixeksponentielle fordelinger sammen med eksempler på deres anvendelse og forskellige relaterede resultater. Disse fordelinger udgør et meget fleksibelt værktøj til modellering af multivariate fænomener. Definitionen synes at være den naturlige definition af multivariate matrixeksponentielle fordelinger. Hovedresultatet er en karakterisering svarende til karakteriseringen af den multivariate normalfordeling. Endelig vises, hvordan klassen af MVME fordelinger forener et antal af tidligere publicerede modeller meget lig den måde PH og ME fordelinger forenede en mængde af tilsyneladende løst forbundne modeller og resultater.

Det arbejde, der beskrives i kapitel 7 åbner op for adskillige ikke-trivielle teoretiske spørgsmål af matematisk art. Hvis nogle af disse udfordringer kan løses tilfredsstillende, vil det bane vejen for et potentielt stort antal anvendelser, og det er meget sandsynligt, at fordelingsklassen kan blive meget nyttig i statistisk analyse.

1 Introduction

Queueing permeates most of man's commercial behaviour. People queue in stores, at banks, and at restaurants. Cars, ships and air-planes queue at roads, ports and runways, while electronic messages queue at transmission lines and in servers, waiting to be processed. The mathematical modelling of queueing phenomena has developed at ever increasing speed since the birth of queueing theory in the early 20th century, usually ascribed to A. K. Erlang, who worked as a mathematician for the Copenhagen telephone Company (KTAS - Kjøbenhavns Telefons Aktie Selskab). Erlang is perhaps the foremost representative among many Danish and Scandinavian scientists who have contributed profoundly to queueing theory and the closely related field of risk theory, with insurance being its main application area. When modelling queueing systems a model is needed for the description of the random arrival stream of demands, in terms of customers in the various forms of people, cars, ships, or electronic messages. Point process theory arose from the field of applied probability to address this need.

The Markovian Arrival Process (MAP) is one of the main concrete manifestations of point process theory. The MAP is an essential building block within matrix analytic methods in queueing theory pioneered by Neuts and coauthors. The theory of matrix analytic methods is appealing from a practical point of view as many systems can be analytically and numerically evaluated using this approach. In this thesis we present contributions to the theoretical development of the field of matrix analytic methods including an extension to a multivariate setting. We further demonstrate the applicability of the theory, giving examples from telecommunications engineering and computer science.

The thesis is based on a number of original contributions and a summary introductory paper. The outline of the summary is as follows.

The class of MAPs and the related class of Phase Type (PH) distributions belong to the slightly larger classes of what have been termed Rational Arrival Processes (RAP) and Matrix Exponential (ME) distributions, respectively. In Chapter 2 we present the basic constructions of phase-type and matrix-exponential distributions along with the Markovian and rational arrival processes. We briefly mention some well-known properties of these constructions while describing our own contributions in more detail. Chapter 3 is devoted to discussion of parameter estimation in the models described in Chapter 2. We give a very brief review of current estimation methods while focusing on our own contributions.

Chapters 4 and 5 contain different aspects of applications. The MAP is a versatile tool in sensitivity analyses of stochastic systems since point process descriptors of a MAP can be evaluated numerically. Sensitivity analyses based on the MAP are described in Chapter 4. That chapter is somewhat more generic in nature than Chapter 5 in which some concrete examples of engineering applications are presented.

In Chapter 6 we present two different ways of proving how the matrix analytic results related to the classical models of phase-type distributions and Markovian arrival processes extend to the case of matrix-exponential distributions and rational arrival processes.

In Chapter 7 we introduce the classes multivariate matrix-exponential and bilateral multivariate matrix-exponential distributions. The chapter starts with a small review of previous work on multivariate phase-type distributions while the rest of the chapter contains recent results of our own research.

The main contributions of the thesis are described in Chapters 4, 6, and 7. Chapter 4 is important from an engineering perspective. The approach described in [8] was somewhat controversial at the time. Measurements in packet based communication networks made some researchers call for a paradigm shift in queueing theory, where models based on Markovian assumptions would be, if not superfluous, then at least of minor importance. The contribution of [8] was to show that the Markovian arrival process could indeed remain a useful tool in modelling modern communication systems. The paper and its preliminary version [7] have been widely cited. Also [4] is important as this paper exemplifies how sensitivity analyses of queueing systems can be carried out using the Markovian arrival process, frequently leading to conclusions of general validity.

The two final chapters, 6 and 7, contain substantial theoretical contributions. The importance of Chapter 6 is at present primarily the mathematical content. It has been satisfying to finally settle the common anticipation that results for PH distributions and MAPs carry over verbatim to the case of ME distributions and RAPs. The method of proof has to rely on new ideas, as the standard probabilistic line of reasoning breaks down in the case of matrix-exponential distributions and rational arrival processes. Two different proof techniques were applied. In [17] a continuous time analysis based on a last exit time approach was applied. The approach taken in [18] was that of an embedded Markov chain with a general state space.

Finally Chapter 7 describes the contributions of [22, 24, 25, 28] containing the definition of the important class of bilateral multivariate matrix-exponential distributions together with examples of their use and various related results. These distributions provide a very flexible tool for modelling multivariate phenomena. The definition seems to be the natural multivariate generalisation of matrix-exponential distributions. The main result is a characterisation theorem similar to the main characterisation theorem of the multivariate normal distribution. Finally we demonstrate how the MVME distribution class unifies a number of previously published models in a way quite similar to the way PH and ME distributions unified a number of seemingly loosely connected models and results. The work described in Chapter 7 opens several non-trivial mathematical and theoretical questions. If just some of these problems can be solved satisfactorily it will pave the way for a huge application potential, and it is very likely that the distributions can and will be useful in statistical analysis too. The research on multivariate distributions lead to [27] describing a closure property of matrix-exponential and phase-type distributions.

In general, results from our own research will be stated as definitions, lemmas, corollaries, and theorems, while other results will be part of the text flow. The notation used in the papers is generally similar to that of the summary, and it is my hope that the slight differences will not reduce the accessibility of the papers.

The thesis is based on the following papers

Primarily related to Chapter 2

[3] Allan T. Andersen, Marcel F. Neuts, and Bo F. Nielsen. PH-Distributions Arising through Conditioning. *Commun. Statist.-Stochastic Models*, 16(1):179–188, 2000.

[27] Mogens Bladt and Bo Friis Nielsen. Moment distributions of phase type. *Stochastic Models*, 27:651–663, 2011.

[69] Bo Friis Nielsen, Uffe Høgsbro Thygesen, L. A. Fredrik Nilsson, and Jan E. Beyer. Higher order moments and conditional asymptotics of the batch Markovian arrival process. *Stochastic Models*, 23(1):1–26, 2007.

Primarily related to Chapter 3

[21] Mogens Bladt, Luz Judith Rodriguez Esparza, and Bo Friis Nielsen.

Fisher information and statistical inference for phase–type distributions. *Journal of Applied Probability*, 48A - A Festschrift for Søren Asmussen:277–293, 2011.

[66] Bo Friis Nielsen and Jan E. Beyer. Estimation of Interrupted Poisson Process parameters from counts. Report No. 21, 2004/2005, fall, Institut Mittag-Leffler, 2005.

Primarily related to Chapter 4

[4] Allan T. Andersen, Marcel F. Neuts, and Bo F. Nielsen. On the time reversal of Markovian Arrival Processes. *Stochastic Models*, 20(3):237–260, 2004.

[5] Allan T. Andersen and Bo F. Nielsen. On the statistical implications of certain random permutations in Markovian Arrival Processes (MAP)s and second order self-similar processes. *Performance Evaluation*, 41:67–82, 2000.

[6] Allan T. Andersen and Bo F. Nielsen. On the use of second order descriptors to predict queueing behaviour of MAPs. *Naval Research Logistics*, 49(4):391–409, 2002.

Primarily related to Chapter 5

[8] Allan T. Andersen and Bo Friis Nielsen. A Markovian approach for modeling packet traffic with long range dependence. *IEEE JSAC*, 16(5):719–732, 1998.

[36] Thomas Kaare Christensen, Bo Friis Nielsen, and Villy Bæk Iversen. Phase-type models of channel holding times in cellular communication systems. *IEEE Trans. on Veh. Technol.*, 53(3):725–733, May 2004.

[67] Bo Friis Nielsen, Flemming Nielson, and Hanne Riis Nielson. Model checking multivariate state rewards. In *Seventh International Conference on the Quantitative Evaluation of Systems*, pages 7–16, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

Primarily related to Chapter 6

[17] Nigel Bean and Bo Friis Nielsen. Quasi-birth-and-death processes with rational arrival process components. *Stochastic Models*, 26(3):309–334, July 2010.

[18] Nigel Bean and Bo Friis Nielsen. Analysis of queues with rational arrival process (RAP) components - a general approach. IMM-Technical Report 5, IMM, 2011.

[68] Bo Friis Nielsen and V. Ramaswami. A Computational Framework for a Quasi Birth and Death Process with a Continuous Phase Variable. In V. Ramaswamiand P.E. Wirth, editor, *Teletraffic Contributions for the Information Age, ITC-15*, page 477–486. ITC, Elsevier, 1997.

Primarily related to Chapter 7

[22] Mogens Bladt, Luz Judith Rodriguez Esparza, and Bo Friis Nielsen. Bilateral matrix–exponential distributions. In G. Latouche, V. Ramaswami, J. Sethuraman, K. Sigman, M.S. Squillante, and D. Yao, editors, *Matrix-Analytic Methods in Stochastic Models*, volume 27. Springer Proceedings in Mathematics and Statistics, 2012.

[24] Mogens Bladt and Bo Friis Nielsen. Multivariate matrix-exponential distributions. In Dario Bini, Beatrice Meini, Vaidyanathan Ramaswami, Marie-Ange Remiche, and Peter Taylor, editors, *Numerical Methods for Structured Markov Chains*, number 07461 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2008. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[25] Mogens Bladt and Bo Friis Nielsen. Multivariate matrix–exponential distributions. *Stochastic Models*, 26(1):1–26, 2010.

[26] Mogens Bladt and Bo Friis Nielsen. On the construction of bivariate exponential distributions with an arbitrary correlation coefficient. *Stochastic Models*, 26(2):295–308, 2010.

[28] Mogens Bladt and Bo Friis Nielsen. On the representation of distributions with rational moment generating functions. IMM-Technical Report 16, IMM, Technical University of Denmark, DK-2800 Kgs. Lyngby, 2012.

2 Background

The memoryless property of the exponential distribution has been of paramount importance for the development of large parts of applied probability. This property ensures that when modelling a lifetime, the distribution of the remaining lifetime stays exponential with the same rate parameter regardless of the current age. For an exponentially distributed random variable X with intensity parameter λ the memoryless property is formally expressed by

$$P(X > x + t | X > t) = P(X > x) = e^{-\lambda x}.$$

The natural choice when modelling the time to decay of radioactive atoms is the exponential distribution due to its memoryless property. The memoryless assumption is also reasonably justified when modelling the lifetime of certain kinds of electronic equipment. In addition, the exponential distribution is frequently well suited for modelling phenomenons where human activity is involved. This is particularly true when modelling the process of telephone call initiations, where a large number of individuals tend to initiate calls independently of each other, with a low rate for each individual.

Most technical and biological systems are, however, characterised by life and process times that cannot be reasonably described by the exponential distribution. The idea of still exploiting the memoryless property by modelling lifetimes and other durations by compositions of exponential random variables is usually ascribed to Erlang, but according to [39, Page 4] the idea was already described by Ellis in 1844. Jensen [48] generalised Erlang's approach by introducing a class of distributions defined as absorption times in Markov chains, which was finally brought to its full potential by Neuts [61]. Neuts termed these distributions of absorption times "distributions of phase-type" (PH distributions). The theory of PH distributions facilitates construction of flexible models using the analytical and mathematical convenience that arises from the memoryless property of the exponential distributions governing the sojourn times in the states of the Markov chain. In the context of phase type distributions states are frequently referred to as phases.

Some applications of PH distributions naturally support decomposition of the lifetime of a phenomenon into phases. One such example is the modelling of the progress of colon cancer which is medically divided into four stages. In many cases there is no such natural interpretation of the phases and the phase-type approach is considered a convenient approximation to the real lifetime distribution, in the same way that a polynomial might approximate a general function. This is supported by the fact that, for any distribution on the non-negative reals, there exists a sequence of PH distributions that converges weakly to the distribution [9, Theorem III.4.2].

In Section 2.1 we define phase-type and matrix-exponential distributions, while the Markovian and rational arrival processes are defined in Section 2.2.

2.1 Phase-type and matrix-exponential distributions

A phase-type (PH) distribution is a distribution that can be interpreted as the distribution of the time to absorption in a finite state Markov chain where one state is absorbing and the other states are transient. The generator of such a Markov chain can be partitioned as

$$\begin{pmatrix} \boldsymbol{S} & \boldsymbol{s} \\ \boldsymbol{0} & 0 \end{pmatrix}.$$

For a discrete PH distribution the zero in the lower right corner is replaced by a 1 to get a probability transition matrix. The initial distribution of the Markov chain is given by the row vector $(\boldsymbol{\alpha}, \alpha_{p+1})$. The $p \times p$ matrix $\boldsymbol{S}$ is a sub-generator, (a sub-probability transition matrix in the discrete case), while the vector $\boldsymbol{s}$ is a column vector of absorption rates. The pair $(\boldsymbol{\alpha}, \boldsymbol{S})$ is called a representation for the distribution. The survival function $G(x)$ of a PH distribution can be expressed as $G(x) = \boldsymbol{\alpha} e^{\boldsymbol{S}x}\boldsymbol{1}$, where $\boldsymbol{1}$ is a column vector of ones of appropriate dimension. The analytical form of the survival function is also valid for a larger class of distributions called Matrix-Exponential (ME) distributions. The class of ME distributions is strictly larger than the class of PH distributions, see e.g. [70] for a thorough discussion and unique classification of PH distributions within the class of ME distributions. The general form of the survival function of an ME distribution is $G(x) = -\boldsymbol{\alpha} e^{\boldsymbol{S}x}\boldsymbol{S}^{-1}\boldsymbol{s}$ with corresponding representation $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{s})$. For a PH representation one must necessarily have $0 \leq \boldsymbol{\alpha}\boldsymbol{1} \leq 1$ and $\boldsymbol{S}\boldsymbol{1} + \boldsymbol{s} = \boldsymbol{0}$. These restrictions do not arise naturally for ME distributions, however, one can without loss of generality assume that a representation for an ME distribution satisfies them. Unless otherwise stated, such representations will be assumed throughout the thesis. In general, the matrix-exponential distributions do not have an interpretation as the distribution of absorption times in finite state Markov chains. The Laplace-Stieltjes transform $H(s)$ of an ME distribution is $H(s) = 1 - \boldsymbol{\alpha}\mathbf{1} + \boldsymbol{\alpha}(\boldsymbol{I} + s(-\boldsymbol{S})^{-1})^{-1}\mathbf{1}$, where the matrix $\boldsymbol{I}$ is an identity matrix of appropriate dimension. The function $H(s)$ is thus a rational function in s. Indeed, an alternative characterisation of matrix-exponential distributions is that a distribution is matrix-exponential if and only if it has a rational Laplace-Stieltjes transform. Very little work has been done on the discrete version of matrix-exponential distributions, called matrix-geometric distributions, but see [46] for examples of genuine matrix-geometric distributions. Here we will focus almost entirely on the continuous case. Main references to the theory of phase-type distributions are [61], giving emphasis to the discrete case, and [55, 63]. In [63] an analytic angle is taken with respect to proofs followed by discussions of the probabilistic interpretation. In [55] the probabilistic arguments are used directly as proofs. The analytic proofs of [63] extend immediately to the matrix-exponential case, such that all closure properties that can be proven analytically for PH distributions also hold for ME distributions. Phase-type distributions, hence also matrix-exponential distributions, are known to be closed under finite convolutions and mixtures, finite order statistics, and random sums where the number of terms in the sum is given by a discrete phase-type distribution. We will now give some additional examples of closure properties for phase-type and matrix-exponential distributions, which are part of our own contributions to the field. As several of these results were proven using the time reversed representation of a PH distribution and time reversal of stationary MAPs a brief introduction to time reversal follows.

Time reversal

Any distribution on the non-negative reals can be used to define a renewal process. This is in particular true for phase-type distributions, where the underlying phase process makes it natural to consider a sequence of inter arrival times as successive visits to an instantaneous state. Hence a phase-type renewal process appears when rather than terminating the process on absorption, it is restarted according to the initial vector. For ease of exposition we will assume that $\boldsymbol{\alpha}\mathbf{1} = 1$. The infinitesimal generator (transition probability) matrix of the Markov chain defined in this way is

$$\boldsymbol{S} + \boldsymbol{s}\boldsymbol{\alpha}. \tag{2.1}$$

The vector $\boldsymbol{\pi} = \boldsymbol{\alpha}(-\boldsymbol{S})^{-1}(\boldsymbol{\pi}\boldsymbol{s})^{-1}$ is the stationary probability vector of this Markov chain generated by the phase-type renewal process. The time reversed representation $\left(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{S}}\right)$ of the representation $(\boldsymbol{\alpha}, \boldsymbol{S})$ is obtained by a standard time reversal operation on the stationary Markov chain

$$\left(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{S}}\right) = \left(\boldsymbol{\pi}\boldsymbol{\Delta}(\boldsymbol{s})(\boldsymbol{\pi}\boldsymbol{s})^{-1}, \boldsymbol{\Delta}(\boldsymbol{\pi})^{-1}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{\Delta}(\boldsymbol{\pi})\right). \tag{2.2}$$

The matrix $\boldsymbol{\Delta}(\boldsymbol{\pi})$ is a diagonal matrix with the entries of $\boldsymbol{\pi}$ in the diagonal.

Conditional distributions of phase-type

Phase-type distributions occur naturally as first exit time distributions from finite sets of states in Markov chains. This is also true in the case where the set can be left in different ways when conditioning on exits to specific states or sets of states. Consider a finite Markov chain with p transient and r absorbing states and partition its generator $\boldsymbol{Q}$ as

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{S} & \boldsymbol{S}_1 \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{S}_1$ is a $p \times r$ matrix. The initial probability vector is now written as $(\boldsymbol{\alpha}, \boldsymbol{0})$. The distribution of the time until one of the r absorbing states is reached is obviously of phase-type. The distributional form of the time to absorption conditioned on absorption in a subset of the r absorbing states was addressed in [5]. Not surprisingly, it turns out that these distributions are also of phase-type. The Laplace-Stieltjes transform of the conditional distribution of the time to absorption in absorbing state j is

$$\Psi_j(s) \;=\; (v_j^*)^{-1}\boldsymbol{\alpha}(sI - \boldsymbol{S})^{-1}\boldsymbol{S}_1(j),$$

for $1 \leq j \leq r$. Here $\boldsymbol{S}_1(j)$ is the jth column of $\boldsymbol{S}_1$ and $v_j^* = \boldsymbol{\alpha}(-\boldsymbol{S})^{-1}\boldsymbol{S}_1(j)$ is the probability of absorption in state j. A corresponding phase-type representation was given as Theorem 2.1 of [3].

Theorem 1 (Theorem 2.1 of [3]) *The function $\Psi_j(s)$ is the Laplace-Stieltjes transform of a* PH-*distribution with representation $(\boldsymbol{\gamma}(j), C) = (\boldsymbol{\gamma}(j), \boldsymbol{\Delta}^{-1}(\boldsymbol{\pi})\boldsymbol{S}'\boldsymbol{\Delta}(\boldsymbol{\pi}))$, with $\boldsymbol{\gamma}(j) = \boldsymbol{\Delta}(\boldsymbol{\pi})\boldsymbol{S}_1(j)(\boldsymbol{\pi}\boldsymbol{S}_1(j))^{-1}$.*

The generalisation to a subset of the r absorbing states is made by replacing the jth column of $\boldsymbol{S}_1$ with the sum of all columns leading to states in the subset. The result was used in [5] and is currently being used in work that is a continuation of [67].

Size-biased distributions

The joint distribution of age and residual lifetime of a phase-type renewal process was studied in [25], leading to the study of the distribution of the related concept of the spread, which is the sum of the age and residual lifetime. The distribution of the spread is an example of a moment distribution, that is, a distribution for a non-negative random variable where the density is proportional to $t^i f(t)$, with $f(t)$ being the density of a random variable. Moment distributions are frequently referred to as size-biased distributions in applications. In [27] we demonstrated that the classes of matrix-exponential and phase-type distributions are also closed under formation of size-biased distributions. For the matrix-exponential distributions this was stated as Theorem 3.1 in [27].

Theorem 2 (Theorem 3.1 in [27]) *Consider a matrix-exponential distribution with representation* $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{s})$ *of dimension* p *such that* $\boldsymbol{s} = -\boldsymbol{S}\mathbf{1}$ *and* $0 \leq \boldsymbol{\alpha}\mathbf{1} \leq 1$. *Then its* n*th moment distribution is also matrix-exponential with representation* $(\boldsymbol{\alpha}_n, \boldsymbol{S}_n, \boldsymbol{s}_n)$, *where*

$$\boldsymbol{\alpha}_n = \left(\frac{\boldsymbol{\alpha}\boldsymbol{S}^{-n}}{\boldsymbol{\alpha}\boldsymbol{S}^{-n}\mathbf{1}}, \mathbf{0}, ..., \mathbf{0} \right) \quad \boldsymbol{S}_n = \begin{pmatrix} \boldsymbol{S} & -\boldsymbol{S} & \mathbf{0} & ... & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S} & -\boldsymbol{S} & ... & \mathbf{0} \\ ... & ... & ... & ... & ... \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{S} \end{pmatrix}, \quad \boldsymbol{s}_n = \begin{pmatrix} 0 \\ 0 \\ .. \\ \boldsymbol{s} \end{pmatrix},$$

where $\boldsymbol{S}_n$ *is an* $(n+1)p \otimes (n+1)p$ *dimensional matrix.*

Even if $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{s})$ is a representation of a phase-type distribution the representation $(\boldsymbol{\alpha}_n, \boldsymbol{S}_n, \boldsymbol{s}_n)$ given in Theorem 2 will generally not be a valid phase-type representation since off-diagonal elements can be negative. Using a probabilistic argument based on time reversal we derived a phase-type representation for the first order moment distribution of a phase-type distribution. This representation was presented as Theorem 3.3 of [27]. It turned out that, using an analytical argument, the result of Theorem 3.3 of [27] could be generalised to give a valid phase-type representation for the nth order moment distribution of a phase-type distribution. The following theorem was initially stated as Theorem 3.5 of [27].

Theorem 3 (Theorem 3.5 of [27]) *Consider a phase-type distribution with representation* $(\boldsymbol{\alpha}, \boldsymbol{S})$. *Then the* n*th order moment distribution is again of* *phase-type with representation* $(\boldsymbol{\alpha}_n^{\bullet}, \boldsymbol{S}_n^{\bullet})$, *where*

$$\begin{aligned}
\boldsymbol{\alpha}_n^{\bullet} &= \left(\frac{\rho_{n+1}}{\rho_n} \boldsymbol{s}' \boldsymbol{\Delta}(\boldsymbol{\pi}_{n+1}), \mathbf{0}, \ldots, \mathbf{0} \right), \\
\boldsymbol{S}_n^{\bullet} &= \begin{bmatrix} \mathbf{C}_{n+1} & \mathbf{D}_{n+1} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_n & \mathbf{D}_n & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_{n-1} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{C}_2 & \mathbf{D}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{C}_1 \end{bmatrix},
\end{aligned}$$

and $\rho_i = \boldsymbol{\alpha}(-\boldsymbol{S})^{-i}\mathbf{1}$ *are the reduced moments with*

$$\boldsymbol{\pi}_i = \rho_i^{-1}\boldsymbol{\alpha}(-\boldsymbol{S})^{-i}, \ \mathbf{C}_i = \boldsymbol{\Delta}(\boldsymbol{\pi}_i)^{-1}\boldsymbol{S}'\boldsymbol{\Delta}(\boldsymbol{\pi}_i), \ \mathbf{D}_i = \frac{\rho_{i-1}}{\rho_i}\boldsymbol{\Delta}(\boldsymbol{\pi}_i)^{-1}\boldsymbol{\Delta}(\boldsymbol{\pi}_{i-1}).$$

For the phase-type case we also gave an alternative forward representation as Corollary 3.6 in [27].

Corollary 4 (Corollary 3.6 in [27]) *The nth order moment distribution of a phase-type distribution with representation* $(\boldsymbol{\alpha}, \boldsymbol{S})$ *has a phase-type representation* $(\boldsymbol{\alpha}_n^{\dagger}, \boldsymbol{S}_n^{\dagger})$ *with*

$$\begin{aligned}
\boldsymbol{\alpha}_n^{\dagger} &= (\rho_n^{-1}\boldsymbol{\alpha}\boldsymbol{\Delta}_n, \mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}) \\
\boldsymbol{S}_n^{\dagger} &= \begin{pmatrix} \boldsymbol{\Delta}_n^{-1}\boldsymbol{S}\boldsymbol{\Delta}_n & \boldsymbol{\Delta}_n^{-1}\boldsymbol{\Delta}_{n-1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta}_{n-1}^{-1}\boldsymbol{S}\boldsymbol{\Delta}_{n-1} & \boldsymbol{\Delta}_{n-1}^{-1}\boldsymbol{\Delta}_{n-2} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Delta}_{n-2}^{-1}\boldsymbol{S}\boldsymbol{\Delta}_{n-2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{S} \end{pmatrix},
\end{aligned}$$

with $\boldsymbol{\Delta}_n = \boldsymbol{\Delta}(\boldsymbol{\rho}_n)$, $\boldsymbol{\rho}_n = (-\boldsymbol{S})^{-n}\mathbf{1}$ *and* $\rho_n = \boldsymbol{\alpha}\boldsymbol{\rho}_n$.

The probabilistic interpretation for the first order moment distributions was also considered in [27].

Size-biased distributions have applications in engineering. In particular, the first order moment distribution is used in financial engineering. The Gini Index and the Lorenz Curve are descriptors calculated from the first order moment distribution that are used to describe inequality in income distributions, where the Gini index is the descriptor used most frequently. These quantities are readily expressed for matrix-exponential distributions through

Theorem 5 (Theorem 4.1 of [27]) *Let F be the distribution function of a matrix-exponential distribution with representation $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{s})$, where $\boldsymbol{s} = -\boldsymbol{S}\mathbf{1}$. Then the Lorenz curve is given by the formula*

$$\gamma : t \to \left(1 - \boldsymbol{\alpha} e^{\boldsymbol{S}t}\mathbf{1}, 1 - \frac{\boldsymbol{\alpha}\boldsymbol{S}^{-1}}{\boldsymbol{\alpha}\boldsymbol{S}^{-1}\mathbf{1}}\left(e^{\boldsymbol{S}t}\mathbf{1} + te^{\boldsymbol{S}t}\boldsymbol{s}\right)\right)$$

and the Gini index G by

$$G = 2(\boldsymbol{\alpha} \otimes \boldsymbol{\alpha}_1)\left(-\left(\boldsymbol{S} \oplus \boldsymbol{S}_1\right)\right)^{-1}(\boldsymbol{s} \otimes \mathbf{1}) - 1.$$

2.2 Markovian and rational arrival processes

The construction in Equation (2.1) can be generalised. The term $\boldsymbol{s}\boldsymbol{\alpha}/\boldsymbol{\alpha}\mathbf{1}$ associated with successive visits to the absorbing state and thus to points occurring in the point process (the renewals) is a rank one matrix of non-negative entries. By allowing for a general non-negative matrix associated with the occurrence of points in the point process one obtains the Markovian Arrival Process - the MAP.

Consider a generator (probability transition) matrix $\boldsymbol{D}$ for a finite dimensional Markov chain and its decomposition into an at most countable set of matrices such that $\boldsymbol{D} = \sum_{i=0}^{\infty} \boldsymbol{D}_i$. If $\boldsymbol{D}_i \geq \mathbf{0}$ for $i \geq 1$, all off-diagonal elements of $\boldsymbol{D}_0$ are non-negative, and $\boldsymbol{D}\mathbf{1} = \mathbf{0}, (\mathbf{1})$ in the continuous (respectively discrete) case, then the entries of the matrices $\boldsymbol{D}_i$ can be interpreted as arrival rates for occurrences of points with characteristic - or mark - i. Such a process is a marked point process, termed a marked MAP or MMAP in [47] when the underlying point process is a MAP as here. Quite frequently the mark i will be associated with the arrival of i homogeneous customers or items, in which case the process is called a Batch Markovian Arrival Process (BMAP), or simply a MAP whenever there is only one non-zero matrix $\boldsymbol{D}_1$ of arrival intensities. The finite dimensional distribution of the first n intervals and types of the marks is given by the joint density

$$f(t_1, i_1, t_2, i_1, \ldots, t_n, i_n) = \boldsymbol{\alpha} e^{\boldsymbol{D}_0 t_1}\boldsymbol{D}_{i_1} e^{\boldsymbol{D}_0 t_2}\boldsymbol{D}_{i_2} \ldots e^{\boldsymbol{D}_0 t_n}\boldsymbol{D}_{i_n}\mathbf{1}. \tag{2.3}$$

Here t_j is the time between the $j - 1$st and jth arrival, and i_j is the type of the jth mark, with the convention that there is an arrival at time 0. For the important special case of a MAP we have no distinction between the different points such that $\boldsymbol{D}_{i_1} = \boldsymbol{D}_1$.

The rational arrival process

The class of rational arrival process (RAP) is defined in [11] as the class of point processes where the prediction process has a version that belongs to a finite dimensional sub-space. The main result of [11] is that RAPs can be equivalently characterised as the point processes with finite dimensional distributions given by Equation (2.3). The process class has also been considered in [60] as a cascading sequence of matrix exponentials but without the complete characterisation of [11]. The RAP generalises the MAP in a way similar to the way the ME distributions generalise PH distributions.

Time reversal

Following [15] or [73], we restate Definition 1 of [4] as

Definition 6 (Definition 1 of [4]) *The time reverse of the* MAP $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ *is the* MAP $(\tilde{\boldsymbol{D}}_0, \tilde{\boldsymbol{D}}_1)$ *with* $\tilde{\boldsymbol{D}}_i = \boldsymbol{\Delta}(\boldsymbol{\pi})^{-1}\boldsymbol{D}_i'\boldsymbol{\Delta}(\boldsymbol{\pi})$*,* $i = 0, 1$ *where* $\boldsymbol{\pi}$ *is the stationary probability distribution satisfying* $\boldsymbol{\pi}\boldsymbol{D} = \boldsymbol{0}$ *and* $\boldsymbol{\Delta}(\boldsymbol{\pi})$ *is a diagonal matrix with the components of* $\boldsymbol{\pi}$ *as diagonal elements.*

MAP properties

Let $N(t)$ be the number of arrivals (counts) in $(0, t]$. It is customary in applied probability to use transform expressions when calculating moments. For the BMAP we have the generating function

$$\mathbb{E}\left(z^{N(t)}\right) = H^\star(z,t) = \boldsymbol{\alpha} e^{\boldsymbol{D}(z)t}\boldsymbol{1}, \tag{2.4}$$

where $\boldsymbol{D}(z) = \sum_{i=0}^{\infty} z^i \boldsymbol{D}_i$. The Index of Dispersion for Counts (IDC) is defined as $IDC(t) = \mathbb{V}ar(N_t)/\mathbb{E}(N_t)$, i.e., as the ratio of the variance of N_t to the corresponding variance which is 1 in the case of a Poisson arrival process [38, p. 72]. The $IDC(t)$ can be calculated by taking first and second derivatives in Equation (2.4). The $IDC(t)$ is most frequently used for time stationary versions of MAPs and BMAPs such that the initial vector $\boldsymbol{\alpha}$ is equal to $\boldsymbol{\pi}$. For a time stationary MAP the $IDC(t)$ is given as (see e.g. [62])

$$\begin{aligned} IDC(t) &= 1 - 2\lambda^\star + \frac{2}{\lambda^\star}\boldsymbol{\pi}\boldsymbol{D}_1(\boldsymbol{\Pi} - \boldsymbol{D})^{-1}\boldsymbol{D}_1\boldsymbol{1} \\ &\quad - \frac{2}{\lambda^\star t}\boldsymbol{\pi}\boldsymbol{D}_1(\boldsymbol{I} - e^{\boldsymbol{D}t})(\boldsymbol{\Pi} - \boldsymbol{D})^{-2}\boldsymbol{D}_1\boldsymbol{1} \end{aligned} \tag{2.5}$$

where $\boldsymbol{\Pi} = \boldsymbol{1}\boldsymbol{\pi}$ and $\lambda^\star = \boldsymbol{\pi}\boldsymbol{D}_1\boldsymbol{1}$. A stationary sequence of inter-arrival times is obtained by initiating the process according to the vector $\boldsymbol{\phi}$ which is obtained from the embedded Markov chain with the value of the state immediately after an arrival. The transition probability matrix $\boldsymbol{P}$ of this embedded Markov chain is given by

$$\boldsymbol{P} = (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1. \tag{2.6}$$

The vector $\boldsymbol{\phi}$ is then obtained by solving $\boldsymbol{\phi} = \boldsymbol{P}\boldsymbol{\phi}$, where some states of $\boldsymbol{P}$ might be ephemeral such that the corresponding entries of $\boldsymbol{\phi}$ are zero.

Now let $S_n = \sum_{i=1}^{n} X_i$ be the sum of the first n inter-arrival intervals in the interval stationary version of the process. The Index of Dispersion for Intervals (IDI) is defined as $IDI(k) = \frac{(\lambda^\star)^2}{k}\mathbb{V}ar(S_k)$, i.e., as the ratio of the variance of S_n to the corresponding variance in case of a Poisson arrival process [38, p. 71].

Theorem 7 (Theorem 1 of [6]) *The IDI for a MAP -* $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ *is*

$$IDI(n) = 2\lambda^\star\boldsymbol{\pi}(\boldsymbol{I} - (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1 + \boldsymbol{\Phi})^{-1}(-\boldsymbol{D}_0)^{-1}\mathbf{1} - 1$$

$$-\tfrac{2}{n}\lambda^\star\boldsymbol{\pi}(\boldsymbol{I} - (-\boldsymbol{D}_0^{-1}\boldsymbol{D}_1)^n)(\boldsymbol{I} - (-\boldsymbol{D}_0)^{-1}\boldsymbol{D}_1 + \boldsymbol{\Phi})^{-2}(-\boldsymbol{D}_0^{-1}\boldsymbol{D}_1)(-\boldsymbol{D}_0)^{-1}\mathbf{1}$$

with $\boldsymbol{\Phi} = \mathbf{1}\boldsymbol{\phi}$.

Using an alternative normalisation in the expression for the IDI one gets the Index of Variation for Intervals (IVI). For the IVI, we normalise by the marginal variance of the process itself. Thus, the IVI is the sequence of dimensionless constants $IVI(n) = \mathbb{V}ar(S_n)[n\mathbb{V}ar(S_1)]^{-1}$. The IVI is a measure of the variability in the S_n that is due to the dependence of the successive intervals. The normalisation with $n\mathbb{V}ar(S_1)$ can sometimes be more natural as the IVI is 1 for a renewal process.

Calculation of matrix exponentials using uniformization

The calculation of the exponential of a matrix is a frequently occurring operation when dealing with ME distributions and RAPs, and of course PH distributions and MAPs. The calculation of the exponential of a generator matrix can be done surprisingly efficiently and in a numerically stable manner using a method called uniformization, even though in general, the calculation of the matrix exponential is challenging. The uniformization formula is

$$e^{\boldsymbol{Q}x} = e^{-\eta x}\sum_{n=0}^{\infty}\frac{(\eta x)^n}{n!}\boldsymbol{K}^i \tag{2.7}$$

where η must be chosen to be at least as large as the largest absolute value on the diagonal of $\boldsymbol{Q}$ and $\boldsymbol{K} = \boldsymbol{I} + \eta^{-1}\boldsymbol{Q}$. This ensures that $\boldsymbol{K}$ is a stochastic matrix, whenever $\boldsymbol{Q}$ is a generator, while $\boldsymbol{K}$ is a sub-stochastic matrix, whenever $\boldsymbol{Q}$ is a sub-generator. The infinite matrix sum involves only bounded non-negative entries, and an exact upper bound for the truncation error can be derived from the Poisson weighting factors $\frac{(\eta x)^n}{n!}e^{-\eta x}$.

The calculation of the probability distribution of $N(t)$ has been addressed in [65] for a BMAP. The matrix $\boldsymbol{P}(n,t)$ with (i,j)th entries $P(N(t)=n, J(t)=j|J(i)=0)$ can be efficiently computed using the uniformization algorithm, with $J(t)$ being the state of the Markov chain at time t. The recursion scheme for a MAP is

$$\begin{aligned}
\boldsymbol{V}(0,0) = \boldsymbol{I} \qquad & \boldsymbol{P}(0,t) \leftarrow \boldsymbol{V}(0,0)b_0 \\
& \text{for } n \geq 1 \\
\boldsymbol{V}(n,0) = \boldsymbol{0} \qquad & \boldsymbol{P}(n,t) \leftarrow \boldsymbol{0} \\
& \text{for } 1 \leq k \leq N \\
\boldsymbol{V}(n,k) = \quad & \boldsymbol{V}(n,k-1)\boldsymbol{K}_0 + \boldsymbol{V}(n-1,k-1)\boldsymbol{K}_1 \qquad (2.8)\\
\boldsymbol{P}(n,t) \leftarrow \quad & \boldsymbol{P}(n,t) + \boldsymbol{V}(n,k)\tfrac{(\eta t)^n}{n!}e^{-\eta t}, \qquad (2.9)
\end{aligned}$$

with $\eta = \max_i |(D_0)_{ii}|$, $\boldsymbol{K}_0 = \frac{1}{\eta}\boldsymbol{D}_0 + \boldsymbol{I}$, and $\boldsymbol{K}_1 = \frac{1}{\eta}\boldsymbol{D}_1$. We will apply this algorithm in Section 3.2 where we will describe its use in relation to an estimation problem described in [66]. In [69] uniformization was used to calculate the *non-central moment matrices* $\boldsymbol{\Theta}_q(t)$, given by their (i,j)-elements. For $t \geq 0$, $q \in \mathbb{N}_0$ we have $[\boldsymbol{\Theta}_q(t)]_{ij} = \mathbb{E}\left(N^q(t)\boldsymbol{\delta}\left(J(t)=j\right)|J(0)=i\right)$, where $\boldsymbol{\delta}\left(A\right)$ is the indicator function of the event A. Corollary 3.4.1 of [69] gave a formula for numerical evaluation of $\boldsymbol{\Theta}_q$.

Corollary 8 (Corollary 3.4.1 of [69]) *The matrices $\boldsymbol{\Theta}_q$ $(q \geq 1)$ are given by*

$$\boldsymbol{\Theta}_q(t) = e^{-\eta t}\sum_{n=1}^{q} \sum_{\substack{\sum_{r=1}^n q_r = q \\ q_r \geq 1}} \sum_{k=0}^{\infty} \frac{q!}{q_1!\cdots q_n!}\frac{(\eta t)^{k+n}}{(k+n)!}\boldsymbol{E}_{q_1\ldots q_n}(k),$$

with

$$\eta \geq \max_i\left(-D_{ii}\right),\ \boldsymbol{D}_n^* = \sum_{i=0}^{\infty} i^n \boldsymbol{D}_i,\ \boldsymbol{K} = \boldsymbol{I} + \frac{1}{\eta}\boldsymbol{D},\ \boldsymbol{K}_n^* = \frac{1}{\lambda}\boldsymbol{D}_n^*,$$

$$\boldsymbol{E}_{\emptyset}(k) = \boldsymbol{K}^k, \qquad \boldsymbol{E}_{q_1\ldots q_n}(0) = \prod_{r=1}^{n}\boldsymbol{K}_{q_r}^*,$$

$$\boldsymbol{E}_{q_1\ldots q_n}(k+1) = \boldsymbol{E}_{q_1\ldots q_n}(k)\boldsymbol{K} + \boldsymbol{E}_{q_1\ldots q_{n-1}}(k+1)\boldsymbol{K}_{q_n}^*.$$

Uniformization was also used in [21, 36] which will be described in Chapters 3 and 5.

3 Estimation

Phase-type distributions and Markovian arrival processes, to a lesser extent matrix-exponential distributions and rational arrival processes, have been used extensively for modelling stochastic systems, particularly in queueing contexts, but also in the context of risk models and for modelling financial systems. Models and applications based on data are significantly less common, this is probably due to several factors. Most noticeable among these is an inherent problem of the models stemming from their flexibility; the number of free parameters is large contrary to the principle of parsimony which is hailed in statistics. A different, yet related, reason is that the theory on statistical inference for these models is still at a relatively undeveloped stage which manifests itself in the sparsity of readily available and reliable software packages.

In this chapter we describe our contributions to the estimation area. The original work was [66] on estimation of parameters in the Interrupted Poisson Process (IPP) [53] with applications to Fisheries Science to be described in Section 3.2. Some of the ideas of that paper were followed in [21] which compares the EM algorithm with a quasi-Newton method with explicit calculation of the gradient for the optimisation of the likelihood function. The paper also provides an explicit algorithm for the calculation of the Fisher information matrix. The interpretation of the Fisher information matrix is useful in cases using non-redundant canonical forms that remove the problem of over-parameterisation.

3.1 Estimation in phase-type models

Early works on estimation in PH distributions were based on more or less heuristic approaches minimising various criteria [2, 30, 34] or moment matching [49, 50, 78]. Initial work [31, 32, 33] based on the maximum likelihood principle was followed by the now dominant approach [12] using the EM algorithm. Estimation based on a Bayesian approach using the Markov Chain

Monte Carlo method has been reported in [23] and [59]. In [23] the problem of non-uniqueness of PH representations is avoided by estimating functionals of PH distributions rather than the parameters of the distributions. Very little work has been reported on the uncertainty of parameter estimates. One exception is [43] that reports uncertainty estimates but apparently from a purely numerical perspective using standard software.

The approach taken in [12] is to view the problem of estimating parameters in phase-type distributions as a missing data problem. The missing information is the realised path behaviour of the underlying Markov chains associated with the different times of absorption. The EM algorithm [40] is particularly well suited in such settings and in [12] formulae for the E and M steps are given. The matrix $\boldsymbol{C}(y)$ given by the integral in Equation (3.1) is used in the E step of the algorithm to calculate the expected time spent in each state given the observed value of the absorption time y and the current values $(\boldsymbol{\alpha}, \boldsymbol{S})$ of the parameters. In [12] the matrix-exponential and the integral given in Equation (3.1) were evaluated solving systems of linear differential equations using a Runge-Kutta method. An alternative way is to use uniformization as in Equation (2.7). We took this approach in [21], while a similar approach has been taken by [51] for the case of estimating parameters in the BMAP. Our method is slightly different from that of [51] although the basic idea is the same. The evaluation using uniformization is given by

$$\boldsymbol{C}(y) = \int_0^y e^{\boldsymbol{S}(y-u)} \boldsymbol{s}\boldsymbol{\alpha} e^{\boldsymbol{S}u} \mathrm{d}u = e^{-\eta y} \sum_{s=0}^{\infty} \frac{(\eta y)^{s+1}}{(s+1)!} \boldsymbol{K}_{\boldsymbol{C}}(s), \tag{3.1}$$

which was Equation 1 of [21] where $\boldsymbol{K}_{\boldsymbol{C}}(s) = \sum_{j=0}^{s} \boldsymbol{K}^j \frac{1}{\eta} \boldsymbol{s}\boldsymbol{\alpha} \boldsymbol{K}^{s-j}$. The matrices $\boldsymbol{K}_{\boldsymbol{C}}(s)$ may be calculated recursively. For large values of the argument y the matrix function $\boldsymbol{C}(y)$ can be evaluated using

$$\boldsymbol{C}(x+y) = e^{\boldsymbol{S}x} \boldsymbol{C}(y) + \boldsymbol{C}(x) e^{\boldsymbol{S}y}.$$

This formula can also be used to calculate $\boldsymbol{C}(x + \Delta x)$, using previous terms, improving the efficiency considerably.

One of the strengths of the uniformization method is the exact upper bound that can be given on the absolute truncation error since the weighting factors can be interpreted as the probability mass function of the Poisson distribution. A similar exact upper bound on the truncation error can be given when determining an upper limit for the truncation of the sum involved in calculating $\boldsymbol{C}(y)$. To see this, we will consider the first moment distribution of the Poisson distribution given by

$$q_i = \frac{i \cdot \lambda^i}{\lambda i!} e^{-\lambda}, \quad i = 0, 1, 2, \ldots, , \qquad \left(\text{or} \;\; q_i = \frac{\lambda^{i-1}}{(i-1)!} e^{-\lambda}, \quad i = 1, 2, \ldots \right).$$

Thus the Poisson distribution is in a sense closed under size biasing albeit shifted to the right. All row sums of $\boldsymbol{K_C}(s)$ are bounded by $s+1$ and we can obtain the lower bound for the truncation limit from the size biased distribution of the Poisson distribution, which happens to be the standard uniformization truncation limit plus one.

The Newton-Raphson method described in [21] was based on an explicit calculation of the gradient. Equation (4) of [21] has the same structure as (3.1) and the determination of the truncation level from the Poisson distribution applies here as well. The two methods had equal performance in the cases we investigated.

Finally in [21], similar numerical techniques were applied to calculate the Fisher information matrix for both the EM method and the Newton-Raphson approach. Phase type distributions with upper bidiagonal representation of the sub-generator $\boldsymbol{S}$ have unique canonical forms avoiding the usual over-parameterisation of PH distributions. One specific type of upper diagonal representation is commonly referred to as a Coxian representation. The Fisher information was calculated for such Coxian PH representations to evaluate the uncertainty of parameter estimates.

3.2 Estimation of MAP parameters from counting information

In [51] the approach of [12] was generalised to the setting of fitting data to observed inter-arrival times of a BMAP. Previous work in this direction is [75, 76, 77]. However, when point processes are observed, it is frequently the case that only information on counts obtained during specific time intervals is available, rather than the more detailed information of arrival instances. We investigated one such case in [66]. The feeding pattern of predatory fish is not well understood, yet under the assumption of constant digestion times, the number of fish in the stomach of a predatory fish can roughly be considered as the number of prey items caught during the previous time interval of a length corresponding to the digestion period. Other similar contexts include observation of packet counts during intervals of fixed lengths in communication systems. The predation process was modelled as an interrupted Poisson process, where predators forage either in a patchy environment of prey encountering prey with a rate of λ or move between patches without feeding opportunities. The rate ω_1 of leaving a patch and the patch encounter rate ω_2 are both constant in the IPP scenario [19]. The IPP is a PH renewal process that can be expressed as a MAP($\boldsymbol{D}_0, \boldsymbol{D}_1$) with

$$\boldsymbol{D}_0 = \begin{bmatrix} -(\lambda+\omega_1) & \omega_1 \\ \omega_2 & -\omega_2 \end{bmatrix}, \qquad \boldsymbol{D}_1 = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.2}$$

We estimated the parameters of the model by optimising the log-likelihood function $l(\boldsymbol{\theta}, \boldsymbol{x})$ where $\boldsymbol{\theta} = (\lambda, \omega_1, \omega_2)$. With n_x denoting the number of fish with exactly x prey items in the stomach the optimisation problem can be stated as

$$\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}; \boldsymbol{x}) = \max_{\boldsymbol{\theta}} \sum_{x=0}^{x_{max}} n_x \log\left(\boldsymbol{\pi}\boldsymbol{P}(x, t_0)\mathbf{1}\right),$$

where $\boldsymbol{P}(x, t_0)$ is calculated using Equation (2.9). The gradient of $l(\boldsymbol{\theta}; \boldsymbol{x})$ is given by

$$\frac{\partial l(\boldsymbol{\theta}; \boldsymbol{x})}{\partial \theta_i} = \frac{\partial \boldsymbol{\pi}}{\partial \theta_i} \sum_{x=0}^{x_{max}} \frac{n_x}{\boldsymbol{\pi}\boldsymbol{P}(x,t)\mathbf{1}} \boldsymbol{P}(x,t)\mathbf{1} + \boldsymbol{\pi} \sum_{x=0}^{x_{max}} \frac{n_x}{\boldsymbol{\pi}\boldsymbol{P}(x,t)\mathbf{1}} \frac{\partial \boldsymbol{P}(x,t)}{\partial \theta_i}\mathbf{1}.$$

By differentiation in Equation (2.8) and (2.9) we obtain $\frac{\partial \boldsymbol{P}(x,t)}{\partial \theta_i}$ through

$$\begin{aligned}
\boldsymbol{V}(0,0)'_i &= \mathbf{0} \\
\boldsymbol{V}(n,k)'_i &= \boldsymbol{V}(n,k-1)'_i\boldsymbol{K}_0 + \boldsymbol{V}(n,k-1)\boldsymbol{K}_0{}'_i \\
&\quad +(\boldsymbol{V}(n-1,k-1)'_i\boldsymbol{K}_1 + \boldsymbol{V}(n-1,k-1)\boldsymbol{K}_1{}'_i)\delta_{n>0} \\
\boldsymbol{P}(n,t)'_i &\leftarrow \boldsymbol{P}(n,t)'_i + b_n{}'_i\boldsymbol{V}(n,k) + b_n\boldsymbol{V}(n,k)'_i \ ,
\end{aligned}$$

where $\delta(n > 0)$ is 1 when $n > 0$ and 0 when $0 \geq n$. The partial derivatives of $\boldsymbol{K}_0$ and $\boldsymbol{K}_1$ are calculated by basic rules of differentiation. In [66] the algorithm was tested on an exhaustive simulation study. The study showed that reliable estimation results can be expected already for moderate sample sizes, whenever the parameter values of the IPP are not too extreme.

In addition, the estimation algorithm was applied in [66] for a data set of cods feeding on capelin. Data was given in the form of number of capelin found in the stomachs of cods partitioned according to the length distribution of the cods. Table 3.1 presents the estimation results. Goodness-of-fit tests for the applicability of the IPP model were accepted at the 5% significance level for all length classes, except class 3. The p-value for the class 3 goodness of fit test was 1%. All fits were significantly better than those obtained with a Poisson distribution as evaluated by the difference in log-likelihood.

Class	LH$_{IPP}$	LH$_{PP}$	MLE			
			$\hat{\omega}_1(\hat{\sigma}_{\hat{\omega}_1})$	$\hat{\omega}_2(\hat{\sigma}_{\hat{\omega}_2})$	$\hat{\lambda}(\hat{\sigma}_{\hat{\lambda}})$	$\hat{\lambda}^*$
1	-136.6	-154.7	1.54(1.75)	0.97(0.55)	3.70(1.45)	1.4
2	-214.0	-290.5	13.26(31.02)	1.68(0.52)	28.40(52.36)	3.2
3	-298.2	-517.0	5.14(3.14)*	1.26(0.23)*	27.18(10.49)*	5.3
4	-288.7	-558.4	3.38(1.50)	1.23(0.23)	29.59(6.79)	7.9
5	-256.0	-595.0	2.27(0.76)	1.22(0.23)	36.20(4.78)	12.7
6	-205.2	-542.1	3.32(1.40)	1.25(0.26)	52.54(10.59)	14.4

Table 3.1: Estimation results for cod-capelin data: The value of the IPP log-likelihood function at maximum (LH$_{IPP}$), the value of the Poisson log-likelihood function, the maximum likelihood estimates ($\hat{\lambda}$, $\hat{\omega}_1$, $\hat{\omega}_2$), the estimated fundamental rate ($\hat{\lambda}^*$). * symbolises a rejected goodness of fit test result.

$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$	$\hat{\omega}_1$	$\hat{\omega}_2$
$\hat{\sigma}_{\hat{\lambda}_1}$	$\hat{\sigma}_{\hat{\lambda}_2}$	$\hat{\sigma}_{\hat{\lambda}_3}$	$\hat{\sigma}_{\hat{\lambda}_4}$	$\hat{\sigma}_{\hat{\lambda}_5}$	$\hat{\sigma}_{\hat{\lambda}_6}$	$\hat{\sigma}_{\hat{\omega}_1}$	$\hat{\sigma}_{\hat{\omega}_2}$
5.33	12.14	21.40	30.09	44.54	53.98	3.53	1.27
1.00	1.97	3.15	4.43	7.34	7.99	0.87	0.12

Table 3.2: Estimation result in the reduced model - parameters (first line) - and their estimated standard deviation (second line).

The similarity of the ω_2 estimates over length classes (Table 3.1) is striking suggesting that the model provides some valid biological information. The biological interpretation of the model suggests that a reasonable hypothesis is $\omega_{i1} = \omega_1$ and $\omega_{i2} = \omega_2$. Formally, the sum of the log-likelihood values is 1398.7 for the reduced model with $\omega_{i1} = \omega_1$ and $\omega_{i2} = \omega_2$ to be compared with 1401.2 for the model with different values of the ωs for each length class. The test statistic is 5 which is clearly insignificant when compared to a $\chi^2(10)$ distribution. The parameter estimates obtained in this reduced model are given in Table 3.2.

4 Sensitivity analyses

One of the main attractions of the Markovian Arrival process is that it offers the opportunity to perform numerical calculations relatively easily. Frequently stochastic models are hard to analyse analytically when one goes beyond standard features such as steady state probabilities for simple queueing systems. Hence in order to gain insight into models and systems one must resort to computer experimentation using simulation or well-designed numerical experiments whenever the latter is feasible. Even though simulation might be faster in quite a few cases it is more satisfactory to make explicit calculations where potential errors can be ascribed solely to implementation, truncation, or to finite precision calculations. In many cases it is possible to give error bounds when working with Markovian arrival processes and phase-type distributions. In this section we will describe some studies where the numerical tractability of the Markovian models has served as a means to obtain generic insights into queueing systems.

4.1 Predictive power on queueing from second order properties

Throughout the 80s and 90s quite a few studies were published approximating arrival processes by second order properties only, or fitting processes from data using moment estimates, again generally based only on second order information. In [6] we addressed the effect on queueing behaviour when varying third order properties of arrival processes, either fixing second order properties of counts or second order properties of intervals. Not surprisingly, the study demonstrated that queueing behaviour could vary substantially even in simple cases of models with few free parameters. The IDC for a MAP is given by Equation (2.6). It is possible to construct a sequence of two-state MAPs that all have the same rate and IDC. Each two state MAP of [6] was a Switched Poisson Process (SPP) [81], which is a superposition of an IPP and a Poisson process. The state corresponding to the active state of

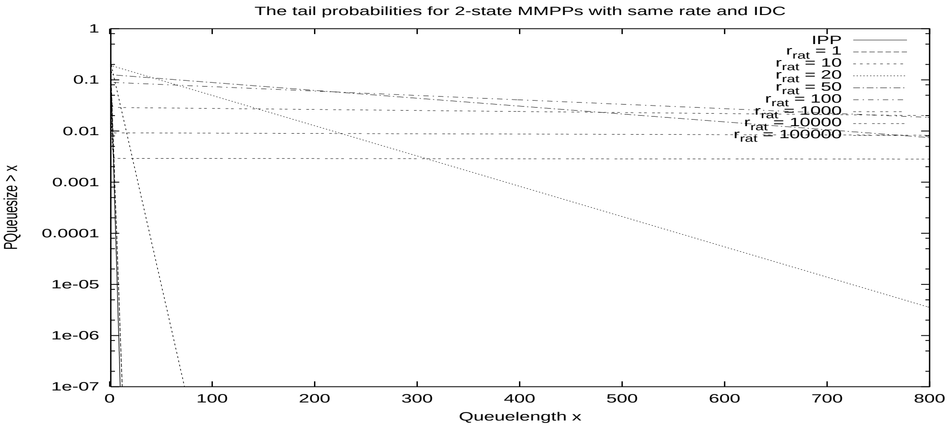

Figure 4.1: **(Figure 1 in [6]) Probability that the queue length exceeds x for a number of SPPs with fixed rate and IDC**

the IPP provides a regime with high activity while the state corresponding to the passive state of the IPP provides a regime with low activity. This construction with a sequence of SPPs was used to make Figure 4.1 (Figure 1 in [6]). The figure demonstrates that the queue length cannot be adequately predicted from first and second order properties of the counting properties. Correspondingly, one can construct a sequence of two-state MAPs with fixed rate and IDI, where the IDI can be calculated using Theorem 7. Figure 4.2 (Figure 4 in [6]) is based on this idea. It is evident that second order properties of the interval process alone do not suffice to give reasonable prediction of queueing behaviour either.

The study raised the question whether simultaneously fixing second order properties of both the stationary versions of the counting and the interval processes would give a good prediction of queueing behaviour. This question was settled in the negative in [4]. The key observation to reach that conclusion was that a Markovian arrival process and the time reversed version of it have identical first and second order descriptors for the time stationary and interval stationary properties. We hence constructed a MAP that, used as an input process to a queue, gave rise to a somewhat different queueing behaviour than what would be obtained using the time reversed version. Before giving the queueing examples at the end of the section we demonstrate that a MAP and its time reversed version have the same first and second order properties of both the interval and counting processes. Theorem 3.1 of [4] stated that the marginal distribution of counts in intervals of fixed lengths

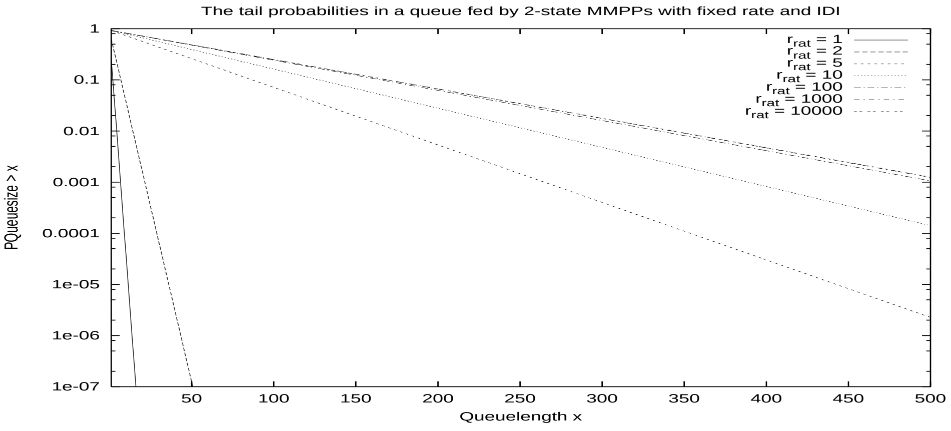

Figure 4.2: **(Figure 4 in [6])Probability that the queue length exceeds x for a number of SPPs with fixed rate and IDI**

agree between a MAP and its reverse.

Theorem 9 (Theorem 3.1 of [4]) *The marginal distributions of $N(t)$ for the time stationary version of a* MAP *and its reverse are identical. In terms of the transforms,*

$$H^\star(z,t) = \tilde{H}^\star(z,t),$$

where $H^\star(z,t) = \boldsymbol{\pi}\, exp((\boldsymbol{D}_0 + z\boldsymbol{D}_1)t)\mathbf{1}$ *and* $\tilde{H}^\star(z,t) = \boldsymbol{\pi}\, exp((\tilde{\boldsymbol{D}}_0 + z\tilde{D}_1)t)\mathbf{1}$.

The marginal distributions of $N(t)$ are identical for the interval stationary versions of a MAP *and its reverse. In terms of the transforms,*

$$J^\star(z,t) = \tilde{J}^\star(z,t),$$

where $J^\star(z,t) = \boldsymbol{\phi} exp((\boldsymbol{D}_0 + z\boldsymbol{D}_1)t)\mathbf{1}$ *and* $\tilde{J}^\star(z,t) = \tilde{\boldsymbol{\phi}} exp((\tilde{\boldsymbol{D}}_0 + z\tilde{\boldsymbol{D}}_1)t)\mathbf{1}$.

From that result it follows immediately that second order properties of the counting process also agree between a MAP and its reverse. This was stated as Corollary 3.2 of [4].

Corollary 10 (Corollary 3.2 of [4]) *The following descriptors agree for a* MAP *and for its reverse:*

1. *The variance time curve, the IDC, and the peakedness functional [41],*

2. *The square wave power spectral density (SQSD).*

The result for interval properties, similar to Theorem 9 for counting properties, was stated as Theorem 3.3 in [4].

Theorem 11 (Theorem 3.3 in [4]) *For all $n \geq 1$, the distributions of the time between the kth and the $(k+n)$th arrival are identical for an interval stationary point process and for its reverse. Equivalently, the distributions of S_n and $\tilde{S}_n$ are identical for all $n \geq 1$.*

Also in this case the properties regarding second order properties follow immediately, stated as Corollary 3.4 of [4].

Corollary 12 (Corollary 3.4 of [4]) *The following descriptors agree for a* MAP *and the corresponding reversed* MAP*:*

1. *The marginal distributions of the inter-arrival times.*
2. *The variance of S_n and $\tilde{S}_n$, and therefore also the IDIs and IVIs.*

We give an important definition following the lines of [58]. The definition was Definition 2 of [4].

Definition 13 (Definition 2 of [4]) *Two point processes are stochastically equivalent (SE) if for any $n \geq 1$, the joint distributions of the first n intervals agree.*

This definition leads naturally to the following definition of a class of MAPs (Definition 3 of [4]).

Definition 14 (Definition 3 of [4]) *A* MAP *that is stochastically equivalent with its reverse is called* reversible.

The existence of reversible MAPs should be immediate. Examples include

Theorem 15 (Theorem 4.1 of [4]) *The following are special cases of reversible* MAP*s:*

- *Phase-type renewal processes*
- *Two-state MAPs*

Another immediate result is (Theorem 4.2 of [4]).

Theorem 16 (Theorem 4.2 of [4]) *The superposition of the reversed versions of two independent* MAP*s is identical to the reverse of the superposition of these* MAP*s.*

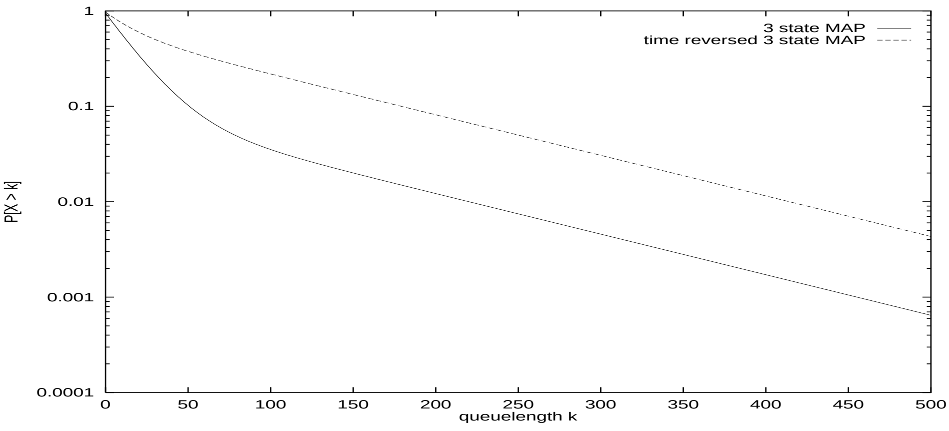

Figure 4.3: **(Figure 2 of [4])Tail behaviour in MAP/M/1 queue induced by a 3-state MAP respectively its reverse (load = 0.5)**

Consequently to create a non-reversible MAP one would need a state space of dimension at least 3. The construction used in [4] to demonstrate differences in queueing behaviour between a MAP and its reverse was a 3-state process. For the 3-state MAP, we selected the parameter values $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 100, D_{0,12} = 5 \cdot 10^{-4}, D_{0,23} = 5 \cdot 10^{-4}$, and $D_{0,31} = 1$. That MAP and its reverse were used as input in a $MAP/M/1$ queue. The MAP cycles through the three phases in such a way that the intensity increases and then drops sharply. Conditioned on state 1 there is no input to the queue and the queue length decreases according to a Poisson process; conditioned on being in state 2 the queue is just stable with usual random fluctuations, while conditioned on state 3 the queue is highly unstable. In general the queue length will be moderate when state 3 is entered, and a short period of overload can be compensated by the following longer stay in state 1. For the reversed process, however, the short period with overload is followed by a longer period with only very little excess capacity to get rid of the unfinished work that remains from the sojourn in state 3, and the queue will recover slowly. Choosing the mean of the exponential server to obtain a traffic intensity of 0.5, we computed the tail probabilities of the steady-state queue lengths. These are shown in Figure 4.3. The difference in queueing behaviour is striking, finally settling that combined second order properties of counts and intervals by no means can be used as predictors of queueing behaviour. Some suggestions for descriptors that might add predictive power were also presented in [4] but the conclusion must be that more specific studies of data related to the particular application in mind need to be performed.

4.2 Sensitivity analyses of alternating processes

The work reported in this section is yet unpublished work jointly performed with Marcel Neuts. Our starting point will be that of an arrival process in an environment alternating between two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ of states. Both sets are finite with p_1 and p_2 states respectively. The sojourn times of visits to $\mathcal{S}_1$ and $\mathcal{S}_2$ will be phase-type distributed with representation $(\boldsymbol{\alpha}_1, \boldsymbol{S}_1)$ and $(\boldsymbol{\alpha}_2, \boldsymbol{S}_2)$ respectively. The Markov chain describing the successive alterations will be used as modulator for an ON-OFF type process. We have Poisson arrivals with intensity λ_1 whenever the alternating process is in $\mathcal{S}_1$ and Poisson arrivals with intensity λ_2 whenever the alternating process is visiting $\mathcal{S}_2$. This model can be formulated as a MAP. When each subset consists of only one state our model reduces to that of a Switched Poisson Process (SPP). Further, if one of the arrival rates is zero the model reduces to that of an IPP, see Equation (3.2). The parameter matrices $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ of the MAP are appropriately partitioned by

$$\boldsymbol{D}_0 = \begin{bmatrix} \boldsymbol{S}_1 & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_2 \end{bmatrix} - \boldsymbol{D}_1 \qquad \text{with } \boldsymbol{D}_1 = \begin{bmatrix} \lambda_1 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \lambda_2 \boldsymbol{I} \end{bmatrix}.$$

The dimension of $\boldsymbol{S}_{ij}$ is $p_i \times p_j$.

The essential part of this section is on alternative choices for the matrices $\boldsymbol{S}_{12}$ and $\boldsymbol{S}_{21}$ which preserve the marginal distribution of the sojourn times in $\mathcal{S}_1$ and $\mathcal{S}_2$ while allowing for dependence between these sojourn times. The motivation for such an analysis is that in general it would be much easier to establish the marginal distributions of the sojourn times in each of the two sets, than to get information on the dependence structure, while the dependence structure might still have importance for the system under investigation. Specifically, we will address the problem of choosing $\boldsymbol{S}_{12}$ and $\boldsymbol{S}_{21}$ such that the correlation between these sojourn times is either maximised or minimised. In the case of general $\boldsymbol{S}_{12}$ and $\boldsymbol{S}_{21}$ the successive sojourn times in the two sets will still be phase-type distributed with a possibly complicated dependence structure. The marginal distributions of the two sojourn times are given by PH distributions with representations $(\boldsymbol{\phi}_1, \boldsymbol{S}_1)$ and $(\boldsymbol{\phi}_2, \boldsymbol{S}_2)$ respectively. To find the two vectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ we consider the MAP where events correspond to the times of entering each of the sets. This MAP has parameter matrices $(\boldsymbol{E}_0, \boldsymbol{E}_1)$ given by

$$\boldsymbol{E}_0 = \begin{bmatrix} \boldsymbol{S}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_2 \end{bmatrix}, \qquad \boldsymbol{E}_1 = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{0} \end{bmatrix}.$$

The embedded Markov chain of the state immediately after an event is given by Equation (2.6) and we get

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{0} & (-\boldsymbol{S}_1)^{-1}\boldsymbol{S}_{12} \\ (-\boldsymbol{S}_2)^{-1}\boldsymbol{S}_{21} & \boldsymbol{0} \end{bmatrix}.$$

The event stationary vector $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ corresponding to the events of entering subset $\mathcal{S}_1$ is the unique solution to the system of equations

$$\boldsymbol{\phi} = \boldsymbol{\phi}\boldsymbol{P}$$

which partitions into

$$\begin{aligned} \boldsymbol{\phi}_1 &= \boldsymbol{\phi}_2(-\boldsymbol{S}_2)^{-1}\boldsymbol{S}_{21}, \\ \boldsymbol{\phi}_2 &= \boldsymbol{\phi}_1(-\boldsymbol{S}_1)^{-1}\boldsymbol{S}_{12}. \end{aligned}$$

As $\boldsymbol{\phi}_i$ should be proportional to $\boldsymbol{\alpha}_i$ we see that we need to choose $\boldsymbol{S}_{21}$ and $\boldsymbol{S}_{12}$ such that the following two equations are satisfied

$$\begin{aligned} \boldsymbol{\alpha}_1 &= \boldsymbol{\alpha}_2(-\boldsymbol{S}_2)^{-1}\boldsymbol{S}_{21}, \qquad (4.1) \\ \boldsymbol{\alpha}_2 &= \boldsymbol{\alpha}_1(-\boldsymbol{S}_1)^{-1}\boldsymbol{S}_{12}. \end{aligned}$$

We always have a feasible solution corresponding to the case of independent sojourn times

$$\boldsymbol{S}_{12} = -\boldsymbol{S}_1\boldsymbol{1}\boldsymbol{\alpha}_2 \qquad \boldsymbol{S}_{21} = -\boldsymbol{S}_2\boldsymbol{1}\boldsymbol{\alpha}_1. \qquad (4.2)$$

Since for fixed parameters $(\boldsymbol{\alpha}_1, \boldsymbol{S}_1)$ and $(\boldsymbol{\alpha}_2, \boldsymbol{S}_2)$ the means and variances are fixed, maximisation or minimisation of correlations reduces to the problem of maximising or minimising the first cross moment, or the expected product, of the generic sojourn times X_1 and X_2. The joint Laplace-Stieltjes transform of the two sojourn times is given by

$$\begin{aligned} H(s_1, s_2) = \mathbb{E}\left(e^{-s_1X_1 - s_2X_2}\right) &= \frac{1}{2}\boldsymbol{\alpha}_1(s_1\boldsymbol{I} - \boldsymbol{S}_1)^{-1}\boldsymbol{S}_{12}(s_2\boldsymbol{I} - \boldsymbol{S}_2)^{-1}(-\boldsymbol{S}_2)\boldsymbol{1} \\ &+ \frac{1}{2}\boldsymbol{\alpha}_2(s_1\boldsymbol{I} - \boldsymbol{S}_2)^{-1}\boldsymbol{S}_{21}(s_2\boldsymbol{I} - \boldsymbol{S}_1)^{-1}(-\boldsymbol{S}_1)\boldsymbol{1} \end{aligned}$$

which upon differentiation with respect to s_1 and s_2, and letting s_1, s_2 tend to 0 gives the expression for the first cross moment

$$\mathbb{E}(X_1X_2) = \frac{1}{2}\boldsymbol{\alpha}_1(-\boldsymbol{S}_1)^{-2}\boldsymbol{S}_{12}(-\boldsymbol{S}_2)^{-1}\boldsymbol{1} + \frac{1}{2}\boldsymbol{\alpha}_2(-\boldsymbol{S}_2)^{-2}\boldsymbol{S}_{21}(-\boldsymbol{S}_1)^{-1}\boldsymbol{1}.$$

It is this expression we want to maximise or minimise, subject to the two sets of linear constraints given in (4.1). We find it convenient to introduce the following re-parameterisation $\boldsymbol{R}_1 = (-\boldsymbol{S}_1)^{-1}\boldsymbol{S}_{12}$ and $\boldsymbol{R}_2 = (-\boldsymbol{S}_2)^{-1}\boldsymbol{S}_{21}$. With this reformulation we can restate our problem in the case of maximisation as

$$\max_{\boldsymbol{R}_1,\boldsymbol{R}_2} \boldsymbol{\alpha}_1(-\boldsymbol{S}_1)^{-1}\boldsymbol{R}_1(-\boldsymbol{S}_2)^{-1}\mathbf{1} + \boldsymbol{\alpha}_2(-\boldsymbol{S}_2)^{-1}\boldsymbol{R}_2(-\boldsymbol{S}_1)^{-1}\mathbf{1}$$
$$\boldsymbol{R}_1 \geq \mathbf{0},\ \boldsymbol{R}_1\mathbf{1} = \mathbf{1},\ \boldsymbol{\alpha}_1\boldsymbol{R}_1 = \boldsymbol{\alpha}_2,\ -\boldsymbol{S}_1\boldsymbol{R}_1 \geq \mathbf{0},$$
$$\boldsymbol{R}_2 \geq \mathbf{0},\ \boldsymbol{R}_2\mathbf{1} = \mathbf{1},\ \boldsymbol{\alpha}_2\boldsymbol{R}_2 = \boldsymbol{\alpha}_1,\ -\boldsymbol{S}_2\boldsymbol{R}_2 \geq \mathbf{0}.$$

The matrix $\boldsymbol{R}_1$ is a $p_1 \times p_2$ transition probability matrix whose (i,j)th element gives the probability, that if the first phase-type distribution exits from state i then the second phase-type distribution will start in state j of that distribution. This matrix and the corresponding matrix $\boldsymbol{R}_2$ of probabilities constitute the decision variables of our optimisation problem. Thus we have formulated the problem as a standard linear programming problem with easily interpretable decision variables, coefficients, and constraints. The constraints of the problem do not guarantee that the Markov chain is irreducible. In the case of a reducible Markov chain the optimisation problem gives an upper, respectively lower, bound on the correlation.

As an example we will consider a case where the sojourn time distributions in each regime are given by mixtures of four Erlang$_2$ distributions with means of 2, 20, 200, and 2000, and with mixture probabilities of 0.85, 0.12, 0.025, and 0.005. The mean and standard deviation of the sojourn time distribution are 19.1 and 1451 respectively. With these values the supremum of the correlation of the sojourn times becomes 0.6627 and the infimum of the correlation becomes -0.0094. The minimum in the optimisation problem is attained by an irreducible MAP, while the parameter matrices that solves the optimisation problem corresponds to a reducible MAP. To construct an irreducible MAP a small perturbation using independence between the two sojourn times as expressed in Equation (4.2) is added to get $(1-\epsilon)\boldsymbol{R}_{12}+\epsilon\mathbf{1}\boldsymbol{\alpha}_2$ and $(1-\epsilon)\boldsymbol{R}_{21}+\epsilon\mathbf{1}\boldsymbol{\alpha}_1$ respectively. The parameter ϵ can be chosen sufficiently small such that the correlation of the sojourn times with four significant digits is 0.6627. The arrival rate λ in the active regime is set to 4. The MAP with these parameters is fed to a single server queue with an Erlang$_2$ distribution of mean 0.4. The effect of the optimisation of correlation is demonstrated in Figure 4.4 and Figure 4.5.

High correlation is beneficial to the queue as long periods with overload tend to be followed by long silent periods. The correlation of the minimum is not that far from 0 but nevertheless contributes to a more variable queue as long periods with overload tend to be followed by small silent periods

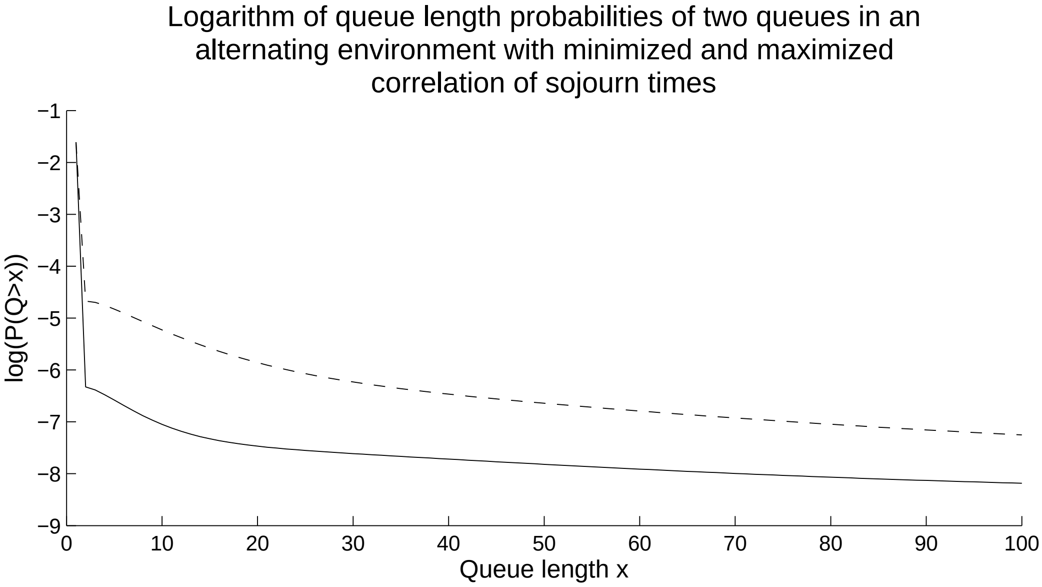

Figure 4.4: Log of queue occupancy probabilities. The upper dashed curve corresponds to the case, where the correlation between the sojourn times has been minimised, while the lower solid curve corresponds to the case where the correlation has been maximised.

such that a queueing event can survive a silent period with even more queue buildup in a possible second long period with overload.

4.3 Effects from random permutations of point processes

Another sensitivity study is [5]. The Poisson process has been applied successfully since the days of Erlang for the modelling of fresh traffic for traditional telephone systems. As telecommunications became digitised, packetised and in particular automised a significant part of traffic became only indirectly created by human activity. The variation in data traffic caused by for instance very long data files resulted in traffic patterns with a variability that far exceeded that of the Poisson process. In the early 90s a number of studies were published documenting this huge variability in packetised traffic processes, particularly the now famous Bellcore traces [57]. It was claimed that this huge variability called for a paradigm shift in queueing theory, see e.g. [45].

Among the studies claiming the need for such paradigm shift was [42] that

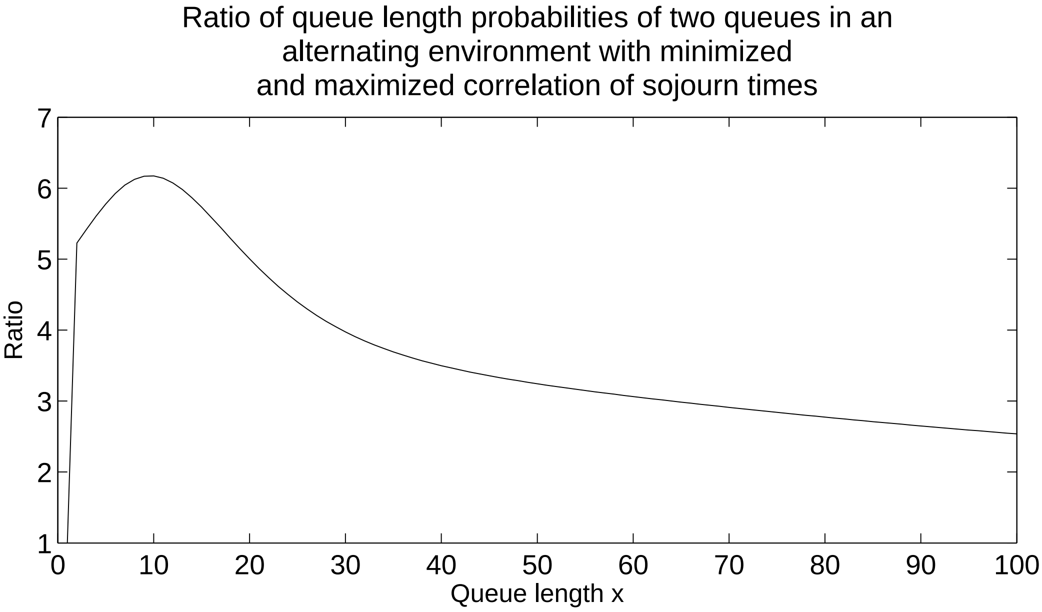

Figure 4.5: Ratio of the queue length probabilities presented in Figure 4.4 between two MAPs with maximised and minimised correlations of sojourn times.

became quite influential. The study concluded that long term rather than short term correlations in traffic processes gave rise to highly varying queue length behaviour in communications systems. The conclusion was reached by making simulations using random permutations referred to as internal and external shuffling. For both shuffling schemes a sample of measurements will be divided in blocks of a certain size m. Internal shuffling consists of random permutations of the samples within each block while keeping the sequence of the blocks, while external shuffling consists of randomly permutating the blocks while keeping the sequence within each block unchanged. In [5] we demonstrated that internal and external shuffling have a more subtle influence on the correlation structure of a process than what was communicated based on intuitive arguments in [42]. This can be seen in Figure 4.6 (Figure 1 of [5]). Internal shuffling merely averages the correlations of small lags rather than destroying them, thus leaving the process somewhat less altered than stated in [42]. The external shuffling not only removed long correlations completely but at the same time reduced the short range correlations drastically.

Even though the most important conclusions of [5] do not rely on MAP formalism we have chosen to incorporate the paper as part of the thesis, as it fits naturally with the other contributions on sensitivity. In [5] explicit con-

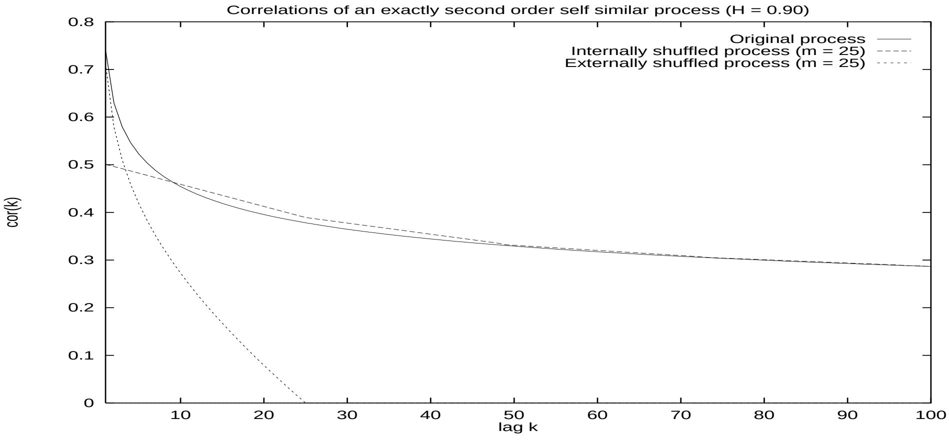

Figure 4.6: **(Figure 1 of [5])Correlations of a shuffled exactly second order self similar process
(H = 0.75 and m = 25)**

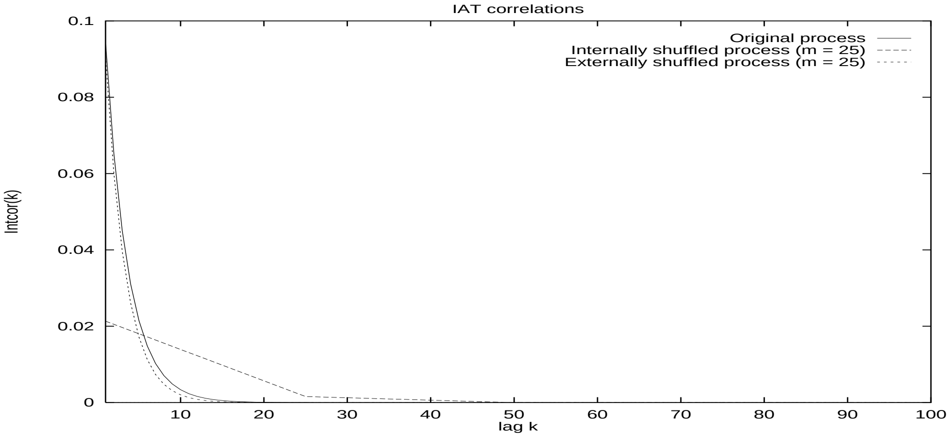

Figure 4.7: **(Figure 5 of [5])Interval correlations in a MAP with no correlation in counts**

structions are given for the external and internal shuffling of a MAP (RAP) but these constructions were not necessary to obtain the main conclusions of the paper. One example of the applicability of MAPs is presented in Figure 4.7. The figure shows the effect on the correlations in the interval process when shuffling is performed on a MAP with no correlation in the counting process.

5 Application examples

Most of the contributions described in this thesis relate to theoretical results and, in particular, enhancements to the general theory. Nevertheless, a primary motivation for our research has always been to demonstrate the applicability and potential of the Markovian arrival process with extensions for the design and analysis of technical systems, as well as for applications in science. In this chapter we will focus on some contributions of this kind of our own, acknowledging that many other significant contributions to the field exist. While Chapter 4 on sensitivity analyses presents examples using the framework to get generic insight into queueing models, in this chapter we will present some concrete examples of applications.

As telecommunications has historically been a main application of queueing theory, and applied probability in general, it should come as no surprise that applications in this field abound. Queueing theory was in fact developed from engineering applications by Erlang. Sections 5.1 and 5.2 contain applications in communications engineering. Recently, phase-type models have also appeared in computer science applications, one of which is described in Section 5.3.

5.1 Processes with excessive variability

The measurements documenting the high variability of packetised traffic as described in Section 4.3 caused many researchers to claim that traffic models based on Markovian assumptions would become partly obsolete. In [8](a preliminary version was published as [7]) we investigated the possibility of modelling arrival processes with extremely high variability using more traditional models like the MAP. The model was inspired by self-similarity as a superposition of a number d of independent SPPs, see Page 36, with logarithmically varying timescales. The algorithm, which is a heuristic, for parameter selection is designed to ensure that the autocorrelation of the counting process behaves like a power law function over a number of time scales. The

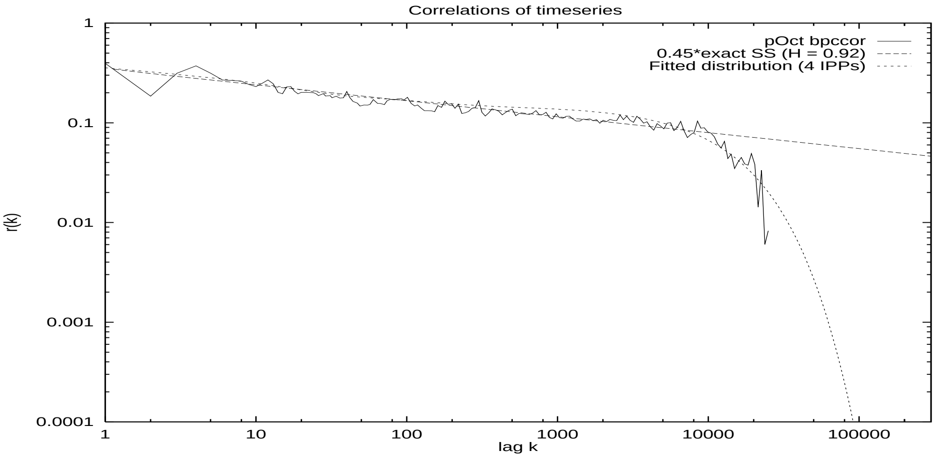

Figure 5.1: **(Figure 6 of [8])Correlations in counting process for pOct.TL and fitted model**

algorithm consists of three major steps.

1. The calculation of the logarithmic spacing between time constants of individual sources.
2. Calculation of the magnitude of the arrival intensity in the high regime for each SPP source relative to the source with the highest time constant.
3. Calculation of the arrival intensity in the high regime for the SPP source with the highest time constant and calculation of the sum of the arrival intensities in the low regime of all the sources.

The result of the fitting procedure with respect to the second order properties of counts is illustrated in Figures 5.1 and 5.2. The three curves in the figures represent the empirical correlation of one of the Bellcore datasets [57], the correlation of a MAP constructed by the heuristic to emulate the Bellcore data, and the correlation of a slight modification of a second order self similar-process. The paper ([8]) illustrated that the Markovian arrival process is fully sufficient as an engineering tool for the modelling of packet arrival processes, even in an environment of high variability and long range correlation. As evident from the results presented in Section 4.1, successfully fitting second order properties of counts is not necessarily sufficient to obtain good prediction of queueing behaviour. Even the correlation in the

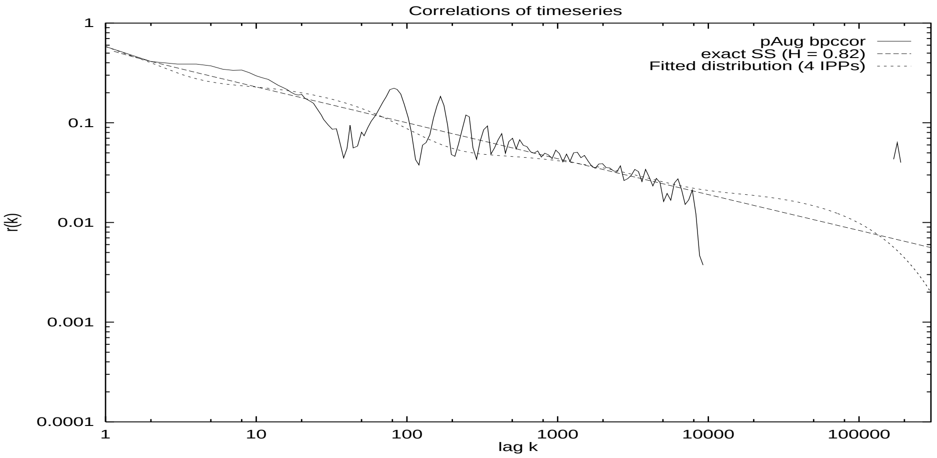

Figure 5.2: **(Figure 11 of [8])Correlations in counting process for pAug.TL and fitted model**

interval process followed a slightly different pattern than the Bellcore dataset as can be seen in Figures 5.3 and 5.4. The queueing behaviour evaluated when using the MAP as an arrival process deviated somewhat from the queueing behaviour obtained from simulations using the Bellcore data [8]. In fact this work partly inspired us to investigate which properties were most decisive for queueing behaviour leading to [6] and later [4], as described in Section 4.1. The model of [8] includes sufficient flexibility to optimise parameter selection with respect to additional criteria like properties of the interval process, but that line of research was not pursued extensively in [8].

5.2 Channel holding times in mobile networks

Another application is the modelling of channel holding times in mobile systems [36] under the influence of roaming customers. A preliminary version appeared as [35]. The call holding time is the entire duration of a call, while the cell residence time is the time a customer spends in a specific cell of the mobile communication system. The channel holding time is the part of the duration of a call that can be ascribed to a specific cell. Due to the mobility of customers and randomly varying signal strengths, the holding times of channels in mobile communication systems differ from the call holding times. In [36] we analysed this problem under the assumption of phase-type distributed call holding and cell residence times. While previous work had been

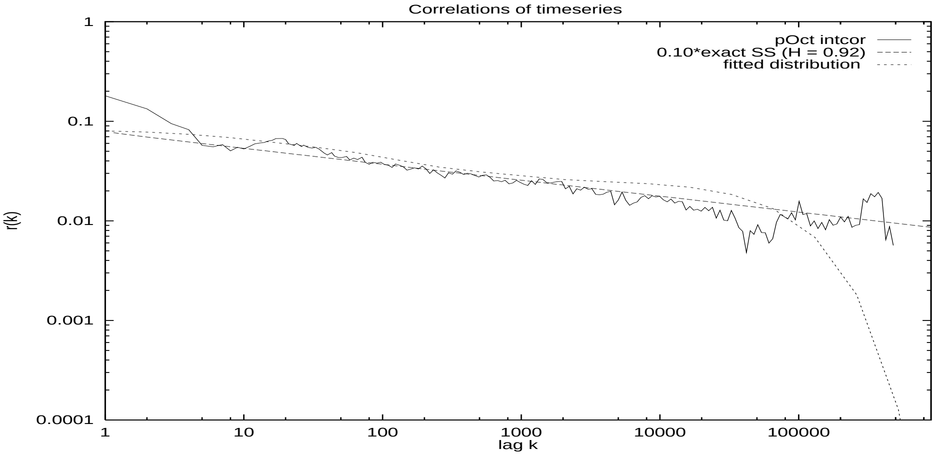

Figure 5.3: **(Figure 7 of [8])Correlations in interval process for pOct.TL and fitted model**

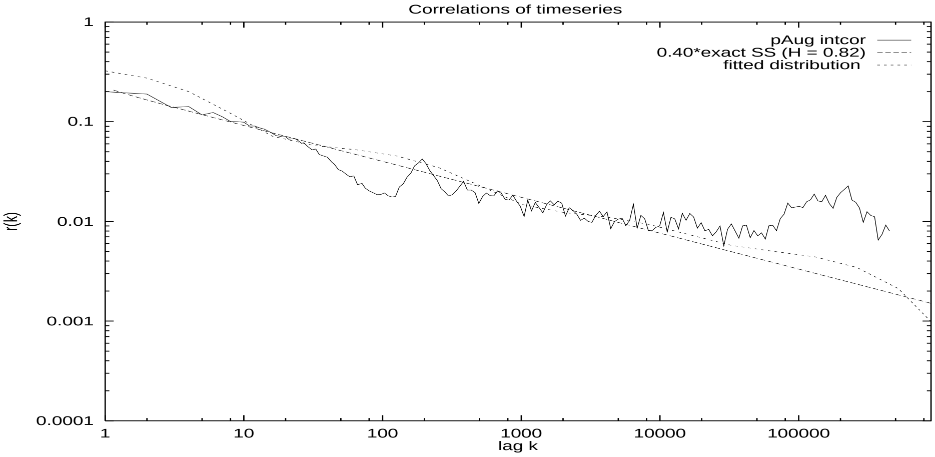

Figure 5.4: **(Figure 12 of [8])Correlations in interval process for pAug.TL and fitted model**

performed modelling residence times or channel holding times using various specific distributions [44, 72, 79], the contribution of [36] was to present a unified treatment to the modelling of channel holding times. The main result of the paper can be stated as

Theorem 17 *Suppose the call holding times are given by a phase-type distribution with representation ($\boldsymbol{\beta}, \boldsymbol{T}$), while the cell residence times are given by phase-type distributions with representations ($\boldsymbol{\alpha}_0, \boldsymbol{S}_0$) and ($\boldsymbol{\alpha}, \boldsymbol{S}$) for the first and successive cell residence times respectively. Then channel holding times are phase-type distributed and a representation can be chosen as* $(\boldsymbol{\gamma}, \boldsymbol{L})$ *with*

$$\begin{aligned}
\gamma &= \left(\boldsymbol{\alpha}_0 \otimes \frac{\boldsymbol{\beta}}{\boldsymbol{\beta K}_0(\boldsymbol{I}-\boldsymbol{K})^{-1}\mathbf{1}+1}, \boldsymbol{\alpha} \otimes \frac{\boldsymbol{\beta K}_0\left(\boldsymbol{I}-\boldsymbol{K}\right)^{-1}}{\boldsymbol{\beta K}_0\left(\boldsymbol{I}-\boldsymbol{K}\right)^{-1}\mathbf{1}+1}\right), \\
\boldsymbol{L} &= \begin{bmatrix} \boldsymbol{S}_0 \oplus \boldsymbol{T} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{S} \oplus \boldsymbol{T} \end{bmatrix}.
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{K}_0 &= \sum_{k=0}^{\infty}\sum_{l=0}^{k} \frac{k!}{l!(k-l)!}\left(\frac{\boldsymbol{I}+\eta_0^{-1}\boldsymbol{T}}{2}\right)^{l} \boldsymbol{\alpha} \left(\frac{\boldsymbol{I}+\eta_0^{-1}\boldsymbol{S}_0}{2}\right)^{k-l} \frac{\eta_0^{-1}\boldsymbol{s}_0}{2}, \\
\boldsymbol{K} &= \sum_{k=0}^{\infty}\sum_{l=0}^{k} \frac{k!}{l!(k-l)!}\left(\frac{\boldsymbol{I}+\eta^{-1}\boldsymbol{T}}{2}\right)^{l} \boldsymbol{\alpha} \left(\frac{\boldsymbol{I}+\eta^{-1}\boldsymbol{S}}{2}\right)^{k-l} \frac{\eta^{-1}\boldsymbol{s}}{2},
\end{aligned}$$

with $\eta_0 = max\{|S_{0,ii}| : 1 \leq i \leq n_0, |T_{ii}| : 1 \leq i \leq m\}$ *and* $\eta = max\{|S_{ii}| : 1 \leq i \leq n, |T_{ii}| : 1 \leq i \leq m\}$.

The phase-type representation of the channel holding times was given as formulae (22) and (23) of [36], while the formulae for $\boldsymbol{K}_0$ and $\boldsymbol{K}$ are (11) and (12) of that paper. The matrices $\boldsymbol{K}_0$ and $\boldsymbol{K}$ can alternatively be determined from (9) and (10) of [36]. The uniformization technique was applied to evaluate expressions involving matrix exponentials as in Corollary 8 and Equation (3.1).

The conditional distributions of the channel holding time conditioned on whether termination was due to a call termination or a call hand over was derived using Theorem 17. The representations of these phase-type distributions were given as $(\boldsymbol{\gamma}_V, \boldsymbol{R})$ and $(\boldsymbol{\gamma}_W, \boldsymbol{R})$ expressed by formulae (24), (25), and (27) of [36].

It is natural to allow for correlated residence times. This was modelled in [36] using a MAP model for the residence times. In this case the phase-type representation $(\boldsymbol{\gamma}, \boldsymbol{Q})$ for the channel holding time had $\boldsymbol{\gamma}$ and $\boldsymbol{Q}$ given by formula (35) and (36) of [36].

5.3 Multivariate modelling of abstract computer systems

An important aspect of formal verification of computer systems is model checking pioneered by Clarke, Emerson, and Sifakis [37, 71]. A recent reference including a description of probabilistic model checking is [14]. Model checking is based on a formal model description language, with semantics, which, in many cases of probabilistic or stochastic model checking, will be that of a discrete or continuous time Markov chain. A logical language is used to establish queries about the system in question. The validity of the logical queries is verified by the actual model checking with algorithms typically developed in the Performance Evaluation field.

In [67] we contributed to the model checking toolbox by incorporating models with rewards of different types. The rewards considered are state rewards earned proportionally to the sojourn times spent in the transient states of an absorbing Markov chain. The focus is thus on rewards earned during the sojourns of the transient states in a phase-type distribution rather than the time to absorption per se, although the absorption time is easily incorporated using a reward of one in all states. A multivariate vector of cumulated rewards is obtained by having different reward variables with different reward rates. We will use this construction extensively in Section 7.1.

The syntax for the model description language was given in Definition 1 of [67].

Definition 18 (Definition 1 of [67]) *The language* MRP *of Markov Reward Expressions consists of definitions D with process expressions E as an auxiliary syntactic category:*

$$\begin{aligned}
E &::= \lambda.E \mid E+E \mid \mathsf{0} \mid X \mid D \\
D &::= [X_1\{ann_1\}[rew_1] := E_1; \cdots ; X_n\{ann_n\}[rew_n] := E_n]_i \\
ann &::= a_1, \cdots , a_n \\
rew &::= Y_1 : r_1;\ \cdots ;\ Y_n : r_n
\end{aligned}$$

The semantics of the language is such that D and E statements are used to express the underlying Markov chain, while *rew* statements express the reward variables and the corresponding reward rates for each state. The semantics was formalised in Definition 2 of [67] as a Labelled Continuous Time Markov Reward Chain (abbreviated $\mathsf{CTMC_{LSR}}$).

The E and D expressions can be expressed in a fully explicit form such that

$$X_i\{ann_i\}[rew_i] := \sum_j \lambda_{ij}.X_j.$$

These statements define the states and the generator matrix $\boldsymbol{Q}$ of a continuous time Markov chain such that $Q_{ij} = \lambda_{ij}$ for $i \neq j$. The diagonal elements Q_{ii} ensure that the row sums of $\boldsymbol{Q}$ are zero. The rewards $[rew_i]$ encompass the rows of a matrix $\boldsymbol{R}$ of reward rates while the annotations $\{ann_i\}$ are state labels not usually used in performance evaluation but central to the computer science applications. These state labels contribute relevant model information, examples could involve which part of a computer system would be active in different states like printing or file transfers.

An important aspect of model checking is the possibility to ask questions about the probabilistic behaviour of the model in a stringent and well defined way using a logical language. These inquiries typically relate to the annotations. The logical language used for formulating queries to be model checked was defined in Definition 4 of [67]. The first two lines of the definition are quite standard in stochastic logics while the two last, as repeated here,

$$\begin{array}{lcl}
\Upsilon & ::= & c \;\mid\; \Upsilon_1 \divideontimes \Upsilon_2 \;\mid \Upsilon^{1/h} \;\mid\; \mathsf{P}[\phi] \;\mid\; \mathsf{S}[\Phi] \\
 & \mid & \mathsf{R}_Y[\mathsf{C}^{\leq t}] \;\mid\; \mathsf{R}_Y[\mathsf{F}\,\Phi] \;\mid\; \mathsf{R}_f[\mathsf{I}^{=t}] \;\mid\; \mathsf{R}_Y[\mathsf{S}] \\
 & \mid & \mathsf{E}_f[\phi] \\
f & ::= & Y \mid c \mid f + f \mid f * f
\end{array}$$

contain some additional expressiveness. Particularly the Υ and f statements support the possibility of constructing algebraic expressions of the reward variables Y_j. The following natural lemma, Lemma 2 of [67], provides the link between the $\mathsf{CSL}_{\mathsf{MSR}}$ language for logical enquiries and the model checking.

Lemma 19 (Lemma 2 of [67]) *Any moment expression f can be written in the following normal form*

$$f := \Sigma_i \; c_i \left(\Pi_{j=1}^m Y_j^{h_{ij}} \right)$$

where the powers h_{ij} indicate the order of the moments of the random variables.

Thus evaluating algebraic expressions of moments reduces to the evaluation of expressions of the form

$$\mathbb{E} \left(\Pi_{j=1}^m Y_j^{h_{ij}} \right).$$

This evaluation of higher order moments and cross moments was performed using Theorem 32 on Page 69 which we will discuss in more detail in Section 7.1.

6 Extensions of classical results

As mentioned previously, one of the main attractions in modelling arrival processes and service processes with Markovian arrival processes and phase-type distributions is the analytical and particularly the numerical tractability of the resulting queueing models. This is most notable for queues of the GI/M/1 type, M/G/1 type, and Quasi-Birth-and-Death (QBD) processes as termed by Neuts [63, 64]. These two books by Marcel Neuts, the book by Latouche and Ramaswami [55], and the book by Bini, Latouche and Meini [20] treat the theory accessibly and comprehensively. The Markov chains describing the queues are two dimensional with an integer component $L \in \mathbb{N}$ called the *level* and a component $J \in \mathbb{J}$ called the *phase*, where $\mathbb{J}$ is some general space. In the classical framework $\mathbb{J}$ is finite. One would expect that the matrix-analytic results for the QBD process would still be valid in cases where the arrival process of the queue is a rational arrival process rather than a Markovian arrival process, and the service time distribution is matrix-exponential rather than phase-type. This was stipulated in the conclusions of [10] as well as of [11]. In [17] we showed that this is indeed true and provided an alternative proof in [18].

We will give a brief summary of the main results of QBD theory that are needed to put the contributions of [17] and [18] in perspective. This will be done in Section 6.1. The approach taken in [17] was to properly adjust the most recent arguments from [55, 74], to prove the validity of the matrix geometric solution in the general setting of a Quasi-birth-and-death process with rational arrival process (RAP) components. This is described in Section 6.2. The pathwise arguments used in the traditional analysis no longer apply. Rather, one needs to consider a process with less information, where only level changes are observed. The information on phases carried over at level changes can be interpreted as posterior probabilities of phases. These probabilities, in turn, can be viewed as weights of measures. It turns out that with this modification one can still apply the part of the argument of [74] related to last entrance times. In the general setting of a QBD with RAP components the entries of the weight vector can be negative, and the interpretation of the entries is then only that of a weight of a measure. This approach was used in [17] to prove the validity of the matrix geometric solution for models with RAP components. In Section 6.3 we will briefly sketch the setting of a structured Markov chain of the GI/M/1 type with a general phase space for a discrete time Markov chain as described in [80]. This theory is then applied in Section 6.4 to a Markov chain with a general state space where the transition kernels can be expressed using orthonormal functions. Finally, in Section 6.5 we provide an alternative to the proof presented in Section 6.2 by considering the embedded Markov chain at level transitions. This is a Markov chain of the type described in Section 6.3 and the theory of [80] can be applied. It is not straightforward to find the measure of the phase so the method is modified to deal with certain operators of the measure. For queues with RAP components we need the expected value of the phase at level changes. This approach was introduced in [18]. The formulation using operators on measures actually includes the approach of Section 6.4 such that the results of Sections 6.4 and 6.5 can be obtained in a unified framework.

6.1 Quasi-Birth-and-Death Processes

A Markov chain in continuous (or discrete time) is called a Quasi-Birth-and-Death (QBD) process if its infinitesimal generator $\boldsymbol{Q}$ (transition probability) matrix can be structured as

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{B}_1 & \boldsymbol{B}_0 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \boldsymbol{B}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \boldsymbol{0} & \boldsymbol{0} & \dots \\ \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \boldsymbol{0} & \dots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots\vdots\vdots & \end{pmatrix}. \tag{6.1}$$

Here the dimension of $\boldsymbol{B}_1$ is $p_0 \times p_0$ and the dimension of $\boldsymbol{A}_1$ is $p \times p$. States with index in $1, \dots, p_0$ are said to belong to level 0, while states with index in $[p_0 + (\ell - 1)p + 1, p_0 + \ell p]$ where ℓ is a positive integer are said to belong to level ℓ. The most prominent example is that of a $MAP/PH/1$ queue, that is a queue with a Markovian arrival process as input and phase-type distributed service times. The key result is that the steady state vector $\boldsymbol{\pi}$ partitioned according to levels can be expressed by

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_1 \boldsymbol{R}^{i-1}, \tag{6.2}$$

where the matrix $\boldsymbol{R}$ is the minimal non-negative solution to the equation

$$\boldsymbol{R}^2 \boldsymbol{A}_0 + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{A}_2 = \boldsymbol{0}$$

(in the continuous case). The matrix $\boldsymbol{R}$ in turn is connected to another matrix $\boldsymbol{G}$ that solves the dual equation

$$\boldsymbol{A}_0 + \boldsymbol{A}_1\boldsymbol{G} + \boldsymbol{A}_2\boldsymbol{G}^2 = \boldsymbol{0}$$

via $\boldsymbol{R} = \boldsymbol{A}_0(-\boldsymbol{A}_1 - \boldsymbol{A}_0\boldsymbol{G})^{-1}$ (in the continuous case). Both $\boldsymbol{R}$ and $\boldsymbol{G}$ have probabilistic interpretations. The vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ are found from boundary conditions and normalisation. It is customary to calculate $\boldsymbol{G}$ rather than $\boldsymbol{R}$ using the logarithmic reduction algorithm described in [56] or some variant thereof. This approach provides an accuracy check as the row sums of $\boldsymbol{G}$ are one whenever the QBD process is stable. The traditional proofs of these results are probabilistic in nature using the interpretation of the matrices $\boldsymbol{A}_i$ as intensity or probability transition matrices.

6.2 Last entrance time approach

The starting point of the analysis is to interpret a QBD process as a random walk on levels. The sojourn times of successive visits to the different levels follow phase-type distributions with generator matrix $\boldsymbol{A}_1$ ($\boldsymbol{B}_1$ in the case of level 0) and initial probability vector given as the unit vector with 1 in the position corresponding to the entering state. Suppose that it is only possible to observe level changes, and not the actual phase entered. Then the sojourn time in each level would be phase-type distributed with the same generator with an initial probability vector that would depend on the value of the initial vector when the previous level was entered, the sojourn time in that level, and whether the transition was to a level below or above. The QBD process with rational arrival process components generalises this idea. The QBD process with rational arrival process (RAP) components as defined and analysed in [17] is constructed by allowing for matrix-exponential sojourn times in levels. Given an initial weight vector of the matrix-exponential distribution for a visit to a level, the initial vector of the subsequent level visit is a function of the initial weight vector, the random sojourn time spent in the level, and whether the next level to be visited is one above or one below the level being left. The set of values that can be attained by the weight vector $\boldsymbol{A}(t)$ is denoted by $\mathbb{A}$, which is a convex compact subset of $\mathbb{R}^p$ for some $p \in \mathbb{N}$.

We consider the Markov process $X(t) = (L(t), \boldsymbol{A}(t))$, where $L(t)$ is the level taking values in $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$ with $\mathbb{N}$ being the set of natural numbers, and $\boldsymbol{A}(t) \in \mathbb{A}$. A key idea in the standard analysis of the QBD process is to consider certain taboo probabilities of the process. This idea can be applied with equal power for the QBD process with RAP components. Rather than considering the taboo probabilities of the Markov chain we consider the taboo of a related yet different process $\{\boldsymbol{B}(t)\}$. The process $\{\boldsymbol{B}(t)\}_{t\geq 0}$, taking values in $\mathbb{A}$, is the phase vector of weights of the censored process consisting of level m only, measured in the local time of level m, and with level $m-1$ taboo. The time spent in level m before return to level $m-1$ is phase-type distributed in the standard QBD case. This generalises to the time spent being matrix-exponentially distributed in the extended case and can be expressed as the lifetime of the $\{\boldsymbol{B}(t)\}$ process. This was stated as Theorem 9 of [17].

Theorem 20 (Theorem 9 of [17]) *The total lifetime $\ell_m(\infty)$ of $\{\boldsymbol{B}(t)\}_{t\geq 0}$ is ME distributed, that is*

$$P(\ell_m(\infty) > t|\boldsymbol{B}(0) = \boldsymbol{a}) = \boldsymbol{a}e^{\boldsymbol{U}t}\boldsymbol{1},$$

for some matrix $\boldsymbol{U}$.

The matrix $\boldsymbol{G}$ has an element-wise probabilistic interpretation in the standard QBD setting. Here the (i,j)th element gives the probability that the first return to level $m-1$ from state i in level m happens in state j. The probabilistic meaning of the matrix $\boldsymbol{G}$ in the QBD process with RAP components is as an operator on the row vector of weights. This interpretation was stated as Theorem 11 in [17].

Theorem 21 (Theorem 11 in [17]) *Let τ_n be the first passage time to level n. For all $\boldsymbol{a} \in \mathbb{A}$, we have*

$$\Psi(\boldsymbol{a}) = \mathbb{E}\left[\boldsymbol{A}(\tau_{n-1})I(\tau_{n-1} < \infty)|X(0) = (n, \boldsymbol{a})\right] = \boldsymbol{a}\boldsymbol{G},$$

for a unique matrix $\boldsymbol{G}$. Further, $\boldsymbol{a}\boldsymbol{G} \in \mathbb{A}$, for all $\boldsymbol{a} \in \mathbb{A}$.

In the classical theory, the probabilistic arguments relate the matrices $\boldsymbol{U}$ and $\boldsymbol{G}$ intrinsically as $\boldsymbol{U} = \boldsymbol{A}_1 + \boldsymbol{A}_0\boldsymbol{G}$ and $\boldsymbol{G} = (-\boldsymbol{U})^{-1}\boldsymbol{A}_2$. The validity of these expressions was established as Corollary 13 and Lemma 14 of [17]. Finally, the matrix $\boldsymbol{R}$ also has to be interpreted as an operator working on (mean) weights of measures. This was stated as Theorem 17 of [17].

Theorem 22 (Theorem 17 of [17]) *Assume that $X(\cdot)$ is an ergodic Markov process.*

1. *Let the vectors $\boldsymbol{\pi}_n, n \geq 0$, denote $\lim_{t\to\infty} \mathbb{E}\left[\boldsymbol{A}(t)I(L(t) = n)|X(0) = (j, \boldsymbol{a})\right]$, then*

   $$\boldsymbol{\pi}_{n+1} = \boldsymbol{\pi}_n\boldsymbol{R} \qquad \text{for all } n \geq 1,$$

   with

   $$\boldsymbol{R} = \boldsymbol{A}_0(-\boldsymbol{U})^{-1}.$$

2. *The vectors* $\boldsymbol{\pi}_0$ *and* $\boldsymbol{\pi}_1$ *satisfy*

$$\boldsymbol{\pi}_1\left(\boldsymbol{A}_1+\boldsymbol{B}_2(-\boldsymbol{B}_1)^{-1}\boldsymbol{B}_0+\boldsymbol{R}\boldsymbol{A}_2\right)=0, \qquad \boldsymbol{\pi}_0=\boldsymbol{\pi}_1\boldsymbol{B}_2(-\boldsymbol{B}_1)^{-1},$$

subject to

$$\boldsymbol{\pi}_1\left(\boldsymbol{B}_2(-\boldsymbol{B}_1)^{-1}\mathbf{1}+(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e}\right)=1.$$

6.3 Tweedie's operator geometric results

Equation (6.2) holds for a generalisation of (6.1), which has has been termed $GI/M/1$-type by Neuts [63]. When the phase space is some general set the transition probability matrix of a discrete time Markov chain is replaced by a transition kernel. If such a kernel has a $GI/M/1$-type structure then there exists a stationary measure with an operator geometric form similar to the classical matrix geometric result (6.2). In this section we describe this setting of a discrete time Markov chain on a general state space where the kernel has a $GI/M/1$-type structure. The framework is a Markov chain on the state space $\mathbb{N}_0\times\mathbb{J}$, where $\mathbb{N}_0=\{0\}\cup\mathbb{N}$ with $\mathbb{N}$ being the set of natural numbers, and $\mathbb{J}$ is a general measurable space equipped with a sigma-algebra $\mathcal{J}$. As in Section 6.2, the two-dimensional Markov chain X_n has an integer component L_n called the level and a component $\boldsymbol{J}$ with values in $\mathbb{J}$ called the phase. The $GI/M/1$-type kernels of the Markov chain $X_n=(L_n,\boldsymbol{J}_n)$ considered by Tweedie [80] are expressed by

$$\hat{P}(x,J)=\begin{bmatrix} \hat{B}_1(x,J) & \hat{A}_0(x,J) & 0 & 0 & \dots \\ \hat{B}_2(x,J) & \hat{A}_1(x,J) & \hat{A}_0(x,J) & 0 & \dots \\ \hat{B}_3(x,J) & \hat{A}_2(x,J) & \hat{A}_1(x,J) & \hat{A}_0(x,J)\dots & \\ \vdots & \vdots & \vdots & & \end{bmatrix}, \tag{6.3}$$

where

$$\begin{aligned} \hat{A}_i(x,J) &= P\left(L_n=L_{n-1}+1-i,\boldsymbol{J}_n\in J|\boldsymbol{J}_{n-1}=x\right), \\ \hat{B}_i(x,J) &= P\left(L_n=0,\boldsymbol{J}_n\in J|L_{n-1}=i-1,\boldsymbol{J}_{n-1}=x\right). \end{aligned}$$

The invariant measure of X_n is of the form $\boldsymbol{\nu}(\cdot)=(\nu_0(\cdot),\nu_1(\cdot),\ldots)$ ([80, Theorem 2]) with

$$\nu_{i+1}(J)=\int_{\mathbb{J}}\nu_i(\mathrm{d}x)\hat{S}(x,J), \tag{6.4}$$

where the operator $\hat{S}(x,J)$ is the minimal non-negative solution to

$$\hat{S}(x,J)=\sum_{j=0}^{\infty}\int_{\mathbb{J}}\hat{S}^j(x,\mathrm{d}y)\hat{A}_j(y,J), \tag{6.5}$$

and the operator $\hat{S}^j(x,J)$ is the jth iterate

$$\hat{S}^j(x,J) = \int_{\mathbb{J}} \hat{S}^{j-1}(x,\mathrm{d}y)\hat{S}(y,J).$$

Equation (6.4) generalises Equation (6.2). In the Markov chain setting of Neuts the operator equation (6.5) reduces to a matrix equation

$$\sum_{i=0}^{\infty} \boldsymbol{R}^i \boldsymbol{A}_i = \boldsymbol{R}. \tag{6.6}$$

The stationary measure $\nu_0(\cdot)$ at level zero, subject to normalisation, can be found from

$$\nu_0(J) = \sum_{j=1}^{\infty} \int_{\mathbb{J}} \int_{\mathbb{J}} \nu_0(\mathrm{d}x)\hat{S}^{j-1}(x,\mathrm{d}y)\hat{B}_j(y,J), \tag{6.7}$$

where, from Proposition 1 of [80], $\sum_{j=1}^{\infty} \int_{\mathbb{J}} \hat{S}^{j-1}(x,\mathrm{d}y)\hat{B}_j(y,\mathbb{J}) = 1$ for all $x \in \mathbb{J}$.

Cases with more complex boundary behaviour are modelled by replacing the kernel $\hat{A}_0$ in the row block corresponding to level 0 with $\hat{B}_0$. It is straightforward that (6.4) is still valid for $i \geq 1$ while (6.7) needs minor adjustments and an additional equation is needed.

$$\begin{aligned}
\nu_0(J) &= \int_{\mathbb{J}} \nu_0(\mathrm{d}x)\hat{B}_1(x,J) \\
&\quad + \sum_{j=1}^{\infty} \int_{\mathbb{J}} \int_{\mathbb{J}} \nu_1(\mathrm{d}x)\hat{S}^{j-1}(x,\mathrm{d}y)\hat{B}_{j+1}(y,J), && (6.8) \\
\nu_1(J) &= \int_{\mathbb{J}} \nu_0(\mathrm{d}x)\hat{B}_0(x,J) \\
&\quad + \sum_{j=1}^{\infty} \int_{\mathbb{J}} \int_{\mathbb{J}} \nu_1(\mathrm{d}x)\hat{S}^{j-1}(x,\mathrm{d}y)\hat{A}_j(y,J). && (6.9)
\end{aligned}$$

In Sections 6.4 and 6.5 we will consider two applications of this theory.

6.4 Kernels with orthonormal bases

In [68] we demonstrated the analytical simplifications that occur whenever the kernels $\hat{A}_i$ and $\hat{B}_i$ are in a function space $\mathcal{S}$ with an orthonormal base. One standard example of such a space is $L^2([0,1] \times [0,1])$.

We assume that we have a set of basis functions $\psi_i(x),\ i = 0, 1, 2, ...,$ satisfying

$$\int_0^1 \psi_i(x)\psi_j(x)dx = \delta(i = j),$$

where $\delta(i = j)$ is 1 when $i = j$ and 0 when $i \neq j$, such that $\mathcal{S}$ consists of all kernels $\hat{\alpha}(x, y)$ which have a density $\hat{a}(x, y)$ that can be expressed in the form

$$\hat{a}(x, y) \ = \ \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} A_{ij}\psi_i(x)\psi_j(y),$$

where the term on the right is absolutely convergent for all x, y, and the matrix $\boldsymbol{A} = [A_{ij}]$ defines the kernel $\hat{a}$.

In [68] the analysis was done for transition kernels (6.3) with QBD structure, simplifying the construction. We have the following result for the computation of the stationary densities of a QBD structure,

Theorem 23 (Theorem 3.2 in [68]) *Suppose the kernels a_i are in the space $\mathcal{S}$ with corresponding matrix representation $\boldsymbol{A}_i$, i=0,1,2. Then the density σ of the kernel $\hat{S}$ belongs to $\mathcal{S}$ and has matrix representation $\boldsymbol{S}$, where*

$$\boldsymbol{S} \ = \ \sum_{k=0}^{\infty} \boldsymbol{A}_{0,k} \ \Pi_{j=k-1}^{0} \boldsymbol{A}_{2,j},$$

with

$$\begin{aligned}
\boldsymbol{A}_{0,0} &= \boldsymbol{A}_0(\boldsymbol{I} - \boldsymbol{A}_1)^{-1} \\
\boldsymbol{A}_{2,0} &= \boldsymbol{A}_2(\boldsymbol{I} - \boldsymbol{A}_1)^{-1} \\
\boldsymbol{A}_{i,k+1} &= [\boldsymbol{A}_{i,k}]^2(\boldsymbol{I} - \boldsymbol{A}_{0,k}\boldsymbol{A}_{2,k} - \boldsymbol{A}_{2,k}\boldsymbol{A}_{0,k})^{-1}.
\end{aligned}$$

The $\boldsymbol{I}$ in the above equations denotes the identity matrix, and the inverse is the matrix inverse defined by $(\boldsymbol{I} - \boldsymbol{A})^{-1} \ = \ \sum_{k=0}^{\infty} \boldsymbol{A}^k$. Also, an empty matrix product is defined to be the identity matrix.

Finally the validity of Equation (6.2) in the setting of a QBD transition kernel with orthonormal bases was stated as Theorem 3.3 in [68].

Theorem 24 (Theorem 3.3 in [68]) *Assume that the conditions of Theorem 23 holds. Then the stationary distributions ν_n have densities f_n with representation $f_n(y) \ = \ \sum_i F_{ni}\psi_i(y)$, where the vector F_n of coefficients F_{ni} is such that*

$$F_n \ = \ F_0 \ S^n$$

and F_0 is an invariant vector of the matrix $B + SA_2$, where B is the matrix associated with the kernel β.

Two applications were given in [68], one small analytical example using Legendre polynomials and one with a trigonometric basis which was truncated for numerical evaluations.

6.5 Embedded Markov chain approach to the GI/RAP/1 queue

The definition of a RAP, based on finite dimensionality, makes it tempting to assume that the method of Tweedie should be applicable for the QBD process with RAP components similarly to the case with orthonormal basis functions. Because of the finite dimensionality of the RAP one would even expect the resulting matrix equations to involve only matrices of finite dimensions leading to the matrix equations like (6.6) being of finite dimension. In this section we demonstrate that this is indeed a viable approach for the generalisation of the classical matrix analytic results and consequently offers an alternative approach to that of Section 6.2. By considering the QBD process with RAP components embedded at level changes we obtain a discrete time Markov chain on a general state space with a structure that makes the operator geometric results in Section 6.3 applicable. As the results of Section 6.2 can already be shown to hold in the more general setting of queues with GI/M/1 and M/G/1 structures, we will first derive results for the GI/RAP/1 queue, and later specialise them to the QBD process with RAP components. One important modification to the framework of Section 6.3 is needed, as Tweedie's results do not apply directly in these two settings. The vectors $\boldsymbol{\pi}_i$ are the expected value of $\boldsymbol{A}(t)\boldsymbol{\delta}\left(L(t)=i\right)$ under the stationary measure $\boldsymbol{\nu}$ and not the measure itself. The measure could have a somewhat more complicated structure. Thus to generalise the matrix-analytic method of Neuts it suffices to find the expectation of $\boldsymbol{A}(t)$ under the stationary measure rather than the measure itself. The concept of operator linearity was introduced in [18] to handle this.

Operator linearity

This paragraph is a technical primer to establish the setting for the results to follow. It is taken more or less verbatim from [18].

We consider a set $\mathbb{J}$ equipped with a σ-algebra $\mathcal{J}$ and denote the set of finite signed measures on $(\mathbb{J},\mathcal{J})$ by $\mathbb{M}$. Of particular importance is the subset $\mathbb{M}_p$ of $\mathbb{M}$ of measures with total variation at most 1. Next we define the set of operators (kernels) that take an element φ of $\mathbb{M}_p$ to $\mathbb{M}_p$ and denote that set by $\mathbb{P}$, such that $\boldsymbol{\Pi} \in \mathbb{P} : \mathbb{M}_p \to \mathbb{M}_p$. The operator $\boldsymbol{\Pi}$ is defined through its kernel $\hat{\Pi}(x, J)$, $\boldsymbol{\Pi}(\varphi)(J) = \int_{\mathbb{J}} \varphi(\mathrm{d}x)\hat{\Pi}(x, J)$, where $x \in \mathbb{J}$ and $J \in \mathcal{J}$, $\hat{\Pi}(x, J)$ is such that $\hat{\Pi}(x, \cdot)$ is a measure for each $x \in \mathbb{J}$, and $\hat{\Pi}(\cdot, J)$ is measurable in $\mathbb{J}$ for fixed J.

We then define the set $\mathbb{G}$ of linear continuous operators on $\mathbb{M}^\star \subset \mathbb{M}$ taking values in some real or complex, normed vector space $\mathbb{V}$ with a countable basis, such that $\Gamma \in \mathbb{G} : \mathbb{M}^\star \to \mathbb{V}$. Of course Γ might be defined and linear for all $\varphi \in \mathbb{M}$, in which case we can take $\mathbb{M}^\star = \mathbb{M}$. Thus an operator Γ is a descriptor that extracts some characteristic from a measure $\mu \in \mathbb{M}^\star$. We will take special interest in the restriction of Γ to $\mathbb{M}_p^\star = \mathbb{M}_p \cap \mathbb{M}^\star$ of measures of total variation at most one. The definition of operator linearity was given as Definition 1 of [18].

Definition 25 (Definition 1 of [18]) *An element $\boldsymbol{\Pi} \in \mathbb{P}$ is said to be Γ-linear with respect to $\mathbb{M}_p^\star \subset \mathbb{M}_p$ if $\boldsymbol{\Pi} : \mathbb{M}_p^\star \to \mathbb{M}_p^\star$ and if $\Gamma(\boldsymbol{\Pi}(\varphi)) = \Gamma(\varphi)\boldsymbol{P}$, for all $\varphi \in \mathbb{M}_p^\star$, for a unique matrix $\boldsymbol{P}$. Whenever $\mathbb{M}_p^\star = \mathbb{M}_p$ we simply say that $\boldsymbol{\Pi}$ is Γ-linear.*

The concept of operator linearity was then applied to the Markov chains of GI/M/1 type assuming operator linearity of all kernels involved.

For the GI/M/1 queue, the matrix sequence $\boldsymbol{R}_{i+1} = \sum_{k=0}^{\infty} \boldsymbol{R}_i^k \boldsymbol{A}_k$ can be shown to converge to the minimal non-negative $\boldsymbol{R}$ that solves Equation (6.6) ([63]). In [80] this generalises to the operator sequence $\hat{S}_{i+1}(x, J) = \sum_{k=0}^{\infty} \int_{\mathbb{J}} \hat{S}_i^k(x, \mathrm{d}y)\hat{A}_k(y, J)$ converging to $\hat{S}$. In [18] we showed that all terms in the sequence $\hat{S}_i$ are operator linear provided that all the $\hat{A}_k$ kernels of the queue are operator linear. The result was stated as Lemma 4 of [18].

Lemma 26 (Lemma 4 of [18]) *If for all $k \geq 0$ the $\hat{A}_k$ are Γ-linear with respect to $\mathbb{M}_p^\star$ with matrix $\boldsymbol{A}_k$, then all elements of the sequence $\hat{S}_i$ are Γ-linear with respect to $\mathbb{M}_p^\star$. The matrices $\boldsymbol{S}_i$ corresponding to $\hat{S}_i$ are given by the (equivalent) matrix sequence*

$$\boldsymbol{S}_0 = \boldsymbol{0}, \qquad \boldsymbol{S}_{i+1} = \sum_{k=0}^{\infty} \boldsymbol{S}_i^k \boldsymbol{A}_k, \quad i \geq 0.$$

Finally we state the main result (Corollary 5 of [18]) converting the operator-geometric result of Tweedie [80] into a matrix-geometric expression under the operation of Γ as $\boldsymbol{S}_i \to \boldsymbol{S}$. The concept of operator linearity changes the operator Equation (6.5) into a matrix equation as stated in Theorem 5 of [18].

Theorem 27 (Theorem 5 of [18]) *If for all $k \geq 0$ the $\hat{A}_k$ are Γ-linear with respect to $\mathbb{M}_p^\star$ with matrix $\boldsymbol{A}_k$, and X_n is positive recurrent, then the operator $\hat{S}$ is Γ-linear with respect to $\mathbb{M}_p^\star$ with matrix $\boldsymbol{S}$ which is a solution to*

$$\boldsymbol{S} = \sum_{k=0}^{\infty} \boldsymbol{S}^k \boldsymbol{A}_k.$$

Having established the matrix associated with the operator linear kernel $\hat{S}$, Corollary 5 of [18] established how to calculate the operator of the stationary measure without needing to calculate the measure itself.

Corollary 28 (Corollary 5 of [18]) *Assume that for all i $\hat{B}_i$ and $\hat{A}_i$ are Γ-linear. Let $\boldsymbol{\nu} = (\nu_0, \nu_2, \nu_2, \dots)$ be the stationary measure determined by equations (6.4), (6.8), and (6.9). Then $\nu_0 \in \mathbb{M}_{p0}^\star$, $\nu_i \in \mathbb{M}_p^\star$ for $i \geq 1$, and we have $\Gamma(\nu_{i+1}) = \Gamma(\nu_i)\boldsymbol{S}$, for $i \geq 1$, with $\Gamma_0(\nu_0)$ and $\Gamma(\nu_1)$ given by $\Gamma_0(\nu_0) = \Gamma_0(\nu_0)\boldsymbol{B}_1 + \Gamma(\nu_1)\sum_{j=1}^{\infty} S^{j-1}B_{j+1}$, and $\Gamma(\nu_1) = \Gamma_0(\nu_0)\boldsymbol{B}_0 + \Gamma(\nu_1)\sum_{j=1}^{\infty} S^{j-1}A_j$.*

The following corollary, initially stated as Corollary 8 of [18], demonstrated how to obtain numerical values efficiently.

Corollary 29 (Corollary 8 of [18]) *When $\hat{A}_i = 0$ for $i > 2$ and $\hat{B}_i = 0$ for $i > 1$ then the logarithmic reduction algorithm of Latouche and Ramaswami [56] applies verbatim to the matrices $\boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{B}_0$ and $\boldsymbol{B}_1$, associated with Γ-linearity.*

As the work described in Section 6.4 deals with the density of the measure rather than the measure itself that development could be seen as an application of operator linearity.

The GI/RAP/1 queue

We now apply the concept of operator linearity to the GI/RAP/1 queue embedded at arrival epochs. The renewal process feeding the queue is generated by the distribution $F(\cdot)$ while the service process is generated by a RAP with parameter matrices $(\boldsymbol{D}_0, \boldsymbol{D}_1)$. The first coordinate, the level, L_n of $X_n = (L_n, \boldsymbol{J}_n)$, is the number of customers in the queue at the nth arrival and $\boldsymbol{J}_n$ is the phase vector taking values in $\mathbb{A}$. The transition probability law of that Markov chain is given by

$$\hat{P}(\boldsymbol{j}, J) = \begin{pmatrix} \hat{B}_0(\boldsymbol{j}, J) & \hat{A}_0(\boldsymbol{j}, J) & 0 & 0 & \dots \\ \hat{B}_1(\boldsymbol{j}, J) & \hat{A}_1(\boldsymbol{j}, J) & \hat{A}_0(\boldsymbol{j}, J) & 0 & \dots \\ \hat{B}_2(\boldsymbol{j}, J) & \hat{A}_2(\boldsymbol{j}, J) & \hat{A}_1(\boldsymbol{j}, J) & \hat{A}_0(\boldsymbol{j}, J) & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}$$

with

$$\begin{aligned}\hat{A}_i(\boldsymbol{j}, J) &= \int_0^\infty \hat{P}_i(\boldsymbol{j}, t; J)\mathrm{d}F(t), i \geq 0\\ \hat{B}_i(\boldsymbol{j}, J) &= \int_0^\infty \hat{Q}_{i+1}(\boldsymbol{j}, t; J)\mathrm{d}F(t), i \geq 0.\end{aligned}$$

Here $\hat{P}_i(\boldsymbol{j}, t; J)$ denotes the probability that the RAP$(\boldsymbol{D}_0, \boldsymbol{D}_1)$ has had exactly i events at time t, is currently serving a customer, and the phase vector $\boldsymbol{J}$ is in the set $J \in \mathcal{J}$, given that it had the value $\boldsymbol{j} \in \mathbb{A}$ immediately after last arrival; while $\hat{Q}_i(\boldsymbol{j}, t; J)$ denotes the probability that the RAP$(\boldsymbol{D}_0, \boldsymbol{D}_1)$ has had exactly i events at time t, and at the expiry of the ith event the phase vector is in the set $J \in \mathcal{J}$ and then remains there as the queue is empty, given it started in $x \in \mathbb{A}$. The role of Γ will here be taken as the expectation operator of the phase vector of the RAP, which clearly exists and is in $\mathbb{A}$ for any measure on $\mathbb{A}$, as $\mathbb{A}$ is compact and convex.

The operators $\hat{A}_i$ and $\hat{B}_i$ are expectation-linear, which is shown by first showing that the operators $\hat{P}_i(\boldsymbol{j}, t; J)$ and $\hat{Q}_i(\boldsymbol{j}, t; J)$ are expectation linear for all i and t. We have from the definition of the RAP$(\boldsymbol{D}_0, \boldsymbol{D}_1)$ that

$$\hat{P}_0(\boldsymbol{j}, t; J) = \boldsymbol{j}e^{\boldsymbol{D}_0 t}\boldsymbol{e}\delta\left(\frac{\boldsymbol{j}e^{\boldsymbol{D}_0 t}}{\boldsymbol{j}e^{\boldsymbol{D}_0 t}\boldsymbol{e}} \in J\right),$$

where δA is 1 when $\frac{\boldsymbol{j}e^{\boldsymbol{D}_0 t}}{\boldsymbol{j}e^{\boldsymbol{D}_0 t}\boldsymbol{e}} \in J$ and 0 when $\frac{\boldsymbol{j}e^{\boldsymbol{D}_0 t}}{\boldsymbol{j}e^{\boldsymbol{D}_0 t}\boldsymbol{e}} \notin J$.

$$\begin{aligned}\hat{P}_1(\boldsymbol{j}, t; J) &= \int_0^t \boldsymbol{j}e^{\boldsymbol{D}_0 t_1}\boldsymbol{D}_1 e^{\boldsymbol{D}_0(t-t_1)}\boldsymbol{e}\boldsymbol{\delta}\left(\frac{\boldsymbol{j}e^{\boldsymbol{D}_0 t_1}\boldsymbol{D}_1 e^{\boldsymbol{D}_0(t-t_1)}}{\boldsymbol{j}e^{\boldsymbol{D}_0 t_1}\boldsymbol{D}_1 e^{\boldsymbol{D}_0(t-t_1)}\boldsymbol{e}} \in J\right)\mathrm{d}t_1,\\ \hat{Q}_1(\boldsymbol{j}, t; J) &= \int_0^t \boldsymbol{j}e^{\boldsymbol{D}_0 t_1}\boldsymbol{D}_1\boldsymbol{e}\boldsymbol{\delta}\left(\frac{\boldsymbol{j}e^{\boldsymbol{D}_0 t_1}\boldsymbol{D}_1}{\boldsymbol{j}e^{\boldsymbol{D}_0 t_1}\boldsymbol{D}_1\boldsymbol{e}} \in J\right)\mathrm{d}t_1,\end{aligned}$$

and for $i \geq 2$,

$$\begin{aligned}\hat{P}_i(\boldsymbol{j}, t; J) &= \int_0^t \int_{\mathbb{A}} \hat{P}_1(\boldsymbol{j}, t_1; \mathrm{d}\boldsymbol{y})\hat{P}_{i-1}(\boldsymbol{y}, t - t_1; J)\mathrm{d}t_1,\\ \hat{Q}_i(\boldsymbol{j}, t; J) &= \int_0^t \int_{\mathbb{A}} \hat{P}_1(\boldsymbol{j}, t_1; \mathrm{d}\boldsymbol{y})\hat{Q}_{i-1}(\boldsymbol{y}, t - t_1; J)\mathrm{d}t_1.\end{aligned}$$

The forms of $\hat{P}_i(\boldsymbol{j}, t; J)$ and $\hat{Q}_i(\boldsymbol{j}, t; J)$ lead to

Lemma 30 (Lemma 7 in [18]) *The operators $\hat{P}_i(\boldsymbol{j}, t; J)$, for all $i \geq 0$, and $\hat{Q}_i(\boldsymbol{j}, t; J)$, for all $i \geq 1$, are expectation-linear, that is*

$$\int_{\mathbb{A}} \boldsymbol{y}\hat{P}_i(\boldsymbol{j}, t; d\boldsymbol{y}) = \boldsymbol{j}\boldsymbol{P}_i(t)$$

for the set of matrices $\boldsymbol{P}_i(t)$, $i \geq 0$, given by

$$\boldsymbol{P}_i(t) = \begin{cases} e^{\boldsymbol{D}_0 t}, & i = 0, \\ \int_0^t e^{\boldsymbol{D}_0 t_1} \boldsymbol{D}_1 e^{\boldsymbol{D}_0 (t-t_1)} dt_1, & i = 1, \\ \int_0^t \boldsymbol{P}_1(t_1) \boldsymbol{P}_{i-1}(t-t_1) dt_1, & i > 1, \end{cases}$$

and

$$\int_{\mathbb{A}} \boldsymbol{y} \hat{Q}_i(\boldsymbol{j}, t; d\boldsymbol{y}) = \boldsymbol{j} \boldsymbol{Q}_i(t)$$

for the set of matrices $\boldsymbol{Q}_i(t)$, $i \geq 1$, given by

$$\boldsymbol{Q}_i(t) = \begin{cases} \int_0^t e^{\boldsymbol{D}_0 t_1} \boldsymbol{D}_1 dt_1, & i = 1, \\ \int_0^t \boldsymbol{P}_1(t_1) \boldsymbol{Q}_{i-1}(t-t_1) dt_1, & i > 1. \end{cases} \tag{6.10}$$

From this lemma the expectation linearity of $\hat{A}_i(\boldsymbol{j}, J)$ and $\hat{B}_i(\boldsymbol{j}, J)$ was immediate, which was Corollary 8 in [18].

Corollary 31 (Corollary 8 in [18]) *The operators $\hat{A}_i(\boldsymbol{j}, J)$ and $\hat{B}_i(\boldsymbol{j}, J)$, for $i \geq 0$, are expectation-linear with matrices $\boldsymbol{A}_i = \int_0^\infty \boldsymbol{P}_i(t) dF(t)$ and $\boldsymbol{B}_i = \int_0^\infty \boldsymbol{Q}_{i+1}(t) dF(t)$, respectively.*

Thus this corollary establishes that we can apply Theorem 27 and Corollary 28 to the $GI/RAP/1$ queue, effectively obtaining exactly the same non-linear matrix equation as in [63]. We can also use Lemma 26 to determine the required solution, say $\boldsymbol{R}$, to that equation.

The application of the method to the QBD process with RAP components was given in Lemmas 9 and 10 of [18], given the alternative proof that the matrix geometric formula is also valid for queues with RAP components though a slight reinterpretation of the results is necessary.

7 Multivariate distributions

Multivariate distributions arise naturally as the joint distributions of a finite number of inter arrival times in point processes. The finite dimensional distributions of MAPs and RAPs is given by Equation 2.3 with $\boldsymbol{D}_{i_1} = \boldsymbol{D}_1$. In this chapter we describe work on multivariate distributions that include the finite dimensional distributions of RAPs as a special case. Once the theory of multivariate matrix-exponential distributions is well established, its applications might well go beyond the point process and queueing theory contexts. As an example, the joint distribution of different observations of the same phenomenon will be a product of the marginal distributions whenever the individual observations are independent. However, the joint distribution will be non-trivial in cases where the observations cannot be considered independent. Many modern data sets are huge and typically inhomogeneous, each observation including several variables. To treat such data primarily two approaches dominate. One approach is using a strictly parametric approach with well defined statistical tests and properties based on the assumption of data being adequately described by a multivariate Gaussian distribution. Another approach is to use non-parametric methods, many of which are highly applicable but also in general quite heuristic in nature. We believe that the distributions, to be described in this chapter, could fill a gap between these two approaches. The multivariate matrix-exponential distributions offer a semi-parametric alternative providing for parameter reductions on a more rigid basis than that offered by the non-parametric methods but with assumptions less restrictive than those of the multivariate Gaussian distribution. In Section 7.1 we describe some previous developments on multivariate phase-type distributions. In Section 7.2 various explicit examples of multivariate distributions are discussed in the framework of the MPH* distributions described in Section 7.1. Finally, Section 7.3 presents multivariate matrix-exponential distributions in their full generality including the main characterisation result, that a distribution is multivariate matrix-exponential if and only if all univariate projections of the random vector follow univariate matrix-exponential distributions. The characterisation result holds true when considering bilateral multivariate matrix-exponential distributions.

7.1 The class of MPH* distributions

The first formulation of multivariate phase-type distributions was presented in [13]. Here the authors introduced the notion of first hitting times of several overlapping absorbing sets. An absorbing set is a set of states that, once entered, will not be left. The first entrance times to each of these sets then constitute the different component random variables of the multivariate phase-type distributed vector. Without loss of generality, the intersection of the absorbing sets can be taken as a singleton that corresponds to the absorbing state of a standard univariate phase-type distribution. This class of multivariate phase-type distributions was termed MPH. Although initially tempting, this definition places non-trivial restrictions on the sub-generator $\boldsymbol{S}$ that cannot be put in an elegant way. One advantage, however, of the MPH definition is that it is possible to give an explicit expression for the joint distribution of the components of an MPH distributed random vector due to the forward nature of the sub-generator. The expression is by no means simple and requires detailed analysis of the sub-generator along with its relation to the various absorbing sets. In [54] a slight reinterpretation of the univariate phase-type distributions opened up for a more general and more compact formulation of multivariate phase-type distributions. This class was termed MPH* distributions. Consider a standard phase-type sub-generator $\boldsymbol{S}$. The time to absorption X_a is then the sum of the cumulated sojourn times Z_i in each of the individual transient states of the Markov chain so that $X_a = \sum_{i=1}^p Z_i$. Now suppose that rather than just summing these cumulated sojourn times each of the cumulated sojourn times is multiplied with a constant r_i and then added to get the random variable $X = \sum_{i=1}^p r_i Z_i$. The random variable X would then again be phase-type distributed, which is easy to see if $r_i > 0$ for all i and requires some work to see if r_i is allowed to be 0 for some i [25, 54]. It is then natural to construct several variables X_j using different weighting or reward factors r_{ij} such that $X_j = \sum_{i=1}^p r_{ij} Z_i$ define a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. The acronym MPH* was introduced to define this class, which is parameterised by $\boldsymbol{\alpha}$, $\boldsymbol{S}$, and the matrix $\boldsymbol{R}$ of the reward factors r_{ij}. We write $\boldsymbol{X}$ $\sim$MPH$^*(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ when $\boldsymbol{X}$ is MPH* distributed with representation $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$. The Laplace-Stieltjes transform of an MPH$^*(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ distribution can be expressed as

$$\mathbb{E}\left(e^{-\langle \boldsymbol{X}, \boldsymbol{s}\rangle}\right) = \alpha_{p+1} + \boldsymbol{\alpha}\left((-\boldsymbol{S})^{-1}\boldsymbol{\Delta}(\boldsymbol{R}\boldsymbol{s}) + \boldsymbol{I}\right)^{-1}\boldsymbol{1}. \tag{7.1}$$

We define a multivariate distribution to belong to the class MME* with representation $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ whenever its Laplace-Stieltjes transform can be expressed as in Equation (7.1).

A recursive formula for the calculation of moments and cross-moments is given in [54]. In [25], Theorem 4.2, we gave a closed form expression.

Theorem 32 (Theorem 4.2 of [25]) *The cross–moments* $\mathbb{E}\left(\prod_{i=1}^{n} X_i^{k_i}\right)$*, where* $\boldsymbol{X}$ *follows an MME** *distribution with representation* $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$*, and where* $k_i \in \mathbb{N}$*, are given by*

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i^{k_i}\right) = \boldsymbol{\alpha} \sum_{\ell=1}^{k!} \prod_{i=1}^{k} (-\boldsymbol{S})^{-1} \boldsymbol{\Delta}(\boldsymbol{r}_{\sigma_\ell(i)}) \boldsymbol{1}.$$

Here $k = \sum_{i=1}^{n} k_i$*,* $\boldsymbol{r}_j$ *is the* j*th column of* $\boldsymbol{R}$ *and* σ_ℓ *is one of the* $k!$ *possible ordered permutations of the derivatives with respect to* s_j *in Equation (7.1), with* $\sigma_\ell(i)$ *being the value among* $1 \ldots n$ *at the* i*'th position of that permutation.*

7.2 Explicit distributions

There is a significant amount of work on multivariate exponential and gamma distributions. Here we will only focus on those that have a rational Laplace-Stieltjes transform, thus excluding gamma distributions with non-integer shape parameter. Our main reference for the various multivariate exponential and gamma distributions is [52]. We will briefly survey the many different contributions emphasising the underlying ideas and our own contributions. A comprehensive treatment will be given in a forthcoming monograph coauthored with Mogens Bladt [29]. All bivariate distributions with rational Laplace-Stieltjes transform we have encountered belong to the MPH* (MME*) class while several even belong to the MPH class. Most distributions of higher order belong to the MPH* class. However, in [25] Theorem 4.3, we gave an example of a trivariate distribution which does not have an MPH* representation of minimal order.

There are basically four generic ways that have been used to construct the different multivariate distributions. In addition, there are a number of related distributions that have attained some popularity without having exponentially or gamma distributed marginals.

Sharing of exponential phases

Erlang distributions, that is gamma distributions with integer shape parameter, can be interpreted as the distribution of the sum of independent exponential random variables. This provides a way of constructing dependent Erlang distributed random variables by letting some of the exponential random variables contribute to two or more components of a multivariate random vector with Erlang distributed marginals.

We illustrate the idea by the case of a bivariate random vector where each component has an Erlang distribution with a shape parameter of 2. The MPH$^*(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ representation of this distribution can be given as

$$\boldsymbol{\alpha} = (1, 0, 0), \qquad \boldsymbol{S} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} \mu_1^{-1} & \mu_2^{-1} \\ \mu_1^{-1} & 0 \\ 0 & \mu_2^{-1} \end{pmatrix}.$$

Here the first exponential phase contributes to both components while the second and third contribute to only one of the two components. This construction is known under the names of McKays bivariate gamma distribution, Cheriyan and Ramabhadran's bivariate gamma distribution, Prèkopa and Szàntai's multivariate gamma distribution, and Cheriyan and Ramabhadran's multivariate gamma distribution.

Exponential distributions expressed as a geometric mixture of Erlang distributions

An exponential distribution can be expressed as a geometric mixture of Erlang distributions, a result that can be used to show that the time a typical customer spends in the M/M/1 queue is exponentially distributed. The model is double stochastic. First one picks the shape parameter of the Erlang distribution as a geometrically distributed random variable N, and then an Erlang$_N$ variable is generated. A multivariate vector with exponential marginals can then be constructed by using the same shape parameter N for all components of the random vector, where each component follows an Erlang distribution with shape parameter N. In the bivariate case this can be accomplished using the MPH$^*(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ representation

$$\boldsymbol{\alpha} = (1, 0), \qquad \boldsymbol{S} = \begin{pmatrix} -1 & 1 \\ p & -1 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} \mu_1^{-1} & 0 \\ 0 & \mu_2^{-1} \end{pmatrix}.$$

The form of $\boldsymbol{\alpha}$ and $\boldsymbol{S}$ ensures that the two states of $\boldsymbol{S}$ have the same geometrically distributed number of visits, while the form of $\boldsymbol{R}$ ensures that the total sojourn time of state i determines the ith component of the bivariate vector. The construction is originally due to Kibble and has reappeared under different names such as Jensen's, Gaver's, and Downton-Moran's distribution. Erlang distributed marginals are obtained by adding random vectors with exponential marginals.

Decomposition of the exponential distribution

The distributions, described in this section, are based on the observation that an exponential random variable with intensity μ can be expressed as the sum of an exponential variable with intensity $\lambda \geq \mu$ and a term which is an indicator variable of probability $\frac{\lambda-\mu}{\lambda}$ multiplied with an exponential variable with intensity μ. If the decomposition is applied recursively to the term that is multiplied with the indicator variable one gets the geometric mixture of Erlang distributions in the limit. A simple bivariate distribution using the decomposition is given by the MPH$^*(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ representation

$$\boldsymbol{\alpha} = (1, 0), \qquad \boldsymbol{S} = \begin{pmatrix} -1 & 1-p \\ 0 & -1 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} \mu_1^{-1} & p\mu_2^{-1} \\ 0 & \mu_2^{-1} \end{pmatrix}.$$

The first component of the bivariate vector is given by the sojourn time in state 1, while the second component is constructed using decomposition with $\mu = \mu_2$ and $\lambda = \frac{\mu_2}{p}$.

Examples of the application of this construction are the Marshall-Olkin distribution, Olkin and Tong's Multivariate Exponential, Raftery's bivariate and Multivariate Exponential, and Dussauchoy and Berland's distribution. Multivariate Erlang distributions can be constructed by summing exponential random vectors as for geometric mixtures of Erlang distributions.

Farlie-Gumbel-Morgenstern distributions

The Farlie-Gumbel-Morgenstern construction [52] is a general construction that can be used to construct a bivariate distribution, where the marginals are given by their distribution functions $F_i, i = 1, 2$. The joint distribution of X_1, X_2 according to this construction is given by

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)\left(1 + \rho\left(1 - F_1(x_1)\right)\left(1 - F_2(x_2)\right)\right),$$

where $-1 \leq \rho \leq 1$.

In Lemma 4.1 of [24] we showed that the Morgenstern copula can be seen as a proper mixture of the two first order statistics.

Lemma 33 (Lemma 4.1 of [24]) *Let $F_i^{\min}(x) = 1 - (1 - F_i(x))^2$ and $F_i^{\max}(x) = F_i^2(x)$ such that $F_i^{\min}(x)$ and $F_i^{\max}(x)$ are cumulative distribution functions of the minimum respectively maximum of two independent random variables distributed according to $F_i(x)$. Then the bivariate Morgenstern distribution $F(x_1, x_2)$ based on $F_1(x_1)$ and $F_2(x_2)$ is*

$$\begin{aligned} F(x_1, x_2) \quad = \quad & \frac{1+\rho}{4} F_1^{\max}(x_1) F_2^{\max}(x_2) + \frac{1-\rho}{4} F_1^{\max}(x_1) F_2^{\min}(x_2) + \\ & \frac{1-\rho}{4} F_1^{\min}(x_1) F_2^{\max}(x_2) + \frac{1+\rho}{4} F_1^{\min}(x_1) F_2^{\min}(x_2). \end{aligned}$$

The probabilistic interpretation of Lemma 33 lead us to the MME* representations of bivariate distributions with matrix-exponential or phase-type distributed marginals of Morgenstern type as presented in Theorem 4.1 of [24].

Theorem 34 (Theorem 4.1 of [24]) *The bivariate Farlie-Gumbel-Morgenstern distribution formed from two matrix-exponential distributions with marginal representation of F_i given by $(\boldsymbol{\alpha}_i, \boldsymbol{S}_i, -\boldsymbol{S}_i\mathbf{1}), i = 1, 2$ is in MME*. An MME* representation $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ is*

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1 \otimes \boldsymbol{\alpha}_1, \mathbf{0}, \mathbf{0}, \mathbf{0})$$

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_1 \oplus \boldsymbol{S}_1 & \frac{1}{2}\left(\boldsymbol{s}_1 \oplus \boldsymbol{s}_1\right) & \frac{1-\rho}{4}\left(\boldsymbol{s}_1 \oplus \boldsymbol{s}_1\right)\mathbf{1}\tilde{\boldsymbol{\alpha}}_2^{(M,m)} & \frac{1+\rho}{4}\left(\boldsymbol{s}_1 \oplus \boldsymbol{s}_1\right)\mathbf{1}\tilde{\boldsymbol{\alpha}}_2^{(m)} \\ \mathbf{0} & \boldsymbol{S}_1 & \frac{1+\rho}{2}\boldsymbol{s}_1\tilde{\boldsymbol{\alpha}}_2^{(M,m)} & \frac{1-\rho}{2}\boldsymbol{s}_1\tilde{\boldsymbol{\alpha}}_2^{(m)} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Delta}_1^{-1}\boldsymbol{S}_2^T\boldsymbol{\Delta}_1 & \boldsymbol{\Delta}_1^{-1}(\boldsymbol{s}_2 \oplus \boldsymbol{s}_2)^T\boldsymbol{\Delta}_2 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\boldsymbol{S}}_2^{(m)} \end{bmatrix},$$

$$\boldsymbol{R} = \begin{bmatrix} \mathbf{1}_{m_1} \otimes \mathbf{1}_{m_1} & 0 \\ \mathbf{1}_{m_1} & 0 \\ 0 & \mathbf{1}_{m_2} \otimes \mathbf{1}_{m_2} \\ 0 & \mathbf{1}_{m_2} \end{bmatrix}$$

with

$$\boldsymbol{\pi}_2 = \boldsymbol{\alpha}_2\left(-\boldsymbol{S}_2\right)^{-1}/\boldsymbol{\alpha}_2\left(-\boldsymbol{S}_2\right)^{-1}\boldsymbol{e} \quad , \qquad \tilde{\boldsymbol{\alpha}}_2 = \boldsymbol{\pi}_2 \bullet \boldsymbol{s}_2/\boldsymbol{\pi}_2\boldsymbol{s}_2 \quad ,$$

$$\boldsymbol{S}_2^{(M)} = \begin{bmatrix} \boldsymbol{S}_2 \oplus \boldsymbol{S}_2 & \boldsymbol{s}_2 \oplus \boldsymbol{s}_2 \\ \mathbf{0} & \boldsymbol{S}_2 \end{bmatrix},$$

$$\zeta_2^{(m)} = \left(\boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_2\right)\left(\boldsymbol{S}_2 \oplus \boldsymbol{S}_2\right)^{-1}\mathbf{1}, \qquad \zeta_2^{(M)} = \left(\boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_2, \mathbf{0}\right)\left(-\boldsymbol{S}_2^{(M)}\right)^{-1}\mathbf{1},$$

$$\boldsymbol{\pi}_2^{(M)} = \left(\zeta_2^{(M)}\right)^{-1}\left(\boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_2, \mathbf{0}\right)\left(-\boldsymbol{S}_2^{(M)}\right)^{-1} = \left(\frac{\zeta_2^{(m)}}{\zeta_2^{(M)}}\boldsymbol{\pi}_2^{(m)}, \boldsymbol{\pi}_2^{(M,m)}\right) \quad ,$$

$$\boldsymbol{\pi}_2^{(m)} = \left(\zeta_2^{(m)}\right)^{-1}\left(\boldsymbol{\alpha}_2 \otimes \boldsymbol{\alpha}_2\right)\left(-\boldsymbol{S}_2 \oplus \boldsymbol{S}_2\right)^{-1} \quad , \quad \tilde{\boldsymbol{\alpha}}_2^{(m)} = \left(\zeta_2^{(m)}\right)^{-1}\boldsymbol{\pi}_2^{(m)}\bullet(\boldsymbol{s}_2 \oplus \boldsymbol{s}_2),$$

$$\boldsymbol{\Delta}_1 = \boldsymbol{\Delta}\left(\boldsymbol{\pi}_2^{(M,m)}\right), \qquad \boldsymbol{\Delta}_2 = \boldsymbol{\Delta}\left(\boldsymbol{\pi}_2^{(m)}\right),$$

$$\tilde{\boldsymbol{\alpha}}_2^{(M)} = \left(\zeta_2^{(M)}\right)^{-1}\left(\mathbf{0}, \boldsymbol{\pi}_2^{(M,m)} \bullet \boldsymbol{s}_2\right) = \left(\mathbf{0}, \tilde{\boldsymbol{\alpha}}_2^{(M,m)}\right)$$

$$\tilde{\boldsymbol{S}}_2^{(m)} = \boldsymbol{\Delta}_2^{-1}\left(\boldsymbol{S}_2 \oplus \boldsymbol{S}_2\right)^T\boldsymbol{\Delta}_2,$$

$$\begin{aligned}\tilde{\boldsymbol{S}}_2^{(M)} &= \boldsymbol{\Delta}\left(\boldsymbol{\pi}_2^{(M)}\right)^{-1}\left[\begin{array}{cc}(\boldsymbol{S}_2 \oplus \boldsymbol{S}_2)^T & \boldsymbol{0} \\ (\boldsymbol{s}_2 \oplus \boldsymbol{s}_2)^T & \boldsymbol{S}_2^T\end{array}\right]\boldsymbol{\Delta}\left(\boldsymbol{\pi}_2^{(M)}\right) \\ &= \left[\begin{array}{cc}\tilde{\boldsymbol{S}}_2^{(m)} & \boldsymbol{0} \\ \boldsymbol{\Delta}_1^{-1}(\boldsymbol{s}_2 \oplus \boldsymbol{s}_2)^T\boldsymbol{\Delta}_2 & \boldsymbol{\Delta}_1^{-1}\boldsymbol{S}_2^T\boldsymbol{\Delta}_2\end{array}\right].\end{aligned}$$

Lemma 33 immediately leads to the idea of constructing multivariate distributions as mixtures of order statistics. That idea was used in [26] in combination with the decomposition of exponentials to express a new type of bivariate exponential distribution. We first state a couple of lemmas regarding univariate exponential distributions using decomposition.

Lemma 35 (Lemma 4.2 of [26]) *Let $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p_1})$ be any initial vector and let*

$$\boldsymbol{S}_1 = \begin{pmatrix} -\lambda_{p_1} & \lambda_{p_1} - \lambda & 0 & \ldots & 0 \\ 0 & -\lambda_{p_1-2} & \lambda_{p_1-2} - \lambda & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & 0 & -\lambda \end{pmatrix}, \tag{7.2}$$

where it is assumed that $\lambda < \lambda_i$ for all i. Then PH$(\boldsymbol{\alpha}_1, \boldsymbol{S}_1)$ is equivalent to an exponential distribution with rate λ.

The time reversed representation obtained from Equation (2.2) for $\boldsymbol{\alpha}_1 = (1, 0, \ldots, 0)$ was stated as Lemma 4.3 of [26].

Lemma 36 (Lemma 4.3 of [26]) *Let $\boldsymbol{\alpha}_2 = (\alpha_{21}, ..., \alpha_{2p_2})$ with*

$$\alpha_{2i} = \frac{\lambda}{\lambda_i}\prod_{j=1}^{p_2-i}\frac{\lambda_{p_2-j+1} - \lambda}{\lambda_{p_2-j+1}}$$

and

$$\boldsymbol{S}_2 = \begin{pmatrix} -\lambda & \lambda & 0 & \ldots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\lambda_{p_2} \end{pmatrix}, \tag{7.3}$$

where $\lambda < \lambda_i$ for all i. Then PH$(\boldsymbol{\alpha}_2, \boldsymbol{S}_2)$ represents an exponential distribution with rate λ.

The idea of [26] was to construct a bivariate exponential distribution belonging to MPH* by combining these two representations of an exponential distribution. Whenever $\lambda_i = i\lambda$ the representation with the sub-generator $\boldsymbol{S}_2$ of Equation (7.3) is a uniform mixture of the first n order statistics of the exponential distribution. As the representation with the sub-generator $\boldsymbol{S}_1$ of Equation (7.2) is the time reversed representation, one gets the ith order statistic whenever the generator is left from state i. In [26] the dimensions p_1 and p_2 of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ were equal as this streamlines the presentation, but this is not required. An MPH* representation $(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$ of the bivariate distribution is

$$\boldsymbol{\alpha} = (1, 0, \ldots, 0), \qquad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_1 & \lambda \boldsymbol{P} \\ \mathbf{0} & \boldsymbol{S}_2 \end{pmatrix}, \qquad \boldsymbol{R} = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix},$$

where $\boldsymbol{P}$ has to satisfy $\sum_{i=1}^{p_1} P_{ij} = 1/p_1$ and $\sum_{j=1}^{p_2} P_{ij} = 1/p_2$. We will now assume $p_1 = p_2 = p$. By choosing $\boldsymbol{P} = \boldsymbol{I}$ one pairs the ith order statistic of the first exponential with the $(p - i + 1)$st order statistic of the other, while choosing $\boldsymbol{P}$ as the anti-diagonal pairs the ith order statistic from both exponential distributions.

These kinds of distributions are capable of exhibiting arbitrary correlations, which was expressed as Theorem 4.4 of [26].

Theorem 37 (Theorem 4.4 of [26]) *Let $\rho \in (1 - \frac{\pi^2}{6}, 1)$. Then we can construct a two–dimensional exponential vector (X_1, X_2) with correlation coefficient ρ in the following way. If $\rho > 0$ we choose $p \in \mathbb{N}$ such that $\rho_{\max} = \mathbf{corr}(X_1, X_2) \geq \rho$ and*

$$\mathbf{P} = \frac{\rho}{\rho_{\max}} \{\delta(i - 1 = p - j)\}_{i,j} + \left(1 - \frac{\rho}{\rho_{\max}}\right) \frac{1}{p} \mathbf{1}\mathbf{1}',$$

and if $\rho < 0$ we choose $p \in \mathbb{N}$ such that $\rho_{\min} = \mathbf{corr}(X_1, X_2) \leq \rho$ and

$$\mathbf{P} = \frac{\rho}{\rho_{\min}} \{\delta(i = j)\}_{i,j} + \left(1 - \frac{\rho}{\rho_{\min}}\right) \frac{1}{p} \mathbf{1}\mathbf{1}',$$

where $\delta(i = j)$ is 1 when $i = j$ and 0 when $i \neq j$ as in Equation (6.10).

The result of Lemma 33 was also given in [16] where the generalisation to arbitrary order statistics was considered too.

Miscellaneous distributions

The four methods described up to now can be combined in different ways to obtain multivariate Erlang distributions. An explicit example is Sarmanov's bivariate gamma and, for specific parameter values, Dussauchoy and Berland's distribution.

Some distributions with rational Laplace-Stieltjes transform without exponential or Erlang distributed marginals have appeared. Freund's bivariate and multivariate distributions are probably the most frequently used of these. A minor variant of Freund's distribution has been introduced as Friday and Patil's bivariate exponential distribution, the name obscuring the fact that these distributions do not in general have exponential marginals.

Parameter selection using linear programming

The linear optimisation approach of Section 4.2 can be applied equally well to parameter selection in bivariate distributions, particularly for the selection of $\boldsymbol{P}$. In Section 4.2 maximisation or minimisation of the correlation led to a linear programming problem. Here we present a couple of other objective functions that can be optimised with linear programming.

We consider the two PH representations $(\boldsymbol{\alpha}_1, \boldsymbol{S}_1)$ and $(\boldsymbol{\alpha}_2, \boldsymbol{S}_2)$ with joint density

$$\boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t_1} \boldsymbol{V} e^{\boldsymbol{S}_2 t_2}(-\boldsymbol{S}_2)\mathbf{1}$$

where $\boldsymbol{\alpha}_1(-\boldsymbol{S}_1)^{-1}\boldsymbol{V} = \boldsymbol{\alpha}_2$ and $\boldsymbol{V}\mathbf{1} = -\boldsymbol{S}_1\mathbf{1}$. We have

$$\begin{aligned} P(\min(X,Y) > t) &= \int_t^\infty \int_t^\infty \boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t_1} \boldsymbol{V} e^{\boldsymbol{S}_2 t_2}(-\boldsymbol{S}_2)\mathbf{1} \mathrm{d}t_2 \mathrm{d}t_1 \\ &= \boldsymbol{\alpha}_1 \int_t^\infty e^{\boldsymbol{S}_1 t_1}\mathrm{d}t_1 \boldsymbol{V} \int_t^\infty e^{\boldsymbol{S}_2 t_2}\mathrm{d}t_2(-\boldsymbol{S}_2)\mathbf{1} \\ &= \boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t}(-\boldsymbol{S}_1)^{-1}\boldsymbol{V} e^{\boldsymbol{S}_2 t}\mathbf{1}. \end{aligned}$$

The expected value of that minimum is

$$\mathbb{E}(\min(X,Y)) = \int_0^\infty \boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t}(-\boldsymbol{S}_1)^{-1}\boldsymbol{V} e^{\boldsymbol{S}_2 t}\mathbf{1}\mathrm{d}t.$$

The probability that Y is greater than X is

$$\begin{aligned} P(Y > X) &= \int_0^\infty \int_{t_1}^\infty \boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t_1} \boldsymbol{V} e^{\boldsymbol{S}_2 t_2}(-\boldsymbol{S}_2)\mathbf{1}\mathrm{d}t_2\mathrm{d}t_1 = \int_0^\infty \boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t}\boldsymbol{V} e^{\boldsymbol{S}_2 t}\mathbf{1}\mathrm{d}t \\ &= \boldsymbol{\alpha}_1(\eta_1+\eta_2)^{-1}\sum_{i=0}^\infty\sum_{j=0}^\infty \binom{i+j}{i}\left(\frac{\eta_1}{\eta_1+\eta_2}\right)^i\left(\frac{\eta_2}{\eta_1+\eta_2}\right)^j \boldsymbol{K}_1^i\boldsymbol{V}\boldsymbol{K}_2^j\mathbf{1} \end{aligned}$$

where we have used uniformization as in Equation (3.1).

Correspondingly,

$$\begin{aligned} P(\max(X,Y) \le t) &= \boldsymbol{\alpha}_1 \int_0^t e^{\boldsymbol{S}_1 t_1}\mathrm{d}t_1 \boldsymbol{V} \int_0^t e^{\boldsymbol{S}_2 t_2}\mathrm{d}t_2(-\boldsymbol{S}_2)\mathbf{1} \\ &= \boldsymbol{\alpha}_1\left(\boldsymbol{I} - e^{\boldsymbol{S}_1 t}\right)(-\boldsymbol{S}_1)^{-1}\boldsymbol{V}\left(\boldsymbol{I} - e^{\boldsymbol{S}_2 t}\right))\mathbf{1} \\ &= \boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t}(-\boldsymbol{S}_1)^{-1}\boldsymbol{V} e^{\boldsymbol{S}_2 t}\mathbf{1} + 1 - \boldsymbol{\alpha}_1 e^{\boldsymbol{S}_1 t}\mathbf{1} - \boldsymbol{\alpha}_2 e^{\boldsymbol{S}_2 t}\mathbf{1} \\ &= P(X \le t) + P(Y \le t) - P(\min(X,Y) \le t), \end{aligned}$$

and

$$\mathbb{E}(\max(X,Y)) = \mathbb{E}(X) + \mathbb{E}(Y) - \mathbb{E}(\min(X,Y))$$

The objective function as well as the restrictions are linear in all of the cases above, thus choosing $\boldsymbol{V}$ by minimisation or maximisation is a linear programming problem.

7.3 Multivariate Matrix Exponential Distributions

As mentioned previously, in [25] Theorem 4.3 we demonstrated that it is not always possible to find an MPH* representation of minimal order for a multivariate distribution with a rational Laplace-Stieltjes transform. Mathematically this is somewhat unsatisfactory for the MPH* class and raised the question whether there is a more general class of distributions that should be considered as the natural generalisation of phase-type and matrix-exponential distributions. In Definition 4.1 of [25] we introduced such a class as

Definition 38 (Definition 4.1 of [25]) *A non-negative random vector $\boldsymbol{X}$ $= (X_1, ..., X_n)$ of dimension n is said to have multivariate matrix-exponential distribution if the joint Laplace-Stieltjes transform $H(\boldsymbol{s}) = \mathbb{E}\left[\exp(-\langle \boldsymbol{X}, \boldsymbol{s}\rangle)\right]$, where $\boldsymbol{s} = (s_1, \ldots, s_n)$, is a multi-dimensional rational function, that is, a fraction between two multi-dimensional polynomials. This class of distributions is denoted by MVME.*

This seems to be the most obvious and reasonable way to generalise the univariate matrix-exponential distributions into multivariate ones. This is even more appealing due to the following characterisation theorem, Theorem 4.1 also from [25].

Theorem 39 (Theorem 4.1 of [25]) [1] *A vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ follows a multivariate matrix-exponential distribution if and only if $\langle \boldsymbol{X}, \boldsymbol{a}\rangle = \sum_{i=1}^{n} a_i X_i$ has a univariate matrix-exponential distribution for all non-negative vectors $\boldsymbol{a} \neq \boldsymbol{0}$.*

Inspired by Theorem 39 we introduced a class - MVPH - of multivariate phase-type distributed random variables in Definition 4.2 of [25].

[1] Embarrassingly, Theorem 39 was not properly proven in [25], as we had been careless with the proof of Lemma 4.1, as the sets C_i were not properly defined. This was corrected in the related proof of Lemma 1 in [22].

Definition 40 (Definition 4.2 of [25]) *A vector $\boldsymbol{X} = (X_1, ..., X_n)$ has a multivariate phase–type distribution (MVPH) if $\langle \boldsymbol{X}, \boldsymbol{a} \rangle$ has a (univariate) phase–type distribution for all non–negative $\boldsymbol{a} \neq \boldsymbol{0}$.*

An alternative characterisation of the MVPH class is open at this point. It is tempting to conjecture that certain conditions on the poles of the denominator of the transform and a requirement for a positive joint density as in [70] will suffice.

Due to the nature of phase-type distributions as modelling absorption times in Markov chains they have been used for modelling phenomena that are inherently non-negative such as inter-arrival times and service times in queueing systems. The reward interpretation given by the MPH* class, however, opens the possibility for modelling phenomena that can attain values in general real spaces. This was done in the univariate case in [1] and extended to the multivariate case in [22], where the equivalent to Theorem 39 was shown also to hold in this more general case as Theorem 4 in [22]. These distributions are termed bilateral phase-type (BPH) distributions [1] and multivariate bilateral matrix-exponential (MBME* corresponding to the MME* class and MVBME corresponding to the MVME class) distributions [22]. An example of a distribution with rational moment-generating function is the Wishart function. This might prove useful when the potential for the application of these distributions is explored, as the Wishart distribution appears as the distribution for empirical covariance matrices in multivariate statistics. In [22] we also showed that multivariate distributions with rational moment generating function occur naturally when analysing certain state-dependent multivariate diffusions at absorption times in a related finite state Markov chain. In fact, the result is slightly more general. Here we state it as a theorem

Theorem 41 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_\ell)$ be an MVME distributed random vector and consider another multidimensional vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ such that*

$$X_j = \sum_{i=1}^{\ell} B_{ij}, \quad j = 1, \ldots, k$$

where $\boldsymbol{B}_i = (B_{i1}, \ldots, B_{ik}) \sim N_k(Y_i \boldsymbol{r}(i), Y_i \boldsymbol{\Sigma}(i))$, with $\boldsymbol{r}(i) = (r_1(i), \ldots, r_k(i))$ and $\boldsymbol{\Sigma}(i)$ is a covariance matrix, $i = 1, \ldots, \ell$. Then $\boldsymbol{X}$ has a rational (multidimensional) moment-generating function, i.e. $\boldsymbol{X}$ belongs to the class of Bilateral Multivariate Matrix-Exponential distributions (MVBME).

Corollary 42 (Equation (18) of [22]) *If $\boldsymbol{Y}$ of Theorem 41 is such that $\boldsymbol{Y} \sim$ MME*$(\boldsymbol{\alpha}, \boldsymbol{S}, \boldsymbol{R})$, then the moment generating $M(\boldsymbol{s})$ function of $\boldsymbol{X}$ of Theorem 41 is given by*

$$M(\boldsymbol{s}) = \boldsymbol{\alpha}\left(\boldsymbol{I} - \boldsymbol{S}^{-1}\boldsymbol{\Delta}(\boldsymbol{R}\boldsymbol{\theta})\right)^{-1}\mathbf{1}.$$

Here $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_\ell)'$ with $\theta_i = \boldsymbol{s}\boldsymbol{r}_i + \frac{1}{2}\boldsymbol{s}\boldsymbol{\Sigma}(i)\boldsymbol{s}'$.

Corollary 42 was stated as Equation (18) in [22]. The proof of the corollary in [22] also proves Theorem 41.

The question whether the class MVBME is strictly larger than the MBME* class was solved in the affirmative using corollary 42. Consider two independent Brownian motions $B_1(t)$ and $B_2(t)$ with zero drift and diffusion coefficients $\sigma_1 > 0$ and $\sigma_2 > 0$ respectively. Hence $B_i(t) \sim \mathrm{N}(0, \sigma_i^2 t)$, $i = 1, 2$. Let T be exponentially distributed with intensity $\lambda > 0$ and define $\boldsymbol{Y} = (B_1(T), B_2(T))$.

Theorem 43 (Theorem 2.1 in [28]) *The distribution of $\boldsymbol{Y}$ is a bivariate bilateral matrix–exponential distribution which cannot be written on the MBME* form.*

8 Conclusion

In this thesis we have given an overview of our contributions to the field of matrix analytic methods in queueing theory. The theory is based on the modelling blocks of Markovian arrival processes and phase-type distributions, and their analytic extensions rational arrival processes and matrix-exponential distributions.

Although the field is reasonably mature, apparently there are still theoretical developments to be made even in the more basic part of the theory as demonstrated with the results on size-biased distributions described in Section 2.1.

Among the most important parts of the contributions described in the thesis is the work on sensitivity analyses described in Chapter 4, including the work on modelling processes with excessive variability described in Section 5.1. The latter has had substantial impact in the field of performance evaluation. These contributions are directly applicable with strong engineering aspects.

We see the work described in Sections 6.2 and 6.5 on extending the classical matrix analytic results to the general setting of queues with RAP components as an important theoretical achievement. From a mathematical point of view, it has been satisfying to finally settle in the affirmative that the main classical results on queue length distributions hold in the full generality of the RAP component framework and that existing algorithms and tools can be used without modification. For now these contributions might primarily be of theoretical importance, but new developments might lead to their use in applications, for instance leading to substantial dimensionality reductions in the matrix equations involved in the solution of queueing problems. Work in this direction has already been reported by several researchers.

Finally, we believe that there is a huge application potential for MVME distributions in such diverse areas as hydrology, road traffic modelling, process algebras and their associated logics, and medicine. Section 7.2 should demonstrate the potential as it demonstrates how the theory unifies earlier contributions to these diverse application fields. The main obstacle being that the overwhelming flexibility of MVME distributions turns into a challenge when performing statistical estimation. However, we believe that the steady increase in computing power, and the increasing need for semi-parametric models less restrictive than the normal distribution will create space for a rich development of MVME theory with applications. Thus the work described in Chapter 7 has strong theoretical as well as practical implications.

There are several open interesting and important theoretical problems. The most challenging, which might not be analytically solvable, is the question whether a given set of an initial vector and a matrix specifies a (matrix-exponential) distribution. Although necessary conditions exist, so far it has not been possible to derive sufficient ones. It is likely that a possible solution of the problem would at the same time give the corresponding solution for determining whenever two matrices characterise a RAP. Among other challenges is an understanding of the full class of MVME and MVBME distributions, and to obtain efficient ways of calculating the cumulative distribution function for the classes of MBME* distributions including the different sub-classes like the important one of MPH*.

References

[1] Soohan Ahn and V. Ramaswami. Bilateral phase type distributions. *Stochastic Models*, 21:239–259, 2005.

[2] Tayfur Altiok. On the phase-type approximations of general distributions. *IIE Transactions*, 17(2):110–116, 1985.

[3] Allan T. Andersen, Marcel F. Neuts, and Bo F. Nielsen. PH-Distributions Arising through Conditioning. *Commun. Statist.-Stochastic Models*, 16(1):179–188, 2000.

[4] Allan T. Andersen, Marcel F. Neuts, and Bo F. Nielsen. On the time reversal of Markovian Arrival Processes. *Stochastic Models*, 20(3):237–260, 2004.

[5] Allan T. Andersen and Bo F. Nielsen. On the statistical implications of certain random permutations in Markovian Arrival Processes (MAP)s and second order self-similar processes. *Performance Evaluation*, 41:67–82, 2000.

[6] Allan T. Andersen and Bo F. Nielsen. On the use of second order descriptors to predict queueing behaviour of MAPs. *Naval Research Logistics*, 49(4):391–409, 2002.

[7] Allan T. Andersen and Bo Friis Nielsen. An application of superpositions of two-state Markovian sources to the modelling of self-similar behaviour. In *INFOCOM'97*, volume 1, pages 196–204. IEEE, 1997.

[8] Allan T. Andersen and Bo Friis Nielsen. A Markovian approach for modeling packet traffic with long range dependence. *IEEE JSAC*, 16(5):719–732, 1998.

[9] Søren Asmussen. *Applied Probability and Queues*. Springer–Verlag, New York, second edition, 2003.

[10] Søren Asmussen and Mogens Bladt. Renewal theory and queueing algorithms for matrix-exponential distributions. In S. R. Chakravarthy and A. S. Alfa, editors, *Matrix-Analytic Methods in Stochastic Models*, pages 313–341, New York, 1996. Marcel Dekker.

[11] Søren Asmussen and Mogens Bladt. Point processes with finite-dimensional probabilities. *Stochastic Processes and their Applications*, 82(1):127–142, 1999.

[12] Søren Asmussen, Olle Nerman, and Marita Olsson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[13] David Assaf, Naftali A. Langberg, Thomas H. Savits, and Moshe Shaked. Multivariate phase-type distributions. *Operations Research*, 32(3):688–702, May-June 1984.

[14] Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT Press, 2008.

[15] Andrea Baiocchi. Analysis of the loss probability of the MAP/G/1/K queue part I: Asymptotic theory. *Commun. Statist.-Stochastic Models*, 10(4):867–893, 1994.

[16] Rose Baker. An order-statistics-based method for constructing multivariate distributions with fixed marginals. *Journal of Multivariate Analysis*, 99:2312–2327, 2008.

[17] Nigel Bean and Bo Friis Nielsen. Quasi-birth-and-death processes with rational arrival process components. *Stochastic Models*, 26(3):309–334, July 2010.

[18] Nigel Bean and Bo Friis Nielsen. Analysis of queues with rational arrival process (RAP) components - a general approach. IMM-Technical Report 5, IMM, 2011.

[19] Jan E. Beyer and Bo Friis Nielsen. Predator foraging in patchy environments: the interrupted Poisson proces (IPP) model unit. *DANA*, 11(2):65–130, 1996.

[20] Dario A. Bini, Guy Latouche, and Beatrice Meini. *Numerical Methods for Structured Markov Chains*. Oxford University Press, 2005.

[21] Mogens Bladt, Luz Judith Rodriguez Esparza, and Bo Friis Nielsen. Fisher information and statistical inference for phase–type distributions. *Journal of Applied Probability*, 48A - A Festschrift for Søren Asmussen:277–293, 2011.

[22] Mogens Bladt, Luz Judith Rodriguez Esparza, and Bo Friis Nielsen. Bilateral matrix–exponential distributions. In G. Latouche, V. Ramaswami, J. Sethuraman, K. Sigman, M.S. Squillante, and D. Yao, editors, *Matrix-Analytic Methods in Stochastic Models*, volume 27, pages 41–56. Springer Proceedings in Mathematics and Statistics, 2012.

[23] Mogens Bladt, Antonio Gonzalez, and Steffen L. Lauritzen. The estimation of phase-type related functionals using Markov chain Monte Carlo methods. *Scand. Actuarial J.*, 4:280–300, 2003.

[24] Mogens Bladt and Bo Friis Nielsen. Multivariate matrix-exponential distributions. In Dario Bini, Beatrice Meini, Vaidyanathan Ramaswami, Marie-Ange Remiche, and Peter Taylor, editors, *Numerical Methods for Structured Markov Chains*, number 07461 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2008. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[25] Mogens Bladt and Bo Friis Nielsen. Multivariate matrix–exponential distributions. *Stochastic Models*, 26(1):1–26, 2010.

[26] Mogens Bladt and Bo Friis Nielsen. On the construction of bivariate exponential distributions with an arbitrary correlation coefficient. *Stochastic Models*, 26(2):295–308, 2010.

[27] Mogens Bladt and Bo Friis Nielsen. Moment distributions of phase type. *Stochastic Models*, 27:651–663, 2011.

[28] Mogens Bladt and Bo Friis Nielsen. On the representation of distributions with rational moment generating functions. IMM-Technical Report 16, IMM, Technical University of Denmark, DK-2800 Kgs. Lyngby, 2012.

[29] Mogens Bladt and Bo Friis Nielsen. *Matrix Analytic Methods in Applied Probability.* Springer-Verlag, 2013. In preparation.

[30] A. Bobbio, A. Cumani, A. Premoli, and O. Saracco. Modelling and identification of non-exponential distributions by homogeneous markov processes. In *6.th. Advances in Reliab. Technol. Symp.*, pages 373–392, Bradford, 1980.

[31] Andrea Bobbio and Aldo Cumani. ML estimation of the parameters of a PH distribution in triangular canonical form. R.T. 393, Istituto Elettrotecnico Nazionale Galileo Ferraris, 1990.

[32] Andrea Bobbio and Aldo Cumani. ML estimation of the parameters of a PH distribution in triangular canonical form. In G. Balbo and G. Serazzi, editors, *Computer Performance Evaluation*, pages 33–46. Elsevier Science Publishers B.V., 1992.

[33] Andrea Bobbio and Miklos Telek. A benchmark for PH estimation algorithms: Results for acyclic-PH. *Commun. Statist.-Stochastic Models*, 10(3):661–677, 1994.

[34] W. Bux and U. Herzog. The phase concept: Approximation of measured data and performance analysis. In K.M. Chandy and M. Reiser, editors, *Computer Performance*, pages 23–38. North Holland Publishing Company, 1977.

[35] Thomas Kaare Christensen, Bo Friis Nielsen, and Villy Bæk Iversen. Distribution of channel holding times in cellular communication systems. In J. Moreira de Souza, Nelson L.S. da Fonseca, and E.A. de Souza e Silva, editors, *Teletraffic Engineering in the Internet Era*, pages 471–480, Salvador da Bahia, Brazil, 2001. ITC, Elsevier Science Publishers.

[36] Thomas Kaare Christensen, Bo Friis Nielsen, and Villy Bæk Iversen. Phase-type models of channel holding times in cellular communication systems. *IEEE Trans. on Veh. Technol.*, 53(3):725–733, May 2004.

[37] Edmund M. Clarke and E. Allen Emerson. *Logic of Programs*, volume 131, chapter Design and Synthesis of Synchronization Skeletons Using Branching-Time Temporal Logic, pages 52–71. Springer-Verlag, 1981.

[38] D.R. Cox and P.A.W. Lewis. *The Statistical Analysis of Series of Events.* Methuen, London, 1966.

[39] D.J. Daley and D. Vere-Jones. *An Inroduction to the Theory of Point Processes. Elementary Theory and Methods*, volume I of *Springer Series in Statistics.* Springer-Verlag, second edition, 2002.

[40] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *J. Royal Statist. Soc.*, 39:1–38, 1977.

[41] A. E. Eckberg. Generalized peakedness of teletraffic processes. In *Proceedings Eleventh International Teletraffic Conference (ITC 11), Kyoto, Japan*, pages Paper 4.4B–3, 1985.

[42] Ashok Erramilli, Onuttom Narayan, and Walter Willinger. Experimental Queueing Analysis with Long-Range Dependent Packet Traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, April 1996.

[43] M.J. Faddy. Phase-Type Distributions for Failure Times. *Mathl. Comput. Modelling*, 22(10-12):63–70, 1995.

[44] Yuguang Fang. Hyper-Erlang Distribution Model and its Application in Wireless Mobile Networks. *Wireless Networks*, 7:211–219, 2001.

[45] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz. The chaning nature of network traffic: Scaling phenomena. *Computer Communication Review*, 28(2):5–29, 1998.

[46] Sietske Greeuw. On the relation between matrix-geometric and discrete phase-type distributions. Master's thesis, Institute of Mathematical Modelling, The Technical University of Denmark, 2009. IMM-M.Sc.-2009-37.

[47] Qi-Ming He and Marcel Neuts. Markov chains with marked transitions. *Stochastic Processes and their Applications*, 74:37–52, 1998.

[48] Arne Jensen. Distribution patterns composed of a limited number of exponential distributions. In *Den 11. skandinaviske matematikerkongres*, pages 209–215, Trondheim, August 1949.

[49] Mary A. Johnson and Michael R. Taaffe. Matching moments to phase distributions: Mixtures of erlang distributions of common order. *Commun. Statist.-Stochastic Models*, 5(4):711–743, 1989.

[50] Mary A. Johnson and Michael R. Taaffe. Matching moments to phase type distributions: Nonlinear programming approaches. *Commun. Statist.-Stochastic Models*, 6(2):259–281, 1990.

[51] Alexander Klemm, Christoph Lindemann, and Marco Lohmann. Modeling IP traffic using the batch Markovian arrival process. *Performance Evaluation*, 54:149–173, 2003.

[52] Samuel Kotz, N. Balakrishnan, and Norman L. Johnson. *Continuous Multivariate Distributions.* John Wiley and Sons, 2000.

[53] Anatol Kuczura. The interrupted Poisson process as an overflow process. *The Bell System Technical Journal*, 52(3):437–448, March 1973.

[54] V. G. Kulkarni. A new class of multivariate phase type distributions. *Operations Research*, 37(1):151–158, January-February 1989.

[55] Guy Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM Publications, 1999.

[56] Guy Latouche and V. Ramaswami. A logarithmic reduction algorithm for quasi-birth-death processes. *J.Appl.Prob.*, 30:650–674, 93.

[57] Will W. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, February 1994.

[58] Dan Liu and Marcel F. Neuts. Counter-Examples Involving Markovian Arrival Processes. *Commun. Statist.-Stochastic Models*, 7(3):499–509, 1991.

[59] C.A. McGrory, A.N. Pettitt, and M.J. Faddy. A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis*, 53(12):4311–4321, October 2009.

[60] K Mitchell. Constructing a correlated sequence of matrix exponentials with invariant first-order properties. *Oper. Res. Lett.*, 28(1):27–34, 2001.

[61] M. F. Neuts. Probability distrubutions of phase type. In *Liber Amicorum Professor Emeritus H. Florin*, pages 173–206, Department of Mathematics, University of Louvian, Belgium, 1975.

[62] Marcel F. Neuts. A versatile Markovian point process. *J.Appl.Prob.*, 16:764–779, 1979.

[63] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models an Algorithmic Approach*, volume 2 of *John Hopkins Series in Mathematical Sciences.* The John Hopkins University Press, 1981.

[64] Marcel F. Neuts. *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, volume 5 of *A Series of Textbook and Reference Books.* Marcel Dekker, Inc., 1989.

[65] Marcel F. Neuts and Jian-Min Li. An algorithm for the $P(N, t)$ matrices of a continuous BMAP. In S. R. Chakrarvarthy and A. S. Alfa, editors, *Matrix analytic methods in stochastic models*, pages 7–19, New York, jul 1996. Marcel Dekker.

[66] Bo Friis Nielsen and Jan E. Beyer. Estimation of Interrupted Poisson Process parameters from counts. Report No. 21, 2004/2005, fall, Institut Mittag-Leffler, 2005.

[67] Bo Friis Nielsen, Flemming Nielson, and Hanne Riis Nielson. Model checking multivariate state rewards. In *Seventh International Conference on the Quantitative Evaluation of Systems*, pages 7–16, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

[68] Bo Friis Nielsen and V. Ramaswami. A Computational Framework for a Quasi Birth and Death Process with a Continuous Phase Variable. In V. Ramaswamiand P.E. Wirth, editor, *Teletraffic Contributions for the Information Age, ITC-15*, page 477–486. ITC, Elsevier, 1997.

[69] Bo Friis Nielsen, Uffe Høgsbro Thygesen, L. A. Fredrik Nilsson, and Jan E. Beyer. Higher order moments and conditional asymptotics of the batch Markovian arrival process. *Stochastic Models*, 23(1):1–26, 2007.

[70] Colm Art O'Cinneide. Characterization of phase-type distributions. *Commun. Statist.-Stochastic Models*, 6(1):1–57, 1990.

[71] J. P. Queille and J. Sifakis. *International Symposium on Programming*, volume 137 of *LNCS*, chapter Specification and Verification of Concurrent Systems in CESAR, pages 337–351. Springer, 1982.

[72] M. Rajaratnam and F. Takawira. Asymptotic approximation for handoff traffic characterisation in cellular networks under non-classical arrival and service time distributions. In *ITC Specialist Seminar on Mobile Systems and Mobility*, pages 95–106. ITC, 2000.

[73] V. Ramaswami. A duality theorem for the matrix paradigms in queueing theory. *Commun. Statist.-Stochastic Models*, 6(1):151–161, 1990.

[74] V. Ramaswami. A tutorial overview of matrix analytic methods: with some extensions & new results. In S. R. Chakravarthy and A. S. Alfa, editors, *Matrix Analytic Methods*, pages 261–295, New York, 1996. Marcel Dekker.

[75] Tobias Rydén. Parameter estimation for Markov modulated Poisson processes. *Commun. Statist.-Stochastic Models*, 10(4):795–829, 1994.

[76] Tobias Rydén. An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis*, 21:431–447, 1996.

[77] Tobias Rydén. Statistical estimation for Markov-modulated Poisson processes and Markovian arrival processes. In G. Latouche and P. G. Taylor, editors, *MAM3*, pages 329–350, Leuven, jul 2000. Notable Publications.

[78] Leonhard Schmickler. MEDA: Mixed Erlang distributions as phase-type representations of empirical distribution functions. *Commun. Statist.-Stochastic Models*, 8(1):131–156, 1992.

[79] B. H. Soong and J. A. Barria. A Coxian model for channel holding time distribution for teletraffic mobility modeling. *IEEE Communications Letters*, 4(12):402–404, December 2000.

[80] R.L. Tweedie. Operator-geometric stationary distributions for markov chains, with applications to queueing models. *Adv. Appl. Prob.*, 14:368–391, 1982.

[81] M. H. van Hoorn and L. P. Seelen. The $SPP/G/1$ queue: Single server queue with a switched Poisson process as input process. *OR Spektrum*, 5:205–218, 1983.