SORT 34 (2) July-December 2010, 101-144

Markovian arrivals in stochastic modelling: a survey and some new results

Jesús R. Artalejo, Antonio Gómez-Corral

Faculty of Mathematics, Complutense University of Madrid, Madrid 28040, Spain

Qi-Ming He

Department of Management Sciences, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

Abstract

This paper aims to provide a comprehensive review on *Markovian arrival processes* (MAPs), which constitute a rich class of point processes used extensively in stochastic modelling. Our starting point is the versatile process introduced by Neuts (1979) which, under some simplified notation, was coined as the *batch Markovian arrival process* (BMAP). On the one hand, a general point process can be approximated by appropriate MAPs and, on the other hand, the MAPs provide a versatile, yet tractable option for modelling a bursty flow by preserving the Markovian formalism. While a number of well-known arrival processes are subsumed under a BMAP as special cases, the literature also shows generalizations to model arrival streams with marks, non-homogeneous settings or even spatial arrivals. We survey on the main aspects of the BMAP, discuss on some of its variants and generalizations, and give a few new results in the context of a recent state-dependent extension.

MSC: 60Jxx, 60G55

Keywords: Markovian arrival process, batch arrivals, marked process, phase-type distribution, BSDE approach

1. Introduction

The *versatile Markovian point process* introduced by Neuts (1979) was the seminal work, in conjunction with the *phase* (PH) type distribution, for getting beyond two common and extended assumptions in stochastic modelling, namely: (a) the exponential

Received: April 2010

distribution and the *Poisson process* (PP), which are the key tools for constructing Markovian models; and (b) the independence and equidistribution of the successive inter-arrival intervals, which are inherent features of the PP and the renewal processes. Later, it was proved that the *batch Markovian arrival process* (BMAP) is equivalent to the versatile Neuts process. Since the former presents a more transparent notation, at present it is widely accepted to refer to the BMAP rather than to the Neuts process.

The popularity of the BMAP and other Markovian arrival processes comes from the following important features:

(i) They provide a natural generalization of the PP and the renewal processes.
(ii) They take into account the correlation aspect, which arises naturally in many applications where the arrival flow is bursty.
(iii) They preserve the tractable Markovian structure.

As a result, the use of Markovian arrival processes in combination with the impetus provided by the modern computational advances explains the spectacular growth of applications to queueing, inventory, reliability, manufacturing, communication systems, and risk and insurance problems.

The use of BMAPs and PH distributions in stochastic modelling readily leads to the so called matrix-analytic formalism where scalar quantities are replaced by matrices. The main resulting structured Markov chains have been extensively studied; see the monographs by Bini *et al.* (2005), Latouche and Ramaswami (1999), Li (2010) and Neuts (1981,1989). Qualitatively, the consideration of the BMAP for modelling the arrival input greatly enhances the versatility of the stochastic model. For practical use, presenting the model under a suitable structured matrix form makes it easy to be studied in a unified manner and in an algorithmically tractable way. However, it should be pointed out that the cost lies in the risk of finding computational problems derived from an excessive dimensionality caused by the matrix formalism.

This survey paper is aimed on providing information on Markovian arrival processes, putting emphasis on the discussion of extensions and variants of the BMAP, as well as on the wide use of this class of processes in applications. Following the leads in this paper and the guidance provided by the bibliographical notes, readers can get access to the background materials where technical details and proofs are available.

This survey is organized as follows. In Section 2, we first introduce the BMAP and the continuous PH distribution. A number of important particular cases, the basic properties and descriptors of the BMAP, as well as some applications in queueing, reliability and inventory models are presented in subsequent sections. In Section 3, we consider a number of generalizations and variants of the BMAP including the discrete counterpart (D-BMAP), the *marked Markovian arrival process* (MMAP), the *HetSigma* approach, the Markov-additive processes of arrivals and the *block-structured state-dependent event* (BSDE) approach. The consideration of these extensions and variants enriches the methodology and enhances the versatility of the arrival processes

in different directions. Based on the fact that the BSDE approach allows us to deal with modulated non-homogeneous settings, but keeping the dimensionality of the underlying matrices tractable, Section 4 applies this approach to the SIS epidemic model. Some new results concerning with the extinction time and the correlation between events are obtained. We conclude the survey with a few bibliographical notes. A glossary of notation is presented in Appendix.

2. The BMAP

The PP is the basic renewal process where inter-renewal times are exponentially distributed. The PH distribution and the BMAP can be thought of as the natural generalizations of the exponential distribution and the PP, respectively. They are both based on the method of stages, which was introduced by A.K. Erlang and extensively generalized by M.F. Neuts. On the other hand, the PH distribution and the BMAP can be viewed as particular cases of the matrix-exponential distribution and the rational arrival process; the interested reader is referred to the papers by Asmussen and Bladt (1999), and Nielsen *et al.* (2007).

Although our main interest is put on the BMAP and its extensions, the PH distribution is used many times along the paper. Thus, before focussing on a description of the BMAP, we briefly introduce the continuous PH distribution.

The class of probability distributions of PH type provides a simple framework to demonstrate how one may extend many results on exponential distributions to more complex models, but without losing computational tractability. The key idea is to exploit the fact that many distributions derived from the exponential law can be formulated as the distribution of the time till absorption in suitably defined Markov processes. This allows one to deal with PH distributions by appealing to the simple dependence structure underlying Markov processes.

To define a PH distribution we consider an absorbing Markov chain on the state space $\{0,1,\ldots,n\}$ with initial probability vector $(1-\boldsymbol{\tau}\mathbf{e}_n,\boldsymbol{\tau})$ and infinitesimal generator

$$\begin{pmatrix} 0 & \mathbf{0}_n \\ \mathbf{t} & \mathbf{T} \end{pmatrix},$$

where $\mathbf{t}=-\mathbf{T}\mathbf{e}_n$. Then, a PH distribution corresponds to the distribution of the time L until absorption into the state 0. Thus, we have the following expressions for the distribution function, the density function and the moments:

$$\begin{aligned} F(x) &= 1-\boldsymbol{\tau}\exp\{\mathbf{T}x\}\mathbf{e}_n, \quad x\geq 0,\\ f(x) &= \boldsymbol{\tau}\exp\{\mathbf{T}x\}\mathbf{t}, \quad x\geq 0,\\ E\left[L^k\right] &= k!\boldsymbol{\tau}\left(-\mathbf{T}^{-1}\right)^k\mathbf{e}_n, \quad k\geq 1. \end{aligned}$$

An important question to be examined is when the absorption occurs in a finite interval almost surely. By using the above expression for the distribution function, it is readily verified that $F(\infty) = 1$ if and only if the matrix $\mathbf{T}$ is non-singular. Furthermore, this is certain if and only if states in $\{1,\ldots,n\}$ are all transient.

For practical use, the class of PH distributions provides ease in conditioning arguments, results in a Markovian structure of models involving exponential assumptions and leads to significant simplifications in various integral and differential equations arising in their analysis. An excellent summary of closure properties can be found in Asmussen (2000), Latouche and Ramaswami (1999, Section 2.6) and Neuts (1981, Chapter 2). Among these, we emphasize three properties. First, this class is dense, in the sense of weak convergence, in the class of all distributions on $[0,\infty)$. Second, sums and mixtures of a finite number of independent PH random variables are PH random variables. Third, all order statistics of a set of independent PH random variables are themselves PH random variables.

The PP has served as the main arrival flow for many years and generalizations have frequently concentrated on renewal processes. Their simplifying feature is the independence and equidistribution of successive inter-renewal intervals. Thus, in queueing and other applications (see Neuts (1992)), the class of renewal processes is not flexible enough and, in particular, arrivals that tend to occur in bursts cannot be modelled in this way.

We present here the BMAP, which is thought to be a fairly general point process where the correlation aspect is not ignored. It is, in general, a non-renewal process having the feature of making many analytic properties explicit or at least computationally tractable. The key idea is to generate counting processes by modelling the transitions of a Markov chain; see also Rudemo (1973).

We begin with a constructive description of the BMAP. The BMAP is a bivariate Markov process $\{(N(t),J(t));t \geq 0\}$ on $\mathscr{S} = \mathbb{N} \times \{1,\ldots,m\}$, where $N(t)$ represents the number of arrivals up to time t, while the states of the background Markov chain $\{J(t);t \geq 0\}$ are called phases. Let us assume that $m < \infty$ and denote by $\mathbf{D}$ the infinitesimal generator of the background Markov chain, which is assumed to be irreducible. At the end of a sojourn time in $(n,i) \in \mathscr{S}$, which is exponentially distributed with parameter λ_i, there occurs a transition to another or (possibly) the same phase state. That transition may or not correspond to an arrival epoch. Specifically, with probability $P_{ij}(k)$, it corresponds to a transition to state j with a batch arrival of size k, for $k \geq 1$, and similarly, with probability $P_{ij}(0)$, the transition corresponds to no arrival and state of the underlying Markov chain is j, for $j \neq i$. Therefore, $J(t)$ can go from state i to state i only through an arrival and

$$\sum_{j=1,\ j\neq i}^{m} P_{ij}(0) + \sum_{j=1}^{m}\sum_{k=1}^{\infty} P_{ij}(k) = 1, \quad 1 \leq i \leq m.$$

Define the matrices $\mathbf{D}_k = (d_{ij}(k))$ with entries $d_{ii}(0) = -\lambda_i$, $d_{ij}(0) = \lambda_i P_{ij}(0)$, for $j \neq i$, and $d_{ij}(k) = \lambda_i P_{ij}(k)$, for $k \geq 1$, from which it is clear that $\mathbf{D} = \sum_{k=0}^{\infty} \mathbf{D}_k$. The particular

choice $\mathbf{D}_0 \neq \mathbf{D}$ and $\mathbf{D}_k = \mathbf{0}_{m\times m}$, for $k \geq 2$, means single arrivals and yields the *Markovian arrival process* (MAP). In this formulation, the introduction of phases is the key to get dependent non-exponential inter-arrival time distributions, and correlated batch sizes.

Our preceding construction shows that the bivariate process $\{(N(t), J(t)); t \geq 0\}$ has the structured infinitesimal generator

$$\mathbf{Q} = \begin{pmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \cdots \\ & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots \\ & & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ & & & \ddots & \ddots \end{pmatrix}.$$

The sequence of matrices $\{\mathbf{D}_k; k \geq 0\}$ contains all information for $\mathbf{Q}$ and thus is usually called the characteristic sequence of a BMAP. Although we often ignore the determination of $J(0)$, a complete specification requires specification of the distribution of $J(0)$. We may do this in terms of a row vector $\boldsymbol{\alpha}$ with ith entry given by $P(J(0) = i)$, for $1 \leq i \leq m$.

By assuming $\mathbf{D}_0$ to be non-singular, the inter-arrival times are finite, with probability one. An additional assumption is that the vector $\mathbf{d} = \overline{\mathbf{D}}_1 \mathbf{e}_m$ is finite, where $\overline{\mathbf{D}}_1 = \sum_{k=1}^{\infty} k\mathbf{D}_k$. This condition is equivalent to require that $E[N(t)] < \infty$ over finite intervals. The fundamental arrival rate is then defined by $\lambda = \boldsymbol{\theta}\mathbf{d}$, where $\boldsymbol{\theta}$ is the unique positive probability vector satisfying $\boldsymbol{\theta}\mathbf{D} = \mathbf{0}_m$ and $\boldsymbol{\theta}\mathbf{e}_m = 1$, and consequently it amounts to the expected number of single arrivals per unit of time in the stationary version of a BMAP.

This family of counting processes has received several names in the literature. The currently used term batch Markovian arrival process evolved from versatile Markovian point process (see Neuts (1979)) and *Neuts process* (see Ramaswami (1980)) to *non-renewal arrival process* (see Lucantoni *et al.* (1990)), until it was settled down at batch Markovian arrival process by Lucantoni (1991). Lucantoni (1991) also introduced a simple matrix representation for the BMAP, which made it easy to interpret parameters of Markovian arrivals and to use this class of arrival processes in stochastic modelling.

We next present two alternative definitions of the BMAP and a few examples of BMAPs with special characteristics.

Remark 2.1 The BMAP can be thought of as a semi-Markovian arrival process. Define the sequence $\{(J_n, K_n, \tau_n); n \geq 0\}$, where J_n is the phase of $\{J(t); t \geq 0\}$ right after the nth batch arrival, K_n is the size of the nth batch, and τ_n is the inter-arrival time between the $(n-1)$st and the nth arrival events. Then, $\{(J_n, K_n, \tau_n); n \geq 0\}$ satisfies

$$\begin{aligned} P(J_n = j, K_n = k, \tau_n \leq x | J_{n-1} = i) &= \left(\int_0^x \exp\{\mathbf{D}_0 u\} du \mathbf{D}_k \right)_{ij} \\ &= \left((\mathbf{I}_m - \exp\{\mathbf{D}_0 x\}) \left(-\mathbf{D}_0^{-1}\right) \mathbf{D}_k \right)_{ij}, \end{aligned}$$

for $1 \leq i, j \leq m$, $k \geq 1$ and $x \geq 0$.

Remark 2.2 Equivalently, we may present a definition of the BMAP based on PPs. Let m be a finite positive integer, $\{\alpha_i; 1 \le i \le m\}$ be non-negative numbers satisfying $\sum_{i=1}^{m} \alpha_i = 1$, and $\{d_{ij}(0); 1 \le i, j \le m, j \ne i\}$ and $\{d_{ij}(k); 1 \le i, j \le m\}$, for $k \ge 1$, be non-negative numbers. Assume that $-d_{ii}(0) > 0$, where

$$-d_{ii}(0) = \sum_{j=1,\ j\neq i}^{m} d_{ij}(0) + \sum_{j=1}^{m}\sum_{k=1}^{\infty} d_{ij}(k), \quad 1 \le i \le m.$$

The bivariate process $\{(N(t), J(t)); t \ge 0\}$ can be defined as follows:

(i) Define independent PPs with parameters $d_{ij}(0)$, for $1 \le i, j \le m$ and $j \ne i$, and $d_{ij}(k)$, for $1 \le i, j \le m$ and $k \ge 1$. If $d_{ij}(k) = 0$, then the corresponding PP has no event.
(ii) Determine $J(0)$ by the probability distribution $\{\alpha_i; 1 \le i \le m\}$. Set $N(0) = 0$.
(iii) If $J(t) = i$, for $1 \le i \le m$, we let $J(t)$ and $N(t)$ remain the same until the first event occurs in the set of PPs with rates $d_{ij}(0)$, for $1 \le i, j \le m$ and $j \ne i$, and $d_{ij}(k)$, for $1 \le i, j \le m$ and $k \ge 1$. If the next event comes from the PP of rate $d_{ij}(0)$, then $J(t)$ changes from phase i to phase j and $N(t)$ does not change at this epoch, for $1 \le j \le m$ and $j \ne i$. On the contrary, if the next event comes from the PP of rate $d_{ij}(k)$, then the phase variable $J(t)$ transits from phase i to phase j, and $N(t)$ is increased by k units at this epoch, for $1 \le j \le m$ and $k \ge 1$; in this case, a batch of k units is associated with the event.

For use in simulations, it is easy to generate realizations of a BMAP from the dynamics described in Remark 2.2. The visualization of simulated paths of a BMAP, and their effect as input streams to queues, is an excellent way for practitioners to appreciate the versatility of this class of point processes; see Figure 1 in Example 2.1.

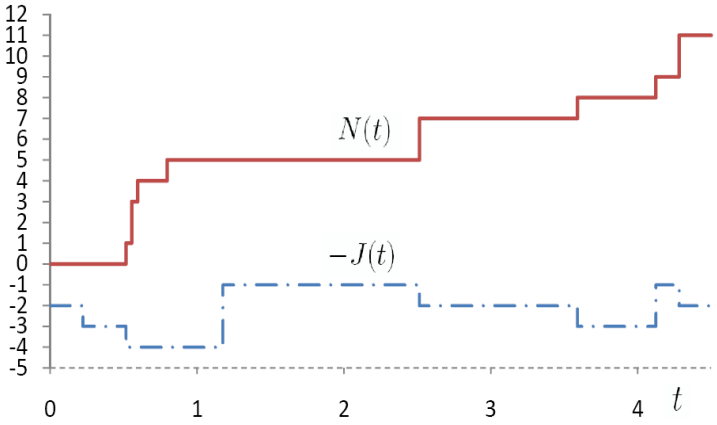

Figure 1: *A simulated sample path of a BMAP.*

Example 2.1 Consider a BMAP with non-null characteristic matrices

$$\mathbf{D}_0 = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 1 & 0 & -5 & 0 \\ 2 & 0 & 0 & -10 \end{pmatrix}, \ \mathbf{D}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}, \ \mathbf{D}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 5 \end{pmatrix}.$$

Figure 1 shows a typical sample path of the bivariate process $\{(N(t), J(t)); t \geq 0\}$.

The following three choices of the BMAP are related to special characteristics:

(i) Bursty arrivals

$$\mathbf{D}_0 = \begin{pmatrix} -50 & 0 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{D}_1 = \begin{pmatrix} 49 & 1 \\ 0 & 0 \end{pmatrix}.$$

A widely accepted definition of burstiness does not exist; instead, several different measures can be used. In this paper, we assume the definition given by Neuts (1993). Qualitatively, the process is bursty as, over intervals of significant length, the actual number of arrivals is far in excess or far below the average. Positive autocorrelation between inter-arrival times explains, to a large extend, traffic burstiness. Obviously, the PP has independent inter-arrival times so it is not the appropriate model in case of bursty traffic.

(ii) Cyclic arrivals

$$\mathbf{D}_0 = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}, \quad \mathbf{D}_1 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{D}_2 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}.$$

In this case, batches of size 1 and batches of size 2 arrive cyclically.

(iii) Bursty vs smooth

$$\mathbf{D}_0 = \begin{pmatrix} -1 & 0 \\ 0 & -50 \end{pmatrix}, \quad \mathbf{D}_1 = \begin{pmatrix} 0 & 0 \\ 1 & 49 \end{pmatrix}, \quad \mathbf{D}_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

The process related to batches of size 1 is bursty, while for batches of size 2 the process is smooth.

In Subsection 2.1, we give a few examples to illustrate the variety of models subsumed under the matrix formulation of a BMAP as special cases. Subsection 2.2 begins by introducing the time-dependent distribution of the bivariate process $\{(N(t), J(t)); t \geq 0\}$. We then examine basic properties that make the BMAP a versatile class for modelling purposes. We present in Subsection 2.3 some interesting descriptors. Our focus in Subsection 2.4 is on four examples showing the interest of the BMAP in different applications, such as reliabilty, queueing and inventory problems.

2.1. Particular cases

We describe in this subsection several special cases of the BMAP. We begin by listing a selected sample of processes obtained as particular cases of the MAP.

(i) *Poisson process*. The PP of rate $\lambda > 0$ corresponds to the simple scalar case where $m = 1$, $\mathbf{D}_0 = -\lambda$ and $\mathbf{D}_1 = \lambda$.

(ii) *Markov modulated Poisson process* (MMPP). The MMPP is a PP whose rate varies according to a finite Markov chain serving as a random environment. Let $\mathbf{Q}_a$ be its underlying infinitesimal generator. The arrival rate is $\delta_i > 0$ when the random environmental state is i. Then, the MMPP is a MAP with $\mathbf{D}_0 = \mathbf{Q}_a - \mathbf{\Lambda}$ and $\mathbf{D}_1 = \mathbf{\Lambda}$, where $\mathbf{\Lambda} = \text{diag}(\delta_1, \ldots, \delta_m)$.

(iii) *PH renewal process*. This is a renewal process in which the inter-renewal times follow a PH distribution with representation $(\boldsymbol{\tau}, \mathbf{T})$. Thus, we have the correspondence $\mathbf{D}_0 = \mathbf{T}$ and $\mathbf{D}_1 = \mathbf{t}\boldsymbol{\tau}$.

(iv) *A sequence of PH inter-arrival times governed via a Markov chain*. This process is also named *PH semi-Markov process*; see Latouche and Ramaswami (1999). Consider l PH distributions with representations $(\boldsymbol{\tau}_i, \mathbf{T}_i)$ of order n_i, for $1 \leq i \leq l$ and $\sum_{i=1}^{l} n_i = m$. The successive inter-arrival distributions are selected from these PH distributions according to a discrete Markov chain with one-step transition probability matrix $\mathbf{P}_a = (p_{ii'})$ of dimension l. We then have $\mathbf{D}_0 = \text{diag}(\mathbf{T}_1, \ldots, \mathbf{T}_l)$ and $\mathbf{D}_1 = (d_{ii'}(1))$, where $d_{ii'}(1) = t_i p_{ii'} \tau_{i'}$, for $1 \leq i, i' \leq l$, with $\mathbf{t} = (t_i)$ and $\boldsymbol{\tau} = (\tau_i)$. The choice $l = 2$ and $p_{12} = p_{21} = 1$ leads to an *alternating PH renewal process*.

It should be noted that the PH renewal process can be viewed as the trivial special case of (iv), where all the PH distributions are chosen to be identical. More interesting is the *Markov switched Poisson process* (MSPP) obtained by choosing the PH distributions as exponential distributions of rate $\delta_i > 0$; see Chakravarthy (2001). We also remark that the modulation in the MSPP is of a discrete nature and it occurs at arrival epochs, whereas the modulation of the MMPP is performed in continuous time.

We now give some examples where arrivals occur properly in batches.

(v) *Compound Poisson process* (CPP). The classical scalar PP with batch arrivals of rate $\lambda > 0$ and jump size distribution $\{g_k; k \geq 1\}$ is a BMAP with $m = 1$, $\mathbf{D}_0 = -\lambda$ and $\mathbf{D}_k = \lambda g_k$, for $k \geq 1$.

(vi) *MAP with i.i.d. batch arrivals*. A MAP with independent and identically distributed batch arrivals amounts to a BMAP with $\mathbf{D}_0 = \mathbf{D}_0^a$ and $\mathbf{D}_k = g_k \mathbf{D}_1^a$, for $k \geq 1$, where the pair $(\mathbf{D}_0^a, \mathbf{D}_1^a)$ is the representation of the underlying MAP of order m. This example shows a choice of the BMAP where the batch size does not depend on phase transitions.

(vii) *Bacth PH semi-Markov process*. This process is the batch version of (iv) in which $d_{ii'}(k) = g_k t_i p_{ii'} \tau_{i'}$, for $k \geq 1$. A *batch Markov switched Poisson process* (BMSPP) follows by reducing the PH distribution to the exponential case.

(viii) *Batch PP with correlated batch arrivals*. This is a CPP where the jump size distribution is selected according to a Markov chain with one-step transition probability matrix $\mathbf{P}_a$ of dimension m. The resulting BMAP has matrices $\mathbf{D}_0 = -\lambda \mathbf{I}_m$ and $\mathbf{D}_k = (d_{ii'}(k))$, where $d_{ii'}(k) = \lambda g_{ik} p_{ii'}$, for $1 \leq i, i' \leq m$ and $k \geq 1$. The notation g_{ik} stands for the probability that a batch of size k arrives when the phase state is i.

We notice that the auxiliary transition matrix is used in the MSPP to modulate arrival rates. However, the role of $\mathbf{P}_a$ in the batch PP with correlated arrivals is to modulate jump sizes.

2.2. Basic properties of the BMAP

We are next interested in the counting component $N(t)$ of the BMAP, the superposition and thinning mechanisms, the local poissonification of a MAP and the denseness property.

2.2.1. The counting function

Consider the matrices $\mathbf{P}(n,t)$, for $n \geq 0$ and $t \geq 0$, with (i,j)th element

$$P_{ij}(n,t) = P\left(N(t) = n, J(t) = j | N(0) = 0, J(0) = i\right), \quad 1 \leq i, j \leq m.$$

From the Kolmogorov forward equations of the process $\{(N(t), J(t)); t \geq 0\}$, we obtain

$$\frac{d\mathbf{P}(n,t)}{dt} = \sum_{k=0}^{n} \mathbf{P}(k,t)\mathbf{D}_{n-k}, \quad n \geq 1, t \geq 0,$$

and the initial condition $\mathbf{P}(0,0) = \mathbf{I}_m$.

The corresponding matrix generating function $\mathbf{P}^*(z,t) = \sum_{n=0}^{\infty} z^n \mathbf{P}(n,t)$, for $|z| \leq 1$ and $t \geq 0$, is given by the exponential matrix

$$\mathbf{P}^*(z,t) = \exp\{\mathbf{D}^*(z)t\},$$

with $\mathbf{D}^*(z) = \sum_{k=0}^{\infty} z^k \mathbf{D}_k$, for $|z| \leq 1$. The numerical computation of $\mathbf{P}(n,t)$ can be based on the uniformization method; see Neuts and Li (1997).

By routine calculations, we can find that the first moment matrix $\mathbf{M}_1(t)$ and the column vector $\mathbf{M}_1(t)\mathbf{e}_m$ are given by

$$\mathbf{M}_1(t) = \left.\frac{\partial \mathbf{P}^*(z,t)}{\partial z}\right|_{z=1} = \sum_{n=1}^{\infty} \frac{t^n}{n!} \sum_{k=0}^{n-1} \mathbf{D}^k \overline{\mathbf{D}}_1 \mathbf{D}^{n-1-k},$$

$$\mathbf{M}_1(t)\mathbf{e}_m = \sum_{n=1}^{\infty} \frac{t^n}{n!} \mathbf{D}^{n-1} \overline{\mathbf{D}}_1 \mathbf{e}_m.$$

By using the above expression, it can be shown (see Neuts (1989)) that the Palm function $E[N(t)]$ is given by

$$E[N(t)] = \lambda t + \boldsymbol{\alpha}\left(\exp\{\mathbf{D}t\} - \mathbf{I}_m\right)\left(\mathbf{D} - \mathbf{e}_m\boldsymbol{\theta}\right)^{-1}\overline{\mathbf{D}}_1\mathbf{e}_m, \quad t \geq 0.$$

Since $\boldsymbol{\alpha}\exp\{\mathbf{D}t\}$ converges to $\boldsymbol{\theta}$ as $t \to \infty$ (see Latouche and Ramaswami (1999)), we find that $\lim_{t\to\infty} E[N(t)]/t = \lambda$, so λ is the expected number of arrivals per unit time.

If the initial phase vector is $\boldsymbol{\theta}$ (i.e., we set $\boldsymbol{\alpha} = \boldsymbol{\theta}$), the Palm function reduces to $E[N(t)] = \lambda t$. For the variance of the number of arrivals in $(0,t]$ and the covariance of the counts, we refer to the results summarized in Subsection 2.3; see also Narayana and Neuts (1992).

2.2.2. Superposition and thinning

The class of BMAPs is closed under superposition. For simplicity, we consider two independent BMAPs $\{(N_i(t), J_i(t)); t \geq 0\}$ with characteristic sequences $\{\mathbf{D}_k^i; k \geq 0\}$ of order m_i, for $i \in \{1,2\}$, but the construction can be readily extended to an arbitrary number of BMAPs. Then, the resulting superposition process $\{(N(t), J(t)); t \geq 0\}$ is a BMAP with matrices $\{\mathbf{D}_k^1 \oplus \mathbf{D}_k^2; k \geq 0\}$. We notice that the count $N(t)$ is defined by $N_1(t) + N_2(t)$ and the phase process $J(t)$ has the form $(J_1(t), J_2(t))$.

Thinning is a mechanism to split or remove a part of the arrivals generated by the BMAP. As a result, thinning can be thought of as an operation opposite to the superposition. One way to single out arrivals from the original BMAP flow is just to discard any individual arrival with probability p independently of the rest of arrivals. The resulting BMAP has a matrix representation $\{\mathbf{D}_k^T; k \geq 0\}$, where

$$\mathbf{D}_0^T = \mathbf{D}_0 + \sum_{j=1}^{\infty} p^j \mathbf{D}_j,$$

$$\mathbf{D}_k^T = \sum_{j=k}^{\infty} \binom{j}{k} p^{j-k}(1-p)^k \mathbf{D}_j, \quad k \geq 1.$$

Another more sophisticated way to understand the thinning is associated with the arrivals of a BMAP and a clock with a PH distribution with representation $(\boldsymbol{\tau}, \mathbf{T})$. An auxiliary state 0 indicates that the PH clock is active, so that during this period the BMAP arrivals are not registered. As soon as the clock expires, the process turns to the auxiliary state 1 and the next arrival is registered. Immediately after one arrival is

registered, the PH clock is restarted. This description leads to a BMAP with matrices

$$\mathbf{D}_0^T = \begin{pmatrix} \mathbf{D} \oplus \mathbf{T} & \mathbf{I}_m \otimes \mathbf{t} \\ \mathbf{0}_{m \times mn} & \mathbf{D}_0 \end{pmatrix} \qquad \mathbf{D}_k^T = \begin{pmatrix} \mathbf{0}_{mn} & \mathbf{0}_{mn \times m} \\ \mathbf{D}_k \otimes \boldsymbol{\tau} & \mathbf{0}_{m \times m} \end{pmatrix}, \quad k \geq 1.$$

Decomposition of BMAPs provides another related operation. We may decompose a BMAP into n types of arrivals by considering independent markings with probabilities p_i, for $1 \leq i \leq n$, where $\sum_{i=1}^{n} p_i = 1$. Then, the split process $\{(N_i(t), J(t)); t \geq 0\}$ is a BMAP with $\mathbf{D}_0^i = \mathbf{D}_0 + (1-p_i)\overline{\mathbf{D}}_0$, $\mathbf{D}_k^i = p_i \mathbf{D}_k$, for $k \geq 1$ and each $1 \leq i \leq n$, where $\overline{\mathbf{D}}_0 = \mathbf{D} - \mathbf{D}_0$.

2.2.3. Local poissonification of a MAP

The local poissonification (see Neuts *et al.* (1992)) is an approach to quantifying the burstiness of a stationary point process. The events in successive intervals of length a are independently and uniformly redistributed over those intervals. The resulting local poissonification process mimics the behaviour of a PP over each interval.

For the MAP, the local poissonification construction can be tractably investigated by using matrix-analytic methods. To construct the stationary local poissonification of the MAP, we first choose the phase according to the vector $\boldsymbol{\theta}$ and a grid of points, regularly placed at a distance a. Then, the time origin is chosen randomly in one of the resulting intervals. Denote by $N_a(t)$ the counting process of the poissonification in any interval of length t.

The Palm function of $N_a(t)$ is $E[N_a(t)] = \lambda t$, for $t \geq 0$, thus showing that the poissonification preserves the fundamental rate of the original MAP. On the other hand, the variance of the count $N_a(t)$ is given by

$$\begin{aligned} Var(N_a(t)) = \lambda t + (V^0(a) - \lambda a) \left(\left(\frac{t}{a}\right)^2 - \frac{1}{3}\left(\frac{t}{a}\right)^3 + \frac{1}{3}\left(\frac{t-a}{a}\right)^3 V(t-a) \right) \\ + \frac{1}{3}\sum_{k=0}^{\infty} \rho_{k+1}(a) \left(\left(\frac{t-ka}{a}\right)^3 V(t-ka) - 2\left(\frac{t-(k+1)a}{a}\right)^3 V(t-(k+1)a) \right. \\ \left. + \left(\frac{t-(k+2)a}{a}\right)^3 V(t-(k+2)a) \right), \end{aligned}$$

where $V(x) = 1$ if $x \geq 0$, and it equals 0 otherwise, whereas $V^0(a)$ and $\rho_k(a)$ denote respectively the variance of the number of events in $(0,a]$ and the covariance of the counts in the intervals $(0,a]$ and $(ka,(k+1)a]$, in the stationary given MAP; see Subsection 2.3.1.

A number of computationally implementable descriptors include the dispersion function and the exponential peakedness (see Subsections 2.3.1 and 2.3.3), as well as

the distribution of the interval length. The latter is defined as the probability distribution of the interval between an arbitrary point and the next event in the poissonification of the stationary MAP. Its Laplace-Stieltjes transform $\varphi_a(s)$ is given by

$$\varphi_a(s) = 1 - \frac{s}{\lambda} + \frac{s^2 a}{\lambda}\left(\boldsymbol{\theta}\mathbf{L}_a^1(s)\mathbf{e}_m + \boldsymbol{\theta}L_a^0(s)\left(\mathbf{I}_m - e^{-sa}\exp\{\mathbf{D}_0 a\}\right)^{-1}\mathbf{L}_a^0(s)\mathbf{e}_m\right),$$

where the matrices $\mathbf{L}_a^0(s)$ and $\mathbf{L}_a^1(s)$ are defined by

$$\mathbf{L}_a^0(s) = \int_0^1 \exp\{\mathbf{D}^*(u)a\}e^{-sa(1-u)}du,$$

$$\mathbf{L}_a^1(s) = \int_0^1 u\exp\{\mathbf{D}^*(u)a\}e^{-sa(1-u)}du.$$

The mean μ_a and the variance σ_a^2 of the inter-arrival time are given by

$$\mu_a = \frac{1}{\lambda},$$

$$\sigma_a^2 = \frac{2a}{\lambda}\left(\boldsymbol{\theta}\mathbf{L}_a^1(0)\mathbf{e}_m + \boldsymbol{\theta}\mathbf{L}_a^0(0)\left(\mathbf{I}_m - \exp\{\mathbf{D}_0 a\}\right)^{-1}\mathbf{L}_a^0(0)\mathbf{e}_m\right) - \frac{1}{\lambda^2}.$$

2.2.4. Denseness property

Asmussen and Koole (1993) prove that a general class of *marked point processes* (MPP) can be approximated by appropriate MAPs. The MPP can be considered either at an arbitrary time or at selected discrete epochs. In the latter case the MPP is represented as a bivariate process $\{(T_n, Y_n); n \geq 0\}$, where the random variables T_n denote inter-arrival times and the marks Y_n are allowed to vary in $(0,\infty)$. In the arbitrary time version, an MPP is viewed as a point process taking values on the state space $[0,\infty)\times(0,\infty)$. A class of *Markovian arrival streams* (MAS) is also defined to approximate the given MPP. In a MAS there exists a finite state space of phases modulated by two matrices playing the same role that $\mathbf{D}_0$ and $\mathbf{D}_1$ in the MAP. When an arrival occurs, a mark is assigned according to a distribution B_{ij} on $(0,\infty)$. The mark depends on the current phase i and the destination phase j. If all B_{ij} are degenerate at 1, then the MAS agrees with the MAP.

The main result in Asmussen and Koole (1993) establishes that the class of MASs is dense in the class of MPPs in both time scales. The convergence must be viewed in distribution. However, related results for stationary processes and convergence of the moments also hold. It is interesting to remark that the convergence result does not hold when the class of MASs is replaced by MMPPs.

The above property is the analogue of the denseness property of PH distributions in the set of all probability distributions on $[0,\infty)$; see Neuts (1989). The proof follows from the fact that any probability distribution on $[0,\infty)$ may be suitably approximated by

a discrete distribution with a finite support, which is indeed a discrete PH distribution; see Latouche and Ramaswami (1999, Section 2.5) and Neuts (1981, Section 2.2).

2.3. Some interesting descriptors

The quantification of the main quality characteristics of the BMAP is of primarily theoretical and practical utility. This important objective is reached through the consideration of a variety of computationally implementable descriptors.

We distinguish three categories of descriptors for BMAPs: (a) descriptors associated with the counting function, (b) descriptors associated with inter-arrival times, and (c) other descriptors.

2.3.1. Descriptors associated with the counting function

To begin with, we recall that expressions for the fundamental arrival rate λ and the expected number of arrivals $E[N(t)]$ were already given in preceding subsections. Other descriptors related to the counting function are

(i) *The variance of the number of arrivals*. Given the initial distribution $\boldsymbol{\theta}$, we have

$$\begin{aligned} Var(N(t)) = &\left(\lambda_2 - 2\lambda^2 - 2\boldsymbol{\theta}\overline{\mathbf{D}}_1(\mathbf{D}-\mathbf{e}_m\boldsymbol{\theta})^{-1}\overline{\mathbf{D}}_1\mathbf{e}_m\right)t \\ &+ 2\boldsymbol{\theta}\overline{\mathbf{D}}_1(\mathbf{D}-\mathbf{e}_m\boldsymbol{\theta})^{-1}\left(\exp\{\mathbf{D}t\}-\mathbf{I}_m\right)(\mathbf{D}-\mathbf{e}_m\boldsymbol{\theta})^{-1}\overline{\mathbf{D}}_1\mathbf{e}_m, \end{aligned}$$

where $\lambda_2 = \boldsymbol{\theta}\overline{\mathbf{D}}_2\mathbf{e}_m$ and $\overline{\mathbf{D}}_2 = \sum_{k=1}^{\infty} k^2\mathbf{D}_k$.

(ii) *The dispersion function*. It is defined as

$$F_d(t) = \frac{Var(N(t))}{E[N(t)]}.$$

We observe that the dispersion function is a minor variant of the coefficient of variation, which is defined as the ratio between the standard deviation and the expectation. The dispersion function is also known as the index of dispersions for the counts; see Chakravarthy (2001).

(iii) *The covariance and the correlation of the counts*. Given the positive real numbers t, u, r and s, we construct the time intervals $(t,t+u]$ and $(t+u+r,t+u+r+s]$. The stationary versions of the covariance $\varphi(u,s,r)$ and the correlation $\rho(u,s,r)$ in these intervals are given by

$$\begin{aligned} \varphi(u,s,r) &= \boldsymbol{\theta}\overline{\mathbf{D}}_1(\mathbf{D}-\mathbf{e}_m\boldsymbol{\theta})^{-1}(\exp\{\mathbf{D}u\}-\mathbf{I}_m)\exp\{\mathbf{D}r\}(\exp\{\mathbf{D}s\}-\mathbf{I}_m) \\ &\quad\times(\mathbf{D}-\mathbf{e}_m\boldsymbol{\theta})^{-1}\overline{\mathbf{D}}_1\mathbf{e}_m - \lambda^2 us, \\ \rho(u,s,r) &= \frac{\varphi(u,s,r)}{\sqrt{Var(N(u))Var(N(s))}}. \end{aligned}$$

Those readers interested in the derivation of the above formulas are referred to the papers by Narayana and Neuts (1992), and Neuts *et al*. (1992).

2.3.2. Descriptors associated with inter-arrival times

Assume that $J(0)$ has a distribution $\boldsymbol{\alpha}$. The random vector $(\tau_1,\ldots,\tau_n)$ of inter-arrival times follows a multivariate continuous PH distribution (see Kulkarni (1989)). Therefore, the nth inter-arrival time τ_n has a PH distribution with representation

$$\left(\boldsymbol{\alpha}\left(\left(-\mathbf{D}_0^{-1}\right)\overline{\mathbf{D}}_0\right)^{n-1},\mathbf{D}_0\right).$$

Then, it is immediate to obtain the expressions for the mean and the variance in the list below.

(i) *The mean of* τ_n

$$E[\tau_n]=\boldsymbol{\alpha}\left(\left(-\mathbf{D}_0^{-1}\right)\overline{\mathbf{D}}_0\right)^{n-1}\left(-\mathbf{D}_0^{-1}\right)\mathbf{e}_m,\quad n\geq 1.$$

(ii) *The variance of* τ_n

$$Var(\tau_n)=2\boldsymbol{\alpha}\left(\left(-\mathbf{D}_0^{-1}\right)\overline{\mathbf{D}}_0\right)^{n-1}\left(-\mathbf{D}_0\right)^{-2}\mathbf{e}_m-\left(\boldsymbol{\alpha}\left(\left(-\mathbf{D}_0^{-1}\right)\overline{\mathbf{D}}_0\right)^{n-1}\left(-\mathbf{D}_0^{-1}\right)\mathbf{e}_m\right)^2,\quad n\geq 1.$$

(iii) *The coefficient of variation*

$$cv(\tau_n)=\frac{\sqrt{Var(\tau_n)}}{E[\tau_n]},\quad n\geq 1.$$

(iv) *The covariance and the correlation between* τ_1 *and* τ_n

$$\begin{aligned}\varphi(\tau_1,\tau_n)&=\boldsymbol{\alpha}\left(-\mathbf{D}_0^{-1}\right)\left(\left(-\mathbf{D}_0^{-1}\right)\overline{\mathbf{D}}_0\right)^{n-1}\left(-\mathbf{D}_0^{-1}\right)\mathbf{e}_m\\&\quad-\left(\boldsymbol{\alpha}(-\mathbf{D}_0^{-1})\mathbf{e}_m\right)\left(\boldsymbol{\alpha}\left((-\mathbf{D}_0^{-1})\overline{\mathbf{D}}_0\right)^{n-1}(-\mathbf{D}_0^{-1})\mathbf{e}_m\right),\quad n\geq 1,\\\rho(\tau_1,\tau_n)&=\frac{\varphi(\tau_1,\tau_n)}{\sqrt{Var(\tau_1)Var(\tau_n)}}.\end{aligned}$$

Setting $\boldsymbol{\alpha}=\hat{\lambda}^{-1}\boldsymbol{\theta}\overline{\mathbf{D}}_0$, we obtain simplified expressions for the mean $\mu=\hat{\lambda}^{-1}$, the variance $\sigma^2=2\mu\boldsymbol{\theta}\left(-\mathbf{D}_0^{-1}\right)\mathbf{e}_m-\mu^2$ and the correlation

$$\rho(\tau_1,\tau_n)=\frac{\mu\boldsymbol{\theta}\left(\left(-\mathbf{D}_0^{-1}\right)\overline{\mathbf{D}}_0\right)^{n-1}\left(-\mathbf{D}_0^{-1}\right)\mathbf{e}_m-\mu^2}{\sigma^2},$$

where $\hat{\lambda}$ is the batch arrival rate defined by $\hat{\lambda} = \boldsymbol{\theta}\overline{\mathbf{D}}_0\mathbf{e}_m$; see Neuts (1995). Thus, $\boldsymbol{\alpha} = \hat{\lambda}^{-1}\boldsymbol{\theta}\overline{\mathbf{D}}_0$ represents the stationary distribution of the phase right after the arrival of a batch.

Example 2.2 We illustrate here the computation of the inter-arrival descriptors for the BMAP described in Example 2.1. The stationary probability vector $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta} = (8/17, 5/17, 2/17, 2/17)$. Then, the arrival rates $\lambda_k = \boldsymbol{\theta}\mathbf{D}_k\mathbf{e}_4$ of batches of size k, for $k \in \{1,2\}$, are given by $\lambda_1 = 1.0$ and $\lambda_2 = 1.17647$, while the batch and the total arrival rates are given by $\hat{\lambda} = \lambda_1 + \lambda_2$ and $\lambda = \lambda_1 + 2\lambda_2$, respectively.

By taking $\boldsymbol{\alpha} = \hat{\lambda}^{-1}\boldsymbol{\theta}\overline{\mathbf{D}}_0$, we easily obtain the values $E[\tau_1] = 0.45945$, $Var(\tau_1) = 0.48619$, $\varphi(\tau_1, \tau_5) = 0.00832$ and $\rho(\tau_1, \tau_5) = 0.01711$.

2.3.3. Other descriptors

(i) *Peakedness*. The peakedness functional is a second order descriptor used in communication engineering. It is a functional of the holding time distribution defined as the ratio between the variance and the expectation of the number of busy servers in a queue with infinite servers and independent, identically distributed service times, which is feeded by a certain arrival process. The particular case where the service times are exponentially distributed with rate $\mu > 0$ is called the exponential peakedness.
Eckberg (1983) has shown that the exponential peakedness $z_{exp}(\mu)$ and the Laplace-Stieltjes transform $\phi_{arr}(s)$ of the expected number of arrivals in $(0,t]$, starting from an arbitrary arrival, are related by the formula

$$z_{exp}(\mu) = 1 + \phi_{arr}(\mu) - \frac{\lambda}{\mu}.$$

Following Neuts *et al.* (1992), we observe that the exponential peakedness for the MAP is obtained from the explicit formulas for the kth factorial moments of the number of customers in the $MAP/M/\infty$ queue, which are given by

$$\mathbf{f}_k = k!\boldsymbol{\theta}\mathbf{D}_1(\mu\mathbf{I}_m - \mathbf{D})^{-1}\mathbf{D}_1(2\mu\mathbf{I}_m - \mathbf{D})^{-1}\cdots\mathbf{D}_1(k\mu\mathbf{I}_m - \mathbf{D})^{-1}, \quad k \geq 1.$$

Thus, we have

$$z_{exp}(\mu) = \frac{\mathbf{f}_2\mathbf{e}_m + \mathbf{f}_1\mathbf{e}_m - (\mathbf{f}_1\mathbf{e}_m)^2}{\mathbf{f}_1\mathbf{e}_m}.$$

For the exponential peakedness of the local poissonification of the MAP, we refer the reader to Neuts *et al.* (1992).

(ii) *Index of burstiness*. The term burstiness is referred to an arrival process whose flow exhibits short intervals with a large number of arrivals separated by long

intervals with few arrivals. In order to quantify burstiness, Neuts (1993) proposed to thinning the original arrival process with the help of an auxiliary labeling process.

Assume that the arrival process is a *Markov renewal process* (MRP) whose Markov renewal sequence has a kernel $\mathbf{H}(x) = (h_{ij}(x))$, where the transition probabilities $p_{ij} = h_{ij}(\infty)$ take values on the finite set $\{1, \ldots, r\}$; see Kulkarni (1995). We choose the labeling process to be a stationary MAP independent of the MRP. A point of the MRP is registered if and only if it is immediately preceded by an arrival of the labeling MAP. If the fundamental rate λ decreases, typically only a few arrivals of the MRP are registered. More importantly, the MRP arrivals occurring in intense short runs are most likely to be unregistered. Thus, the proposed labeling mechanism removes the bursts of the MRP.

Suppose that, in the stationary version of the MRP, arrivals occur at rate δ. Let $\boldsymbol{\pi}$ be the invariant distribution of the stochastic matrix $\mathbf{H}(\infty) = (p_{ij})$. Then, we define the index $\chi(p)$ of burstiness by

$$\chi(p) = \frac{1}{\delta}\kappa^{-1}(p), \quad 0 \le p \le 1,$$

where $\kappa^{-1}(p)$ is the inverse function of $\kappa(\lambda)$ defined by

$$\kappa(\lambda) = 1 - \int_0^\infty \boldsymbol{\theta} \exp\{\mathbf{D}_0 u\}\mathbf{e}_m d\,(\boldsymbol{\pi}\mathbf{H}(u)\mathbf{e}_r)\,.$$

Thus, $\delta\chi(p)$ is interpreted as the rate of the MAP labeling process for which a fraction p of the arrivals of the MRP are registered.

In Neuts (1993), the analysis is even extended to investigate correlations and run distributions.

We conclude this subsection by illustrating the calculation of $\chi(p)$ for the *interrupted Poisson process* (IPP).

Example 2.3 An IPP is a bursty MAP with $m = 2$ and matrices

$$\mathbf{D}_0 = \begin{pmatrix} -(\lambda_a + \delta_1) & \delta_1 \\ \delta_2 & -\delta_2 \end{pmatrix}, \qquad \mathbf{D}_1 = \begin{pmatrix} \lambda_a & 0 \\ 0 & 0 \end{pmatrix}.$$

This means that a PP of rate λ_a can be interrupted with probability $\delta_1(\lambda_a + \delta_1)^{-1}$. If this occurs, then an interruption period (exponentially distributed of rate δ_2) takes place.

Assume that the MAP labeling process is Poisson of rate λ. By using the fact that the IPP is equivalent to a certain hyperexponential renewal process (see Milne (1982)), it is easy to find that

$$\kappa(\lambda) = 1 - \frac{\lambda_a(\lambda+\delta_2)}{\lambda^2+\lambda(\lambda_a+\delta_1+\delta_2)+\lambda_a\delta_2}.$$

By normalizing the fundamental rate of the IPP to be one, we obtain the following expression for the index of burstiness:

$$\chi(p) = \frac{p\rho^{-1} - \bar{p}\sigma + \sqrt{(p\rho^{-1}-\bar{p}\sigma)^2 + 4p\bar{p}\sigma}}{2\bar{p}}, \quad 0 < p < 1,$$

where $\bar{p} = 1-p$, $\sigma = \delta_1 + \delta_2$ and $\rho = \delta_2/\sigma$.

2.4. Some applications

The next examples in queueing, reliability and inventory models are intended to help the reader acquire some feeling for the range of applications of the BMAP and its variants. By means of them, we briefly motivate the use of structured Markov chains; see Bini *et al.* (2005), Latouche and Ramaswami (1999), Li (2010) and Neuts (1981,1989).

2.4.1. The BMAP/G/1 queue

Consider a single-server queue whose arrival process is a BMAP with sequence $\{\mathbf{D}_k; k \geq 0\}$. Let the service times have an arbitrary probability distribution function $H(x)$.

We may find many similarities between the $BMAP/G/1$ and the $M/G/1$ queues. To begin with, we construct an embedded Markov chain $\{(Q_n, J_n); n \geq 0\}$ at the times of service completions by defining the pair (Q_n, J_n) as the queue length and the phase of the BMAP immediately after the nth service completion. Define the matrices

$$\mathbf{A}_n = \int_0^\infty \mathbf{P}(n,u)dH(u), \quad n \geq 0,$$

$$\mathbf{B}_n = \sum_{k=1}^{n+1} \int_0^\infty \exp\{\mathbf{D}_0 u\}du\mathbf{D}_k \int_0^\infty \mathbf{P}(n+1-k,v)dH(v)$$

$$= -\mathbf{D}_0^{-1} \sum_{k=1}^{n+1} \mathbf{D}_k \mathbf{A}_{n+1-k}, \quad n \geq 0.$$

The matrix $\mathbf{A}_n = (a_{ij}(n))$ consists of the conditional probabilities that n customers arrive during a service time starting from phase i and finishing at phase j of the BMAP. We can therefore describe some of the transition probabilities for the embedded Markov chain by

$$P(Q_1 = l+n-1, J_1 = j | Q_0 = l, J_0 = i) = a_{ij}(n), \quad n \geq 0, 1 \leq i, j \leq m,$$

independently of $l \geq 1$. It can be readily verified that the matrix generating function $\mathbf{A}^*(z) = \sum_{n=0}^{\infty} z^n \mathbf{A}_n$ is given by

$$\mathbf{A}^*(z) = \int_0^{\infty} \exp\{\mathbf{D}^*(z)u\} dH(u).$$

Similarly, the matrix $\mathbf{B}_n = (b_{ij}(n))$ contains the probabilities that first a batch of k customers arrives and then $n+1-k$ additional customers arrive during the subsequent service time, for $1 \leq k \leq n+1$. Note that this situation occurs whenever a service completion leaves the queue empty. Hence, we can write down

$$P(Q_1 = n, J_1 = j | Q_0 = 0, J_0 = i) = b_{ij}(n), \quad n \geq 0, 1 \leq i, j \leq m.$$

As a result, the one-step transition probability matrix of $\{(Q_n, J_n); n \geq 0\}$ is given by

$$\mathbf{P} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \dots \\ \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_3 & \dots \\ & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \dots \\ & & \mathbf{A}_0 & \mathbf{A}_1 & \dots \\ & & & \ddots & \ddots \end{pmatrix}.$$

A matrix of this structured form is said to be of $M/G/1$-type (see Neuts (1989)), which underlines the similarity to the univariate embedded Markov chain of the $M/G/1$ queue.

The $BMAP/G/1$ was first analyzed in Ramaswami (1980), where the BMAP was used under its older, more complicated notation. An outline of Ramaswami's results under the present matrix formulation, along with some new results, are presented in Lucantoni (1991). For a historical survey on the model, see Lucantoni (1993).

2.4.2. The D-BMAP/D/1/K queue

Consider a discrete-time queue in which arrivals are generated by M independent input sources. Incoming arrivals are queued in a shared buffer of capacity K, with $K < M$. The time needed to serve an arrival is selected as time unit and named slot. Each input source in a slot takes either ON state or OFF state. When an input source is in ON state, one arrival is generated with probability g. If the source is in OFF state, then no arrival is generated. Suppose also that any OFF (or ON) source in a time slot changes to the ON (or OFF) state with probability p (or q) in the next slot. This superposition of sources can be modelled as a *discrete-time batch Markovian arrival process* (D-BMAP); see Subsection 3.1.

Let Q_n and J_n be the queue length and the number of ON sources (phase) at the nth slot. Then, the sequence $\{(Q_n, J_n); n \geq 1\}$ is a discrete-time Markov chain on the state space $\{0, 1, \dots, K\} \times \{0, 1, \dots, M\}$ with one-step transition probability matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_{K-1} & \sum_{k=K}^{M} \mathbf{D}_k \\ \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_{K-1} & \sum_{k=K}^{M} \mathbf{D}_k \\ & \mathbf{D}_0 & \mathbf{D}_1 & \cdots & \mathbf{D}_{K-2} & \sum_{k=K-1}^{M} \mathbf{D}_k \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & & \mathbf{D}_0 & \sum_{k=1}^{M} \mathbf{D}_k \end{pmatrix},$$

where the matrices $\mathbf{D}_k$ have the following elements:

$$d_{ii'}(k) = \binom{i}{k} g^k (1-g)^{i-k} f_{ii'}, \quad 0 \le k \le i,$$

and $f_{ii'}$, for $0 \le i, i' \le M$, is given by

$$f_{ii'} = \sum_{j=0}^{i} \binom{i}{j} q^j (1-q)^{i-j} \binom{M-i}{i'+j-i} p^{i'+j-i} (1-p)^{M-i'-j}.$$

The binomial term in $d_{ii'}(k)$ is the probability of k arrivals in the current slot, given that the number of ON sources is i. On the other hand, $f_{ii'}$ is the probability that in the next slot there will be i' ON sources, given that in the current slot there are i.

The structure of $\mathbf{P}$ shows that $\{(Q_n, J_n); n \ge 1\}$ is a finite Markov chain of $M/G/1$-type. This structured Markov chain, but involving a more sophisticated sequence $\{\mathbf{D}_k; k \ge 0\}$, is the analytical model used by Blondia and Casals (1992) for a statistical multiplexer whose input consists of the superposition of *variable bit rate* (VBR) sources.

2.4.3. A reliability system subject to failures

Consider a system subject to internal and external failures. An internal failure causes a fatal failure of the system and implies that the system must be replaced. External failures affect the system in two ways: some of them cause damage that can be repaired, whereas others cause fatal failure and consequently the system must be replaced. Assume that the replacement and repair operations are instantaneous.

In practice, it is frequent that a system can bear only a certain number of failures, in such a way that when the next failure occurs it is replaced. Let $k \ge 1$ be the maximum number of imperfect repairs that the system can undergo. At an arbitrary time, the state of the system can be described by means of the number $K(t)$ of imperfect repairs suffered by the system in process at time t. The random variable $K(t)$ takes values in the set $\{0, 1, \ldots, k\}$ and, in particular, it records the state 0 if the system in process at time t is new.

Montoro-Cazorla and Pérez-Ocón (2006) use a matrix-analytic approach when the lifetime of the system due to wear out follows a PH distribution, with representation

$(\boldsymbol{\tau}, \mathbf{T})$ of order n. Arrivals of external failures are modelled by a MMAP (see Subsection 3.2) with two types of marks referring to external failures with minimal repair and external failures causing a replacement. In the characteristic matrices $\{\mathbf{D}_0, \mathbf{D}_1, \mathbf{D}_2\}$ of dimension m, the matrix $\mathbf{D}_1$ refers to the occurrence of an external failure with minimal repair, and $\mathbf{D}_2$ refers to a failure that causes the replacement of the system. The matrix $\mathbf{D}_0$ records those changes that do not imply any failure.

Then, a Markovian description of the system state follows from the Markov chain $\{(K(t), J_l(t), J_a(t)); t \geq 0\}$, where $J_l(t)$ and $J_a(t)$ denote the lifetime phase and the phase of the arrival process, respectively, at time t. This is a Markov chain on the space state $\{0, 1, \ldots, k\} \times \{1, \ldots, n\} \times \{1, \ldots, m\}$ and infinitesimal generator

$$\mathbf{Q} = \begin{pmatrix} (\mathbf{T} + \mathbf{t}\boldsymbol{\tau}) \oplus \mathbf{D}_0 + \mathbf{e}_n\boldsymbol{\tau} \otimes \mathbf{D}_2 & \mathbf{I}_n \otimes \mathbf{D}_1 & & & \\ \mathbf{t}\boldsymbol{\tau} \otimes \mathbf{I}_m + \mathbf{e}_n\boldsymbol{\tau} \otimes \mathbf{D}_2 & \mathbf{T} \oplus \mathbf{D}_0 & \mathbf{I}_n \otimes \mathbf{D}_1 & & \\ \vdots & & \ddots & \ddots & \\ \mathbf{t}\boldsymbol{\tau} \otimes \mathbf{I}_m + \mathbf{e}_n\boldsymbol{\tau} \otimes \mathbf{D}_2 & & & \mathbf{T} \oplus \mathbf{D}_0 & \mathbf{I}_n \otimes \mathbf{D}_1 \\ \mathbf{t}\boldsymbol{\tau} \otimes \mathbf{I}_m + \mathbf{e}_n\boldsymbol{\tau} \otimes (\mathbf{D}_1 + \mathbf{D}_2) & & & & \mathbf{T} \oplus \mathbf{D}_0 \end{pmatrix}.$$

Therefore, the structural form of $\mathbf{Q}$ yields a finite Markov chain of $GI/M/1$-type; see Neuts (1981).

2.4.4. A multi-location inventory system

The next example (see Ching (1997)) is an inventory system in a multi-location situation under continuous review and one-for-one replenishment.

Consider a multi-location inventory system consisting of K locations that replenish their stocks from a common main depot. For the ith location, the inventory system is modelled by the $M/M/s_i/q_i$ queue with arrival rate λ_i and exponentially distributed lead times of each server with parameter μ_i. The overflow process of demand of the ith location can be approximated by a two-state MMPP with underlying matrices

$$\mathbf{Q}_{ia} = \begin{pmatrix} -\sigma_{i1} & \sigma_{i1} \\ \sigma_{i2} & -\sigma_{i2} \end{pmatrix}, \qquad \boldsymbol{\Lambda}_i = \begin{pmatrix} \lambda_i & 0 \\ 0 & 0 \end{pmatrix}.$$

The first state is equivalent to the event $\{$*the ith location is full*$\}$, and the second one amounts to the event $\{$*the ith location is not yet full*$\}$. Note that, in the former case, the maximum level of backlogs is attained and, consequently, a further demand will overflow to the main depot whenever the queue remains full. In the latter case, a further demand will be acceptable. Based on the stationary distribution of the $M/M/s_i/q_i$ queue, the parameters σ_{i1} and σ_{i2} are approximated as $\sigma_{i1} = s_i\mu_i$ and $\sigma_{i2} = b_i s_i \mu_i/(1 - b_i)$, where b_i denotes the blocking probability at the ith location

$$b_i = \sum_{j=-q_i}^{s_i} \prod_{k=1}^{s_i-j} \frac{\lambda_i}{\mu_i \min(k, s_i)}.$$

Therefore, we may regard the $MMPP/M/s/q$ queue describing the inventory system at the main depot as a finite Markov chain $\{(Q(t), J(t)); t \geq 0\}$ on the state space $\{-q, \ldots, s\} \times \{1, \ldots, 2^K\}$, where $Q(t)$ is the inventory level at the depot and $J(t)$ is the phase of the underlying Markov chain with infinitesimal generator $\mathbf{Q}_a = \mathbf{Q}_{1a} \oplus \ldots \oplus \mathbf{Q}_{Ka}$. Negative values for the inventory level $Q(t)$ amount to backlog.

The infinitesimal generator $\mathbf{Q}$ of $\{(Q(t), J(t)); t \geq 0\}$ has the following structured form:

$$\begin{pmatrix}
\mathbf{Q}_a - \mathbf{\Lambda} & \mathbf{\Lambda} & & & & & & \\
\mu \mathbf{I}_{2^K} & \mathbf{Q}_a - \mathbf{\Lambda} - \mu \mathbf{I}_{2^K} & \mathbf{\Lambda} & & & & & \\
 & \ddots & \ddots & \ddots & & & & \\
 & & s\mu \mathbf{I}_{2^K} & \mathbf{Q}_a - \mathbf{\Lambda} - s\mu \mathbf{I}_{2^K} & \mathbf{\Lambda} & & & \\
 & & & s\mu \mathbf{I}_{2^K} & \mathbf{Q}_a - \mathbf{\Lambda} - s\mu \mathbf{I}_{2^K} & \mathbf{\Lambda} & & \\
 & & & & \ddots & \ddots & & \\
 & & & & & s\mu \mathbf{I}_{2^K} & \mathbf{Q}_a - s\mu \mathbf{I}_{2^K}
\end{pmatrix},$$

where $\mathbf{\Lambda} = \mathbf{\Lambda}_1 \oplus \ldots \oplus \mathbf{\Lambda}_K$.

The stationary distribution of $\mathbf{Q}$ can be readily derived from the general theory of finite QBD processes; see e.g. Latouche and Ramaswami (1999, Chapter 10). For more information on finite QBD processes arising in manufacturing problems, the reader is referred to the monograph by Ching (2001).

3. Variants and extensions of the BMAP

In this section we collect several generalizations and variants of the BMAP. We start in Subsection 3.1 by presenting the D-BMAP; that is, the discrete-time analogue of the BMAP. The use of discrete-time models is motivated by many applications in communication systems where the basic units are digital. The consideration of Markov arrival processes with marked transitions opens new directions to investigate stochastic models with multiple types of items, fluid input, spatial arrivals, etc. In Subsection 3.2 we follow the original formulation by He and Neuts (1998) to introduce the MMAP. The HetSigma approach summarized in Subsection 3.3 provides a versatile way to get joint modulation of the arrival and service processes. In Subsection 3.4, under the title Markov-additive arrival processes, we briefly introduce some generalized arrival processes which allow the counting/marked and background processes to take values on more general spaces. The time-inhomogeneous case and the possibility of incorporating spatial features can also be subsumed under appropriate versions of the

Markov-additive umbrella. Finally, in Subsection 3.5 we deal with the BSDE approach which has been recently presented by Artalejo and Gómez-Corral (2010) as a tool for constructing Markov modulated stochastic models taking into account the reduction of dimensionality inherent to the matrix formulation.

3.1. The D-BMAP

The D-BMAP was introduced by Blondia and Casals (1992) as the discrete-time analogue of the BMAP. They showed that many useful discrete-time arrival processes can be obtained as particular cases of the D-BMAP and how this versatile arrival pattern can be used as *asynchronous transfer mode* (ATM) source model.

The key point in the constructive description of the D-BMAP is the consideration of finite matrices $\{\mathbf{D}_k; k \geq 0\}$, which govern phase transitions and batch sizes. Suppose that at time k the phase in progress is i, for $1 \leq i \leq m$. At the next time epoch $k+1$, a transition to another or the same phase takes place and a batch arrival may occur or not. More concretely, the elements $d_{ij}(0)$ of matrix $\mathbf{D}_0$ give the probabilities that the phase goes to state j with no arrival, given that the initial phase is i. On the other hand, the elements $d_{ij}(k)$ of $\mathbf{D}_k$ denote that, in the next time unit, there is a transition from phase i to phase j with a batch of size $k \geq 1$. We notice that

$$\sum_{j=1}^{m} \sum_{k=0}^{\infty} d_{ij}(k) = 1, \quad 1 \leq i \leq m.$$

We also assume that the matrix $\mathbf{I}_m - \mathbf{D}_0$ is non-singular, so the D-BMAP has an arrival with probability one.

With the help of $\{\mathbf{D}_k; k \geq 0\}$, we formally define the D-BMAP as the bivariate process $\{(N_k, J_k); k \geq 0\}$, where $\{J_k; k \geq 0\}$ is the background phase Markov chain and N_k denotes the counting variable. The one-step transition probability matrix of the D-BMAP is given by

$$\mathbf{P} = \begin{pmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \cdots \\ & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots \\ & & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ & & & \ddots & \ddots \end{pmatrix}.$$

A number of well-known processes are obtained by choosing appropriately the sequence of matrices $\{\mathbf{D}_k; k \geq 0\}$. The list includes the *Bernoulli arrival process*, the *Markov modulated Bernoulli process*, the *batch Bernoulli process with correlated arrivals* and many other processes which, in general, can be considered as the discrete counterparts of those particular cases of the BMAP listed in Subsection 2.1. For further details of other special cases of the D-BMAP, we refer to the papers by Chakravarthy (2001,2010).

We also remark that, like in the continuous-time BMAP, many interesting properties (such as counting, descriptors, superpositions, etc.) can be investigated. Since arguments are similar, these results will not be presented here, but we refer to the paper by Chakravarthy (2010) for a summary of basic results for the D-BMAP.

In what follows, we focus on the class of *platoon arrival processes* (PAP).

Example 3.1 The following description of the PAP is based on the paper by Alfa and Neuts (1995), who used the PAP to model vehicular traffic. Recently, Breuer and Alfa (2005) used a terminating D-MAP to generalize the concept of PAP.

The PAP is a discrete-time arrival process composed of platoons. Suppose that the number of arrivals in a platoon is a discrete PH of order d with representation $(\boldsymbol{\delta}, \mathbf{D})$ and absorption vector $\mathbf{d}$. Moreover, we assume that $p_1 = \delta_0 = 1 - \boldsymbol{\delta}\mathbf{e}_d > 0$ is the probability of a platoon consisting of a single vehicle (i.e., the probability of starting in the absorbing state) and $p_k = \boldsymbol{\delta}\mathbf{D}^{k-2}\mathbf{d}$, for $k \geq 2$, is the probability of having k arrivals in the platoon. In a first general approach, intraplatoon intervals separating two arrivals in the same platoon, have the probability mass function $\{p_1(k); k \geq 1\}$. On the other hand, the interplatoon interval separating the last arrival in a platoon and the first one of the immediately following platoon, have the probability mass function $\{p_2(k); k \geq 1\}$.

Let S_n be the nth arrival epoch and suppose that Y_n records the phase of the discrete PH distribution observed at time S_n+, whose representation is given by $(\boldsymbol{\delta}, \mathbf{D})$. Then, the PAP is the MRP associated with the Markov renewal sequence $\{(Y_n, S_n); n \geq 0\}$, whose kernel is described by the matrices

$$\mathbf{H}(j) = \begin{pmatrix} \delta_0 p_2(j) & \boldsymbol{\delta} p_2(j) \\ \mathbf{d} p_1(j) & \mathbf{D} p_1(j) \end{pmatrix}, \quad j \geq 1.$$

For practical purposes, the MRP formalism can be simplified by assuming that the intraplatoon intervals and the interplatoon intervals are distributed as discrete PH distributions with representations $(\boldsymbol{\alpha}_i, \mathbf{T}_i)$ with m_i phases and absorption vectors $\mathbf{t}_i$, for $i \in \{1,2\}$, respectively. The vectors $\boldsymbol{\alpha}_i$, for $i \in \{1,2\}$, are now assumed to be probability vectors. Thus, the PAP can be now seen as a D-MAP with matrices $\mathbf{D}_0$ and $\mathbf{D}_1$ given by

$$\mathbf{D}_0 = \begin{pmatrix} \mathbf{T}_2 & \mathbf{0}_{m_2 \times dm_1} \\ \mathbf{0}_{dm_1 \times m_2} & \mathbf{I}_d \otimes \mathbf{T}_1 \end{pmatrix}, \qquad \mathbf{D}_1 = \begin{pmatrix} \delta_0 \mathbf{t}_2 \boldsymbol{\alpha}_2 & \boldsymbol{\delta} \otimes \mathbf{t}_2 \boldsymbol{\alpha}_1 \\ \mathbf{d} \otimes \mathbf{t}_1 \boldsymbol{\alpha}_2 & \mathbf{D} \otimes \mathbf{t}_1 \boldsymbol{\alpha}_1 \end{pmatrix},$$

where the underlying states (i, j) denote the phase of the discrete PH law with representation $(\boldsymbol{\delta}, \mathbf{D})$ and the phase of the (interplatoon or intraplatoon) interval in process.

3.2. The marked Markovian arrival process

The MMAP can be viewed as a multi-class extension of the BMAP. Although the analysis can be presented both in discrete- and continuous-time, we restrict our exposition

to the latter case. Similar to the BMAP, the MMAP definition is based on a background Markov chain $\{J(t); t \geq 0\}$, often called phase chain, with m states, which determines the arrivals of some marks taking values on a set $\mathscr{C}^0$. The set of marks $\mathscr{C}^0$ may have different interpretations, as we show in the sequel.

Let $\mathscr{C}^0$ be a finite or countable set of indices. More specifically, we may assume that a generic element $\mathbf{h}$ of $\mathscr{C}^0$ is a K-tuple $(h_1, \ldots, h_K)$, where $h_k \in \mathbb{N}$, for $1 \leq k \leq K$, and at least one coordinate is strictly positive. Define the non-negative matrices $\mathbf{D}_0$ and $\{\mathbf{D}_\mathbf{h}; \mathbf{h} \in \mathscr{C}^0\}$ of order m. The entries of $\mathbf{D}_0$ describe the motion of the phase Markov chain without any arrival. $\mathbf{D}_0$ is assumed to be a non-singular matrix with negative diagonal elements. The matrices $\mathbf{D}_\mathbf{h}$ are non-negative and give the transition rates of the phase Markov chain with a mark $\mathbf{h}$. Then, $\mathbf{D} = \mathbf{D}_0 + \sum_{\mathbf{h} \in \mathscr{C}^0} \mathbf{D}_\mathbf{h}$ is an infinitesimal generator. The counting process $\{(N_\mathbf{h}(t), J(t)); \mathbf{h} \in \mathscr{C}^0, t \geq 0\}$ is called a MMAP.

Alternatively, we may define the MMAP in terms of PPs. To this end, it is enough to replace the role of the rates $\{d_{ij}(k); 1 \leq i, j \leq m\}$, for $k \geq 1$, in Remark 2.2 by the analogue marked version $\{d_{ij}(\mathbf{h}); 1 \leq i, j \leq m\}$, for $\mathbf{h} \in \mathscr{C}^0$. The semi-Markovian representation in Remark 2.1 for the BMAP also holds for the MMAP.

It is clear that the choice $K = 1$ and $\mathscr{C}^0 = \mathbb{N} - \{0\}$ reduces the MMAP to the BMAP. The case $K = 1$ and $\mathscr{C}^0 = \{1, \ldots, C\}$ determines arrivals of C different types of customers or items; that is, the MMAP is interpreted as a proper multi-class generalization of the BMAP.

The following specifications of the matrices $\mathbf{D}_0$ and $\{\mathbf{D}_\mathbf{h}; \mathbf{h} \in \mathscr{C}^0\}$ show interesting features captured under the MMAP formulation:

(i) A reinterpretation of the batch sizes in terms of different classes of customers allows us to see example (ii) for cyclic arrivals in Section 2 as an arrival process where type-1 and type-2 customers arrive cyclically.

(ii) Individual vs group

$$\mathbf{D}_0 = \begin{pmatrix} -1 & 0 \\ 0 & -2 \end{pmatrix}, \quad \mathbf{D}_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{D}_{2,1} = \begin{pmatrix} 0 & 0.5 \\ 1 & 0 \end{pmatrix}.$$

First, we notice that the marks $\mathscr{C}^0 = \{\{1\}, \{2,1\}\}$ can be put in correspondence with the case $K = 1$ and $\mathscr{C}^0 = \{1, 2\}$. This comment can be readily extended to any arbitrary finite set $\mathscr{C}^0$.

In this arrival process, there are individual arrivals of type-1 and group arrivals where the group consists of one type-2 customer accompanied by a type-1 customer.

(iii) Type-2 follows type-1

$$\mathbf{D}_0 = \begin{pmatrix} -4 & 0 \\ 0 & -5 \end{pmatrix}, \quad \mathbf{D}_1 = \begin{pmatrix} 3 & 1 \\ 0 & 0 \end{pmatrix}, \quad \mathbf{D}_{2,1} = \begin{pmatrix} 0 & 0 \\ 5 & 0 \end{pmatrix}.$$

A group arrival $\{2,1\}$ is always preceded by the arrival of a customer of type-1.

(iv) Orders within batches

$$\mathbf{D}_0 = \begin{pmatrix} -15 & 0 \\ 0 & -10 \end{pmatrix}, \quad \mathbf{D}_{\{112\}} = \begin{pmatrix} 14 & 0 \\ 0 & 9 \end{pmatrix}, \quad \mathbf{D}_{\{121\}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The marks $\{112\}$ and $\{121\}$ are associated with group arrivals of size 3. Each group consists of two type-1 customers and one customer of type-2. The orders in which individuals are scheduled within a group do matter, so the two marks are distinguished.

Among the descriptors of the MMAP, we stress the interest in the counting functions. The generating function of $\mathbf{N}(t) = (N_1(t), \ldots, N_K(t))$ is given by

$$\mathbf{P}^*(\mathbf{z},t) = \sum_{\mathbf{n}} \mathbf{z}^{\mathbf{n}} \mathbf{P}(\mathbf{n},t) \; = \; \exp\{\mathbf{D}^*(\mathbf{z})t\},$$

where $\mathbf{n} = (n_1, \ldots, n_K)$ with $n_i \geq 0$, for $1 \leq i \leq K$, and $\mathbf{P}(\mathbf{n},t)$ is the matrix with elements $P_{ij}(\mathbf{n},t) = P(\mathbf{N}(t) = \mathbf{n}, J(t) = j | \mathbf{N}(0) = \mathbf{0}_K, J(t) = i)$, while $\mathbf{z}^{\mathbf{n}} = z_1^{n_1} \cdots z_K^{n_K}$ and $\mathbf{D}^*(\mathbf{z}) = \mathbf{D}_0 + \sum_{\mathbf{h} \in \mathscr{C}^0} \mathbf{z}^{\mathbf{h}} \mathbf{D}_{\mathbf{h}}$, for $|z_k| \leq 1$ and $1 \leq k \leq K$.

Now the covariances and correlations between $\{N_{\mathbf{h}}(t); t \geq 0\}$, for $\mathbf{h} \in \mathscr{C}^0$, can be explicitly expressed; see He and Neuts (1998).

For easiness, we assume $\mathscr{C}^0 = \{1,2\}$; i.e., we have two types of arrivals.

Given any initial probability distribution $\boldsymbol{\alpha}$ for the phase Markov chain, we have

$$E[N_{\mathbf{h}}(t)] = \lambda_{\mathbf{h}} t + \boldsymbol{\alpha} (\exp\{\mathbf{D}t\} - \mathbf{I}_m) (\mathbf{D} - \mathbf{e}_m \boldsymbol{\theta})^{-1} \mathbf{D}_{\mathbf{h}} \mathbf{e}_m, \quad \mathbf{h} \in \mathscr{C}^0, \; t \geq 0,$$

where $\boldsymbol{\theta}$ is the stationary distribution of $\mathbf{D}$ and $\lambda_{\mathbf{h}} = \boldsymbol{\theta} \mathbf{D}_{\mathbf{h}} \mathbf{e}_m$ is the fundamental arrival rate of type-$\mathbf{h}$ marks.

If we take $\boldsymbol{\alpha} = \boldsymbol{\theta}$, then

$$\begin{aligned} Var(N_{\mathbf{h}}(t)) = & \left(\lambda_{\mathbf{h}} - 2\lambda_{\mathbf{h}}^2 - 2\boldsymbol{\theta} \mathbf{D}_{\mathbf{h}} (\mathbf{D} - \mathbf{e}_m \boldsymbol{\theta})^{-1} \mathbf{D}_{\mathbf{h}} \mathbf{e}_m \right) t \\ & + 2\boldsymbol{\theta} \mathbf{D}_{\mathbf{h}} (\mathbf{D} - \mathbf{e}_m \boldsymbol{\theta})^{-1} (\exp\{\mathbf{D}t\} - \mathbf{I}_m) (\mathbf{D} - \mathbf{e}_m \boldsymbol{\theta})^{-1} \mathbf{D}_{\mathbf{h}} \mathbf{e}_m, \end{aligned}$$

and the covariance between $N_1(t)$ and $N_2(t)$ is given by

$$\begin{aligned} \varphi(N_1(t), N_2(t)) = & - \left(2\lambda_1 \lambda_2 + \boldsymbol{\theta} \left(\sum_{k=1}^{2} \mathbf{D}_k (\mathbf{D} - \mathbf{e}_m \boldsymbol{\theta})^{-1} \mathbf{D}_{3-k} \right) \mathbf{e}_m \right) t \\ & + \boldsymbol{\theta} \left(\sum_{k=1}^{2} \mathbf{D}_k (\mathbf{D} - \mathbf{e}_m \boldsymbol{\theta})^{-1} (\exp\{\mathbf{D}t\} - \mathbf{I}_m) (\mathbf{D} - \mathbf{e}_m \boldsymbol{\theta})^{-1} \mathbf{D}_{3-k} \right) \mathbf{e}_m. \end{aligned}$$

We illustrate the computation of the counting moments by means of the BMAP considered in Examples 2.1 and 2.2. Obviously, the batch size becomes here the mark in the MMAP terminology.

Example 3.2 If the MMAP with matrices $\{\mathbf{D}_0, \mathbf{D}_1, \mathbf{D}_2\}$ given in Example 2.1 is stationary, for $t = 2.5$, we get

$$E[N_1(t)] = 2.5, \qquad E[N_2(t)] = 2.94117,$$
$$Var(N_1(t)) = 4.24980, \qquad Var(N_2(t)) = 6.17905.$$

The covariance and correlation between $N_1(t)$ and $N_2(t)$ are given by

$$\varphi(N_1(t), N_2(t)) = 3.30791, \qquad \rho(N_1(t), N_2(t)) = 0.64551.$$

The mean and variance of the total number of counts $N(t) = N_1(t) + 2N_2(t)$ are $E[N(t)] = 8.38235$ and $Var(N(t)) = 42.19772$.

A good account of results for other basic properties of the MMAP, including thinning, type of arrivals, peakedness and closure properties, are found in He and Neuts (1998), and He (2010).

3.3. The HetSigma approach

The HetSigma approach (see Chakka and Do (2007)) has been proposed in order to evaluate the performance of queueing models with burstiness and correlation arising from applications to wireless broadband networks. The proposed modulation mechanism could be subsumed under a MMAP pattern. However, the HetSigma approach presents some interesting features which justify its presentation in this specific subsection.

In the HetSigma approach both the arrival and service processes are modulated in continuous-time by a single infinitesimal generator $\mathbf{Q}_{as}$, with m modulating phase states. This assumption includes as a particular case the situation where the arrival and service processes are modulated individually by infinitesimal generators $\mathbf{Q}_a$ and $\mathbf{Q}_s$ with m_a and m_s phases, respectively. This independent modulation case can be converted into a joint modulation by taking $\mathbf{Q}_{as} = \mathbf{Q}_a \oplus \mathbf{Q}_s$ and $m = m_a m_s$.

Arrivals, under each modulating phase i, consist of the superposition of K independent CPPs of positive arrivals and an independent CPP of negative arrivals. More concretely, the $K+1$ CPPs are described in terms of *generalized exponential* (GE) distributions, which govern exponential inter-arrival times with batches having geometric size distribution. For example, during phase i, the stream of negative arrivals follows a GE distribution with representation (ρ_i, δ_i), which means that a negative batch arrives to the system after an exponential time of rate ρ_i, and its size is $k \geq 1$ with probability $(1-\delta_i)\delta_i^{k-1}$. On the other hand, the service facility has c heterogeneous servers. Each

server is labeled and has its own independent GE service time with parameters (μ_{in}, ϕ_{in}), for $1 \leq n \leq c$ and $1 \leq i \leq m$.

The model description must be completed with a number of queueing specifications including the first come first scheduled for service discipline, a switching policy guaranteeing that the servers labeled with lowest indexes are those rendering service, a killing policy which removes customers at the end of the queue when a negative arrival takes place, and other necessary specifications which are described in detail in Chakka and Do (2007).

3.4. Markov-additive processes of arrivals

In this subsection, we follow Pacheco and Prabhu (1995) to introduce the class of *Markov-additive processes of arrivals*. First of all, we remark that the acronym MAP is used in the literature both for the Markovian arrival process introduced in Section 2 and for the Markov-additive processes of arrivals. For the sake of clarity, here we shall denote the latter as MAPA.

A MAPA is a Markov process with two components X and J. In general, X is a non-Markovian component called the additive component since increments of X correspond to arrivals. The Markov component J sometimes represents an environment factor. In other applications, the phenomenon under study leads naturally to the pair (X,J).

The state space assumed in Pacheco and Prabhu (1995) is $\mathscr{S} = \mathbb{R}^r \times E$, where E is a discrete set. Moreover, it is also assumed that (X,J) is a continuous-time process. Then, a process $(X,J) = \{(X(t),J(t)); t \geq 0\}$ on $\mathscr{S}$ is a MAPA if

(i) (X,J) is a Markov process.
(ii) For all $s \geq 0$ and $t \geq 0$, the conditional distribution of $(X(t+s) - X(s), J(t+s))$, given $(X(s),J(s))$, depends only on $J(s)$.

The above definition follows the spirit of Çinlar (1972a,b), who assumed a more general space E. It is convenient to extend E including a special state Δ which indicates the termination of the process (X,J). Some interesting properties including closure properties under linear transformations and linear combinations can be investigated. On the other hand, to study the lack of memory property, inter-arrival times, moments of the number of counts and other structural properties, it is convenient to reduce to the state space $\mathscr{S} = \mathbb{N}^r \times E$. In this context, the dynamics of the MAPA comprise three types of transitions: (a) arrivals without change of state in J; (b) changes of state in J without arrivals; and (c) arrivals with change of state in J.

Secondary recording of the MAPA is a mechanism that generates a secondary arrival process from the original arrival process. This mechanism includes interesting features like thinning and marking.

Closely related to the MAPA is the class of MMAPs defined for the case where E is finite; see Subsection 3.2. The BMAP corresponds to the simple case with $r = 1$ and $E = \{1,\ldots,m\}$.

The contribution by Pacheco and Prabhu (1995) is generalized in Breuer (2003) to cover the inhomogeneous case. The inhomogeneous BMAP is defined as a MAPA (X,J) with additive space $\mathbb{N}$, finite phase space $E=\{1,\ldots,m\}$ and time-inhomogenous structure for the generator functions

$$\mathbf{Q}(t)=\begin{pmatrix} \mathbf{D}_0(t) & \mathbf{D}_1(t) & \mathbf{D}_2(t) & \mathbf{D}_3(t) & \cdots \\ & \mathbf{D}_0(t) & \mathbf{D}_1(t) & \mathbf{D}_2(t) & \cdots \\ & & \mathbf{D}_0(t) & \mathbf{D}_1(t) & \cdots \\ & & & \ddots & \ddots \end{pmatrix},$$

where the (i,j)th entry of $\mathbf{D}_k(t)$ can be interpreted as the infinitesimal transition rate of recording k arrivals during the infinitesimal interval $(t,t+dt]$ while changing from phase i to phase j. Likewise, other interpretations for BMAPs can be adapted to the time-inhomogeneous case. For example, the matrix $\mathbf{D}(t)=\sum_{k=0}^{\infty}\mathbf{D}_k(t)$ is a generator for all $t\geq 0$. If the phase process J has a stationary distribution $\boldsymbol{\theta}$, then starting the phase process in this distribution without prior arrivals yields the following expression for the mean number of arrivals until time t:

$$\int_0^t \boldsymbol{\theta}\sum_{k=1}^{\infty} k\mathbf{D}_k(u)\mathbf{e}_m du.$$

Breuer (2003) also generalizes the notion of characteristic sequence slightly in order to define a class of fluid MAPs. In this generalization, the phase space is finite $E=\{1,\ldots,m\}$ and the additive space is given by $[0,\infty)$. Unlike the additive space $\mathbb{N}$ which allows us to arrange the matrices containing arrival rates in a single sequence, an analogue for the additive space $[0,\infty)$ is a characteristic measure $\boldsymbol{\Delta}$ providing an arrival rate matrix for every Borel-measurable subset of $[0,\infty)$. For the homogeneous fluid MAP, the measure $\boldsymbol{\Delta}$ is specified by the matrices $\boldsymbol{\Delta}(x)$, whose (i,j)th elements are given by the corresponding infinitesimal transition rates $q(i;[0,x]\times\{j\})$, for $x\geq 0$ and $1\leq i,j\leq m$. Thus, the matrix $\boldsymbol{\Delta}(x)$ has an analogous meaning as the matrix $\mathbf{D}_k$ for the BMAP. The infinitesimal generator of J is given by $\mathbf{D}=\lim_{x\to\infty}\boldsymbol{\Delta}(x)$. Let $\boldsymbol{\theta}$ be its stationary probability vector. Then,

$$\int_0^{\infty}\boldsymbol{\theta}u d\boldsymbol{\Delta}(u)\mathbf{e}_m t$$

gives the expected number of arrivals until time t, if the process starts without prior arrivals and in phase equilibrium $\boldsymbol{\theta}$. It can be also shown that $\lim_{t\to\infty}X(t)/t=\int_0^{\infty}\boldsymbol{\theta}u d\boldsymbol{\Delta}(u)\mathbf{e}_m$, almost surely for all initial phase distributions.

The concept of BMAP can be even generalized towards a class of time-space processes, called spatial MAPs; see Breuer (2003, Chapters 7-9), and Breuer and Baum (2005, Chapter 14). This generalization addresses three essential points: (a) the phase

state E is allowed to be general; (b) the generator functions of the spatial MAP may depend on time; and (c) arrivals may assume a location in some space.

Based on an underlying MAPA, Sengupta (1989) defines a bivariate Markov process (X,J) with a special structure, which can be seen as a continuous-time and continuous-space version of the Markov chains of $GI/M/1$-type studied by Neuts (1981). The *Sengupta process* yields a notably simplified characterization of the waiting time and the queue length distributions in the $GI/PH/1$ queue. Specifically, the phase space is finite $E=\{1,\ldots,m\}$, and the additive component X is skip-free to the right, takes values in $[0,\infty)$ and increases at a linear rate of 1, if there is no downward jump. Moreover, changes in the state of the process (X,J) may also occur in one of two ways:

(i) If $(X(t),J(t))=(x,i)$, then (X,J) may change its state to somewhere between $(x-u,j)$ and $(x-u+du,j)$ at a rate of $da_{ij}(u)$, for $u\in[0,x)$ and $1\le i,j\le m$.
(ii) If $(X(t),J(t))=(x,i)$, then it may transit from (x,i) to $(0,j)$ at a rate of $b_{ij}(x)$, for $x>0$ and $1\le i,j\le m$.

The level-dependent rates $a_{ij}(x)$ and $b_{ij}(x)$ satisfy the condition

$$\sum_{j=1}^{m}(a_{ij}(x)+b_{ij}(x))=-d_i,\quad x>0,\ 1\le i\le m,$$

where $-d_i$ is the rate at which the next state change can occur from the initial state (x,i). This equality clearly implies that the probability that the additive component X takes a downward jump of $u\in[0,x)$ units from x, given that a downward jump occurs, does not depend on the initial level x.

For a related work, we also refer to the bivariate Markov process (X,J) analyzed by Tweedie (1982), where the additive component X takes values in $\mathbb{N}$ and the Markov component J takes values on a general set such as an interval of the real line.

3.5. The BSDE approach

The rationale for using Markovian arrival processes and PH distributions has been already discussed in Section 2. However, the price to be paid frequently in practice is a significant burden on computational time and memory needed due to the high dimensionality of the resulting block-structured Markov chains. The complexity of the underlying stochastic models increases drastically in non-homogeneous settings, where an arbitrary, even infinite number of MAPs and/or PH distributions could be involved. The BSDE approach provides a versatile tool to deal with a non-exponential model with correlated flows, but keeping the dimensionality of the block-structured Markov chain tractable.

In the BSDE approach, we are concerned with a multidimensional continuous-time Markov chain $(\mathbf{X},\mathbf{Y})=\{(X_1(t),\ldots,X_k(t),Y_1(t),\ldots,Y_l(t));t\ge 0\}$. We assume

that $(\mathbf{X},\mathbf{Y})$ is regular and time-homogeneous; in applications, it is often assumed to be irreducible. The sub-vector $\mathbf{X}(t)=(X_1(t),\ldots,X_k(t))$ provides a k-dimensional description of the fundamental aspects of the system state at time t. On the other hand, the sub-vector $\mathbf{Y}(t)=(Y_1(t),\ldots,Y_l(t))$ is a l-dimensional phase vector which completes the Markovian system description. The state space of $(\mathbf{X},\mathbf{Y})$ is a discrete set $\mathscr{S}_{(\mathbf{X},\mathbf{Y})}$ with $(k+l)$-dimensional elements.

The sojourn time $E_{(\mathbf{x},\mathbf{y})}$ that the Markov chain remains in the state $(\mathbf{x},\mathbf{y})$ is exponentially distributed with rate $\lambda_{(\mathbf{x},\mathbf{y})}$. For a given state $(\mathbf{x},\mathbf{y})$, the p-dimensional random vector $\mathbf{N}|_{(\mathbf{x},\mathbf{y})}=(N_1,\ldots,N_p)|_{(\mathbf{x},\mathbf{y})}$ counts the events taking place when $E_{(\mathbf{x},\mathbf{y})}$ expires. The case when no event is observed is denoted by $\mathbf{N}|_{(\mathbf{x},\mathbf{y})}=\mathbf{0}_p$, whereas the occurrence of an event of type s is associated with $\mathbf{N}|_{(\mathbf{x},\mathbf{y})}=n\mathbf{e}_p(s)$, where $n\in\mathbb{Z}-\{0\}$. For example, $n>1$ denotes a multiple positive jump, $n=-1$ represents a negative jump, etc.

The fundamental state $\mathbf{x}$ is updated in the light of the observed value of $\mathbf{N}|_{(\mathbf{x},\mathbf{y})}$. More concretely, we assume that the resulting fundamental state $\mathbf{x}'$ is of the form $\mathbf{x}'=f\left(\mathbf{x},\mathbf{N}|_{(\mathbf{x},\mathbf{y})}\right)$, where the fundamental state function f has to be specified for each particular Markov chain $(\mathbf{X},\mathbf{Y})$. We notice that $\mathbf{x}'=\mathbf{x}$ if $\mathbf{N}|_{(\mathbf{x},\mathbf{y})}=\mathbf{0}_p$.

It should be noted that the case $\mathbf{N}|_{(\mathbf{x},\mathbf{y})}=\mathbf{0}_p$ implies that the phase state $\mathbf{y}$ jumps to a new state $\mathbf{y}'\neq\mathbf{y}$. In contrast, the existence of proper events may or not be accompanied by a phase change.

The kernel $\{\mathbf{P}^{\mathbf{n}}_{\mathbf{x}};(\mathbf{x},\mathbf{n})\in\mathscr{S}_{(\mathbf{X},\mathbf{N})}\}$ completes the specification of the BSDE approach. The elements $p^{\mathbf{n}}_{\mathbf{x}}(\mathbf{y};\mathbf{y}')$ of the matrix $\mathbf{P}^{\mathbf{n}}_{\mathbf{x}}$ record the probabilities of generating the event $\mathbf{n}$ and a transition from phase $\mathbf{y}$ to phase $\mathbf{y}'$, given that the system state was $(\mathbf{x},\mathbf{y})$ just before $E_{(\mathbf{x},\mathbf{y})}$ expires. Since $E_{(\mathbf{x},\mathbf{y})}$ is a sojourn time, we notice that $p^{\mathbf{0}_p}_{\mathbf{x}}(\mathbf{y};\mathbf{y})=0$.

Finally, the infinitesimal generator $\mathbf{Q}=(q_{(\mathbf{x},\mathbf{y})(\mathbf{x}',\mathbf{y}')})$ of the Markov chain $(\mathbf{X},\mathbf{Y})$ is given by

$$q_{(\mathbf{x},\mathbf{y})(\mathbf{x}',\mathbf{y}')}=\begin{cases}-\lambda_{(\mathbf{x},\mathbf{y})}, & \text{if } (\mathbf{x}',\mathbf{y}')=(\mathbf{x},\mathbf{y}),\\ \lambda_{(\mathbf{x},\mathbf{y})}p^{\mathbf{n}}_{\mathbf{x}}(\mathbf{y};\mathbf{y}'), & \text{if } \mathbf{x}'=f\left(\mathbf{x},\mathbf{N}|_{(\mathbf{x},\mathbf{y})}\right),\\ 0, & \text{otherwise.}\end{cases}$$

If it is desired, then the BSDE approach can be used to construct only a part of the stochastic model. In fact, the BMAP can be readily obtained as a particular case of the BSDE approach; see Artalejo and Gómez-Corral (2010, Example 2.1). The BSDE approach can be easily adapted to the discrete-time setting. Indeed, the above BSDE construction is inspired in a similar discrete mechanism, called discrete block state-dependent arrival distribution, which was introduced in Artalejo and Li (2010) to generate the arrival input of a certain discrete-time queue.

4. Application of the BSDE approach to epidemic models

In this section, we show how the BSDE approach presented in Subsection 3.5 can be used to extend many stochastic systems that use Markov chains to model a biological population. More concretely, we consider the *state-dependent susceptible-infected-susceptible* (SD-SIS) epidemic model which generalizes the scalar SIS model allowing non-exponential infection and recovery times, as well as the existence of correlation. Once the SD-SIS model is constructed, we focus in Subsection 4.2 on the time until the extinction. In Subsection 4.3, the counterpart of the coefficient of correlation between inter-arrival times in the BMAP (see Subsection 2.3.2) is introduced.

4.1. Construction of the SD-SIS model

Firstly, we recall the scalar SIS model (see also Allen (2003)). Consider a closed population of size K. At time t, the population consists of $I(t)$ infected individuals and $S(t) = K - I(t)$ susceptible individuals. In this context, the process $\{I(t); t \geq 0\}$ is assumed to be a birth-and-death process on the state space $\{0, 1, \ldots, K\}$. Let β and γ denote the contact and recovery rates, respectively. Then, the birth rates are defined by $\lambda_i = \beta i(K-i)/K$, for $0 \leq i \leq K$. These rates correspond to transitions occurring when a susceptible individual becomes infected in agreement with the current contacts between $I(t)$ and $S(t)$. On the other hand, the death rates $\mu_i = \gamma i$, for $1 \leq i \leq K$, are associated with the recovery of infected individuals.

The construction of the SD-SIS model is based on a BSDE approach with $k = 1$ and $l = p = 2$. The fundamental state $\mathbf{x} = i$ represents the number of infected individuals, whereas the phase state $\mathbf{y} = (m, n)$ consists of the infection and recovery phases in process at time t. The state space $\mathscr{S}_{(\mathbf{X},\mathbf{Y})}$ is given by

$$\mathscr{S}_{(\mathbf{X},\mathbf{Y})} = \{\overline{0}\} \cup \{(i, m, n); 1 \leq i \leq K, 1 \leq m \leq M, 1 \leq n \leq N\}.$$

We notice that the epidemic ends as soon as there are no infected individuals in the population. Thus, we consider an absorbing macrostate $\overline{0}$ with rate $\lambda_{\overline{0}} = 0$. The individuals do not develop immunity after they recover. As a result, the Markov chain $(\mathbf{X}, \mathbf{Y})$ is reducible and the absorption occurs in a finite time with probability one. The events are associated with infections (i.e., single positive jumps) and recoveries (i.e., single negative jumps). It means that the SD-SIS model can be viewed as a particular case of a finite *state-dependent quasi-birth-and-death* (SD-QBD) process; see Artalejo and Gómez-Corral (2010, Section 3).

Then, the infinitesimal generator $\mathbf{Q}$ of the SD-SIS model has the following non-homogeneous block-tridiagonal structure:

$$\mathbf{Q} = \begin{pmatrix} 0 & \mathbf{0}_g & & & & & \\ \mathbf{q}_{10} & \mathbf{Q}_{11} & \mathbf{Q}_{12} & & & & \\ & \mathbf{Q}_{21} & \mathbf{Q}_{22} & \mathbf{Q}_{23} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \mathbf{Q}_{K-1,K-2} & \mathbf{Q}_{K-1,K-1} & \mathbf{Q}_{K-1,K} \\ & & & & \mathbf{Q}_{K,K-1} & \mathbf{Q}_{KK} \end{pmatrix},$$

where the blocks $\mathbf{Q}_{ii'}$ are square matrices of dimension $g = MN$, for $1 \leq i, i' \leq K$. The column vector $\mathbf{q}_{10}$ describes the motion from states $(1, m, n)$ to the absorbing state 0, for $1 \leq m \leq M$ and $1 \leq n \leq N$.

For the derivation of the blocks $\mathbf{Q}_{ii'}$, we need to introduce families of rate matrices $\{\overline{\mathbf{A}}_i^k; 1 \leq i \leq K-1\}$ and $\{\overline{\mathbf{D}}_i^k; 1 \leq i \leq K\}$, for $k \in \{0,1\}$. The elements $\vec{a}_i^k(m;m')$ are defined by

$$\begin{aligned} \vec{a}_i^0(m;m) &= -\lambda_{(i,m)}^A, \\ \vec{a}_i^0(m;m') &= \lambda_{(i,m)}^A a_i^0(m;m'), \quad m' \neq m, \\ \vec{a}_i^1(m;m') &= \lambda_{(i,m)}^A a_i^1(m;m'). \end{aligned}$$

We observe that $\lambda_{(i,m)}^A$ denotes the rate of the exponential sojourn time E_{im}^A, which ends either when an infection takes place (with or without phase change) or simply when the infection phase is changed (no arrival case). If $i = K$, then the whole population is infected, so we have $\lambda_{(K,m)}^A = 0$. In contrast, $\lambda_{(i,m)}^A > 0$ for $1 \leq i \leq K-1$. The kernel probabilities $a_i^k(m,m')$ are the probabilities of $k \in \{0,1\}$ infections (i.e., positive jumps) and a transition from phase m to phase m', given that $\mathbf{x} = i$. The description of the rate matrices $\overline{\mathbf{D}}_i^k$ is similar and thus it is omitted. By assuming independence between E_{im}^A and the analogue recovery sojourn time E_{in}^D, we have that $\lambda_{(i,m,n)} = \lambda_{(i,m)}^A + \lambda_{(i,n)}^D > 0$.

Under the above BSDE specifications, we finally obtain the following non-zero blocks

$$\begin{aligned} \mathbf{q}_{10} &= \left(\mathbf{I}_M \otimes \overline{\mathbf{D}}_1^1\right) \mathbf{e}_g, \\ \mathbf{Q}_{i,i-1} &= \mathbf{I}_M \otimes \overline{\mathbf{D}}_i^1, \quad 2 \leq i \leq K, \\ \mathbf{Q}_{ii} &= \overline{\mathbf{A}}_i^0 \oplus \overline{\mathbf{D}}_i^0, \quad 1 \leq i \leq K-1, \\ \mathbf{Q}_{KK} &= \mathbf{I}_M \otimes \overline{\mathbf{D}}_K^0, \\ \mathbf{Q}_{i,i+1} &= \overline{\mathbf{A}}_i^1 \otimes \mathbf{I}_N, \quad 1 \leq i \leq K-1. \end{aligned}$$

We now turn our attention to the dimensionality problem. The objective is to deal with a particularization of the rate matrices such that the formulation remains sufficiently tractable, yet enough versatile for computational purposes. To reach this objective, we consider the choice

$$\overline{\mathbf{A}}_i^0 = \frac{\lambda_i}{\lambda}\mathbf{D}_0^A, \qquad \overline{\mathbf{A}}_i^1 = \frac{\lambda_i}{\lambda}\mathbf{D}_1^A, \quad 1 \le i \le K-1,$$

$$\overline{\mathbf{D}}_i^0 = \frac{\mu_i}{\mu}\mathbf{D}_0^D, \qquad \overline{\mathbf{D}}_i^1 = \frac{\mu_i}{\mu}\mathbf{D}_1^D, \quad 1 \le i \le K,$$

where $(\mathbf{D}_0^A, \mathbf{D}_1^A)$ and $(\mathbf{D}_0^D, \mathbf{D}_1^D)$ denote the characteristic matrices of two auxiliary MAPs of orders M and N, respectively. Their corresponding fundamental rates are λ and μ.

Since λ_i and μ_i are the birth-and-death rates of the scalar SIS model, we obtain a BSDE formulation that, given that the current number of infected individuals equals i, the expectations until the next infection and recovery epochs match the corresponding expected values in the scalar SIS model.

4.2. Extinction in the SD-SIS model

The extinction time quantifies the spread of the epidemic on the population and describes the time until the end of the epidemic process. Thus, the time to extinction is an important measure of the persistence of an infection. There exists a vast literature studying the extinction time of stochastic biological models. In this subsection, we extend the study to the SD-SIS model.

We distinguish between a conditional version of the extinction time given an initial state and an unconditional version properly defined. The conditional extinction time $L_{(i,m,n)}$ is defined as the absorption time in 0, given that the initial state of the SD-SIS model is $(\mathbf{x},\mathbf{y}) = (i,m,n)$. Let $\varphi_{(i,m,n)}(s)$ be its Laplace-Stieltjes transform. The vectors $\boldsymbol{\varphi}_i(s) = (\varphi_{(i,1,1)}(s), \ldots, \varphi_{(i,M,N)}(s))'$, for $1 \le i \le K$, and $\boldsymbol{\varphi}(s) = (\boldsymbol{\varphi}_1(s), \ldots, \boldsymbol{\varphi}_K(s))'$ comprise the Laplace-Stieltjes transforms according to the levels determined by the number of infected individuals.

By introducing an initial distribution $\boldsymbol{\tau}$ on the state space $\mathscr{S}_{(\mathbf{X},\mathbf{Y})}$, we arrive to the unconditional version L of the extinction time. From the general theory for continuous-time Markov chains (see e.g. Kulkarni (1995), and Latouche and Ramaswami (1999)), we know that L follows a PH distribution of order Kg with representation $(\boldsymbol{\tau}, \mathbf{M})$, where $\mathbf{M}$ is the submatrix of $\mathbf{Q}$ corresponding to the set of transient states $\mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\overline{0}\}$.

Since the set $\mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\overline{0}\}$ is irreducible, the existence of the inverse $\mathbf{M}^{-1}$ is guaranteed. We may also observe that the starting point of the density function is given by $f_L(0) = -\boldsymbol{\tau}\mathbf{M}\mathbf{e}_{Kg} = \boldsymbol{\tau}_1\mathbf{q}_{10}$, where $\boldsymbol{\tau}_1$ is the sub-vector of $\boldsymbol{\tau}$ containing the initial probabilities $\tau_{(1,m,n)}$ of the level $i = 1$.

Coming back to the unconditional version, we notice that the vector $\boldsymbol{\varphi}(s)$ satisfies the block-tridiagonal system

$$(\mathbf{M} - s\mathbf{I}_{Kg})\,\boldsymbol{\varphi}(s) = -\begin{pmatrix} \mathbf{q}_{10} \\ \mathbf{0}'_{(K-1)g} \end{pmatrix}.$$

By using Euler and Post-Widder algorithms, we can numerically invert the above expression to get the conditional density functions $f_{L_{(i,m,n)}}(x)$ and, consequently, the unconditional density $f_L(x)$; see Cohen (2007).

Finally, we observe that the conditional moments $m^k_{(i,m,n)} = E[L^k_{(i,m,n)}]$, for $(i,m,n) \in \mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\bar{0}\}$ and $k \geq 1$, can be computed from the formula

$$\mathbf{m}^k = k! \left(-\mathbf{M}^{-1}\right)^k \mathbf{e}_{Kg}, \quad k \geq 1,$$

or, alternatively, from the recursive expressions

$$\begin{aligned} \mathbf{m}^0 &= \mathbf{e}_{Kg}, \\ \mathbf{m}^k &= -k\mathbf{M}^{-1}\mathbf{m}^{k-1}, \quad k \geq 1, \end{aligned}$$

where $\mathbf{m}^k$ denotes the column vector of dimension Kg containing the moments $m^k_{(i,m,n)}$ in lexicographic order.

The unconditional time to extinction depends on the initial distribution $\boldsymbol{\tau}$. In epidemiology, it is often known that a certain epidemic has been evolving for a long time and that it has not reached the extinction yet. However, it may be very difficult to know the exact distribution $\boldsymbol{\tau}$. In this case, the use of the quasi-stationary distribution is especially interesting. The starting point is the conditional probabilities

$$u_{(i,m,n)}(t) = P\left((\mathbf{X}(t),\mathbf{Y}(t)) = (i,m,n) | L > t\right) = \frac{p_{(i,m,n)}(t)}{1 - p_{\bar{0}}(t)},$$

for $(i,m,n) \in \mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\bar{0}\}$, where $p_{(i,m,n)}(t)$ and $p_{\bar{0}}(t)$ are the transient probabilities of the Markov chain $(\mathbf{X},\mathbf{Y})$.

Suppose that the Markov chain starts with the initial distribution $\tau_{(i,m,n)} = P((\mathbf{X}(0),\mathbf{Y}(0)) = (i,m,n))$, for $(i,m,n) \in \mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\bar{0}\}$. If there exists a starting distribution $\tau_{(i,m,n)} = u_{(i,m,n)}$, such that $u_{(i,m,n)}(t) = u_{(i,m,n)}$, for all $t \geq 0$, then $\mathbf{u} = (u_{(i,m,n)})$ is called a quasi-stationary distribution. Moreover, there also exists a limiting interpretation which states that $\lim_{t\to\infty} u_{(i,m,n)}(t) = u_{(i,m,n)}$, independently of the initial distribution.

In our case, the set $\mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\bar{0}\}$ is finite and irreducible. Then, the quasi-stationary distribution $\mathbf{u}$ amounts to the left eigenvector associated with the eigenvalue with maximal real part of the matrix $\mathbf{M}$; see Darroch and Seneta (1967). This result gives a method for numerical computation.

In what follows, we set $\boldsymbol{\tau} = \mathbf{u}$ and generalize the existing approach for the study of the extinction time $L_{\mathbf{u}}$ in the scalar SIS model (see Norden (1982)) to the SD-SIS model.

By differentiating $u_{(i,m,n)}(t)$ with respect to t, we obtain

$$u'_{(i,m,n)}(t) = \frac{p'_{(i,m,n)}(t)}{1 - p_{\bar{0}}(t)} + \frac{p_{(i,m,n)}(t)\, p'_{\bar{0}}(t)}{(1 - p_{\bar{0}}(t))^2}, \quad (i,m,n) \in \mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\bar{0}\}.$$

By combining the above formula and the Kolmogorov forward equation for the absorbing state $\bar{0}$, we find that

$$u'_{(i,m,n)}(t) = \frac{p'_{(i,m,n)}(t)}{1-p_{\bar{0}}(t)} + \frac{p_{(i,m,n)}(t)}{1-p_{\bar{0}}(t)} \sum_{n=1}^{N} \bar{d}_1^1(n;\cdot) u_{(1,\cdot,n)}(t), \quad (i,m,n) \in \mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\bar{0}\},$$

where $\bar{d}_1^1(n;\cdot) = \sum_{n'=1}^{N} \bar{d}_1^1(n;n')$ and $u_{(1,\cdot,n)}(t) = \sum_{m=1}^{M} u_{(1,m,n)}(t)$, for $1 \le n \le N$.

Now, we appeal to the fact that the initial distribution is $\mathbf{u}$ and we thus put $u'_{(i,m,n)}(t) = 0$. Hence, for each $(i,m,n) \in \mathscr{S}_{(\mathbf{X},\mathbf{Y})} - \{\bar{0}\}$, we get the differential equation

$$\begin{aligned} p'_{(i,m,n)}(t) &= -p_{(i,m,n)}(t) \sum_{n=1}^{N} \bar{d}_1^1(n;\cdot) u_{(1,\cdot,n)}, \\ p_{(i,m,n)}(0) &= u_{(i,m,n)}, \end{aligned}$$

which yields the solution

$$p_{(i,m,n)}(t) = u_{(i,m,n)} \exp\left\{ -t \sum_{n=1}^{N} \bar{d}_1^1(n;\cdot) u_{(1,\cdot,n)} \right\}.$$

Finally, for $p_{\bar{0}}(t)$, we now have $p'_{\bar{0}}(t) = \sum_{n=1}^{N} \bar{d}_1^1(n;\cdot) p_{(1,\cdot,n)}(t)$, with $p'_{\bar{0}}(0) = 0$, so that

$$P(L_{\mathbf{u}} \le t) = p_{\bar{0}}(t) \;=\; 1 - \exp\left\{ -t \sum_{n=1}^{N} \bar{d}_1^1(n;\cdot) u_{(1,\cdot,n)} \right\}, \quad t \ge 0.$$

This establishes that the time to extinction, when the initial distribution is the quasi-stationary distribution, has an exponential distribution with rate $1/E[L_{\mathbf{u}}] = \sum_{n=1}^{N} \bar{d}_1^1(n;\cdot)\, u_{(1,\cdot,n)}$.

The following example illustrates the influence of the characteristic matrices and the correlation in the distribution of $L_{\mathbf{u}}$.

Example 4.1 We consider the following three choices for the characteristic matrices $(\mathbf{D}_0^A, \mathbf{D}_1^A)$ and $(\mathbf{D}_0^D, \mathbf{D}_1^D)$:

(i) *Exponential kernel*. We take $M = N = 1$, $\mathbf{D}_0^A = \mathbf{D}_0^D = -1$ and $\mathbf{D}_1^A = \mathbf{D}_1^D = 1$.
(ii) *Erlang-hyperexponential kernel*. We take $M = 3$, $N = 2$ and

$$\mathbf{D}_0^A = \begin{pmatrix} -3 & 3 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & -3 \end{pmatrix}, \qquad \mathbf{D}_1^A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 3 & 0 & 0 \end{pmatrix},$$

$$\mathbf{D}_0^D = \begin{pmatrix} -1.9 & 0 \\ 0 & -0.19 \end{pmatrix}, \qquad \mathbf{D}_1^D = \begin{pmatrix} 1.71 & 0.19 \\ 0.171 & 0.019 \end{pmatrix}.$$

(iii) *MAP-MAP kernel*. We take $M = N = 3$ and

$$\mathbf{D}_0^A = \begin{pmatrix} -1.00221 & 1.00221 & 0 \\ 0 & -1.00221 & 0 \\ 0 & 0 & -225.75 \end{pmatrix}, \quad \mathbf{D}_1^A = \begin{pmatrix} 0 & 0 & 0 \\ 0.99219 & 0 & 0.01002 \\ 2.2575 & 0 & 223.4925 \end{pmatrix},$$

$$\mathbf{D}_0^D = \begin{pmatrix} -0.87478 & 0.87478 & 0 \\ 0 & -0.87478 & 0 \\ 0 & 0 & -94.76811 \end{pmatrix}, \quad \mathbf{D}_1^D = \begin{pmatrix} 0 & 0 & 0 \\ 0.78730 & 0 & 0.08748 \\ 7.28985 & 0 & 87.47826 \end{pmatrix}.$$

For the above three scenarios, the fundamental rates associated with infection and recovery characteristic matrices are $\lambda = \mu = 1.0$. We notice that scenarios (i) and (ii) are associated with renewal processes and, on the contrary, scenario (iii) has positive correlated infection and recovery times. The values of the coefficients of correlation are 0.48890 and 0.43482, respectively.

Table 1: *$E[\mathbf{u}]$, $\sigma(\mathbf{u})$ and $E[L_{\mathbf{u}}]$ for three scenarios.*

	Scenario (i)	**Scenario (ii)**	**Scenario (iii)**
$E[\mathbf{u}]$	64.48076	60.04070	38.91698
$\sigma(\mathbf{u})$	11.87236	20.12606	44.42737
$E[L_{\mathbf{u}}]$	2094831.60843	1140.40538	7.75147

For a population size $K = 200$ and the rates $\beta = 1.5$ and $\gamma = 1.0$, we summarize in Table 1 the main statistical descriptors; that is, the mean and the standard deviation of $\mathbf{u}$, and the expected value $E[L_{\mathbf{u}}]$.

In Figure 2, we turn our attention to the probability distribution function $P(L_{\mathbf{u}} \leq t)$. In this case, we deal with scenario (iii) with $K = 200$, $\gamma = 1.0$ and $\beta \in \{0.5, 1.0, 1.5\}$.

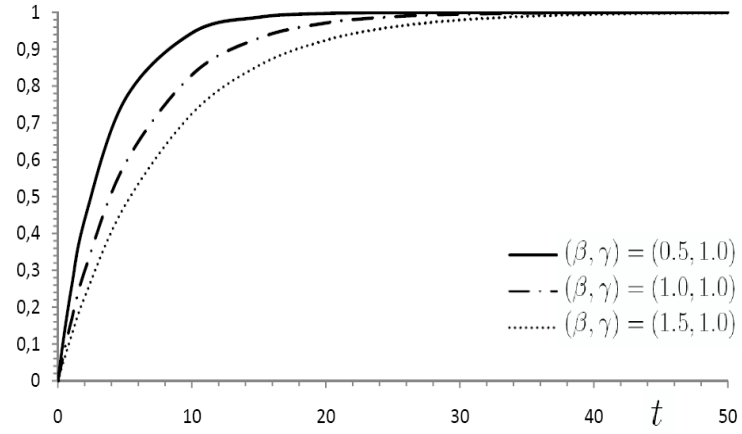

Figure 2: *The probability distribution function $P(L_{\mathbf{u}} \leq t)$.*

In the light of the numerical results, the conclusion is that the influence of the scenario is significant. In other words, the underlying distribution and the correlation are important features which cannot be ignored.

4.3. Correlation between successive events

In this subsection we define a coefficient of correlation between two successive events of the SD-SIS process. Assume that the initial distribution is $\boldsymbol{\tau}$ and denote the first two inter-event intervals as X and Y. To avoid trivialities, we also assume that $K \geq 2$.

First of all, we observe that the one-step transition probability matrix governing the embedded Markov chain at event epochs is given by

$$\mathbf{P} = \begin{pmatrix} 1 & & & & & \\ (-\mathbf{Q}_{11}^{-1})\mathbf{q}_{10}\ \mathbf{0}_{g\times g} & & (-\mathbf{Q}_{11}^{-1})\mathbf{Q}_{12} & & & \\ & \ddots & & \ddots & & \ddots \\ & & (-\mathbf{Q}_{K-1,K-1}^{-1})\mathbf{Q}_{K-1,K-2} & \mathbf{0}_{g\times g} & & (-\mathbf{Q}_{K-1,K-1}^{-1})\mathbf{Q}_{K-1,K} \\ & & & (-\mathbf{Q}_{KK}^{-1})\mathbf{Q}_{K,K-1} & & \mathbf{0}_{g\times g} \end{pmatrix}.$$

To construct a coefficient of correlation, we must guarantee the existence of at least two events before the process reaches its extinction. Thus, if $i = 1$, we correct matrix $\mathbf{P}$ by imposing that the next event is an infection. This modification only affects to the blocks associated with the level $i = 1$ of $\mathbf{Q}$, which are now given by $\mathbf{q}_{10}^{c} = \mathbf{0}_g'$, $\mathbf{Q}_{12}^{c} = \mathbf{Q}_{12}$ and $\mathbf{Q}_{11}^{c} = \mathbf{Q}_{11} + \mathrm{diag}(\mathbf{e}_g(1)\mathbf{q}_{10}, \ldots, \mathbf{e}_g(g)\mathbf{q}_{10})$. As a result, the second row of the corrected matrix $\mathbf{P}^{c}$ becomes $(\mathbf{0}_g', \mathbf{0}_{g\times g}, (-\mathbf{Q}_{11}^{c})^{-1}\mathbf{Q}_{12}, \mathbf{0}_{g\times g}, \ldots)$, while the rest of row blocks does not vary.

In calculating the correlation between X and Y, we shall need the marginal density functions of X and Y, and the joint density function of (X,Y). It is easy to show that they are as follows:

$$f_X(x) = \boldsymbol{\tau}(1)\exp\{\mathbf{Q}_{11}^{c}x\}(-\mathbf{Q}_{11}^{c})\mathbf{e}_g + \sum_{i=2}^{K}\boldsymbol{\tau}(i)\exp\{\mathbf{Q}_{ii}x\}(-\mathbf{Q}_{ii})\mathbf{e}_g, \quad x \geq 0,$$

$$f_Y(y) = \sum_{i=1}^{K}\overline{\boldsymbol{\tau}}(i)\exp\{\mathbf{Q}_{ii}x\}(-\mathbf{Q}_{ii})\mathbf{e}_g, \quad y \geq 0,$$

$$\begin{aligned} f_{(X,Y)}(x,y) &= \boldsymbol{\tau}(1)\exp\{\mathbf{Q}_{11}^{c}x\}\mathbf{Q}_{12}\exp\{\mathbf{Q}_{22}y\}(-\mathbf{Q}_{22})\mathbf{e}_g \\ &\quad + \sum_{i=2}^{K-1}\boldsymbol{\tau}(i)\exp\{\mathbf{Q}_{ii}x\}\mathbf{Q}_{i,i+1}\exp\{\mathbf{Q}_{i+1,i+1}y\}(-\mathbf{Q}_{i+1,i+1})\mathbf{e}_g \\ &\quad + \sum_{i=2}^{K}\boldsymbol{\tau}(i)\exp\{\mathbf{Q}_{ii}x\}\mathbf{Q}_{i,i-1}\exp\{\mathbf{Q}_{i-1,i-1}y\}(-\mathbf{Q}_{i-1,i-1})\mathbf{e}_g, \quad x \geq 0, y \geq 0, \end{aligned}$$

where the vector $\overline{\boldsymbol{\tau}} = (\overline{\boldsymbol{\tau}}(1), \ldots, \overline{\boldsymbol{\tau}}(K))$ is given by

$$\begin{aligned}
\overline{\boldsymbol{\tau}}(1) &= \boldsymbol{\tau}(2)\left(-\mathbf{Q}_{22}^{-1}\right)\mathbf{Q}_{21},\\
\overline{\boldsymbol{\tau}}(2) &= \boldsymbol{\tau}(1)\left(-\mathbf{Q}_{11}^{c}\right)^{-1}\mathbf{Q}_{12} + \boldsymbol{\tau}(3)\left(-\mathbf{Q}_{33}^{-1}\right)\mathbf{Q}_{32},\\
\overline{\boldsymbol{\tau}}(i) &= \boldsymbol{\tau}(i-1)\left(-\mathbf{Q}_{i-1,i-1}^{-1}\right)\mathbf{Q}_{i-1,i} + (1-\delta_{iK})\boldsymbol{\tau}(i+1)\left(-\mathbf{Q}_{i+1,i+1}^{-1}\right)\mathbf{Q}_{i+1,i}, \quad 3 \le i \le K.
\end{aligned}$$

The vector $\overline{\boldsymbol{\tau}}$ can be readily obtained by noticing that $(\mathbf{0}_g, \overline{\boldsymbol{\tau}}) = (\mathbf{0}_g, \boldsymbol{\tau})\mathbf{P}^c$.

From the density functions, it is straightforward to find the first two moments of X and Y, as well as the cross expectation $E[XY]$. They are given by

$$\begin{aligned}
E[X] &= \boldsymbol{\tau}(1)\left(-\mathbf{Q}_{11}^{c}\right)^{-1}\mathbf{e}_g + \sum_{i=2}^{K}\boldsymbol{\tau}(i)\left(-\mathbf{Q}_{ii}^{-1}\right)\mathbf{e}_g,\\
E[X^2] &= 2\left(\boldsymbol{\tau}(1)\left(-\mathbf{Q}_{11}^{c}\right)^{-2}\mathbf{e}_g + \sum_{i=2}^{K}\boldsymbol{\tau}(i)\left(-\mathbf{Q}_{ii}^{-1}\right)^{2}\mathbf{e}_g\right),\\
E[Y] &= \sum_{i=1}^{K}\overline{\boldsymbol{\tau}}(i)\left(-\mathbf{Q}_{ii}^{-1}\right)\mathbf{e}_g,\\
E[Y^2] &= 2\sum_{i=1}^{K}\overline{\boldsymbol{\tau}}(i)\left(-\mathbf{Q}_{ii}^{-1}\right)^{2}\mathbf{e}_g,\\
E[XY] &= \boldsymbol{\tau}(1)\left(-\mathbf{Q}_{11}^{c}\right)^{-2}\mathbf{Q}_{12}\left(-\mathbf{Q}_{22}^{-1}\right)\mathbf{e}_g\\
&\quad + \sum_{i=2}^{K}\boldsymbol{\tau}(i)\left(-\mathbf{Q}_{ii}^{-1}\right)^{2}\left(\mathbf{Q}_{i,i-1}\left(-\mathbf{Q}_{i-1,i-1}^{-1}\right) + (1-\delta_{iK})\mathbf{Q}_{i,i+1}\left(-\mathbf{Q}_{i+1,i+1}^{-1}\right)\right)\mathbf{e}_g.
\end{aligned}$$

The combination of the above expressions leads to the desired coefficient of correlation

$$\rho(X,Y) = \frac{E[XY] - E[X]E[Y]}{\sqrt{Var(X)Var(Y)}}.$$

The initial distribution can be chosen as $\boldsymbol{\tau} = \mathbf{u}_R$, where $\mathbf{u}_R$ denotes the quasi-stationary distribution of the embedded Markov chain between two regular event epochs, with transition matrix $\mathbf{P}$.

5. Bibliographical notes

Within the list of references we may distinguish between two categories of contributions, depending on whether or not they have been cited throughout the main body of this survey.

Papers and books of the first category have allowed us to review the main aspects of the BMAP and its basic properties, as well as related variants, generalizations and new results in the context of the BSDE approach. The reader has been also addressed to the existing survey papers by Asmussen (2000), Chakravarthy (2001,2010) and Neuts (1992) on the PH distribution and the BMAP and, in a more general setting, to the monographs by Bini *et al.* (2005), Latouche and Ramaswami (1999), Li (2010) and Neuts (1981,1989) which present the main results and algorithms of the matrix-analytic theory.

Regarding to the second category, we associate those papers we do not cite in preceding sections to our desire to present a few selected references dealing with the problem of estimating parameters, multiple types of customers and applications. They are classified as follows:

(i) *Estimation and fitting*

Bodrog *et al.* (2008), Breuer (2002), Breuer and Alfa (2005), Horváth *et al.* (2010), Okamura *et al.* (2009), and Telek and Horváth (2007).

(ii) *Marked arrivals and multiple types of customers*

Alfa *et al.* (2003), He (1996,2000), He and Alfa (2000), Takine and Hasegawa (1994), and Van Houdt and Blondia (2002).

(iii) *Applications*

In queueing and communication systems: Artalejo and Gómez-Corral (2008), Asmussen and Møller (2001), Baek *et al.* (2008), Chakravarthy *et al.* (2006), Choi *et al.* (2004), Daikoku *et al.* (2007), Dudin and Nishimura (1999), He (2001), Kim and Kim (2010), Kim *et al.* (2010), Lambert *et al.* (2006), Li *et al.* (2006), Lucantoni *et al.* (1994), Ost (2001), Shin (2004), Squillante *et al.* (2008), Takine (1999), and Tian and Zhang (2006).

In reliability and maintenance models: Chakravarthy and Gómez-Corral (2009), Frostig and Kenzin (2009), and Montoro-Cazorla and Pérez-Ocón (2008).

In inventory systems: Cheng and Song (2001), He *et al.* (2002), Manuel *et al.* (2007) and Ramaswami (1981).

In risk and insurance problems: Ahn and Badescu (2007), Badescu *et al.* (2007), and Cheung and Landriault (2009).

Since an exhaustive bibliographical work should include several hundreds of papers on the subject in stochastic modelling, we have elaborated the above list only for illustrative purposes.

Acknowledgments

J. R. Artalejo and A. Gómez-Corral were supported by the Government of Spain (Ministry of Science and Innovation) and the European Commission through project MTM2008-01121.

Appendix: Glossary of notation

To begin with, matrices have uppercase letters and vectors lowercase letters. The transpose of $\mathbf{A}$ is written as $\mathbf{A}'$. The matrix $\mathrm{diag}(a_1,\ldots,a_p)$ is the square matrix having elements a_1, ..., a_p along its diagonal and zeros elsewhere.

We denote by $\mathbf{I}_p$ and $\mathbf{0}_{p\times q}$ the identity matrix of order p and the null matrix of dimension $p\times q$, respectively. We let $\mathbf{e}_p$ be the column vector of order p of 1s, and $\mathbf{0}_p$ be the row vector of order p of 0s. The vector $\mathbf{e}_p(j)$ is a column vector of order p such that all entries equal 0, except for the jth one which is equal to 1.

For a square matrix $\mathbf{A}$, the matrix exponential, denoted by $\exp\{\mathbf{A}\}$, is defined by

$$\exp\{\mathbf{A}\}=\sum_{k=0}^{\infty}\frac{1}{k!}\mathbf{A}^k.$$

Consider a matrix $\mathbf{A}=(a_{ij})$ of dimension $p\times q$ and a matrix $\mathbf{B}$ of dimension $r\times s$. The Kronecker product of these matrices, denoted by $\mathbf{A}\otimes\mathbf{B}$, is defined as the structured matrix of dimension $pr\times qs$

$$\mathbf{A}\otimes\mathbf{B}=\begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1q}\mathbf{B}\\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2q}\mathbf{B}\\ \vdots & \vdots & \ddots & \vdots\\ a_{p1}\mathbf{B} & a_{p2}\mathbf{B} & \cdots & a_{pq}\mathbf{B}\end{pmatrix}.$$

Given two square matrices $\mathbf{A}$ and $\mathbf{B}$ of orders p and q, respectively, their Kronecker sum, denoted by $\mathbf{A}\oplus\mathbf{B}$, is defined as the matrix $\mathbf{A}\oplus\mathbf{B}=\mathbf{A}\otimes\mathbf{I}_q+\mathbf{I}_p\otimes\mathbf{B}$.

The Kronecker delta δ_{ij} takes the values 1 if $i=j$, and 0 if $i\neq j$.

References

Ahn, S. and Badescu, A. L. (2007). On the analysis of the Gerber-Shin discounted penalty function for risk processes with Markovian arrivals. *Insurance: Mathematics and Economics*, 41, 234-249.

Alfa, A. S. and Neuts, M. F. (1995). Modelling vehicular traffic using the discrete time Markovian arrival process. *Transportation Science*, 29, 109-117.

Alfa, A. S., Liu, B. and He, Q.-M. (2003). Discrete-time analysis of $MAP/PH/1$ multiclass general pre-emptive priority queue. *Naval Research Logistics*, 50, 662-682.

Allen, L. J. S. (2003). *An Introduction to Stochastic Processes with Applications to Biology*. New Jersey: Prentice Hall.

Artalejo, J. R. and Gómez-Corral, A. (2008). *Retrial Queueing Systems: A Computational Approach*. Berlin: Springer-Verlag.

Artalejo, J. R. and Gómez-Corral, A. (2010). A state-dependent Markov-modulated mechanism for generating events and stochastic models. *Mathematical Methods in the Applied Sciences*, 33, 1342-1349.

Artalejo, J. R. and Li, Q.-L. (2010). Performance analysis of a block-structured discrete-time retrial queue with state-dependent arrivals. *Discrete Event Dynamic Systems*, 20, 325-347.

Asmussen, S. and Koole, G. (1993). Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability*, 30, 365-372.

Asmussen, S. and Bladt, M. (1999). Point processes with finite-dimensional conditional probabilities. *Stochastic Processes and their Applications*, 82, 127-142.

Asmussen, S. (2000). Matrix-analytic models and their analysis. *Scandinavian Journal of Statistics*, 27, 193-226.

Asmussen, S. and Møller, J. R. (2001). Calculation of the steady state waiting time distribution in $GI/PH/c$ and $MAP/PH/c$ queues. *Queueing Systems*, 37, 9-29.

Badescu, A. L., Drekic, S. and Landriault, D. (2007). Analysis of a threshold dividend strategy for a MAP risk model. *Scandinavian Actuarial Journal*, 2007, 227-247.

Baek, J. W., Lee, H. W., Lee, S. W. and Ahn, S. (2008). A factorization property for $BMAP/G/1$ vacation queues under variable service speed. *Annals of Operations Research*, 160, 19-29.

Bini, D. A., Latouche, G. and Meini, B. (2005). *Numerical Methods for Structured Markov Chains*. Oxford: Oxford University Press.

Blondia, C. and Casals, O. (1992). Statistical multiplexing of VBR sources: A matrix-analytic approach. *Performance Evaluation*, 16, 5-20.

Bodrog, L., Horváth, A. and Telek, M. (2008). Moment characterization of matrix exponential and Markovian arrival processes. *Annals of Operations Research*, 160, 51-68.

Breuer, L. (2002). An EM algorithm for batch Markovian arrival processes and its comparison to a simpler estimation procedure. *Annals of Operations Research*, 112, 123-138.

Breuer, L. (2003). *From Markov Jump Processes to Spatial Queues*. Dordrecht: Kluwer Academic Publishers.

Breuer, L. and Alfa, A. S. (2005). An EM algorithm for platoon arrival processes in discrete time. *Operations Research Letters*, 33, 535-543.

Breuer, L. and Baum, D. (2005). *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Dordrecht: Springer.

Chakka, R. and Do, T. V. (2007). The $MM \sum_{k=1}^{K} CPP_k/GE/c/L$ G-queue with heterogenous servers: Steady state solution and an application to performance evaluation. *Performance Evaluation*, 64, 191-209.

Chakravarthy, S. R. (2001). The batch Markovian arrival process: A review and future work. In *Advances in Probability & Stochastic Processes*, A. Krishnamoorthy, N. Raju and V. Ramaswami (eds.), Notable Publications, 21-49.

Chakravarthy, S. R., Krishnamoorthy, A. and Joshua, V. C. (2006). Analysis of a multi-server retrial queue with search of customers from the orbit. *Performance Evaluation*, 63, 776-798.

Chakravarthy, S. R. and Gómez-Corral, A. (2009). The influence of delivery times on repairable k-out-of-N systems with spares. *Applied Mathematical Modelling*, 33, 2368-2387.

Chakravarthy, S. R. (2010). Markovian arrival process. In *Wiley Encyclopedia of Operations Research and Management Science*, J. J. Cochran (ed.), John Wiley and Sons, to appear.

Chen, F. and Song, J.-S. (2001). Optimal policies for multiechelon inventory problems with Markov-modulated demand. *Operations Research*, 49, 226-234.

Cheung, E. C. K. and Landriault, D. (2009). Perturbed MAP risk models with dividend barrier strategies. *Journal of Applied Probability*, 46, 521-541.

Ching, W. K. (1997). Markov-modulated Poisson processes for multi-location inventory problems. *International Journal of Production Economics*, 53, 217-223.

Ching, W. K. (2001). *Iterative Methods for Queuing and Manufacturing Systems*. London: Springer-Verlag.

Choi, B. D., Kim, B. and Zhu, D. (2004). $MAP/M/c$ queue with constant impatient time. *Mathematics of Operations Research*, 29, 309-325.

Çinlar, E. (1972a). Markov additive processes. I. *Z. Wahrscheinlichkeitstheory verw. Geb.*, 24, 85-93.

Çinlar, E. (1972b). Markov additive processes. II. *Z. Wahrscheinlichkeitstheory verw. Geb.*, 24, 95-121.

Cohen, A. M. (2007). *Numerical Methods for Laplace Transform Inversion*. New York: Springer.

Daikoku, K., Masuyama, H., Takine, T. and Takahashi, Y. (2007). Algorithmic computation of the transient queue length distribution in the $BMAP/D/c$ queue. *Journal of the Operational Research Society of Japan*, 50, 55-72.

Darroch, J. N. and Seneta, E. (1967). On quasi-stationary distributions in absorbing continuous-time finite Markov chains. *Journal of Applied Probability*, 4, 192-196.

Dudin, A. N. and Nishimura, S. (1999). A $BMAP/SM/1$ queueing system with Markovian arrival input of disasters. *Journal of Applied Probability*, 36, 868-881.

Eckberg, A. E. (1983). Generalized peakedness of teletraffic processes. In *Proceedings of the Tenth International Teletraffic Congress*, Montreal, Canada, paper no. 4, 4B3.

Frostig, E. and Kenzin, M. (2009). Availability of inspected systems subject to shocks – A matrix algorithmic approach. *European Journal of Operational Research*, 193, 168-183.

He, Q.-M. (1996). Queues with marked customers. *Advances in Applied Probability*, 28, 567-587.

He, Q.-M. and Neuts, M. F. (1998). Markov chains with marked transitions. *Stochastic Processes and their Applications*, 74, 37-52.

He, Q.-M. (2000). Quasi-birth-and-death Markov processes with a tree structure and the $MMAP[K]/PH[K]/N/LCFS$ non-preemptive queue. *European Journal of Operational Research*, 120, 641-656.

He, Q.-M. and Alfa, A. S. (2000). Computational analysis of $MMAP[K]/PH[K]/1$ queues with a mixed FCFS and LCFS service discipline. *Naval Research Logistics*, 47, 399-421.

He, Q.-M. (2001). The versatility of $MMAP[K]$ and the $MMAP[K]/G[K]/1$ queue. *Queueing Systems*, 38, 397-418.

He, Q.-M., Jewkes, E. M. and Buzaccot, J. (2002). Optimal and near-optimal inventory control policies for a make-to-order inventory-production system. *European Journal of Operational Research*, 141, 113-132.

He, Q.-M. (2010). Construction of continuous time Markov arrival processes. *Journal of Systems Science and Systems Engineering*, 19, 351-366.

Horváth, A., Horváth, G. and Telek, M. (2010). A joint moments based analysis of networks of $MAP/MAP/1$ queues. *Performance Evaluation*, 67, 759-788.

Kim, B. and Kim, J. (2010). Queue size distribution in a discrete-time $D-BMAP/G/1$ retrial queue. *Computers & Operations Research*, 37, 1220-1227.

Kim, C. S., Klimenok, V. I., Mushko, V. and Dudin, A. N. (2010). The $BMAP/PH/N$ retrial queueing system operating in Markovian random environment. *Computers & Operations Research*, 37, 1228-1237.

Kulkarni, V. G. (1989). A new class of multivariate phase type distributions. *Operations Research*, 37, 151-158.

Kulkarni, V. G. (1995). *Modelling and Analysis of Stochastic Systems*. London: Chapman & Hall.

Lambert, J., Van Houdt, B. and Blondia, C. (2006). Queues with correlated service and inter-arrival times and their application to optical buffers. *Stochastic Models*, 22, 233-251.

Latouche, G. and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modelling*. Philadelphia: ASA-SIAM.

Li, Q.-L., Ying, Y. and Zhao, Y. Q. (2006). A $BMAP/G/1$ retrial queue with a server subject to breakdowns and repairs. *Annals of Operations Research*, 141, 233-270.

Li, Q.-L. (2010). *Constructive Computation in Stochastic Models with Applications: The RG-factorizations*. Beijing, Berlin Heidelberg: Tsinghua University Press, Springer-Verlag.

Lucantoni, D. M., Meier-Hellstern, K. S. and Neuts, M. F. (1990). A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22, 676-705.

Lucantoni, D. M. (1991). New results on the single server queue with a batch Markovian arrival process. *Stochastic Models*, 7, 1-46.

Lucantoni, D. M. (1993). The $BMAP/G/1$ queue: A tutorial. In *Performance Evaluation of Computer and Communication Systems*, L. Donatiello and R. Nelson (eds.), Lecture Notes in Computer Science, Vol. 729, Springer-Verlag, 330-358.

Lucantoni, D. M., Choudhury, G. L. and Whitt, W. (1994). The transient $BMAP/G/1$ queue. *Stochastic Models*, 10, 145-182.

Manuel, P., Sivakumar, B. and Arivarignan, G. (2007). A perishable inventory system with service facilities, MAP arrivals and PH-service times. *Journal of Systems Science and Systems Engineering*, 16, 62-73.

Milne, C. (1982). Transient behaviour of the interrupted Poisson process. *Journal of the Royal Statistical Society. Series B*, 44, 398-405.

Montoro-Cazorla, D. and Pérez-Ocón, R. (2006). Reliability of a system under two types of failures using a Markovian arrival process. *Operations Research Letters*, 34, 525-530.

Montoro-Cazorla, D. and Pérez-Ocón, R. (2008). A maintenance model with failures and inspection following Markovian arrival processes and two repair modes. *European Journal of Operational Research*, 186, 694-707.

Narayana, S. and Neuts, M. F. (1992). The first two moment matrices of the counts for the Markovian arrival process. *Stochastic Models*, 8, 459-477.

Neuts, M. F. (1979). A versatile Markovian point process. *Journal of Applied Probability*, 16, 764-779.

Neuts, M. F. (1981). *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore: The Johns Hopkins University Press.

Neuts, M. F. (1989). *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. New York: Marcel Dekker, Inc.

Neuts, M. F. (1992). Models based on the Markovian arrival process. *IEICE Transactions on Communications*, E75-B, 1255-1265.

Neuts, M. F., Liu, D. and Narayana, S. (1992). Local poissonification of the Markovian arrival process. *Stochastic Models*, 8, 87-129.

Neuts, M. F. (1993). The burstiness of point processes. *Stochastic Models*, 9, 445-466.

Neuts, M. F. (1995). *Algorithmic Probability: A Collection of Problems*. London: Chapman & Hall.

Neuts, M. F. and Li, J.-M. (1997). An algorithm for the $P(n,t)$ matrices of a continuous BMAP. In *Matrix-analytic Methods in Stochastic Models*, S.R. Chakravarthy and A.S. Alfa (eds.), Lecture Notes in Pure and Applied Mathematics, Vol. 183, Marcel Dekker, 7-19.

Nielsen, B. F., Nilsson, L. A. F., Thygesen, U. H. and Beyer, J. E. (2007). Higher order moments and conditional asymptotics of the batch Markovian arrival process. *Stochastic Models*, 23, 1-26.

Norden, R. H. (1982). On the distribution of the time to extinction in the stochastic logistic population model. *Advances in Applied Probability*, 14, 687-708.

Okamura, H., Dohi, T. and Trivedi, K. S. (2009). Markovian arrival process parameter estimation with group data. *IEEE/ACM Transactions on Networking*, 17, 1326-1339.

Ost, A. (2001). *Performance of Communication Systems: A Model-based Approach with Matrix-geometric Methods.* Berlin: Springer-Verlag.

Pacheco, A. and Prabhu, N. U. (1995). Markov-additive processes of arrivals. In *Advances in Queuing: Theory, Methods and Open Problems*, J.H. Dshalalow (ed.), CRC Press, 167-194.

Ramaswami, V. (1980). The $N/G/1$ queue and its detailed analysis. *Advances in Applied Probability*, 12, 222-261.

Ramaswami, V. (1981). Algorithms for a continuous-review (s,S) inventory system. *Journal of Applied Probability*, 18, 461-472.

Rudemo, M. (1973). Point processes generated by transitions of Markov chains. *Advances in Applied Probability*, 5, 262-286.

Sengupta, B. (1989). Markov processes whose steady state distribution is matrix-exponential with an application to the $GI/PH/1$ queue. *Advances in Applied Probability*, 21, 159-180.

Shin, Y. W. (2004). $BMAP/G/1$ queue with correlated arrivals of customers and disasters. *Operations Research Letters*, 32, 364-373.

Squillante, M. S., Zhang, Y., Sivasubramanian, A. and Gautam, N. (2008). Generalized parallel-server fork-join queues with dynamic task scheduling. *Annals of Operations Research*, 160, 227-255.

Takine, T. and Hasegawa, T. (1994). The workload in the $MAP/G/1$ queue with state-dependent services: Its applications to a queue with preemptive resume priority. *Stochastic Models*, 10, 183-204.

Takine, T. (1999). The nonpreemptive priority $MAP/G/1$ queue. *Operations Research*, 47, 917-927.

Telek, M. and Horváth, G. (2007). A minimal representation of Markov arrival processes and a moments matching method. *Performance Evaluation*, 64, 1153-1168.

Tian, N. and Zhang, Z. G. (2006). *Vacation Queueing Models: Theory and Applications*, International Series in Operations Research & Management, Vol. 93. New York: Springer.

Tweedie, R. L. (1982). Operator-geometric stationary distributions for Markov chains, with application to queueing models. *Advances in Applied Probability*, 14, 368-391.

Van Houdt, B. and Blondia, C. (2002). The delay distribution of a type *k* customer in a first-come-first-served $MMAP[K]/PH[K]/1$ queue. *Journal of Applied Probability*, 39, 213-223.

Discussion of "Markovian arrivals in stochastic modelling: a survey and some new results" by Jesús R. Artalejo, Antonio Gómez-Corral and Qi-Ming He

Rafael Pérez Ocón
Departamento de Estadística e Investigación Operativa
Universidad de Granada, España

Matrix-analytic methods (MAMs) have become an important tool for studying complex systems. They preserve the Markovian structure and present the results in a tractable manner. These methods are based in two fundamental elements: the phase-type distributions (PH-distributions) and the Markovian arrival processes (MAPs). Given the potential of these methods, new results and applications arise frequently, and a survey of these methods is very useful from time to time. The paper initiates considering the batch Markovian arrival processes (BMAPs) and describing their properties. The associated counting processes and the descriptors for quantifying the main quantities are given. These processes are introduced in a methodological way, considering examples and particular cases for a better comprehension of how they operate. The application of the methods in queueing, inventories, and reliability is interesting. Variants of the BMAPs that are proven to be useful in applications, the MMAPs and the MAPAs are presented. The BMAPs occupy a central role in the queueing theory, and it is expected that the study and use of these variants will be increasing with time, not only in queueing, but in others domains of application. This part of the paper resumes and illustrates the properties and applications of these classes of processes. The construction of algorithms and computational programs would complete the present paper; it is a challenge for specialists in these topics.

The introduction of block-structured state-dependent event (BSDE) approach for the treatment of stochastic models is an important contribution. Based in the Markovian structure by means of the introduction of phases, this approach allows constructing stochastic models for complex systems. It can be used in the discrete and continuous cases, and some Markovian stochastic models governed by particular MAPs can be deduced from the BSDE approach. The application of the BSDE to the epidemic models illustrates the power of the method, and contributes to consider non-homogeneous stochastic models, involving non-exponential times and the existence of correlation between successive events. The introduction of the non-homogeneity in the MAMs enlarges the possibility of applications that would be very difficult to do following another methodology. The results are complex, but they can be presented in an algorithmic form as a consequence of the MAMs. The incorporation of a methodology and algorithms to elucidate the structure of the BSDE would be useful in the application of this technique for solving problems in different domains of activity.

In the study of stochastic models three are the elements to be considered: modelling, applications, and inference. Modelling and applications must involve methods to be tractable mathematically. The present survey completes and updates previous ones related to modelling and application. Given the complexity of the methods and the speediness of the applications, this is an excellent paper to know the state of the art of the Markovian arrival processes at the present moment.

Thinking of the applications, the paper can be extended in aspects of inference. Essential for the use of MAPs in practice are the numerical algorithms to fit these processes, and the statistical methods for applying to dataset. In the Bibliographical notes in the paper some references about estimation and fitting are given. Related to the fit of phase-type distributions and to the Markov-modulated Poisson process (MMPP), the paper of Asmussen (1997) shows that the EM algorithm can be successfully applied to maximum-likelihood (ML) estimates in Markov models, even in the case of incomplete data, and computational programs for the treatment of the data are constructed and their properties commented. The paper of Asmussen alludes to the previous one of Ryden (1996), where the problem on identifiability and the order of the involved Markov processes in these two particular cases is presented. An area for future research is the inclusion of problems related to the identifiability of general MAPs into the matrix-analytic methods. This will allow to extend the use of MAPs and solve problems that cannot be addressed with the actual knowledge of the inference about these processes.

Asmussen (1997). Phase-type Distributions and Related Point Processes: Fitting and Recent Advances. In: *Matrix-analytic methods in stochastic models*. Chakravarthy, S. R. and Alfa, A. S. (Eds). Marcel Dekker, New York, 137-149.

Ryden (1996). On identifiability and order of continuous-time aggregated Markov chains, Makov-modulated Poisson processes, and phase-type distributions. *Journal of Applied Probability*, 33, 640-653.

Miklos Telek
Budapest University of Technology and Economics
Department of Telecommunications

The paper mainly presents a survey of Markovian arrival process models. It is always hard to decide the level of knowledge of the aimed audience of a paper or a scientific presentation. I think that the goal of a survey paper should be to introduce the main concepts of a field to those who are not that familiar with them yet. Assuming it is the goal of this paper I recommend to be more detailed and precise with the introduction of the applied concepts, a list of explicit points for considerations are forwarded to the authors.

Section 2 starts with the introduction of BMAPs. It is based on a short summary of PH distributions. I would recommend to unify all PH distribution related content into this part.

In a paper like this I prefer derivations starting from a limited number of initial expression than list of final expressions! The majority of the presented complex expressions on MAP properties can be obtained in simple steps from the joint density functions. I recommend at least indicating how to obtain the presented properties (e.g. on page 113).

The relation of structured Markov processes, like quasi birth death processes (QBD), and those generalization of MAPs which account for the arrival and departure of customers (HetSigma, BSDE) is not expressed in the papers. These models can be viewed as queueing systems resulting structured Markov processes. As a consequence efficient computational methods developed for the analysis of structured Markov processes can be applied for the analysis of these arrival processes. A discussion about this relation would further enhance the paper.

The paper introduces the basic theory of various Markovian arrival processes and presents several examples to indicate the wide spread applicability of this versatile set of models. To make this picture complete it would be interesting to add the basic limitations of these models which needs to be considered when applying them in practice.

Some of these limitations are inherited from PH distributions. The most well know one is about the coefficient of variation of the inter-event time distribution which is greater or equal to $1/n$ when the state space of the modulating process is composed by n state. An other typical feature of these models is the exponential asymptotic decay. It holds for a lot of properties like inter-event time distribution, autocorrelation, lag correlation. Beyond these two most well-known ones a set of further practical limitations are published recently. A summary of these limits would be a nice contribution of the manuscript.

Consequently, real systems with quasi deterministic inter-event times or strange decay behaviour or any other property in conflict with the limits of these models cannot be closely modelled with Markovian arrival models. But fortunately also in these cases, in accordance with the denseness property (Section 2.2.4), a computational complexity – accuracy trade-off can be found by increasing the size of the Markovian model.

Yiqiang Q. Zhao
School of Mathematics and Statistics
Carleton University, Canada

First, I would like to congratulate the authors on this excellent comprehensive review on BMAP. This review paper provides readers with easy access to all the important aspects of the BMAP, from its definition to its basic properties; from its history to its extensions; from theoretical aspects to applications.

Applying BMAP in modelling is popular not only because it is a natural generalization of the Poisson process and captures correlations between arrivals, but also because of the more important fact that the use of BMAPs in modelling often leads to a matrix-structured formalism, to which the powerful matrix-analytic method can be applied.

The variants and generalizations touched on in the review paper have been well chosen by the authors, as they also lead to matrix formulations for which analysis can be carried out in terms of matrix-analytic methods. The contents of Section 4 are interesting, though structurally this section seems sidetracked from the main focus of the review. The variants and generalizations of BMAPs could have also gone in a few different directions. One of such alternatives is a comparison, of modelling properties, of the arrival models discussed in the review paper and other commonly seen arrivals, such as arrivals with long-range dependence, Gaussian queues, periodic arrivals and possibly others.

Markov additive processes deserve special attention among all generalizations of BMAPs. The reason for this goes back to the core of the matrix-analytic method. The quasi-birth-and-death (QBD) process is considered an excellent example for explicitly demonstrating some of the key techniques in the core of the matrix-analytic method, such as duality, probabilistic measures under taboo or censoring technique. A comprehensive summary of QBD processes can be found in Latouche and Ramaswami (1999). These techniques, together with Wiener-Hopf factorizations including RG-factorizations and block-form generating functions (or exponential change of matrix (measure)), lead to a concise treatment of the more general matrix-structured paradigm, the GI/G/1 type of matrices in parallel to that for the QBD process, for example, see Zhao, Li and Braun (1998, 2003). The sequence of the non-boundary matrices in the GI/G/1 paradigm leads to a Markov additive process with finitely many background states. It is of interest to notice that the above mentioned techniques are in fact key general tools and methods for queues in applied probability, for example, see Asmussen (2003).

Standard matrix-analytic methods deal with matrices of finite size, like BMAPs, since the method, in both theoretical and computational aspects, relies on properties of

finite dimensional linear spaces or finite matrices. Attempts to generalize finite matrices to infinite ones have a long history dating back to the early 80s, including Tweedie (1982), Ramaswami and Taylor (1996), and Shi, Guo and Liu (1996), among others. Although basic formalizations stand valid for models with infinite matrices, such as the operator-geometric solution and generalized phase type distributions described by an absorbing Markov chain with infinitely many states, there are two main challenges when finite matrices are extended to infinite ones: (1) many key properties from linear algebra are no longer valid for infinite matrices and instead infinite dimensional linear operators now play a key role; and (2) additional non-trivial efforts should be made to address computational issues of the R- and G-measures since they are no longer finite matrices. Recently, analysis of exact tail asymptotics in the stationary probability distribution for a model whose non-boundary matrices defines an additive process with an infinite background space has been a central topic in terms of (extended) matrix-analytic methods. Tail asymptotics can lead to various performance bounds and accurate approximations. The core of extended matrix-analytic methods consists of the same general tools used in the applied probability mentioned above, such as limit theorems for Markov renewal processes, censoring, RG-factorizations, duality, exponential change of matrix. These tools and properties of Markov additive processes are the key for the success of expanding matrix-analytic methods. References in this direction include Takahashi, Fujimoto and Makimoto (2001), Haque (2003), Kroese, Scheinhardt and Taylor (2004), Miyazawa (2004), Miyazawa and Zhao (2004), and He, Li, and Zhao (2009), among others.

Finally, it was a great pleasure for me to be invited as a discussant for this interesting review paper.

References

Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edition. Springer.

Haque, L. (2003). *Tail behaviour for stationary distributions for two-dimensional stochastic models*. Ph.D. Thesis, Carleton University, Ottawa, ON, Canada.

He, Q., Li, H. and Zhao, Y.Q. (2009). Light-tailed behaviour in QBD process with countably many phases. *Stochastic Models*, 25, 50-75.

Kroese, D.P., Scheinhardt, W.R.W. and Taylor, P.G. (2004). Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. *Annals of Applied Probability*, 14(4), 2057-2089.

Miyazawa, M. (2004). The Markov renewal approach to M/G/1 type queues with countably many background states. *Queueing Systems*, 46, 177-196.

Miyazawa, M. and Zhao, Y.Q. (2004). The stationary tail asymptotics in the GI/G/1 type queue with countably many background states. *Advances in Applied Probability*, 36(4), 1231-1251.

Ramaswami, V. and Taylor, P.G. (1996). Some Properties of the Rate Operators in level dependent Quasi Birth and Death Processes with a Countable Number of Phases. *Stochastic Models*, 12(1), 143-164.

Shi, D.H., Guo J. and Liu, L. (1996). SPH-distributions and the rectangle iterative algorithm. *Matrix-Analytic Methods in Stochastic Models*, in S. Chakravarthy and A.S. Alfa (eds). New York: Marcel Dekker, 207-224.

Takahashi, Y., Fujimoto, K. and Makimoto, N. (2001). Geometric decay of the steady-state probabilities in a quasi-birth-and-death process with a countable number of phases. *Stochastic Models*, 17(1), 1-24.

Zhao, Y.Q., Li, W. and Braun, W.J. (1998). Infinite block-structured transition matrices and their properties. *Advances in Applied Probability*, 30, 365-384.

Zhao, Y.Q., Li, W. and Braun, W.J. (2003). Censoring, factorizations, and spectral analysis for transition matrices with block-repeating entries. *Methodology and Computing in Applied Probability*, **5**, 35-58.

Rejoinder

First of all, we would like to thank the three invited discussants for the time spent commenting on our paper. We appreciate their constructive and insightful comments, which have made valuable contributions to the understanding of various interesting problems.

We now briefly respond to some of their comments.

Comments from Prof. R. Pérez-Ocón

Prof. Pérez-Ocón comments on the important role played by the matrix-analytic formalism and the Markovian arrival processes in stochastic modelling. We thank the discussant for his positive and kind remarks on the recently introduced BSDE approach. At a first glance, the BSDE approach and the matrix-analytic methods present common elements; e.g. structured Markov chains, phase method. Although the BSDE approach is closely related to the methods developed for structured Markov chains, the aim of the BSDE approach is to reduce the cost caused from an excessive dimensionality in the matrix representation, which frequently occurs in non-homogenous settings where an arbitrary number of MAPs and/or PH distributions are simultaneously involved. In this sense, the BSDE approach goes beyond the commonly used matrix-analytic methods. Thus, we completely agree with the remarks of the discussant about the need of developing methodological and algorithmic tools for practical use of the BSDE approach. In particular, efforts leading to a suitable treatment of the positive recurrence of infinite structured non-homogeneous Markov chains would be welcome.

Other relevant points commented by the discussant are the fitting and inference aspects. We touched these matters only in the bibliographical notes, where some selected references were given. We are happy that the discussant is adding basic references that will assist readers who are interested in pursuing this subject further.

Comments from Prof. M. Telek

Prof. Telek pointed out in a separate communication a number of helpful comments to improve the paper presentation. These comments have been partially taken into account. We have also incorporated some additional citations in the text, which should be helpful for those readers desirous of knowing how to derive the presented properties.

In the opinion of the discussant, the HetSigma and the BSDE approaches can be viewed as queueing systems resulting in structured Markov processes. Regarding to the HetSigma approach, Chakka and Do (2007) clearly assert that transitions from a level to any other level are possible. Therefore, the matrix structure is general and the standard matrix-analytic methods cannot be used directly. We stress that our interest in the HetSigma approach comes from the fact that both the arrival and the service processes are modulated by the same Markov process. On the other hand, the BSDE approach is intended to construct either a specific part (i.e., the arrival process) or a whole stochastic model in state-dependent frameworks where neither a well-posed matrix structure or the reducibility of the resulting Markov chain are assumed. In this setting, it is our opinion that the possibility of using the classical matrix-analytic tools is limited. Further methodological and computational efforts are definitively needed, as it was mentioned by Prof. Pérez-Ocón.

The discussant accurately points out some limitations of the PH distribution and consequently of the BMAP, whose distribution of inter-arrival times is of PH type; see Subsection 2.3.2 of the paper. This fact leads to a geometrically decaying correlation structure which makes the MAPs less suitable to model certain correlated input processes. Despite of this difficulty, Markovian arrival processes have been also used to model arrivals with long-range dependence whose autocovariance function decays slower than exponentially; see the references given in our reply to Prof. Zhao.

As a general comment, it should be noticed that catching properly some real inputs with time dependence implies to use MAPs of an excessive large order. This important issue connects with the computational cost inherent to the matrix-analytic formalism. Thus, the use of MAPs in practice is limited by the existing fitting methods. The development of good fitting methods for MAPs is a very interesting research topic, which has received a significant attention during the last years. In addition to the references in Section 5 of our paper (see also the comments by Prof. Pérez-Ocón), we now just add one more recent paper by Casale et al. (2010). In this paper, the MAP fitting is based on the Kronecker product composition method. The paper provides an exhaustive study that includes a discussion on some fundamental difficulties of MAP fitting. In another related work, Bause et al. (2009) provide an experimental comparison between MAPs and ARMA (*auto regressive moving average*) and ARTA (*auto regressive to anything*) based models. The authors conclude that MAP fitting is most demanding in terms of running time.

Comments from Prof. Y.Q. Zhao

Prof. Zhao points out that the paper did not give a complete survey on the possible variants and generalizations of the BMAP. More concretely, the discussant mentions arrivals with long-range dependence, periodic arrivals and Gaussian queues as other alternative arrival processes. There exists a number of papers (e.g. Andersen and

Nielsen (1998), Casale et al. (2008), and Salvador et al. (2004)) where Markovian arrival processes and, specifically, superpositions of MMPPs are used as a very versatile tool to model variable packet traffic exhibiting long-range dependence. The Hurst parameter introduced by Willinger et al. (1995) is frequently used to measure long-range dependence. Periodic arrivals are related to time-inhomogeneous structures; see Section 3.4 in the paper. We agree that periodic arrivals have interest in modelling communication networks. These arrival inputs include, among others, the periodic Poisson process (see Margolius (2007)) and the periodic BMAP (see Breuer (2003)). Despite of the interest in Gaussian sources and Gaussian queues, it is our opinion that they are not commonly analyzed through those techniques belonging to the core of the matrix-analytic methods. We would recommend the book by Mandjes (2007) to the interested readers.

Other important comments from the discussant are regarding to the relevance of a variety of techniques, such as duality, taboo and censoring, and *RG*-factorizations, in the core of the matrix-analytic methods. The discussant accurately makes observations on these techniques as in fact very general and powerful methods for investigating challenging problems including generalization from finite blocks to Markov chains with infinite blocks. Prof. Zhao provides a set of references that deal with this issue, putting emphasis on tail asymptotic results. These comments are more relevant to matrix-analytic methods in general, rather than Markovian arrival processes. We thank Prof. Zhao for this valuable addition.

Finally, we would like to thank once again the discussants. We sincerely hope that our review paper and their comments will be of interest for the audience of this journal. We also take this opportunity to thank the Editor-in-Chief, M. Guillén, and the Executive Editor, P. Puig, for their kind invitation to write the paper and for organizing the stimulating discussion.

Additional references

Andersen, A. T. and Nielsen, B. F. (1998). A Markovian approach for modelling packet traffic with long-range dependence. *IEEE Journal on Selected Areas in Communications*, 16, 719-732.

Bause, F., Buchholz, P. and Kriege, J. (2009). A comparison on Markovian arrival and ARMA/ARTA processes for the modelling of correlated input processes. In *Proceedings of the 2009 Winter Simulation Conference*, M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls (eds.), IEEE Press, 634-645.

Breuer, L. (2003). *From Markov Jump Processes to Spatial Queues*. Dordrecht: Kluwer Academic Publishers.

Casale, G., Mi, N. and Smirni, E. (2008). Versatile models of systems using MAP queueing networks. In *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing (IPDPS 2008)*, IEEE Press, 1-5.

Casale, G., Zhang, E. and Smirni, E. (2010). KPC-Toolbox: Best recipes for automatic trace fitting using Markovian arrival processes. *Performance Evaluation*, 67, 873-896.

Mandjes, M. (2007). *Large Deviations for Gaussian Queues: Modelling Communication Networks*. Chichester: John Wiley & Sons.

Margolius, B. H. (2007). Transient and periodic solution to the time-inhomogeneous quasi-birth death process. *Queueing Systems*, 56, 183-194.

Salvador, P., Pacheco, A. and Valadas, R. (2004). Modelling IP traffic: Joint characterization of packet arrivals and packet sizes using BMAPs. *Computer Networks*, 44, 335-352.

Willinger, W., Taqqu, M. S., Leland, W. E. and Wilson, D. V. (1995). Self-similarity in high-speed packet traffic: Analysis and modelling of Ethernet traffic measurements. *Statistical Science*, 10, 67-85.