Departure process analysis of the multi-type MMAP[K]/PH[K]/1 FCFS queue

G. Horváth[1]

MTA-BME Information systems research group, Magyar Tudosok krt. 2, 1117 Budapest, Hungary

B. Van Houdt

Department of Mathematics and Computer Science, University of Antwerp - iMinds

Abstract

The analysis of the departure process of queues is important in several aspects, for instance, it plays a prominent role in the decomposition based analysis of open queueing networks. While there are several results available for the departure process analysis of MAP driven single-class (or, single-type) queues, there are very few results available for the multi-type variants of these queues.

In this paper we consider the departure process of the multi-type MMAP[K]/PH[K]/1 FCFS queue. We derive the joint Laplace-Stieltjes transform of the lag-n inter-departure times, and provide efficient algorithms to compute the lag-1 joint moments, the lag-n joint means and cross correlations of the inter-departure times.

While the analysis of the departure process is typically performed via the queue length distribution at departure instants, we rely on the age process to derive various properties of the departure process.

Keywords: Departure process, age process, multi-type queues

1. Introduction

The research interest in the analysis of the departure processes of various queues dates back to the appearance of queueing theory itself. The departure process helps to understand the shaping effect of the queue and the impact of the service policy on the incoming traffic. However, the study of the departure process is primarily motivated by the analysis of queueing networks, where the departure process of a queue is the arrival process of a subsequent queue. The

Email addresses: ghorvath@hit.bme.hu (G. Horváth), benny.vanhoudt@ua.ac.be (B. Van Houdt)

[1]This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

exact analysis of open queueing networks is a difficult problem except of some models that are often a bit too restrictive in practice. A popular (approximate) solution technique for open, acyclic queueing networks is based on traffic based decomposition. According to this technique each queue is analyzed in isolation, and the incoming traffic of the subsequent connected queues are approximated based on several properties of the departure process of the queue. The accuracy of the characterization of the departure process has a crucial impact on the accuracy of the queueing network analysis itself.

Various properties of the departure process of MAP driven queues have been determined in an exact manner over the last ten years. In the single-type case (where the customers are treated to be identical), when the service process is a Markovian Arrival Process (MAP) and the waiting room has infinite capacity, it is well known that the departure process is a MAP with an infinite number of phases. Solutions based on the truncation of this process are presented in [1] and [2], that capture the the marginal distribution of the inter-departure times accurately, but not the correlations, and in [3, 4], where a large MAP is constructed that preserves the joint distribution of the inter-departure times up to lag-n. The case with a semi-Markovian service process is considered in [5], where the joint Laplace-Stieltjes transform (LST) of the inter-departure times are provided as well as the generating function of the covariance sequence. [6] and [7] consider a system with BMAP arrivals and independent identically distributed general service times and provide the joint LST of the inter-departure times and the auto-covariance function up to lag-n. A numerically efficient method to compute the auto-covariance function up to lag-n, for n large, was also discussed in [6].

The are far fewer results available for multi-type MAP (or MMAP) driven queues. [8] derives the marginal distribution of the inter-departure times for FCFS and nonpreemptive priority MMAP[2]/G[2]/1 queues. In [9] the lag-1 joint moments of the inter-departure times are considered for priority queues with MMAP input and PH service times, but again limited to two classes.

In this paper we provide the departure process analysis of the MMAP[K]/PH[K]/1 FCFS queue. The principal difficulty we are faced with in this system is that we can not use the typical approach taken in the past as it is based on the knowledge of the queue length behavior at departure instants. In case of the MMAP[K]/PH[K]/1 FCFS queue the queue length distribution does not have a simple closed form solution, it is in fact rather involved to work with. Instead, our solution is based on the age process, that has a simple matrix-exponential distribution for this system.

The paper is organized as follows. Section 2 introduces the notations used in the paper. Section 3 gives a short overview on the age process of the MMAP[K]/PH[K]/1 FCFS queues. The main contribution of the paper is Section 4, where the joint LST of the lag-n inter-departure times is derived. Based on these results, Section 5 provides the lag-1 properties, while Section 6 provides the lag-n joint means and cross covariances of the inter-departure times. Section 7 discusses algorithmic complexity, while Section 8 provides some numerical examples. Finally, conclusions are drawn in Section 9.

2. Notations and basic relations

We consider a queueing system with a single server, infinite waiting room, a marked Markovian arrival process (MMAP, [10]), phase type (PH) distributed service times and a first-come-first-served (FCFS) service discipline.

Let us denote the number of customer types by $K \geq 1$. The MMAP[K] characterizing the arrivals is given by a set of $m_a \times m_a$ matrices $\boldsymbol{D_k}, k = 0, \ldots, K$, where $(\boldsymbol{D_0})_{j,j'}$, for $j \neq j' \in \{1, \ldots, m_a\}$, holds the rate at which the underlying Markov chain jumps from phase j to j' while no arrival occurs and $(\boldsymbol{D_k})_{j,j'}$, for $j, j' \in \{1, \ldots, m_a\}$ and $k = 1, \ldots, K$, the jump rate from phase j to j' accompanied by the arrival of a type k customer. Finally, $(-\boldsymbol{D_0})_{j,j} = \sum_{j' \neq j} (\boldsymbol{D_0})_{j,j'} + \sum_{k=1}^{K} \sum_{j'} (\boldsymbol{D_k})_{j,j'}$, such that $\boldsymbol{D} = \sum_{k=0}^{K} \boldsymbol{D_k}$ is the generator of the underlying Markov process $Z(t)$ with state space $\{1, \ldots, m_a\}$, where as usual we assume that $\boldsymbol{D}$ is irreducible. The mean arrival rate of type k customers is denoted by λ_k and is calculated as $\lambda_k = \theta \boldsymbol{D_k} \mathbb{1}$ with vector θ being the unique solution of $\theta \boldsymbol{D} = 0, \theta \mathbb{1} = 1$ ($\mathbb{1}$ denotes the column vector of ones). The overall arrival rate is $\lambda = \sum_{k=1}^{K} \lambda_k$. Note that the vector θ is the steady state distribution of the MMAP phase. We can also define the steady state distribution at arrival epochs as $\alpha = \theta \sum_{k=1}^{K} \boldsymbol{D_k} / \lambda$. For further use, define $\boldsymbol{P_k} = (-\boldsymbol{D_0})^{-1} \boldsymbol{D_k}$, for $k \in \{1, \ldots, K\}$, where $(-\boldsymbol{D_0})^{-1}$ is well defined due to the irreducibility of $\boldsymbol{D}$ and all the eigenvalues of $\boldsymbol{D_0}$ lie in the open left half plane.

The service time of the type k customers is PH distributed with m_k phases, given by initial vector σ_k and transient generator $\boldsymbol{S_k}, k = 1, \ldots, K$. Vector β_k denotes the steady state phase distribution of the service process, that is, the unique solution of $\beta_k(\boldsymbol{S_k} - \boldsymbol{S_k} \mathbb{1} \sigma_k) = 0, \beta_k \mathbb{1} = 1$. The service rate of type k customers is then $\mu_k = \beta_k(-\boldsymbol{S_k})\mathbb{1}$. The density function of the service time $f_{S_k}(x)$, the LST $f^*_{S_k}(s)$, and its nth moment $E(S_k^n)$ are given by

$$
\begin{aligned}
f_{S_k}(x) &= \sigma_k e^{\boldsymbol{S_k} x} (-\boldsymbol{S_k}) \mathbb{1}, \\
f^*_{S_k}(s) &= \sigma_k (s\boldsymbol{I} - \boldsymbol{S_k})^{-1} (-\boldsymbol{S_k}) \mathbb{1}, \\
E(S_k^n) &= n! \sigma_k (-\boldsymbol{S_k})^{-n} \mathbb{1},
\end{aligned}
$$

where $\boldsymbol{I}$ is the identity matrix.

Furthermore, let $J(t)$ denote the phase of the service process and $C(t)$ the type of the customer in the server at time t (if any). The load ρ of the queue is defined as $\rho = \sum_{k=1}^{K} \lambda_k / \mu_k$ and represents the fraction of time that the server is busy (provided that $\rho < 1$).

In this paper we will make extensive use of the age process. The age process of the MMAP[K]/PH[K]/1 FCFS queue is defined as $\{(A(t), C(t), J(t), Z(t - A(t)), t \geq 0\}$, where $A(t)$ is the age of the customer in the system at time t whenever the server is busy. Notice, the age process keeps track of the MMAP[K] phase immediately after the arrival time of the customer in service. When the server is idle, one could define $A(t) = 0$. However, to study the departure process we will only require the density of the age process just before service completion instants, which can be derived from the process that observes the system only when the server is busy.

3. The distribution of the age process

The class of MMAP[K]/PH[K]/1 queues forms a subclass of the SM[K]/PH[K]/1 queues, the age process of which was analyzed in [11]. We start the discussion with a short summary of the main results in [11].

As we have only one server, the two dimensional process $\{(C(t), J(t)), t \geq 0\}$ describing the customer type in service and the current service phase can be represented by a PH distribution of size $m_s = \sum_{k=1}^{K} m_k$ with the following generator

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S_1} & & & \\ & \boldsymbol{S_2} & & \\ & & \ddots & \\ & & & \boldsymbol{S_K} \end{bmatrix},$$

and the initial vector given that there is a type k customer in the server is

$$\sigma^{(k)} = (\underbrace{0, \dots, 0}_{\sum_{j=1}^{k-1} m_j}, \sigma_k, \underbrace{0, \dots, 0}_{\sum_{j=k+1}^{K} m_j}), \quad k = 1, \dots, K.$$

For further use, let $m = m_a m_s$.

Let us denote the density of the steady state distribution of the age process by $\pi(x) = \{\pi_i(x), i = 1, \dots, m\}$. According to [12, 11], $\pi(x)$ has a matrix-exponential form:

$$\pi(x) = \pi(0) e^{\boldsymbol{T}x}, x \geq 0,$$

where the density function at $x = 0$ is given by

$$\pi(0) = \frac{1}{\rho} \sum_{k=1}^{K} \frac{\lambda_k}{\mu_k} \left((0, \dots, 0, \beta_k, 0, \dots, 0) \otimes \frac{\theta \boldsymbol{D_k}}{\lambda_k} \right) (-\boldsymbol{T}). \tag{1}$$

Note that the right-hand term of the Kronecker product is the phase distribution of the MMAP at the type k arrival epochs.

The matrix $\boldsymbol{T}$ is of size $m \times m$ and it is the minimal solution of the following matrix equation (derived from [12] by simple algebraic manipulations):

$$\boldsymbol{T} = \boldsymbol{S} \otimes \boldsymbol{I} + \underbrace{\int_{x=0}^{\infty} e^{\boldsymbol{T}x} \left(-\boldsymbol{S}\mathbb{1} \otimes \boldsymbol{I} \right) e^{\boldsymbol{D_0}x} dx}_{\boldsymbol{Y_0}} \sum_{k=1}^{K} \left(\sigma^{(k)} \otimes \boldsymbol{D_k} \right). \tag{2}$$

Theorem 4.4 in [11] also indicates that all the eigenvalues of $\boldsymbol{T}$ lie in the open left half plane. We recall the following theorem that implies that $\boldsymbol{Y_0}$ is the unique solution of the following Sylvester matrix equation:

$$\boldsymbol{T}\boldsymbol{Y_0} + \boldsymbol{Y_0}\boldsymbol{D_0} = (\boldsymbol{S}\mathbb{1}) \otimes \boldsymbol{I}. \tag{3}$$

Theorem 1 (Theorem 9.2 [13]). *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are square matrices with eigenvalues in the open left half plane, then the unique solution of the equation $\boldsymbol{AX} + \boldsymbol{XB} = -\boldsymbol{C}$ can be expressed as*

$$\boldsymbol{X} = \int_{x=0}^{\infty} e^{\boldsymbol{A}x}\boldsymbol{C}e^{\boldsymbol{B}x}dx.$$

We could compute $\boldsymbol{Y_0}$ by solving (3) if we first compute $\boldsymbol{T}$, however, by constructing a Markovian fluid queue as in [14], one can show that the $m \times m_a$ matrix $\boldsymbol{Y_0}$ is the minimal non-negative solution to the following algebraic Riccati equation:

$$(-\boldsymbol{S}\mathbb{1}) \otimes \boldsymbol{I} + \boldsymbol{Y_0 D_0} + (\boldsymbol{S} \otimes \boldsymbol{I})\boldsymbol{Y_0} + \boldsymbol{Y_0} \sum_{k=1}^{K} \left(\sigma^{(k)} \otimes \boldsymbol{D_k}\right) \boldsymbol{Y_0} = 0, \tag{4}$$

which does not depend on $\boldsymbol{T}$. This algebraic Riccati equation can be solved iteratively by means of the Structure-preserving Doubling Algorithm (SDA) [15] or the Alternating-Directional Doubling Algorithm (ADDA) [16]. Both these algorithms have quadratic convergence as opposed to the algorithms proposed in [12, 11] to compute $\boldsymbol{T}$, the convergence of which is only linear.

Since we are focusing on the departure process in this paper, we will be interested in the age process embedded at service completion instants most of the time. The density of the age process just before service completion instants $\pi_D(x) = \{(\pi_D(x))_i, i = 1, \ldots, m_a\}$ can be expressed as

$$\pi_D(x) = \frac{1}{\lambda}(\rho\pi(0)e^{\boldsymbol{T}x})(-\boldsymbol{S}\mathbb{1} \otimes \boldsymbol{I}), \quad x > 0, \tag{5}$$

and integrating $\pi_D(x)$ over x gives

$$\int_{x=0}^{\infty} \pi_D(x)\, dx = \sum_{k=1}^{K} \frac{1}{\mu_k} \left((0, \ldots, 0, \beta_k, 0, \ldots, 0)(-\boldsymbol{S}\mathbb{1}) \otimes \frac{\theta \boldsymbol{D_k}}{\lambda} \right) = \alpha, \tag{6}$$

as $\mu_k = \beta_k(-\boldsymbol{S_k})\mathbb{1}$.

Finally, we recall the following theorem, which is due to Theorem 1 in [17].

Theorem 2. *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are square matrices with*

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{C} \\ 0 & \boldsymbol{B} \end{bmatrix},$$

then

$$e^{\boldsymbol{X}t} = \begin{bmatrix} e^{\boldsymbol{A}t} & \int_{a=0}^{t} e^{\boldsymbol{A}a}\boldsymbol{C}e^{\boldsymbol{B}(t-a)}da \\ 0 & e^{\boldsymbol{B}t} \end{bmatrix}.$$

4. The lag-n joint transform

In this section we derive an expression for the joint LST of the 1st and $(n+1)$th inter-departure time. Let T_n denote the nth departure time, with $T_0 = 0$ and set $\tau_n = T_n - T_{n-1}$ for $n \geq 1$. Further let C_n denote the type of the nth departing customer, then $f_{D(n)}^{(k,p)*}(s_1, s_2)$ is defined as

$$f_{D(n)}^{(k,p)*}(s_1, s_2) = \int_{t_1=0}^{\infty}\int_{t_2=0}^{\infty} e^{-s_1 t_1 - s_2 t_2} dP[\tau_1 < t_1, \tau_{n+1} < t_2, C_1 = k, C_{n+1} = p], \tag{7}$$

for $p, k \in \{1, \ldots, K\}$ and $n \geq 1$.

Observe that the inter-departure times are

- either equal to service times (during the busy periods of the queue),
- or, if a customer arrives to an idle queue, the service time of the customer plus the preceding idle time.

Due to this kind of relation between the busy periods and the inter-departure times we introduce several busy period related quantities before providing the solution for (7).

Denote $\boldsymbol{I_i} = [\boldsymbol{I}\ \boldsymbol{0}\ \ldots\ \boldsymbol{0}]$ and $\boldsymbol{J_i} = [\boldsymbol{0}\ \ldots\ \boldsymbol{0}\ \boldsymbol{I}]^T$ such that they have size $m_a \times i \cdot m_a$ and $i \cdot m_a \times m_a$, respectively. Further, let $\boldsymbol{J_{i,j}} = [\boldsymbol{J_j}^T\ \boldsymbol{0}\ \ldots\ \boldsymbol{0}]^T$ such that it is a size $i \cdot m_a \times m_a$ matrix.

We start by defining $(\boldsymbol{M_{k,i}}(t))_{j,j'}$, for $i \geq 1$, $k \in \{1, \ldots, K\}$ and $j, j' \in \{1, \ldots, m_a\}$, as the probability that i customers get served during a busy period that was initiated by a type k arrival, while the service time of the initial type k customer is at most t and the MMAP phase equals j at the start and j' at the end of the busy period. Let $\boldsymbol{M_{k,i}}(t)$ be the matrix with entry (j, j') equal to $(\boldsymbol{M_{k,i}}(t))_{j,j'}$. Let $\boldsymbol{M^*_{k,i}}(s)$ be the LST of $\boldsymbol{M_{k,i}}(t)$ and denote $\boldsymbol{M^*_{k,i}}(0) = \boldsymbol{M_{k,i}}(\infty)$ as $\boldsymbol{M_{k,i}}$.

The following lemma gives an expression for the matrices $\boldsymbol{M^*_{k,i}}(s)$. We note that the final results do note require the computation (nor the inversion) of the size $i \cdot m_a \times i \cdot m_a$ matrices $\boldsymbol{Q_i}$ defined in this lemma.

Lemma 1. *The matrices $\boldsymbol{M^*_{k,i}}(s)$ can be expressed recursively as*

$$\boldsymbol{M^*_{k,i}}(s) = (\sigma_k \otimes \boldsymbol{I_i})((s\boldsymbol{I} - \boldsymbol{S_k}) \oplus \boldsymbol{Q_i})^{-1}(-\boldsymbol{S_k}\mathbb{1} \otimes \boldsymbol{J_i}). \tag{8}$$

where $\boldsymbol{Q_1} = \boldsymbol{D_0}$ and $\boldsymbol{Q_i}$ is the size $i \cdot m_a \times i \cdot m_a$ block triangular block Toeplitz matrix given by

$$\boldsymbol{Q_i} = \begin{bmatrix} \boldsymbol{D_0} & \sum_{q=1}^{K} \boldsymbol{D_q M_{q,1}} & \ldots & \sum_{q=1}^{K} \boldsymbol{D_q M_{q,i-1}} \\ & \ddots & \ddots & \vdots \\ & & \boldsymbol{D_0} & \sum_{q=1}^{K} \boldsymbol{D_q M_{q,1}} \\ & & & \boldsymbol{D_0} \end{bmatrix}. \tag{9}$$

Proof. As $e^{\boldsymbol{A}\otimes\boldsymbol{I}} = e^{\boldsymbol{A}} \otimes \boldsymbol{I}$, $e^{\boldsymbol{I}\otimes\boldsymbol{B}} = \boldsymbol{I} \otimes e^{\boldsymbol{B}}$ and $e^{\boldsymbol{A}+\boldsymbol{B}} = e^{\boldsymbol{A}}e^{\boldsymbol{B}}$ if $\boldsymbol{A}$ and $\boldsymbol{B}$ commute, (8) can be rewritten as

$$\begin{aligned} \boldsymbol{M}^*_{\boldsymbol{k},\boldsymbol{i}}(s) &= (\sigma_k \otimes \boldsymbol{I_i}) \int_{y=0}^{\infty} \left(e^{(\boldsymbol{S_k}-s\boldsymbol{I})y} \otimes \boldsymbol{I} \right) \left(\boldsymbol{I} \otimes e^{\boldsymbol{Q_i}} \right) dy (-\boldsymbol{S_k}\mathbb{1} \otimes \boldsymbol{J_i}) \\ &= \int_{y=0}^{\infty} \left(\sigma_k e^{\boldsymbol{S_k}y} (-\boldsymbol{S_k})\mathbb{1} e^{-sy} \right) \otimes \left(\boldsymbol{I_i} e^{\boldsymbol{Q_i}} \boldsymbol{J_i} \right) dy. \end{aligned}$$

As such it suffices to prove that

$$\boldsymbol{M}^*_{\boldsymbol{k},\boldsymbol{i}}(s) = \int_{y=0}^{\infty} f_{S_k}(y) e^{-sy} \boldsymbol{I_i} e^{\boldsymbol{Q_i}y} \boldsymbol{J_i}\, dy, \tag{10}$$

with $\boldsymbol{Q_i}$ given by (9). The result for $i = 1$ is immediate as $\boldsymbol{I_1} = \boldsymbol{I} = \boldsymbol{J_1}$, $\boldsymbol{Q_1} = \boldsymbol{D_0}$ and

$$\boldsymbol{M_{k,1}}(t) = \int_{y=0}^{t} f_{S_k}(y) e^{\boldsymbol{D_0}y}\, dy,$$

as there should be no arrivals during the service of the type k customer that initiated the busy period.

To establish the general case, we assume that the order of service is preemptive (resume) last-come-first-served instead of first-come-first-served. Notice, the probabilities $(\boldsymbol{M_{k,i}}(t))_{j,j'}$ are not affected by the order of service and therefore the expressions are also valid for the first-come-first-served order considered in this paper.

Assume that the type k customer is in service and that the first arrival that occurs during the service is of type q. Then with probability $(\boldsymbol{M_{q,r_1}})_{j,j'}$ this arrival induces its own sub-busy period during which r_1 customers are served, while the MMAP phase changes from j to j'. Hence, when the type k customer resumes service the MMAP phase equals j'. If another customer arrives while the type k customer is served, this customer will induce another sub-busy period during which r_2 customers are served, etc. Hence, when the initial type k customer gets interrupted for the n-th time, the MMAP phase changes according to the matrix $\sum_{q=1}^{K} \boldsymbol{D_q}\boldsymbol{M_{q,r_n}}$. In order to have exactly i customers served, the sum of all the r_n values should equal $i-1$. Hence, if the service time of the initial type k customer equals y, we therefore find that $(\boldsymbol{I_i}e^{\boldsymbol{Q_i}y}\boldsymbol{J_i})_{j,j'}$ represents the probability that i customers are served during the busy period initiated by the type k customer, while the MMAP phase equals j at the start and j' at the end of the busy period. This suffices to establish (10). □

Define the $(\boldsymbol{Z_{k,i}}(t))_{j,j'}$, for $i \geq 1$, $k \in \{1, \ldots, K\}$ and $j, j' \in \{1, \ldots, m_a\}$, as the probability that the ith customers leaves the server idle at departure time, given that a type k arrival (called the 1st customer) initiated a busy period, the service time of customer 1 is at most t and the MMAP phase equals j at the start of the busy period and j' when the ith customer leaves. Note, the ith customer marks the end of a busy period, but not necessarily the one initiated by customer 1 (unless $i = 1$). Let $\boldsymbol{Z_{k,i}}(t)$ be the matrix with entry (j, j') equal to $(\boldsymbol{Z_{k,i}}(t))_{j,j'}$. Let $\boldsymbol{Z^*_{k,i}}(s)$ be the LST of $\boldsymbol{Z_{k,i}}(t)$ and denote $\boldsymbol{Z^*_{k,i}}(0) = \boldsymbol{Z_{k,i}}(\infty)$ as $\boldsymbol{Z_{k,i}}$.

Lemma 2. *The $m_a \times m_a$ matrices $\boldsymbol{Z^*_{k,i}}(s)$ can be expressed recursively as*

$$\begin{aligned} \boldsymbol{Z^*_{k,1}}(s) &= \boldsymbol{M^*_{k,1}}(s), \\ \boldsymbol{Z^*_{k,i}}(s) &= \boldsymbol{M^*_{k,i}}(s) + \sum_{j=1}^{i-1} \boldsymbol{M^*_{k,j}}(s) \left(\sum_{q=1}^{K} \boldsymbol{P_q} \boldsymbol{Z_{q,i-j}} \right), \end{aligned} \tag{11}$$

for $i \geq 2$.

Proof. The equality $\boldsymbol{Z^*_{k,1}}(s) = \boldsymbol{M^*_{k,1}}(s)$ is immediate as $(\boldsymbol{Z_{k,1}}(t))_{j,j'}$ and $(\boldsymbol{M_{k,1}}(t))_{j,j'}$ represent the same probability. For $i \geq 2$, there are two options: either the busy period initiated by customer 1 ends when the i-th customer leaves (which corresponds to $\boldsymbol{M^*_{k,i}}(s)$) or it ends when the jth customer leaves with $j < i$. In the latter case assume the $j+1$th customer is of type q, then this customer initiates another busy period and we still demand that customer $i > j$ leaves the server idle. Hence, in the latter case we find

$$\boldsymbol{Z_{k,i}}(t) = \sum_{j=1}^{i-1} \boldsymbol{M_{k,j}}(t) \left(\sum_{q=1}^{K} (\boldsymbol{D_0})^{-1} \boldsymbol{D_q} \boldsymbol{Z_{q,i-j}} \right),$$

which implies (11). □

Define the $(\boldsymbol{H_{k,n}}(t,x))_{j,j'}$, for $n \geq 1$, $k \in \{1, \ldots, K\}$ and $j, j' \in \{1, \ldots, m_a\}$, as the following conditional probability: given that an age x customer (labeled customer 0) departs and the MMAP phase at its arrival time was j, $(\boldsymbol{H_{k,n}}(t,x))_{j,j'}$ holds the probability that (a) the next customer (labeled customer 1) is of type k, (b) the inter-departure time between customers 0 and 1 is at most t, (c) customer n leaves the server idle and (d) the MMAP phase equals j' when customer n departs. Let $\boldsymbol{H_{k,n}}(t,x)$ be the matrix with entry (j, j') equal to $(\boldsymbol{H_{k,n}}(t,x))_{j,j'}$. Let $\boldsymbol{H^*_{k,n}}(s,x)$ be the LST of $\boldsymbol{H_{k,n}}(t,x)$ and denote $\boldsymbol{H^*_{k,n}}(0,x) = \boldsymbol{H_{k,n}}(\infty, x)$ as $\boldsymbol{H_{k,n}}(x)$.

Lemma 3. *The $m_a \times m_a$ matrices $\boldsymbol{H^*_{k,n}}(s,x)$ can be computed as*

$$\begin{aligned} \boldsymbol{H^*_{k,n}}(s,x) = e^{\boldsymbol{D_0}x}(s\boldsymbol{I} - \boldsymbol{D_0})^{-1}\boldsymbol{D_k}\boldsymbol{Z^*_{k,n}}(s) + \\ \boldsymbol{I_{n+1}} e^{\boldsymbol{Q^*_{k,n+1}}(s)x} \left[\boldsymbol{J_{n+1}} + \sum_{i=1}^{n-1} \boldsymbol{J_{n+1,i+1}} \left(\sum_{q=1}^{K} \boldsymbol{P_q} \boldsymbol{Z_{q,n-i}} \right) \right], \end{aligned} \tag{12}$$

with

$$\boldsymbol{Q^*_{k,n+1}}(s) = \left[\begin{array}{c|ccc} \boldsymbol{D_0} & \boldsymbol{D_k}\boldsymbol{M^*_{k,1}}(s) & \ldots & \boldsymbol{D_k}\boldsymbol{M^*_{k,n}}(s) \\ \hline & & & \\ & & \boldsymbol{Q_n} & \\ & & & \end{array} \right].$$

Proof. The probabilities $(\boldsymbol{H_{k,n}}(t,x))_{j,j'}$ are not affected by the amount of time that customer 0 had to wait, we may therefore assume that customer 0 initiated a busy period and his service time equals x.

We consider two cases. First, with probability $e^{\boldsymbol{D_0}x}$, there are no arrivals while customer 0 is in the system. In this case the inter-departure time between customer 0 and 1 consists of an idle period plus the service time of customer 1. Hence, by the probabilistic interpretation of $\boldsymbol{Z_{k,n}}(t)$, the first case results in

$$\boldsymbol{H^*_{k,n}}(t,x,\text{cust. 0 leaves the queue empty}) = e^{\boldsymbol{D_0}x}\int_{a=0}^{t} e^{\boldsymbol{D_0}a}\boldsymbol{D_k}\boldsymbol{Z_{k,n}}(t-a)da,$$

which yields the first term appearing in (12).

Second, if there is at least one arrival while customer 0 is in the system, then the inter-departure time between customer 0 and 1 equals the service time of customer 1. Hence, in this case $(\boldsymbol{H_{k,n}}(t,x))_{j,j'}$ is also equal to the following conditional probability: given that a customer (labeled customer 0) initiates a busy period, requires service time x and the MMAP phase at its arrival time was j, $(\boldsymbol{H_{k,n}}(t,x))_{j,j'}$ holds the probability that (a) at least one arrival occurs during the service of customer 0 and the first arrival is of type k (labeled customer 1), (b) the service time of customer 1 is at most t, (c) customer n leaves the server idle and (d) the MMAP phase equals j' when customer n departs. Note, the above probability is not affected by the order of service either. In this case we may therefore think in terms of a preemptive (resume) last-come-first-served system in which customer 0 has service time x. The first arrival during the service of customer 0 must be of type k, its service time should be at most t and the MMAP phase when customer 0 resumes service is determined by $\boldsymbol{D_k M_{k,r_1}}(t)$, provided that customer 1 induces a sub-busy period during which r_1 customers are served. A possible second arrival of some type q will cause the MMAP phase to change according to $\boldsymbol{D_q M_{q,r_2}}$, for some r_2, etc. Notice, in this case there is no restriction on the service time of the type q customer. Hence, for $i = 1,\ldots,n$

$$\boldsymbol{I_{n+1}} e^{\left[\begin{array}{c|ccc} \boldsymbol{D_0} & \boldsymbol{D_k M_{k,1}}(t) & \ldots & \boldsymbol{D_k M_{k,n}}(t) \\ \hline & & & \\ & & \boldsymbol{Q_n} & \\ & & & \end{array}\right] x} \boldsymbol{J_{n+1,i+1}}.$$

is an $m_a \times m_a$ matrix with entry (j,j') equal to the following conditional probability: given that customer 0 initiates a busy period, has a service time of x and the MMAP phase at its arrival time was j, entry (j,j') holds the probability that (a) customer 1 is of type k and arrives while customer 0 is in the system, (b) the service time of customer 1 is at most t, (c) customer i leaves the server idle and (d) the MMAP phase equals j' when customer i departs. If $i = n$ this results in the term containing $\boldsymbol{J_{n+1}}$ in (12). Otherwise, we need another arrival of some type q (labeled customer $i+1$) that initiates a busy period such that customer n leaves the server idle. This explains the terms containing $\sum_{q=1}^{K} \boldsymbol{P_q Z_{q,n-i}}$ in (12), for $i = 1,\ldots,n-1$. $\square$

Define $(v_{k,n}(t))_j$, for $n \geq 1$, $k \in \{1, \dots, K\}$ and $j \in \{1, \dots, m_a\}$, as the probability of the following event: assume we observe the system at an arbitrary departure instant, then the next inter-departure time is at most t and involves a type k customer (labeled customer 1), while customer n leaves the server idle and the phase of the MMAP is j when customer n departs. Let $v_{k,n}(t)$ be the vector with entry j equal to $(v_{k,n}(t))_j$. Let $v^*_{k,n}(s)$ be the LST of $v_{k,n}(t)$.

Finally, let $(v_0)_j$, for $j \in \{1, \dots, m_a\}$, be the probability that the server becomes idle at an arbitrary departure instant while the MMAP phase equals j. Denote v_0 as the vector with entry j equal to $(v_0)_j$.

Lemma 4. *The $1 \times m_a$ vectors $v^*_{k,n}(s)$ can be expressed as*

$$v^*_{k,n}(s) = \frac{\rho\pi(0)}{\lambda} \int_{x=0}^{\infty} e^{\boldsymbol{T}x}(-\boldsymbol{S}\mathbb{1} \otimes \boldsymbol{I})\boldsymbol{H^*_{k,n}}(s,x)dx, \tag{13}$$

while $v_0 = \rho\pi(0)\boldsymbol{Y_0}/\lambda$, where $\pi(0)$ and $\boldsymbol{Y_0}$ are defined by (1) and (2).

Proof. From the probabilistic interpretation of $\boldsymbol{H_{k,n}}(t,x)$ it is clear that

$$v_{k,n}(t) = \int_{x=0}^{\infty} \pi_D(x)\boldsymbol{H_{k,n}}(t,x)dx,$$

where $\pi_D(x)$ is the density of the age process at departure times. The expression in (13) therefore follows from (5). The expression for v_0 is immediate from

$$v_0 = \int_{x=0}^{\infty} \pi_D(x)e^{\boldsymbol{D_0}x}dx,$$

and the definition of $\pi(0)$ and $\boldsymbol{Y_0}$. $\square$

Theorem 3. *The LST of the joint distribution of the 1st and $(n+1)$th inter-departure time with the first one being of type k and the $n+1$th of type p is given by*

$$\begin{aligned} &f^{(k,p)*}_{D(n)}(s_1,s_2) = \\ &\left[(\alpha - v_0)(-\boldsymbol{D_0})^{-1} + v_0(s_1\boldsymbol{I} - \boldsymbol{D_0})^{-1}\right]\boldsymbol{D_k}\boldsymbol{P}^{n-1}\boldsymbol{P_p}\mathbb{1}f^*_{S_k}(s_1)f^*_{S_p}(s_2) \\ &+ v^*_{k,n}(s_1)\left[(s_2\boldsymbol{I} - \boldsymbol{D_0})^{-1} - (-\boldsymbol{D_0})^{-1}\right]\boldsymbol{D_p}\mathbb{1}f^*_{S_p}(s_2). \end{aligned} \tag{14}$$

Proof. We can write the joint LST as the sum of the joint LST in the case that the server is idle at the start of the $(n+1)$th inter-departure time and the joint LST in the case it is not. Due to the probabilistic interpretation of the vector $v^*_{k,n}(s_1)$, the LST for the case where the server is idle at the start of the $(n+1)$th inter-departure time is given by

$$v^*_{k,n}(s_1)(s_2\boldsymbol{I} - \boldsymbol{D_0})^{-1}\boldsymbol{D_p}\mathbb{1}f^*_{S_p}(s_2). \tag{15}$$

The vector v_0 and $\alpha - v_0$ correspond to the cases where the 1st inter-departure time starts with and without an idle period, respectively. Hence, the term

$$\left[(\alpha - v_0)(-\boldsymbol{D_0})^{-1} + v_0(s_1\boldsymbol{I} - \boldsymbol{D_0})^{-1}\right]\boldsymbol{D_k}\boldsymbol{P}^{n-1}\boldsymbol{P_p}\mathbb{1}f^*_{S_k}(s_1)$$

in (14) holds the LST of the first inter-departure time when the $(n+1)$th inter-departure time involves a type p customer, denoted by $f^{(k,p)*}_{D(n)}(s_1, 0)$. This implies that

$$f^{(k,p)*}_{D(n)}(s_1, 0) - v^*_{k,n}(s_1)(-\boldsymbol{D_0})^{-1}\boldsymbol{D_p}\mathbb{1}$$

holds the LST of the first inter-departure time when the $(n+1)$th inter-departure time involves a type p customer in case the server is not idle at the start of the $(n+1)$th inter-departure time. If the server is not idle at the start of the $(n+1)$th inter-departure time, its LST is given by $f^*_{S_p}(s_2)$, as it is equal to the LST of the service time of the $(n+1)$th customer, the type of which is p. Hence, the joint LST in case the server is busy at the start of the $(n+1)$th inter-departure time is given by

$$f^{(k,p)*}_{D(n)}(s_1, 0)f^*_{S_p}(s_2) - v^*_{k,n}(s_1)(-\boldsymbol{D_0})^{-1}\boldsymbol{D_p}\mathbb{1}f^*_{S_p}(s_2). \tag{16}$$

Combining (15) and (16) establishes (14). □

5. The inter-departure time distribution and lag-1 joint moments

In this section we determine an expression for the moments of the inter-departure time distribution as well as the joint lag-1 moments via Theorem 3. Based on the lag-1 moments it is possible to plug the MMAP[K]/PH[K]/1 FCFS queue into the queueing network analysis framework introduced in [18] for single-type MAP driven queues, which was extended in [9] to multi-type MAP driven priority queues.

We start by defining $(v^{(k)}_1)_j$, for $j \in \{1, \ldots, m_a\}$ and $k \in \{1, \ldots, K\}$, as the probability that an arbitrary departing customer leaves a single customer behind, the type of which is k, while the MMAP phase at the departure epoch is j. Denote $v^{(k)}_1$ as the vector with entry j equal to $(v^{(k)}_1)_j$.

Lemma 5. *The $1 \times m_a$ vectors $v^{(k)}_1$ can be computed as $v^{(k)}_1 = \rho\pi(0)\boldsymbol{Y^{(k)}_1}/\lambda$, where the matrices $\boldsymbol{Y^{(k)}_1}$, for $k = 1, \ldots, K$, are the unique solutions to the Sylvester matrix equations*

$$\boldsymbol{T}\boldsymbol{Y^{(k)}_1} + \boldsymbol{Y^{(k)}_1}\boldsymbol{D_0} = -\boldsymbol{Y_0}\boldsymbol{D_k}. \tag{17}$$

Proof. A departure leaves a single (type k) customer behind if the MMAP generates a single (type k) arrival during the sojourn time of the departing customer. By conditioning on the arrival time of this type k customer we get

$$v^{(k)}_1 = \int_{x=0}^{\infty} \pi_D(x) \int_{a=0}^{x} e^{\boldsymbol{D_0}a}\boldsymbol{D_k}e^{\boldsymbol{D_0}(x-a)}da\, dx, \tag{18}$$

which yields (due to Theorem 2)

$$v^{(k)}_1 = \frac{\rho\pi(0)}{\lambda} \int_{x=0}^{\infty} e^{\boldsymbol{T}x}(-\boldsymbol{S}\mathbb{1} \otimes \boldsymbol{I})\boldsymbol{I_2}e^{\begin{bmatrix}\boldsymbol{D_0} & \boldsymbol{D_k} \\ & \boldsymbol{D_0}\end{bmatrix}x}\boldsymbol{J_2}dx,$$

due to (5). Hence, due to Theorem 1, $v_1^{(k)} = \rho\pi(0)\boldsymbol{X^{(k)}}\boldsymbol{J_2}/\lambda$, where the $m\times 2m_a$ matrix $\boldsymbol{X^{(k)}}$ is the unique solution of the Sylvester matrix equation

$$\boldsymbol{T}\boldsymbol{X^{(k)}} + \boldsymbol{X^{(k)}}\begin{bmatrix}\boldsymbol{D_0} & \boldsymbol{D_k}\\ & \boldsymbol{D_0}\end{bmatrix} = \underbrace{(\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})\boldsymbol{I_2}}_{\left[\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I} \quad \boldsymbol{0}\right]}. \tag{19}$$

Due to (3), it is easy to verify that $\boldsymbol{X^{(k)}} = [\boldsymbol{Y_0}\ \boldsymbol{Y_1^{(k)}}]$ satisfies (19) if $\boldsymbol{Y_1^{(k)}}$ is the unique solution of (17). ☐

Theorem 4. *The LST $f_D^*(s)$ of the inter-departure time distribution is given by*

$$f_D^*(s) = \left[(\alpha - v_0)(-\boldsymbol{D_0})^{-1} + v_0(s\boldsymbol{I} - \boldsymbol{D_0})^{-1}\right]\left(\sum_{k=1}^{K}\boldsymbol{D_k}\mathbb{1}f_{S_k}^*(s)\right). \tag{20}$$

The joint LST $f_D^{(k)*}(s_1, s_2)$ of two consecutive inter-departure times where the type of the first customer is k, can be expressed as

$$\begin{aligned} f_D^{(k)*}(s_1,s_2) = {} & \left[(\alpha - v_0)(-\boldsymbol{D_0})^{-1} + v_0(s_1\boldsymbol{I} - \boldsymbol{D_0})^{-1}\right]\boldsymbol{D_k}\cdot \\ & \left(\sum_{p=1}^{K}\boldsymbol{P_p}\mathbb{1}f_{S_p}^*(s_2)\right)f_{S_k}^*(s_1) + \left[v_0(s_1\boldsymbol{I}-\boldsymbol{D_0})^{-1}\boldsymbol{D_k} + v_1^{(k)}\right]\boldsymbol{M_{k,1}^*}(s_1)\cdot \\ & \left((s_2\boldsymbol{I}-\boldsymbol{D_0})^{-1} - (-\boldsymbol{D_0})^{-1}\right)\left(\sum_{p=1}^{K}\boldsymbol{D_p}\mathbb{1}f_{S_p}^*(s_2)\right). \end{aligned} \tag{21}$$

Proof. As $f_D^*(s) = \sum_{p=1}^{K}\sum_{k=1}^{K} f_{D(n)}^{(k,p)*}(s,0)$, (20) follows from (14). To establish (21), it suffices to sum (14) over p for $n=1$ and to note that

$$v_{k,1}^*(s) = (v_0(s\boldsymbol{I}-\boldsymbol{D_0})^{-1}\boldsymbol{D_k} + v_1^{(k)})\boldsymbol{M_{k,1}^*}(s),$$

due to the probabilistic interpretation of v_0, $v_1^{(k)}$, $\boldsymbol{M_{k,1}}(t)$ and $v_{k,1}(t)$. The above equality can also be proven algebraically as follows. Combining (13) and (12) yields

$$\begin{aligned} v_{k,1}^*(s) = {} & \frac{\rho\pi(0)}{\lambda}\int_{x=0}^{\infty} e^{\boldsymbol{T}x}(-\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})e^{\boldsymbol{D_0}x}dx(s\boldsymbol{I}-\boldsymbol{D_0})^{-1}\boldsymbol{D_k}\boldsymbol{M_{k,1}^*}(s) + \\ & \frac{\rho\pi(0)}{\lambda}\int_{x=0}^{\infty} e^{\boldsymbol{T}x}(-\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})\boldsymbol{I_2}e^{\begin{bmatrix}\boldsymbol{D_0} & \boldsymbol{D_k}\boldsymbol{M_{k,1}^*}(s)\\ & \boldsymbol{D_0}\end{bmatrix}x}\boldsymbol{J_2}dx. \end{aligned}$$

Equation (2) and Lemma 4 imply that the first term reduces to

$$v_0(s\boldsymbol{I}-\boldsymbol{D_0})^{-1}\boldsymbol{D_k}\boldsymbol{M_{k,1}^*}(s),$$

while the second equals $\rho\pi(0)\boldsymbol{X}\boldsymbol{J_2}/\lambda$ (due to Theorem 1), with $\boldsymbol{X}$ the unique solution of

$$\boldsymbol{T}\boldsymbol{X} + \boldsymbol{X}\begin{bmatrix}\boldsymbol{D_0} & \boldsymbol{D_k}\boldsymbol{M^*_{k,1}(s)} \\ & \boldsymbol{D_0}\end{bmatrix} = (\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})\boldsymbol{I_2}.$$

It is easy to verify that $\boldsymbol{X} = [\boldsymbol{Y_0} \;\; \boldsymbol{Y_1^{(k)}}\boldsymbol{M^*_{k,1}(s)}]$ provided that

$$\boldsymbol{T}\boldsymbol{Y_1^{(k)}}\boldsymbol{M^*_{k,1}(s)} + \boldsymbol{Y_1^{(k)}}\boldsymbol{M^*_{k,1}(s)}\boldsymbol{D_0} = -\boldsymbol{Y_0}\boldsymbol{D_k}\boldsymbol{M^*_{k,1}(s)}.$$

As the matrix $\boldsymbol{D_0}$ commutes with $\boldsymbol{M^*_{k,1}}(s)$, this equation follows from (17) and we may conclude that $\rho\pi(0)\boldsymbol{X}\boldsymbol{J_2}/\lambda = v_1^{(k)}\boldsymbol{M^*_{k,1}}(s)$ as required. □

The nth moment of the inter-departure times is given by $E(D^n) = (-1)^n\frac{d^n}{ds^n}f^*_D(s)|_{s=0}$. Instead of computing the moments directly, we introduce the so-called reduced moments

$$\hat{E}(D^n) = E(D^n)/n!, \quad \hat{E}(S_k^n) = E(S_k^n)/n!,$$

because they make the forthcoming expressions simpler.

Corollary 1. *The nth reduced moment of the inter-departure time distribution is given by*

$$\hat{E}(D^n) = \sum_{k=1}^{K}\left(\frac{\lambda_k}{\lambda}\hat{E}(S_k^n) + v_0\sum_{\ell=1}^{n}(-\boldsymbol{D_0})^{-\ell-1}\boldsymbol{D_k}\mathbb{1}\hat{E}(S_k^{n-\ell})\right),$$

Proof. As $E(D^n) = (-1)^n\frac{d^n}{ds^n}f^*_D(s)|_{s=0}$, (20) implies

$$\hat{E}(D^n) = \sum_{k=1}^{K}\Bigg((\alpha - v_0)(-\boldsymbol{D_0})^{-1}\boldsymbol{D_k}\mathbb{1}\frac{E(S_k^n)}{n!} + \frac{v_0}{n!}\sum_{\ell=0}^{n}\binom{n}{\ell}(\ell!)(-\boldsymbol{D_0})^{-\ell-1}\boldsymbol{D_k}\mathbb{1}E(S_k^{n-\ell})\Bigg),$$

which establishes the result as

$$\alpha(-\boldsymbol{D_0})^{-1}\boldsymbol{D_k}\mathbb{1} = \theta\left(\sum_{s=1}^{K}\boldsymbol{D_s}\right)(-\boldsymbol{D_0})^{-1}\boldsymbol{D_k}\mathbb{1}/\lambda = \theta\boldsymbol{D_k}\mathbb{1}/\lambda = \lambda_k/\lambda.$$

□

Taking the derivatives of $f_D^{(k)*}(s_1, s_2)$ gives the joint moments of two consecutive inter-departure times. Again, for simplicity we use the reduced moments instead of the standard ones. The (n_1, n_2)th reduced joint moment is denoted by $\hat{\eta}^{(k)}_{n_1,n_2}$ and is obtained from the LST as

$$\hat{\eta}^{(k)}_{n_1,n_2} = \frac{(-1)^{n_1+n_2}}{n_1!n_2!}\frac{\partial^{n_1}}{\partial s_1^{n_1}}\frac{\partial^{n_2}}{\partial s_2^{n_2}}f_D^{(k)*}(s_1, s_2)|_{s_1=0,s_2=0}.$$

Corollary 2. *The (n_1, n_2)th reduced joint moment of the inter-departure times are given by*

$$\hat{\eta}^{(k)}_{n_1,n_2} = \left[\alpha \boldsymbol{P_k} \hat{E}(S_k^{n_1}) + v_0 \sum_{\ell=1}^{n_1} (-\boldsymbol{D_0})^{-\ell} \boldsymbol{P_k} \hat{E}(S_k^{n_1-\ell})\right] \left(\sum_{q=1}^{K} \boldsymbol{P_q} \mathbb{1} \hat{E}(S_q^{n_2})\right)$$
$$+ \left[v_1^{(k)} \bar{\boldsymbol{M}}^{\boldsymbol{n_1}}_{\boldsymbol{k,1}} + v_0 \sum_{\ell=0}^{n_1} (-\boldsymbol{D_0})^{-\ell} \boldsymbol{P_k} \bar{\boldsymbol{M}}^{\boldsymbol{n_1-\ell}}_{\boldsymbol{k,1}}\right] \sum_{d=1}^{n_2} (-\boldsymbol{D_0})^{-d} \left(\sum_{q=1}^{K} \boldsymbol{P_q} \mathbb{1} \hat{E}(S_q^{n_2-d})\right), \tag{22}$$

where $\bar{\boldsymbol{M}}^{\boldsymbol{n}}_{\boldsymbol{k,1}}$ is defined and computed as follows:

$$\bar{\boldsymbol{M}}^{\boldsymbol{n}}_{\boldsymbol{k,1}} = \frac{(-1)^n}{n!} \frac{d^n}{ds^n} \boldsymbol{M}^*_{\boldsymbol{k,1}}(s)|_{s=0} = (\sigma_k \otimes \boldsymbol{I})((-\boldsymbol{S_k}) \oplus \boldsymbol{D_0})^{-n-1} (-\boldsymbol{S_k} \mathbb{1} \otimes \boldsymbol{I}). \tag{23}$$

6. The lag-n joint means

In this section we derive an expression for the lag-n joint means

$$R_n^{(k,p)} = \frac{\partial}{\partial s_1} \frac{\partial}{\partial s_2} f^{(k,p)*}_{D(n)}(s_1, s_2)|_{s_1=0, s_2=0}. \tag{24}$$

At the end of this section we will also use these lag-n joint means to compute the lag-n cross covariances. Note, the approach taken in this section can also be used to obtain higher lag-n joint moments.

Denote $\bar{\boldsymbol{Z}}_{\boldsymbol{k,1}} = -\frac{\partial}{\partial s} \boldsymbol{Z}^*_{\boldsymbol{k,1}}(s)|_{s=0}$ and $\bar{\boldsymbol{M}}_{\boldsymbol{k,i}} = -\frac{\partial}{\partial s} \boldsymbol{M}^*_{\boldsymbol{k,i}}(s)|_{s=0}$, then $\bar{\boldsymbol{Z}}_{\boldsymbol{k,1}} = \bar{\boldsymbol{M}}_{\boldsymbol{k,1}}$ and

$$\bar{\boldsymbol{Z}}_{\boldsymbol{k,i}} = \bar{\boldsymbol{M}}_{\boldsymbol{k,i}} + \sum_{j=1}^{i-1} \bar{\boldsymbol{M}}_{\boldsymbol{k,j}} \sum_{q=1}^{K} \boldsymbol{P_q} \boldsymbol{Z_{q,i-j}},$$

because of Lemma 2 with

$$\bar{\boldsymbol{M}}_{\boldsymbol{k,i}} = (\sigma_k \otimes \boldsymbol{I_i})((-\boldsymbol{S_k}) \oplus \boldsymbol{Q_i})^{-2} (-\boldsymbol{S_k} \mathbb{1} \otimes \boldsymbol{J_i}),$$

due to (8).

Similarly, let us define $\bar{\boldsymbol{H}}_{\boldsymbol{k,n}}(x) = -\frac{\partial}{\partial s} \boldsymbol{H}^*_{\boldsymbol{k,n}}(s, x)|_{s=0}$.

Lemma 6. *The $m_a \times m_a$ matrices $\bar{\boldsymbol{H}}_{\boldsymbol{k,n}}(x)$ can be computed as*

$$\bar{\boldsymbol{H}}_{\boldsymbol{k,n}}(x) = e^{\boldsymbol{D_0} x} [(-\boldsymbol{D_0})^{-1} \boldsymbol{P_k} \boldsymbol{Z_{k,n}} + \boldsymbol{P_k} \bar{\boldsymbol{Z}}_{\boldsymbol{k,n}}] +$$
$$\boldsymbol{I_{n+1}} e^{\bar{\boldsymbol{Q}}_{k,n+1} x} \left[\boldsymbol{J_{n+1}} + \sum_{i=1}^{n-1} \boldsymbol{J_{n+1,i+1}} \left(\sum_{q=1}^{K} \boldsymbol{P_q} \boldsymbol{Z_{q,n-i}}\right)\right], \tag{25}$$

with

$$\bar{\boldsymbol{Q}}_{\boldsymbol{k,n+1}} = \left[\begin{array}{c|ccc} \boldsymbol{D_0} & \boldsymbol{D_k}\bar{\boldsymbol{M}}_{\boldsymbol{k,1}} & \dots & \boldsymbol{D_k}\bar{\boldsymbol{M}}_{\boldsymbol{k,n}} \\ \hline & & \boldsymbol{Q_n} & \end{array}\right],$$

Proof. Based on Theorem 2 we have that

$$\boldsymbol{I_{n+1}} e^{\bar{\boldsymbol{Q}}^*_{k,n+1}(s)x} \boldsymbol{J_{n+1,i+1}} = \int_{a=0}^{x} e^{\boldsymbol{D_0}a} \boldsymbol{D_k} \left[\boldsymbol{M^*_{k,1}}(s) \quad \dots \quad \boldsymbol{M^*_{k,n}}(s)\right] e^{\boldsymbol{Q_n}(x-a)} \boldsymbol{J_{n,i}},$$

for $i = 1, \dots, n$, which results in (25) due to (12) and (8). □

Theorem 5. *The lag-n joint means of the departure times of type k and type p customers is given by*

$$\begin{aligned} R_n^{(k,p)} &= (\alpha E(S_k) + v_0(-\boldsymbol{D_0})^{-1})\boldsymbol{P_k}\boldsymbol{P}^{n-1}\boldsymbol{P_p}\mathbb{1}E(S_p) \\ &+ v_0\left[(-\boldsymbol{D_0})^{-1}\boldsymbol{P_k}\boldsymbol{Z_{k,n}} + \boldsymbol{P_k}\bar{\boldsymbol{Z}}_{\boldsymbol{k,n}}\right](-\boldsymbol{D_0})^{-1}\boldsymbol{P_p}\mathbb{1} \\ &+ \frac{\rho\pi(0)}{\lambda}\left[\mathbf{Y_{k,n}} + \sum_{i=1}^{n-1}\mathbf{Y_{k,i}}\sum_{q=1}^{K}\boldsymbol{P_q}\boldsymbol{Z_{q,n-i}}\right](-\boldsymbol{D_0})^{-1}\boldsymbol{P_p}\mathbb{1}, \end{aligned} \tag{26}$$

where the $m \times m_a$ matrices $\mathbf{Y_{k,i}}$, for $i = 1, \dots, n$ and $k = 1, \dots, K$, are the unique solutions to the Sylvester matrix equations

$$\boldsymbol{T}\mathbf{Y_{k,i}} + \mathbf{Y_{k,i}}\boldsymbol{D_0} = -\mathbf{Y_0}\boldsymbol{D_k}\bar{\boldsymbol{M}}_{\boldsymbol{k,i}} - \sum_{j=1}^{i-1}\mathbf{Y_{k,j}}\sum_{q=1}^{K}\boldsymbol{D_q}\boldsymbol{M_{q,i-j}}. \tag{27}$$

Proof. By applying (24) on (14) we get

$$\begin{aligned} R_n^{(k,p)} &= (\alpha E(S_k) + v_0(-\boldsymbol{D_0})^{-1})\boldsymbol{P_k}\boldsymbol{P}^{n-1}\boldsymbol{P_p}\mathbb{1}E(S_p) \\ &- \frac{\partial}{\partial s}v^*_{k,n}(s)|_{s=0}(-\boldsymbol{D_0})^{-1}\boldsymbol{P_p}\mathbb{1}, \end{aligned} \tag{28}$$

where the derivative of $v^*_{k,n}(s)$ can be expressed from (13) as

$$-\frac{\partial}{\partial s}v^*_{k,n}(s)|_{s=0} = \frac{\rho\pi(0)}{\lambda}\int_{x=0}^{\infty} e^{\boldsymbol{T}x}(-\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})\bar{\boldsymbol{H}}_{\boldsymbol{k,n}}(x)dx.$$

Using (25) yields

$$-\frac{\partial}{\partial s}v^*_{k,n}(s)|_{s=0} = \frac{\rho\pi(0)}{\lambda}$$
$$\Big[\underbrace{\int_{x=0}^{\infty} e^{\boldsymbol{T}x}(-\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})e^{\boldsymbol{D_0}x}dx}_{\boldsymbol{Y_0}}\left[(-\boldsymbol{D_0})^{-1}\boldsymbol{P_k}\boldsymbol{Z_{k,n}}+\boldsymbol{P_k}\bar{\boldsymbol{Z}}_{\boldsymbol{k,n}}\right]+$$
$$\underbrace{\int_{x=0}^{\infty} e^{\boldsymbol{T}x}(-\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})\boldsymbol{I_{n+1}}e^{\bar{\boldsymbol{Q}}_{\boldsymbol{k,n+1}}x}dx}_{\boldsymbol{X_k}}\left[\boldsymbol{J_{n+1}}+\sum_{i=1}^{n-1}\boldsymbol{J_{n+1,i+1}}\left(\sum_{q=1}^{K}\boldsymbol{P_q}\boldsymbol{Z_{q,n-i}}\right)\right]\Big], \tag{29}$$

where $\boldsymbol{X_k}$ is the unique solution of the Sylvester equation

$$\boldsymbol{T}\boldsymbol{X_k}+\boldsymbol{X_k}\bar{\boldsymbol{Q}}_{\boldsymbol{k,n+1}} = (\boldsymbol{S}\mathbb{1}\otimes\boldsymbol{I})\boldsymbol{I_{n+1}}. \tag{30}$$

Let us partition the $m\times(n+1)m_a$ matrix $\boldsymbol{X_k} = [\boldsymbol{X_{k,0}}\ \boldsymbol{X_{k,1}}\ \ldots\ \boldsymbol{X_{k,n}}]$, then by definition of $\bar{\boldsymbol{Q}}_{\boldsymbol{k,n+1}}$ and $\boldsymbol{Q_n}$, we get

$$\boldsymbol{T}\boldsymbol{X_{k,0}}+\boldsymbol{X_{k,0}}\boldsymbol{D_0} = (\boldsymbol{S}\mathbb{1})\otimes\boldsymbol{I},$$
$$\boldsymbol{T}\boldsymbol{X_{k,i}}+\boldsymbol{X_{k,0}}\boldsymbol{D_k}\bar{\boldsymbol{M}}_{\boldsymbol{k,i}}+\sum_{j=1}^{i-1}\boldsymbol{X_{k,j}}\sum_{q=1}^{K}\boldsymbol{D_q}\boldsymbol{M_{q,i-j}}+\boldsymbol{X_{k,i}}\boldsymbol{D_0} = 0,$$

for $i=1,\ldots,n$, which implies that $\boldsymbol{X_{k,0}} = \boldsymbol{Y_0}$, due to (3), and $\boldsymbol{X_{k,i}} = \boldsymbol{Y_{k,i}}$ as defined in (27).

Combining (28) with (29) and noting that $v_0 = \rho\pi(0)\boldsymbol{Y_0}/\lambda$ (by Lemma 4), $\boldsymbol{X_k}\boldsymbol{J_{n+1}} = \boldsymbol{Y_{k,n}}$ and $\boldsymbol{X_k}\boldsymbol{J_{n+1,i+1}} = \boldsymbol{Y_{k,i}}$ proves the theorem. □

We end this section by defining the lag-n cross covariances $C_n^{(k,p)}$ as

$$C_n^{(k,p)} = E\Big((\tau_1-e_k)(\tau_{n+1}-e_p)\Big|C_1=k, C_{n+1}=p\Big), \tag{31}$$

where τ_i represents the i-th inter-departure time (after an arbitrary departure), C_i the type of the i-th customer and e_k is the mean inter-departure time given that it ends by a type k departure. It is not hard to see that e_k can be expressed as

$$e_k = \frac{\lambda}{\lambda_k}(\alpha+v_0(-\boldsymbol{D_0})^{-1})\boldsymbol{P_k}\mathbb{1}.$$

Using these we obtain the following theorem:

Theorem 6. *The lag-n cross covariance of the departure times of type k and* *type p customers is given by*

$$
\begin{aligned}
C_n^{(k,p)} &= e_k e_p - e_p E(S_k) - e_k E(S_p) + \frac{1}{\alpha \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1}} \Bigg[R_n^{(k,p)} - \\
& e_p v_0 (-\boldsymbol{D_0})^{-1} \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1} + e_k v_0 \boldsymbol{P_k} \boldsymbol{Z_{k,n}} (-\boldsymbol{D_0})^{-1} \boldsymbol{P_p} \mathbb{1} - \\
& e_k \frac{\rho \pi(0)}{\lambda} \left(\tilde{\boldsymbol{Y}}_{\boldsymbol{k,n}} + \sum_{i=1}^{n-1} \tilde{\boldsymbol{Y}}_{\boldsymbol{k,i}} \sum_{q=1}^{K} \boldsymbol{P_q} \boldsymbol{Z_{q,n-i}} \right) (-\boldsymbol{D_0})^{-1} \boldsymbol{P_p} \mathbb{1} \Bigg],
\end{aligned} \tag{32}
$$

where the $m \times m_a$ matrices $\tilde{\boldsymbol{Y}}_{\boldsymbol{k,i}}$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, are the unique solutions to the Sylvester matrix equations

$$
\boldsymbol{T} \tilde{\boldsymbol{Y}}_{\boldsymbol{k,i}} + \tilde{\boldsymbol{Y}}_{\boldsymbol{k,i}} \boldsymbol{D_0} = -\boldsymbol{Y_0} \boldsymbol{D_k} \boldsymbol{M_{k,i}} - \sum_{j=1}^{i-1} \tilde{\boldsymbol{Y}}_{\boldsymbol{k,j}} \sum_{q=1}^{K} \boldsymbol{D_q} \boldsymbol{M_{q,i-j}}. \tag{33}
$$

Proof. The cross covariance $C_n^{(k,p)}$ clearly equals

$$
\begin{aligned}
C_n^{(k,p)} &= E(\tau_1 \tau_{n+1} | C_1 = k, C_{n+1} = p) - e_p E(\tau_1 | C_1 = k, C_{n+1} = p) \\
&\quad - e_k E(\tau_{n+1} | C_1 = k, C_{n+1} = p) + e_k e_p.
\end{aligned}
$$

By definition of $f_{D(n)}^{(k,p)*}(s_1, s_2)$ and $R_n^{(k,p)}$, we therefore have

$$
\begin{aligned}
C_n^{(k,p)} P[C_1 = k, C_{n+1} = p] &= R_n^{(k,p)} + e_k e_p P[C_1 = k, C_{n+1} = p] \\
& - e_p \frac{\partial}{\partial s_1} f_{D(n)}^{(k,p)*}(s_1, 0)|_{s_1=0} - e_k \frac{\partial}{\partial s_2} f_{D(n)}^{(k,p)*}(0, s_2)|_{s_2=0}.
\end{aligned} \tag{34}
$$

By means of (14), we find

$$
\frac{\partial}{\partial s_1} f_{D(n)}^{(k,p)*}(s_1, 0)|_{s_1=0} = \alpha \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1} E(S_k) + v_0 (-\boldsymbol{D_0})^{-1} \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1}
$$

and

$$
\frac{\partial}{\partial s_2} f_{D(n)}^{(k,p)*}(0, s_2)|_{s_2=0} = \alpha \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1} E(S_p) + v_{k,n}^*(0) (-\boldsymbol{D_0})^{-1} \boldsymbol{P_p} \mathbb{1}.
$$

When combined with $P[C_1 = k, C_{n+1} = p] = \alpha \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1}$, (34) implies

$$
\begin{aligned}
C_n^{(k,p)} &= e_k e_p - e_p E(S_k) - e_k E(S_p) + \frac{1}{\alpha \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1}} \Bigg[R_n^{(k,p)} - \\
& e_p v_0 (-\boldsymbol{D_0})^{-1} \boldsymbol{P_k} \boldsymbol{P}^{n-1} \boldsymbol{P_p} \mathbb{1} - e_k v_{k,n}^*(0) (-\boldsymbol{D_0})^{-1} \boldsymbol{P_p} \mathbb{1} \Bigg].
\end{aligned}
$$

The equality

$$
v_{k,n}^*(0) = v_0 \boldsymbol{P_k} \boldsymbol{Z_{k,n}} + \frac{\rho \pi(0)}{\lambda} \left(\tilde{\boldsymbol{Y}}_{\boldsymbol{k,n}} + \sum_{i=1}^{n-1} \tilde{\boldsymbol{Y}}_{\boldsymbol{k,i}} \sum_{q=1}^{K} \boldsymbol{P_q} \boldsymbol{Z_{q,n-i}} \right)
$$

can be established in a manner completely analogue to the expression for $\frac{\partial}{\partial s} v^*_{k,n}(s)|_{s=0}$ in the proof of Theorem 5. It is worth noting that (33) is identical to (27) except that $\bar{\boldsymbol{M}}_{\boldsymbol{k,i}}$ is replaced by $\boldsymbol{M}_{\boldsymbol{k,i}}$. $\square$

7. Algorithmic efficiency

This section investigates the computational complexity to determine the lag-1 joint moments and the lag-n joint means by means of Corollary 2 and Theorem 5. While it is not typical to compute a large number of lag-1 joint moments, the lag-n joint means may decay slowly and it may be of interest to compute them up to $n = 1000$ or even further, which results in a high computational complexity if a naive implementation of Theorem 5 is used.

7.1. Solution of the lag-1 joint moments

The calculation of the reduced joint moments $\hat{\eta}^{(k)}_{n_1,n_2}$ for $n_1 = 0, \ldots, N_1$, $n_2 = 0, \ldots, N_2$, $k = 1, \ldots, K$, based on Corollary 2 is computationally not very demanding. The two most expensive steps exist in the computation of the vectors $v^{(k)}_1$, for $k = 1, \ldots, K$, and the matrices $\bar{\boldsymbol{M}}^{\boldsymbol{n}}_{\boldsymbol{k,1}}$, for $k = 1, \ldots, K$ and $n_1 = 0, \ldots, N_1$.

- According to Lemma 5 each of the vectors $v^{(k)}_1$, for $k = 1, \ldots, K$, is the solution of a Sylvester matrix equation of size $m \times m_a$, the time complexity of which is $\mathcal{O}(m^3)$ using the Hessenberg-Schur algorithm in [19]. In fact, as the first step of this algorithm involves the decomposition of the same two matrices $\boldsymbol{T}$ and $\boldsymbol{D_0}$, for $k = 1, \ldots, K$, one finds that the overall complexity reduces to $\mathcal{O}(m^3 + Km^2 m_a)$.
- The matrices $\bar{\boldsymbol{M}}^{\boldsymbol{n}}_{\boldsymbol{k,1}}$ are pre-calculated and stored for $k = 1, \ldots, K$ and $n = 0, \ldots, N_1$. Based on (23), this can be done in a time complexity of $\mathcal{O}(m^3 K N_1)$.

7.2. Solution of the lag-n joint means

The lag-n joint means are calculated by Theorem 5. The naive evaluation of the matrices $\boldsymbol{M}_{\boldsymbol{k,i}}$ and $\bar{\boldsymbol{M}}_{\boldsymbol{k,i}}$ in (26), based on (8), requires the inversion of the matrix $\boldsymbol{S_k} \oplus \boldsymbol{Q_i}$, the size of which is $i \cdot m_a m_k \times i \cdot m_a m_k$. The following theorem avoids the need to work with large matrices and therefore enables us the compute the lag-n joint means for large n values.

Theorem 7. *The $m_a \times m_a$ matrices $\boldsymbol{M}^*_{\boldsymbol{k,i}}(s)$ can be expressed as*

$$\boldsymbol{M}^*_{\boldsymbol{k,i}}(s) = (\sigma_k \otimes \boldsymbol{I}) \boldsymbol{L}^*_{\boldsymbol{k,i}}(s) (-\boldsymbol{S_k} \mathbb{1} \otimes \boldsymbol{I}), \tag{35}$$

where $m_k m_a \times m_k m_a$ matrices $\boldsymbol{L}^*_{\boldsymbol{k,i}}(s)$ are recursively given by

$$\boldsymbol{L}^*_{\boldsymbol{k,1}}(s) = ((s\boldsymbol{I} - \boldsymbol{S_k}) \oplus \boldsymbol{D_0})^{-1}, \tag{36}$$

$$\boldsymbol{L}^*_{\boldsymbol{k,i}}(s) = ((s\boldsymbol{I} - \boldsymbol{S_k}) \oplus \boldsymbol{D_0})^{-1} \sum_{j=1}^{i-1} \left(\boldsymbol{I} \otimes \sum_{q=1}^{K} \boldsymbol{D_q} \boldsymbol{M_{q,j}} \right) \boldsymbol{L}^*_{\boldsymbol{k,i-j}}(s), \quad \textit{for } i > 1. \tag{37}$$

Proof. For $i = 1$ (36) follows directly from (8). For $i > 1$ let us express $\boldsymbol{M^*_{k,i}}(s)$ based on (10). We get

$$\boldsymbol{M^*_{k,i}}(s) = (\sigma_k \otimes \boldsymbol{I}) \underbrace{\int_{y=0}^{\infty} e^{-(s\boldsymbol{I}-\boldsymbol{S_k})y} \otimes \boldsymbol{I_i} e^{\boldsymbol{Q}_i y} \boldsymbol{J_i} dy}_{\boldsymbol{L^*_{k,i}}(s)} (-\boldsymbol{S_k} \mathbb{1} \otimes \boldsymbol{I}).$$

To express $\boldsymbol{L^*_{k,i}}(s)$, we can rewrite $\boldsymbol{I_i} e^{\boldsymbol{Q}_i y} \boldsymbol{J_i}$ via Theorem 2 as

$$\begin{aligned}
\boldsymbol{I_i} e^{\boldsymbol{Q}_i y} \boldsymbol{J_i} &= \int_{a=0}^{\infty} e^{\boldsymbol{D_0} a} \left(\sum_{q=1}^{K} \boldsymbol{D_q} [\boldsymbol{M_{q,1}} \; \dots \; \boldsymbol{M_{q,i-1}}] \right) e^{\boldsymbol{Q_{i-1}}(y-a)} \boldsymbol{J_{i-1}} da \\
&= \int_{a=0}^{\infty} e^{\boldsymbol{D_0} a} \left(\sum_{q=1}^{K} \boldsymbol{D_q} \sum_{j=1}^{i-1} \boldsymbol{M_{q,j}} \boldsymbol{J_{i-1,j}}^T \right) e^{\boldsymbol{Q_{i-1}}(y-a)} \boldsymbol{J_{i-1}} da.
\end{aligned}$$

Due to the block triangular block Toeplitz structure of $\boldsymbol{Q_i}$ we have $\boldsymbol{J_{i-1,j}}^T e^{\boldsymbol{Q_{i-1}}(y-a)} \boldsymbol{J_{i-1}} = \boldsymbol{I_{i-j}} e^{\boldsymbol{Q}_{i-j}(y-a)} \boldsymbol{J_{i-j}}$, for $j = 1, \dots, i-1$. Hence,

$$\begin{aligned}
\boldsymbol{L^*_{k,i}}(s) &= \int_{y=0}^{\infty} e^{-(s\boldsymbol{I}-\boldsymbol{S_k})y} \otimes \int_{a=0}^{y} e^{\boldsymbol{D_0} a} \sum_{j=1}^{i-1} \sum_{q=1}^{K} \boldsymbol{D_q} \boldsymbol{M_{q,j}} \boldsymbol{I_{i-j}} e^{\boldsymbol{Q}_{i-j}(y-a)} \boldsymbol{J_{i-j}} da \, dy \\
&= \underbrace{\int_{a=0}^{\infty} e^{-(s\boldsymbol{I}-\boldsymbol{S_k})a} \otimes e^{\boldsymbol{D_0} a} da}_{(s\boldsymbol{I}-\boldsymbol{S_k} \oplus \boldsymbol{D_0})^{-1}} \sum_{j=1}^{i-1} \left(\boldsymbol{I} \otimes \sum_{q=1}^{K} \boldsymbol{D_q} \boldsymbol{M_{q,j}} \right) \\
&\quad \cdot \underbrace{\int_{y=a}^{\infty} e^{-(s\boldsymbol{I}-\boldsymbol{S_k})(y-a)} \otimes \boldsymbol{I_{i-j}} e^{\boldsymbol{Q}_{i-j}(y-a)} \boldsymbol{J_{i-j}} dy}_{\boldsymbol{L^*_{k,i-j}}(s)} .
\end{aligned}$$

□

Taking the derivatives in (35), (36) and (37) yields efficient recursive expressions to compute $\boldsymbol{M_{k,i}}$ and $\bar{\boldsymbol{M}}_{\boldsymbol{k,i}}$.

Corollary 3. *The $m_a \times m_a$ matrices $\boldsymbol{M_{k,i}}$ and $\bar{\boldsymbol{M}}_{\boldsymbol{k,i}}$ are given by*

$$\begin{aligned}
\boldsymbol{M_{k,i}} &= (\sigma_k \otimes \boldsymbol{I}) \boldsymbol{L_{k,i}} (-\boldsymbol{S_k} \mathbb{1} \otimes \boldsymbol{I}), \\
\bar{\boldsymbol{M}}_{\boldsymbol{k,i}} &= (\sigma_k \otimes \boldsymbol{I}) \bar{\boldsymbol{L}}_{\boldsymbol{k,i}} (-\boldsymbol{S_k} \mathbb{1} \otimes \boldsymbol{I}),
\end{aligned}$$

where the matrices $\boldsymbol{L_{k,i}}$ and $\bar{\boldsymbol{L}}_{\boldsymbol{k,i}}$ are defined recursively as

$$\boldsymbol{L_{k,1}} = (-\boldsymbol{S_k} \oplus \boldsymbol{D_0})^{-1}, \tag{38}$$

$$\boldsymbol{L_{k,i}} = (-\boldsymbol{S_k} \oplus \boldsymbol{D_0})^{-1} \sum_{j=1}^{i-1} \left(\boldsymbol{I} \otimes \sum_{q=1}^{K} \boldsymbol{D_q} \boldsymbol{M_{q,j}} \right) \boldsymbol{L_{k,i-j}}, \quad \textit{for } i > 1, \tag{39}$$

$$\bar{\boldsymbol{L}}_{\boldsymbol{k,1}} = (-\boldsymbol{S_k} \oplus \boldsymbol{D_0})^{-2}, \tag{40}$$

$$
\begin{aligned}
\bar{\boldsymbol{L}}_{\boldsymbol{k},\boldsymbol{i}} &= (-\boldsymbol{S}_{\boldsymbol{k}} \oplus \boldsymbol{D}_{\boldsymbol{0}})^{-1} \sum_{j=1}^{i-1} \left(\boldsymbol{I} \otimes \sum_{q=1}^{K} \boldsymbol{D}_{\boldsymbol{q}} \boldsymbol{M}_{\boldsymbol{q},\boldsymbol{j}} \right) \bar{\boldsymbol{L}}_{\boldsymbol{k},\boldsymbol{i}-\boldsymbol{j}} \\
&\quad + (-\boldsymbol{S}_{\boldsymbol{k}} \oplus \boldsymbol{D}_{\boldsymbol{0}})^{-2} \sum_{j=1}^{i-1} \left(\boldsymbol{I} \otimes \sum_{q=1}^{K} \boldsymbol{D}_{\boldsymbol{q}} \boldsymbol{M}_{\boldsymbol{q},\boldsymbol{j}} \right) \boldsymbol{L}_{\boldsymbol{k},\boldsymbol{i}-\boldsymbol{j}}, \quad \textit{for } i > 1.
\end{aligned} \tag{41}
$$

With these results in mind the computational complexity to calculate $R_n^{(k,p)}$, for $n = 1, \dots, N$, $k = 1, \dots, K$ and $q = 1, \dots, K$ consists of the following main steps:

- Computing the matrices $\boldsymbol{M}_{\boldsymbol{k},\boldsymbol{i}}$ and $\bar{\boldsymbol{M}}_{\boldsymbol{k},\boldsymbol{i}}$, for $k = 1, \dots, K$ and $i = 1, \dots, N$, using Corollary 3 requires $\mathcal{O}(N^2 m_a^3 \sum_{k=1}^{K} m_k^3)$ time. If necessary, this can be further reduced by exploiting the structure of $\boldsymbol{I} \otimes \sum_{q=1}^{K} \boldsymbol{D}_{\boldsymbol{q}} \boldsymbol{M}_{\boldsymbol{q},\boldsymbol{j}}$ in (39) and (41).
- Computing matrices $\boldsymbol{Z}_{\boldsymbol{k},\boldsymbol{i}}$ and $\bar{\boldsymbol{Z}}_{\boldsymbol{k},\boldsymbol{i}}$, for $k = 1, \dots, K$ and $i = 1, \dots, N$, requires $\mathcal{O}(N^2 m_a^3 K)$ time.
- The overall time complexity of solving the KN Sylvester matrix equations to obtain the matrices $\boldsymbol{Y}_{\boldsymbol{k},\boldsymbol{i}}$, for $k = 1, \dots, K$ and $i = 1, \dots, N$, is $\mathcal{O}(m^3 + KNm^2 m_a)$ using the Hessenberg-Schur algorithm in [19] by noting that the decomposition step is identical for all k and i.

Overall we see that the time complexity is a quadratic function of N, while the memory usage is linear in N.

8. Numerical Examples

For the numerical experiments, we make use of the MMAP[K]/PH[K]/1 queue of Example 6.1 in [20]. The matrices defining the MMAP[K] are given by

$$
\boldsymbol{D}_{\boldsymbol{0}} = \begin{bmatrix} -2 & 1 \\ 0 & -5 \end{bmatrix}, \quad \boldsymbol{D}_{\boldsymbol{1}} = \begin{bmatrix} 0 & 1 \\ 0.1 & 0 \end{bmatrix}, \quad \boldsymbol{D}_{\boldsymbol{2}} = \begin{bmatrix} 0 & 0 \\ 1.9 & 3 \end{bmatrix}. \tag{42}
$$

With these matrices the arrival rate of type 1 (type 2) customers is 0.55 (2.45), respectively. The initial probability vector and transient generator of the PH distributions defining the service times are given by

$$
\sigma_1 = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix}, \quad \boldsymbol{S}_{\boldsymbol{1}} = \begin{bmatrix} -2 & 1.5 \\ 0 & -1 \end{bmatrix}, \tag{43}
$$

$$
\sigma_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \boldsymbol{S}_{\boldsymbol{2}} = \begin{bmatrix} -25 & 5 \\ 0 & -25 \end{bmatrix}, \tag{44}
$$

resulting in mean service times $E(S_1) = 1.2$ and $E(S_2) = 0.048$. The utilization of the queue is thus $\rho = 0.7776$.

8.1. Comparing the cross correlations of the input and the output processes

Let us first investigate the cross correlation of the input traffic of the queue defined by

$$\begin{aligned}
\tilde{\rho}_n^{(k,p)} &= \frac{E((\tau_1^{(a)} - e_k^{(a)})(\tau_{n+1}^{(a)} - e_p^{(a)})|C_1 = k, C_{n+1} = p)}{\sqrt{Var\{\tau_1^{(a)}|C_1 = k\}}\sqrt{Var\{\tau_{n+1}^{(a)}|C_{n+1} = p\}}} \\
&= \Big(\alpha(-\boldsymbol{D_0})^{-1}\boldsymbol{P_k}\boldsymbol{P}^{n-1}(-\boldsymbol{D_0})^{-1}\boldsymbol{P_p}\mathbb{1} - e_p^{(a)}\alpha(-\boldsymbol{D_0})^{-1}\boldsymbol{P_k}\boldsymbol{P}^{n-1}\boldsymbol{P_p}\mathbb{1} \\
&\quad - e_k^{(a)}\alpha\boldsymbol{P_k}\boldsymbol{P}^{n-1}(-\boldsymbol{D_0})^{-1}\boldsymbol{P_p}\mathbb{1} + e_p^{(a)}e_k^{(a)}\alpha\boldsymbol{P_k}\boldsymbol{P}^{n-1}\boldsymbol{P_p}\mathbb{1}\Big) \\
&\quad \cdot \frac{1}{\alpha\boldsymbol{P_k}\boldsymbol{P}^{n-1}\boldsymbol{P_p}\mathbb{1}} \cdot \frac{1}{\sqrt{Var\{\tau_1^{(a)}|C_1 = k\}}\sqrt{Var\{\tau_{n+1}^{(a)}|C_{n+1} = p\}}},
\end{aligned} \tag{45}$$

where $\tau_n^{(a)}$ denotes the inter-arrival time, C_n denotes the type of the nth customer, and $e_k^{(a)}$ is the mean inter-arrival time ending by a type k customer. The variance of the inter-arrival times of type k customers can be obtained as

$$Var\{\tau_1^{(a)}|C_1 = k\} = \frac{2\alpha(-\boldsymbol{D_0})^{-3}\boldsymbol{D_k}\mathbb{1}}{\alpha(-\boldsymbol{D_0})^{-1}\boldsymbol{D_k}\mathbb{1}} - \left(\frac{\alpha(-\boldsymbol{D_0})^{-2}\boldsymbol{D_k}\mathbb{1}}{\alpha(-\boldsymbol{D_0})^{-1}\boldsymbol{D_k}\mathbb{1}}\right)^2. \tag{46}$$

The cross correlations are listed in Table 1. Since type 2 customers arrive only in the second phase, the distribution of the inter-arrival times of a type 2 customer is independent of any subsequent inter-arrival times. Hence, $\tilde{\rho}_n^{(2,p)} = 0$ for $p = \{1, 2\}$ and $n \geq 1$; therefore Table 1 only lists $\tilde{\rho}_n^{(1,1)}$ and $\tilde{\rho}_n^{(1,2)}$. The counter diagonal structure of $\boldsymbol{D_1}$ explains why $\tilde{\rho}_1^{(1,p)}$ is negative. As the eigenvalue of $\boldsymbol{P}$ other than 1 is negative, the sign of the cross correlations alternates with n.

n	$\tilde{\rho}_n^{(1,1)}$	$\tilde{\rho}_n^{(1,2)}$
1	$-1.7864 * 10^{-2}$	$-2.9264 * 10^{-2}$
2	$1.1135 * 10^{-3}$	$6.9050 * 10^{-3}$
3	$-2.5826 * 10^{-4}$	$-1.3331 * 10^{-3}$
4	$5.0054 * 10^{-5}$	$2.6848 * 10^{-4}$
5	$-1.0073 * 10^{-5}$	$-5.3621 * 10^{-5}$
6	$2.0121 * 10^{-6}$	$1.07271 * 10^{-5}$

Table 1: Cross correlations of the MMAP feeding the queue

To investigate the cross correlation of the departure process of the queue, we rely on Theorem 6 and define

$$\rho_n^{(k,p)} = \frac{C_n^{(k,p)}}{\sqrt{Var\{\tau_1|C_1 = k\}}\sqrt{Var\{\tau_{n+1}|C_{n+1} = p\}}},$$

where τ_n represents the inter-departure time. The numerical results are shown in Figure 1 both on linear- and log-scale (the latter one plots the logarithm of the absolute values). The cross correlations of the departure process show a very different behavior than the cross correlations of the input process of the queue, that is, the decay of the cross correlations is much slower, and the alternating sign disappears as well. More specifically, the lag-n cross correlations only become less than 10^{-5} for n close to 1000, which emphases the importance of the computational efficiency of our results.

The results in Figure 1 have also been verified by discrete event simulations. We have also checked the results at the two extreme settings of the service times: as the mean service time approaches to zero, the departure process statistics approach to the statistics of the input MMAP. At the other hand, when the service process is slowed down such that the probability of the idle queue tends to zero the departure times are determined by the service process, that is uncorrelated (given the customer types).

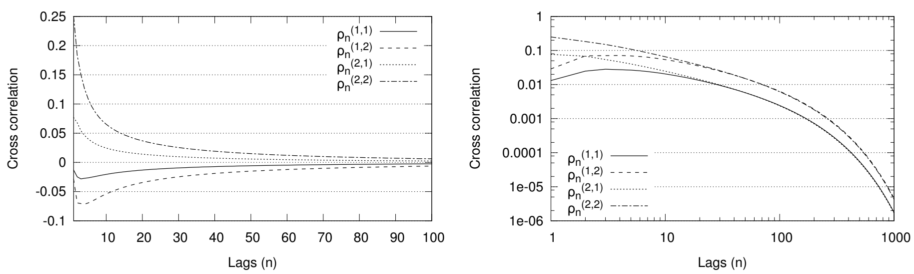

Figure 1: The cross correlations of the departure process of the queue

8.2. Execution time analysis

In this section we analyze how the execution times scale with the size of the input parameters. For this reason both the size of the MMAP and the number of customer types are increased gradually, and the computation time of the departure process statistics are measured.

To increase the size of the MMAP by a factor f we apply the following operation:

$$\boldsymbol{D}_{\boldsymbol{k}}^{\times \boldsymbol{f}} = \underbrace{\boldsymbol{D}_{\boldsymbol{k}} \oplus \ldots \oplus \boldsymbol{D}_{\boldsymbol{k}}}_{f \text{ times } \boldsymbol{D}_{\boldsymbol{k}}}/f, \quad k = 0, \ldots, K, \tag{47}$$

where the division by f ensures that the arrival intensity is maintained.

The number of customer types is increased in the upcoming example as well. Let us denote the MMAP matrices and the parameters of the PH distributed service time distributions having twice as many customer types by $\boldsymbol{D}'_{\boldsymbol{k}}, s'_k, \boldsymbol{S}'_{\boldsymbol{k}}$, respectively. They are obtained as

$$\begin{aligned} \boldsymbol{D'_k} &= \boldsymbol{D_{\lceil k/2 \rceil}}/2 \\ s'_k &= s_{\lceil k/2 \rceil} \\ \boldsymbol{S'_k} &= \boldsymbol{S_{\lceil k/2 \rceil}} \end{aligned} \tag{48}$$

and matrix $\boldsymbol{D_0}$ remains the same.

First the lag-1 joint moments are calculated up to order 10. The measurements have been performed on an average PC with a CPU clocked at 3.4 GHz and with 4 GB of RAM. The results are summarized in Table 2. The execution times are below or around 1 second except the case when the size of the MMAP or the number of customer types equals 32, where the size of the matrices appearing in the formulas grow up to $m = 2048$. Profiling the algorithm revealed that there are two computational bottlenecks: ca. 50% of the execution time is taken by the solution of the K Sylvester equations in (17) needed to compute $\boldsymbol{Y_1^{(k)}}$, and further 35% of the execution time is required to solve matrix $\boldsymbol{Y_0}$ from (4) using the ADDA algorithm. According to the results the algorithm scales well with the size of the MMAP and the number of customer types, but slows down more rapidly when the number of customer types increases.

Cust. Types (K)	Size of the MMAP 2	4	8	16	32
2	0.033784	0.029956	0.034512	0.042347	0.076943
4	0.053699	0.059457	0.069391	0.091846	0.20388
8	0.11395	0.12206	0.15429	0.25445	0.88628
16	0.23886	0.26867	0.41342	1.198	9.2455
32	0.52783	0.73959	1.9283	13.178	119.03

Table 2: Execution times to calculate joint moments up to order 10 (in seconds)

Next, the lag-n cross correlations are calculated up to lag 100 and lag 1000. The execution times are depicted in Table 3. The most time consuming operations are the solution of $\boldsymbol{Y_0}$ and the Sylvester equations (27) needed to compute $\boldsymbol{Y_{k,i}}$. As with the lag-1 joint moments, the computation time of the cross correlations is more sensitive to the number of customer types than to the size of the MMAP.

Cust. Types	up to lag 100			up to lag 1000		
(K)	$m_a = 2$	$m_a = 4$	$m_a = 16$	$m_a = 2$	$m_a = 4$	$m_a = 16$
2	0.51266	0.51714	1.1843	32.413	35.242	83.473
4	1.209	1.4315	5.3499	85.909	94.43	317.26
8	3.7016	4.5306	61.888	255.39	289.28	4829.1

Table 3: Execution times to calculate cross correlations (in seconds)

As a closing remark we note that we did not encounter any numerical problems or instabilities in any of the above examples. To solve the Sylvester matrix equations, we relied on the MATLAB lyap function, which is based on the SB04MD (SLICOT) routine that implements the Hessenberg-Schur algorithm. This implies that the matrices $\boldsymbol{T}$ and $\boldsymbol{D_0}$ were decomposed during each function call to lyap.

8.3. Approximating the departure process by an MMAP

An appealing practical application for the results in our paper is the traffic decomposition based queueing network analysis. In this section we study a tandem queueing network composed of two stations. For the sake of simplicity, the service times of both stations are given by the same parameters as in (48), and the first station is driven by the MMAP given in (47). The idea of the decomposition approach exists in modeling the input traffic of the second queue by means of an MMAP that approximates the departure process of the first queue as much as possible. We are only aware of two methods that can be used to create an MMAP from a set of statistical parameters, and both methods make use of the lag-1 joint moments (which makes sense, as according to [9] the lag-1 joint moments characterize non-redundant MMAPs completely). One of these methods applies (marginal- and joint-) moment matching [9], while the method in [21] is based on fitting.

For this particular example, the moment matching method was unable to construct an MMAP that matches the moments of the departure process of the first queue (as computed by Corollary 1 and 2), thus we had to rely on the fitting method instead[2]. The fitting method has the advantage over the matching approach in that the number of statistical parameters to be fitted can be selected by the user independently of the size of the MMAP. With 2 states, by fitting the first 3 marginal moments and 3x3 class specific lag-1 joint moments we obtained the following MMAP:

$$
\begin{aligned}
\boldsymbol{G_0} &= \begin{bmatrix} -0.97743 & 0.00092327 \\ 0 & -17.621 \end{bmatrix}, \\
\boldsymbol{G_1} &= \begin{bmatrix} 0.053089 & 0.57303 \\ 0 & 0 \end{bmatrix}, \\
\boldsymbol{G_2} &= \begin{bmatrix} 0.016035 & 0.33436 \\ 6.5632 & 11.058 \end{bmatrix}.
\end{aligned}
\tag{49}
$$

Table 4 compares the first three moments of the sojourn time of the second station obtained by the analysis and by discrete event simulation. Even in this small example a simulation run of approximately one minute was required due to the slow decay of the correlations, while the decomposition based analysis with MMAP fitting provided prompt results. As shown in Table 4, the first moments are captured quite accurately, while the relative error grows as higher moments are considered.

[2] The authors would like to thank P. Buchholz for sharing his implementation of the MMAP fitting algorithm developed in [21].

	Simulation	Analysis	Relative error
Class 1, 1st moment	4.05667	4.3797	7.96%
Class 2, 1st moment	3.8297	3.8491	0.5%
Class 1, 2nd moment	31.671	37.036	16.9%
Class 2, 2nd moment	29.833	32.438	8.7%
Class 1, 3rd moment	370.56	468.01	26.3%
Class 2, 3rd moment	348.76	409.84	17.5%

Table 4: Sojourn time moments of the second station, analysis vs. simulation

We also tried to fit larger MMAPs and involve more moments and joint moments into the fitting, but in general it did not improve the accuracy of the results. A possible source of the inaccuracy can be that the fitting algorithm is based exclusively on the lag-1 joint moments and neglects all other statistics completely. Figure 2 confirms that the cross correlations of the departure process and its MMAP differ significantly. We emphasize again that the departure process analysis presented in this paper is exact. We expect that as better and more mature MMAP fitting procedures become available, the decomposition based analysis of multi-type queuing networks becomes more accurate.

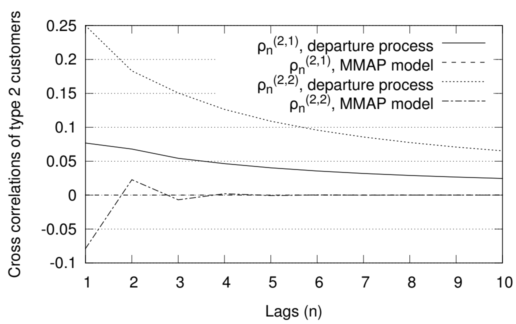

Figure 2: Cross correlations of type 2 customers, departure process and its MMAP model

9. Conclusion

In this paper we derived the joint LST of the lag-n inter-departure times in the MMAP[K]/PH[K]/1 FCFS multi-type queue. The key observation is that the departure process can be derived via the age process, the steady state distribution of which is matrix-exponential. Based on the joint LST we presented efficient algorithms to compute the lag-1 joint moments, the lag-n joint means and cross correlations. Further the approach can be used to calculate other lag-n moment-like expressions as well.

Our results can be used to introduce $\bullet$/PH[K]/1 FCFS nodes in the multi-type queueing network framework of [9], but is equally useful in other decomposition based queueing network approaches that rely on MMAP fitting [21].

Acknowledgements

This paper was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

[1] N. Bean, D. Green, P. Taylor, Approximations to the output process of MAP/PH/1 queues, in: Advances in Matrix Analytic Methods for Stochastic Models, Notable Publications Inc, 1998, pp. 151–170.

[2] R. Sadre, B. Haverkort, Characterizing traffic streams in networks of MAP/MAP/1 queues, in: Proceedings 11th GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB 2001), VDE Verlag, 2001, pp. 195–208.

[3] Q. Zhang, A. Heindl, E. Smirni, Characterizing the BMAP/MAP/1 departure process via the etaqa truncation, Stochastic Models 21 (2-3) (2005) 821–846.

[4] Q. Zhang, A. Heindl, E. Smirni, A. Stathopoulos, Comparison of two output models for the BMAP/MAP/1 departure process, in: Proceedings of QEST, 2009, pp. 143–152.

[5] S. Shioda, Departure process of the MAP/SM/1 queue, Queueing Systems 44 (2003) 31–50.

[6] P.-C. Yeh, J.-F. Chang, Characterizing the departure process of a single server queue from the embedded markov renewal process at departures, Queueing Systems 35 (2000) 381–395.

[7] H.-W. Ferng, J.-F. Chang, Departure processes of BMAP/G/1 queues, Queueing Systems 39 (2001) 109–135.

[8] Q.-M. He, D. A. Stanford, Distributions of the interdeparture times in FCFS and nonpreemptive priority MMAP[2]/G[2]/1 queues, Performance Evalulation 38 (2) (1999) 85–103.

[9] A. Horváth, G. Horváth, M. Telek, A traffic based decomposition of two-class queueing network with priority service, Computer Networks 53 (2009) 1235–1248.

[10] Q. He, M. Neuts, Markov chains with marked transitions, Stochastic Processes and their Applications 74 (1998) 37–52.

[11] Q. He, Analysis of a continuous time SM[K]/PH[K]/1/FCFS queue: Age process, sojourn times, and queue lengths, Journal of Systems Science and Complexity 25 (2012) 133–155.

[12] B. Sengupta, Markov processes whose steady state distribution is matrix exponential with an application to the GI/PH/1 queue, Adv. in Appl. Probab. 21 (1989) 159–180.

[13] R. Bhatia, P. Rosenthal, How and why to solve the operator equation ax-xb = y, Bulletin of the London Mathematical Society 29 (1997) 1–21.

[14] B. Van Houdt, Analysis of the adaptive MMAP[K]/PH[K]/1 queue: a multi-type queue with adaptive arrivals and general impatience, European Journal of Operational Research 220 (3) (2012) 695–704.

[15] C.-H. Guo, B. Iannazzo, B. Meini, On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation, SIAM J. Matrix Anal. Appl. 29 (2007) 1083–1100.

[16] W. Wang, W. Wang, R. Li, Alternating-directional Doubling Algorithm for M-Matrix Algebraic Riccati Equations, SIAM Journal on Matrix Analysis and Applications 33 (1) (2012) 170–194.

[17] C. Van Loan, Computing integrals involving the matrix exponential, Automatic Control, IEEE Transactions on 23 (3) (1978) 395 – 404. doi:10.1109/TAC.1978.1101743.

[18] A. Horváth, G. Horváth, M. Telek, A joint moments based analysis of networks of MAP/MAP/1 queues, Perform. Eval. 67 (9) (2010) 759–778.

[19] G. H. Golub, S. Nash, C. Van Loan, A Hessenberg-Schur method for the problem AX+XB=C, IEEE Transactions on Automatic Control 24 (1979) 909–913.

[20] Q. He, Analysis of a continuous time SM[K]/PH[K]/1/FCFS queue: Age process, sojourn times, waiting times, and queue lengths, Journal of Systems Science and Complexity (JSSC) 25 (2012) 133–155.

[21] P. Buchholz, P. Kemper, J. Kriege, Multi-class markovian arrival processes and their parameter fitting, Performance Evaluation 67 (2010) 1092–1106.