European Journal of Operational Research 255 (2016) 822–833

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Compact Markov-modulated models for multiclass trace fitting

Giuliano Casale[a,*], Andrea Sansottera[b], Paolo Cremonesi[b]

[a] Department of Computing, Imperial College London, 180 Queen's Gate, SW7 2AZ, London, UK.
[b] Politecnico di Milano, DEIB, Via Ponzio 34/5, 20133 Milan, Italy

ARTICLE INFO

Article history:
Received 8 May 2014
Accepted 6 June 2016
Available online 14 June 2016

Keywords:
Counting process
Marked Markov-modulated Poisson process
Trace
Fitting

ABSTRACT

Markov-modulated Poisson processes (MMPPs) are stochastic models for fitting empirical traces for simulation, workload characterization and queueing analysis purposes. In this paper, we develop the first counting process fitting algorithm for the marked MMPP (M3PP), a generalization of the MMPP for modeling traces with events of multiple types. We initially explain how to fit two-state M3PPs to empirical traces of counts. We then propose a novel form of composition, called *interposition*, which enables the approximate superposition of several two-state M3PPs without incurring into state space explosion. Compared to exact superposition, where the state space grows exponentially in the number of composed processes, in interposition the state space grows linearly in the number of composed M3PPs. Experimental results indicate that the proposed interposition methodology provides accurate results against artificial and real-world traces, with a significantly smaller state space than superposed processes.

1. Introduction

The Markov-modulated Poisson process (MMPP) is a general fitting tool for traces with correlated arrivals (Fischer & Meier-Hellstern, 1993) which finds application in modeling network traffic (Okamura, Dohi, & Trivedi, August 2009), disk I/O patterns (Verma & Anand, December 2007), grid and cloud workloads (Li, Muskulus, & Wolters, 2006), and self-adaptive software systems (Perez-Palacin, Merseguer, & Mirandola, 2012). A central property of MMPPs is the composability with other Markov models, such as queueing systems (Horváth, Horváth, & Telek, 2009; Horváth, 2012; Houdt, 2012) or stochastic Petri nets (Perez-Palacin et al., 2012), which enables using non-renewal processes in performance models. This is important because autocorrelation and temporal dependence can greatly affect system performance and therefore need to be taken into account in system modeling (Mi, Zhang, Riska, Smirni, & Riedel, 2007).

The MMPP is a special case of the Markovian Arrival Process (MAP) (Neuts, 1979) in which the departure process is a modulation of Poisson processes and the active process is chosen according to the state of a continuous-time Markov chain. In this paper, we propose a generalization of the MMPP, which we call the *marked MMPP* (M3PP), and develop a scalable methodology for fitting M3PPs to empirical datasets. The M3PP can be regarded as a specialization of the MMAP, the marked extension of the MAP (He & Neuts, 1998). A marked point process associates to each arrival a class label (He & Neuts, 1998; He, 2001), thus it is useful to model traces with events of multiple classes (e.g., read and write requests in disk drives). Marked processes are also important in the analysis of priority queues and multiclass models (Buchholz, Kemper, & Kriege, 2010; Horváth et al., 2009; Horváth, 2012; Houdt, 2012). However, few techniques exist for their fitting and they all focus on matching moments of the inter-arrival time process for marked MAPs (MMAPs) (Buchholz et al., 2010; Horváth et al., 2009). Therefore, no techniques exist yet for fitting marked Markov-modulated processes to count data. Still, to limit monitoring overheads, only count data can be extracted from certain classes of computer and communications systems (e.g., network links). This motivates the investigation into methodologies to fit marked count data.

In this paper, we fill this gap by developing approximate fitting algorithms for the counting process of the M3PP. In particular, we first explain how to *approximately* fit a two-state MMPP counting process and then extend the idea to two-state M3PPs. The proposed approach is applicable to traces with aggregated or coarse-grained measurements, which cannot be analysed using approaches based on inter-arrival times, since these require moments from a trace recording *all* the arrivals. The main drawback is that counting processes are mathematically less tractable than inter-arrival processes, therefore one normally focuses on low-order moments of counts.

* Corresponding author. Tel.: +44 20 759 42920.
E-mail addresses: g.casale@imperial.ac.uk (G. Casale), sansottera@elet.polimi.it (A. Sansottera), paolo.cremonesi@polimi.it (P. Cremonesi).

http://dx.doi.org/10.1016/j.ejor.2016.06.005
0377-2217/

Two-state models are often insufficient to fit complex traces, therefore we also study the approximate fitting of large M3PPs. In the single class setting, a known limitation of MMPPs is the inability to simultaneously fit many statistical descriptors due to the non-linearity of their underlying equations (Bodrog, Heindl, Horváth, & Telek, 2008; Heindl, Horváth, & Gross, 2006; Horváth & Telek, 2009). This has led to the definition of several approaches to fit complex traces by composing multiple small-sized MMPPs or MAPs using Kronecker operators (Andersen & Nielsen, 1998; Casale, Zhang, & Smirni, 2010; Horváth & Telek, 2002). These methods employ composition operators for moment fitting, offering a different trade-off between computational cost and fitting accuracy compared to fitting methods based on the EM algorithm (Breuer, 2002; Horváth & Okamura, 2013; Klemm, Lindemann, & Lohmann, 2003). In particular, the superposition operator allows one to describe a trace by the statistical multiplexing of several MMPPs, at the expense of an exponential growth of the number of states in the resulting process (Sriram & Whitt, 1986). This state space explosion is an obstacle for the application of MMPPs and MAPs to modeling real systems; for example it considerably slows down, or even renders infeasible, the numerical evaluation of queueing models by matrix geometric methods (Bini, Meini, Steffé, Pérez, & Houdt, 2012; Pérez, Velthoven, & Houdt, 2008).

In this paper, we tackle the state space explosion problem of superposition by showing that M3PPs admit a particular form of composition, which we call *interposition*, that enables several MMPPs to share the same state space without mutually affecting the marginal distributions of their counting processes. However, interposition introduces spurious covariances between class arrivals that may be seen as the cost of the state-space reduction. The interposition method defines an original approach to build large Markov-modulated processes, in which a set of J two-state M3PPs is merged into a single M3PP process with just $J+1$ states and without affecting the marginal counting processes of the initial M3PPs. We identify general conditions for interposition to result in a valid MMAP and observe that these conditions can be satisfied by M3PPs, but not by general MMAPs. The ability to interpose processes makes a case for using M3PPs, instead of general MMAPs, for fitting count data. We then define a method to automatically identify the M3PPs to be interposed and a mixed-integer linear program (MILP) that can help in automatically identifying the order of interposition of the M3PPs.

We conclude the paper by reporting fitting experiments for a set of artificial and real-world traces. We show that the proposed M3PP fitting algorithms are widely applicable and run efficiently even in the case of approximate fitting. We also find that interposition is much more scalable than superposition, while retaining comparable accuracy.

Summarizing, our main contributions are as follows:

- Section 3 defines fitting algorithms for the counting process of two-state M3PPs with arbitrary number of classes. Our formulas are in closed-form for exact fitting and use quadratic programming for approximate fitting. As a by-product, our approach also introduces the first infeasibility adjustment methodology for approximate fitting of unmarked MMPPs.
- Section 4 introduces the new notion of interposition. This is an aid to compose multiple M3PPs, while preserving their statistical properties, as we rigorously establish in Theorem 1.
- Section 5 develops a methodology for fitting empirical traces using the interposition operator.

The paper reports in Section 6 an experimental study on random models and a real trace validating the effectiveness of the proposed models and algorithms. In addition to the above, a description of necessary background is given in Section 2. Section 7 concludes the paper.

2. Background

2.1. Model and notation

An m-state MMPP is specified by a continuous-time Markov chain (CTMC) with irreducible infinitesimal generator $\boldsymbol{Q}$, having m states, and rate vector $(\lambda(1), \ldots, \lambda(m))$. When the CTMC is in state j, the MMPP generates arrivals according to a Poisson process with rate $\lambda(j)$. The effect of the modulating action of the underlying CTMC is to enable the modeling of non-Poisson, possibly nonrenewal, arrival processes. We assume the underlying CTMC to be initialized according to its steady-state distribution so that the process is time-stationary. For ease of comparison with the literature, we use throughout the paper the MAP $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ notation, where for a MMPP $\boldsymbol{D}_1 = \mathrm{diag}(\lambda(1), \ldots, \lambda(m))$ and $\boldsymbol{D}_0 = \boldsymbol{Q} - \boldsymbol{D}_1$.

We define a M3PP[K] as a MMPP in which arrivals are marked with one out of K available classes. This may be seen as a marking of the Poisson processes of the MMPP where one decomposes each rate $\lambda(j)$, $1 \le j \le m$, into arrival rates $q_{j,k}\lambda(j)$, $1 \le k \le K$, subject to $\sum_k q_{j,k} = 1$, $q_{j,k} \ge 0$. When an arrival is generated by a Poisson process with rate $q_{j,k}\lambda(j)$ it is said to be of class k. Equivalently, in matrix notation, augmenting a MMPP $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ with marks defines a set of matrices $\boldsymbol{D}_{1,k} = \mathrm{diag}(q_{1,k}\lambda(1), \ldots, q_{m,k}\lambda(m))$, such that $\boldsymbol{D}_1 = \sum_{k=1}^{K} \boldsymbol{D}_{1,k}$. The tuple $(\boldsymbol{D}_0, \boldsymbol{D}_{1,1}, \ldots, \boldsymbol{D}_{1,K})$ is called the *representation* of the M3PP[K]. Also, $(\boldsymbol{D}_0, \boldsymbol{D}_1)$ is the embedded MMPP, which we refer to as the *unmarked process*.

In the rest of the paper we often deal with processes having $m = 2$ states. In this case, for readability, we use λ and λ' in place of $\lambda(1)$ and $\lambda(2)$ and q_k and q'_k in place of $q_{1,k}$ and $q_{2,k}$.

2.2. Problem statement

Let us consider a M3PP[K] and let $n_k(t) \ge 0$ be the number of arrivals of class k at time t after initialization, subject to $n_k(0) = 0$ for all classes. The *counting process* of the M3PP[K] is the CTMC with state $\boldsymbol{n}(t) = (n_1(t), n_2(t), \ldots, n_K(t))$. We shall denote by $n(t) = \sum_{k=1}^{K} n_k(t)$ the aggregate arrival count. Given an initial state probability vector, the evolution over time of this process is characterized by a matrix $\boldsymbol{P}(\boldsymbol{n}, t)$, with element $p_{i,j}(\boldsymbol{n}, t)$ in row i and column j representing the probability that a M3PP initialized in state i is in state j at time t with an arrival count $\boldsymbol{n}(t)$. We refer to $\boldsymbol{P}(\boldsymbol{n}, t)$ as the *counting process matrix*. The counting process evolves according to the Kolmogorov forward equations

$$\frac{\partial \boldsymbol{P}(\boldsymbol{n}, t)}{\partial t} = \boldsymbol{P}(\boldsymbol{n}, t)\boldsymbol{D}_0 + \sum_{k=1}^{K} \boldsymbol{P}(\boldsymbol{n} - \boldsymbol{e}_k, t)\boldsymbol{D}_{1,k}, \tag{1}$$

where $\boldsymbol{e}_k$ is a column vector of zeros except for a one in position k. From this equation, it is simple to derive factorial moment functions, from which moments of the counting process can be obtained in closed form He and Neuts (1998). For example, this method yields the mean arrival count of class k at time t as

$$\mu_k(t) = \mathrm{E}[n_k(t)] = \mu_k t, \qquad \mu_k = \boldsymbol{\pi}\boldsymbol{D}_{1,k}\boldsymbol{1}, \tag{2}$$

where $\boldsymbol{\pi}$ is the stationary distribution of the embedded CTMC, $\boldsymbol{\pi Q} = \boldsymbol{0}$, $\boldsymbol{\pi 1} = 1$, μ_k is the *mean arrival rate* of class k in the time-stationary process, $\boldsymbol{0}$ and $\boldsymbol{1}$ are respectively column vectors of zeros and ones. The *variance* of class k in counts (also called variance-time curve) is (He & Neuts, 1998)

$$\mathrm{Var}[n_k(t)] = \left(\mu_k - 2\mu_k^2 + 2\boldsymbol{c}_k\boldsymbol{D}_{1,k}\boldsymbol{1}\right)t - 2\boldsymbol{c}_k\left(\boldsymbol{I} - e^{\boldsymbol{Q}t}\right)\boldsymbol{d}_k, \tag{3}$$

where $\boldsymbol{c}_k = \boldsymbol{\pi}\boldsymbol{D}_{1,k}(\boldsymbol{1\pi} - \boldsymbol{Q})^{-1}$ and $\boldsymbol{d}_k = (\boldsymbol{1\pi} - \boldsymbol{Q})^{-1}\boldsymbol{D}_{1,k}\boldsymbol{1}$. The above formulas and notations are also valid for the unmarked process ($K = 1$), but in that case we omit the dependence on class k. The *covariance* of counts for classes k and h is given by

$$\mathrm{Cov}[n_k(t), n_h(t)] = \frac{1}{2}(\mathrm{Var}[n_k(t)+n_h(t)] - \mathrm{Var}[n_k(t)] - \mathrm{Var}[n_h(t)]), \tag{4}$$

where $\mathrm{Var}[n_k(t)+n_h(t)]$ can be obtained from (3) by replacing μ_k with $\mu_k+\mu_h$ and $\boldsymbol{D}_{1,k}$ with $\boldsymbol{D}_{1,k}+\boldsymbol{D}_{1,h}$.

The M3PP[K] fitting problem is to find a representation $(\boldsymbol{D}_0, \boldsymbol{D}_{1,1}, \ldots, \boldsymbol{D}_{1,K})$ such that (2)–(4) match, to the best possible extent, the corresponding empirical moments at given timescales t. Such a representation needs to be valid, meaning that all rates need be non-negative real numbers and the generator $\boldsymbol{Q}$ of the embedded CTMC must be valid and irreducible.

2.3. Two-state MMPP fitting

In order to fit a two-state M3PP[K], we propose to separately fit the MMPP for the unmarked process and the class markings of the M3PP. Since MMPP fitting is well-understood, we just summarize the MMPP fitting method proposed in Heffes and Lucantoni (1986). This algorithm receives in input the following descriptors of the counting process:

$$\mu = \frac{\mathrm{E}[n(t)]}{t}, \quad I(t) = \frac{\mathrm{Var}[n(t)]}{\mathrm{E}[n(t)]}, \quad I = \lim_{t\to\infty} I(t),$$
$$\mu^{(3)}(t) = \mathrm{E}\left[(n(t) - \mu t)^3\right],$$

where μ is the rate, $I(t)$ is the *index of dispersion for counts* (Sriram & Whitt, 1986) at a given timescale t, I is the asymptotic index of dispersion value for $t \to \infty$, and $\mu^{(3)}(t)$ is the third central moment of counts at timescale t. The objective is to fit the rates used in the MMPP representation

$$\boldsymbol{D}_0 = \begin{bmatrix} -(\lambda+r) & r \\ r' & -(\lambda'+r') \end{bmatrix}, \qquad \boldsymbol{D}_1 = \mathrm{diag}(\lambda, \lambda'),$$

subject to all the parameters being non-negative and to $\lambda + \lambda' > 0$. Note in particular that we exclude the trivial cases $\lambda = \lambda'$ and $r = r' = 0$, where the MMPP degenerates into a Poisson process, which does not require a two-state model.

Fitting the above descriptors to a two-state MMPP requires to solve for the unknowns r, r', λ, and λ' using a nonlinear system composed by the following four equations (Heffes & Lucantoni, 1986)

$$\mu = \frac{\lambda r' + \lambda' r}{x}, \quad I = 1 + \frac{2(\lambda-\lambda')^2 r r'}{x^2(\lambda r' + \lambda' r)}, \quad \frac{I - I(t)}{I-1} = \frac{1-e^{-xt}}{xt},$$
$$h = (\lambda - \lambda')x', \tag{5}$$

where $x = r + r'$, $x' = r - r'$, t is an arbitrary finite timescale at which we want to fit the index of dispersion, and $h = (g^{(3)}(1,t) - k_1 - k_2 + k_3\mu - k_4\mu x)(k_4 + (k_3/x) - k_5)^{-1}$. The parameters in the last equation are $k_1 = \mu^3 t^3$, $k_2 = 3\mu^2(I-1)t^2$, $k_3 = 3\mu(I-1)/xt$, $k_4 = 3\mu(I-1)te^{-xt}/x^2$, $k_5 = 6\mu(I-1)(1-e^{-xt})/x^3$, and $g^{(3)}(1, t)$ is the third factorial moment of counts at timescale t. We point to Heffes and Lucantoni (1986, Eq. (14d)) for explicit expressions of $g^{(3)}(1, t)$.

Heffes and Lucantoni (1986) propose the following fitting method. First, compute $x = r + r'$, solving at an arbitrary timescale $t = t_1$ the fixed point equation

$$x = \frac{1}{t}\left(\frac{I-1}{I-I(t)}\right)\left(1 - e^{-xt}\right), \tag{6}$$

obtained by rewriting the third equation appearing in (5). Then, compute h at a second arbitrary timescale $t = t_2$. If $h \neq 0$, the fitting formulas are then given by

$$r = \frac{x}{2}\left(1 + \frac{1}{\sqrt{4y+1}}\right), \quad r' = x - r, \quad \lambda' = \mu - \frac{hr'}{x'x}, \quad \lambda = \mu + \frac{hr}{x'x}.$$

where $y = (I-1)\mu x^3(2h^2)^{-1}$. If $h = 0$, the explicit fitting formulas instead become

$$r = r' = \frac{x}{2}, \quad \lambda' = \mu - \frac{1}{2}\sqrt{2(I-1)\mu x}, \quad \lambda = \mu + \frac{1}{2}\sqrt{2(I-1)\mu x}.$$

The main drawback of this method is that the formulas can return negative rates for some combinations of the input parameters. In this case, approximate fitting techniques are required. Yet, to the best of our knowledge these are not available in the literature on MMPP counting process fitting.

3. Fitting the counting process of a two-state M3PP[K]

In this section, we develop a method to fit two-state M3PP[K]s with representation

$$\boldsymbol{D}_0 = \begin{bmatrix} -(\lambda+r) & r \\ r' & -(\lambda'+r') \end{bmatrix}, \quad \boldsymbol{D}_{1,k} = \mathrm{diag}(q_k\lambda, q'_k\lambda'),$$

where $\lambda + \lambda' > 0$, $\sum_{k=1}^K q_k = \sum_{k=1}^K q'_k = 1$, $q_k \geq 0$, $q'_k \geq 0$, for all $1 \leq k \leq K$. As mentioned before, the availability of exact methods to fit unmarked MMPPs justifies an approach where we separately fit the embedded MMPP first, followed by the marking process that defines the M3PP. In practice, this means that we first fit $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ using a MMPP fitting algorithm applied to the unmarked trace. Then we determine the individual $\boldsymbol{D}_{1,k}$ matrices using the marked trace and subject to the condition $\boldsymbol{D}_1 = \sum_{k=1}^K \boldsymbol{D}_{1,k}$, which ensures that the embedded MMPP does not change. In Section 3.1, we discuss how to perform approximate fitting of the unmarked process using an extension of the algorithm by Heffes and Lucantoni. In Section 3.2, we discuss the fitting of the marked process.

3.1. Step 1: approximate MMPP fitting

The fitting algorithm of Heffes and Lucantoni described in Section 2.3 cannot be applied to cases where the input set of descriptors μ, $I(t)$, I and $\mu^{(3)}(t)$ is infeasible for the considered MMPP. We have found this to happen frequently in real traces, and even though one may repeatedly attempt to fit different timescales t_1 and t_2 until finding a feasible set of descriptors, we believe that better solutions are needed and possible. We have not found previous works addressing this problem, at least for fitting counting processes. Our approximation consists in sacrificing the degree of freedom of the third moment of counts $\mu^{(3)}(t)$ to restore feasibility of all second-order descriptors. As before, we focus on cases where the M3PP process is not Poisson, thus we assume $\lambda \neq \lambda'$, $r > 0$, $r' > 0$.

We begin by characterizing the feasibility region of the index of dispersion for a two-state MMPP.

Proposition 1. *In a two-state MMPP with $\lambda \neq \lambda'$, the index of dispersion satisfies $I > I(t) > 1$ for any timescale t.*

Proof. From the second equation in (5), it readily follows that $I > 1$ since $\lambda \neq \lambda'$. Moreover, since $x > 0$, it follows from (6) that $I > I(t)$. Noting that for $x > 0$ it is

$$0 < \frac{1-e^{-xt}}{xt} < 1, \qquad \forall t > 0, x > 0,$$

by (5) and the constraint $I > I(t)$ we conclude that $I(t) > 1$. □

The previous proposition states the well-known fact that MMPPs can only fit traces that have greater variability than a Poisson process. We now show that asking for $r > 0$ and $r' > 0$, to exclude the case where the MMPP is a Poisson process, is implied by the previous statement and thus always verified.

Proposition 2. *For $I > 1$, the conditions $r > 0$ and $r' > 0$ are always satisfied.*

Proof. For the case $h \neq 0$, observe that $y > 0$ if $I > 1$. Therefore $1/\sqrt{4y+1} < 1$ that implies $r > 0$ and $r' > 0$. Since x is positive, it also follows that $0 < r < x$ and $0 < r' < x$.

For the case $h = 0$, the result follows immediately given that $x > 0$ since $r = r' = 0$ would imply a Poisson process that would have $I = 1$ against the assumption. □

Therefore, provided that $I > I(t) > 1$, feasibility follows by ensuring that arrival rates are non-negative and $\lambda + \lambda' > 0$. Using (5), we can reformulate this requirement as:

$$-\frac{\mu x}{r} \leq \frac{h}{x'} \leq \frac{\mu x}{r'}. \tag{7}$$

Note that this expression implicitly characterizes the feasible region for the rate μ and, via the h term, for the third moment $g^{(3)}(1, t)$. If one of these conditions is not met, we propose to relax the fitting method by ignoring the matching of the third moment $g^{(3)}(1, t)$ as follows. Consider first the following lower bound on r.

Proposition 3. *Without loss of generality, assume $\lambda > \lambda'$. Then in a feasible MMPP(2) it is $r \geq u$, where*

$$u = \frac{(I-1)x^2}{2\mu + (I-1)x}. \tag{8}$$

in which $x = r + r'$ is the solution of (6).

Proof. We substitute the first equation in (5) into the second one and find after simple algebra

$$\lambda = \lambda' + \sqrt{\frac{(I-1)\mu x^3}{2rr'}}. \tag{9}$$

Obtaining λ from the first equation in (5) and equating to (9), we get the feasibility constraint

$$\lambda' = \mu - \frac{r'}{x}\sqrt{\frac{(I-1)\mu x^3}{2rr'}} \geq 0. \tag{10}$$

The result follows using $r' = x - r$ and solving for r, which yields the lower bound u. □

Any value $u \leq r < x$ leads to a feasible solution, thus we can for example choose the middle point of this interval $r = u + (u - x)/2$. Our suggestion of the middle point is convenient to keep the formulas in closed-form, however other choices are possible. After choosing r, the other parameters are easily obtained by setting $r' = x - r$ and using (9) and (10) to obtain λ' and λ.

Summarizing, provided that $I > I(t_1) > 1$, the approximate method we have defined fits μ, $I(t_1)$ and I exactly at the expense of sacrificing an infeasible third moment $\mu^{(3)}(t_2)$, where t_1 and t_2 are arbitrary timescales.

3.2. Step 2: fitting the marking process

In the second step of the fitting algorithm we determine the $\boldsymbol{D}_{1,k}$ matrices. After Step 1, all the statistical properties of the unmarked process are constrained by the $\boldsymbol{D}_0$ and $\boldsymbol{D}_1$ matrices of the MMPP, thus the focus of this step is to fit the class-specific properties and the class covariances. We consider both exact and approximate fitting methods.

There are several ways of choosing which empirical descriptors to fit with a M3PP and each choice leads to equations that may differ in tractability compared to other choices. We have experimented with several possibilities, and we have found that fitting a set of central moments, such as mean class rates μ_k, class variances, or covariances, typically leads to a difficult non-convex formulation. Conversely, we have found more efficient to fit a two-state M3PP[K] using the mean class rates and per-class contributions to the asymptotic index of dispersion I. Other efficient approaches are possible, such as fitting mean class rates and relative covariances or fitting mean class rates and differences between the class variances. However, we here focus on the first method, which we believe provides the best combination of efficiency (quadratic programming) and ease of interpretation.

3.2.1. Fitting mean class rates and per-class contributions to the index of dispersion

We begin by considering the more general problem of fitting the mean class rates μ_k and $g_k(t) = \mathrm{Var}[n_k(t)] + \mathrm{Cov}[n_k(t), \sum_{i\neq k} n_i(t)]$. The $g_k(t)$ terms have the simple interpretation of modeling the contribution of class k to the variance-time curve of the unmarked process, since $\sigma(t) = \mathrm{Var}\left[\sum_{k=1}^{K} n_k(t)\right] = \sum_{k=1}^{K} g_k(t)$. We then specialize the method to the index of dispersion, using the fact that $I(t) = \sigma(t)/E[n(t)] = \sum_{k=1}^{K} g_k(t)/E[n(t)]$.

Our goal is to compute the $\boldsymbol{D}_{1,k}$ matrices from the $g_k(t)$ values, given $\boldsymbol{D}_0$ and $\boldsymbol{D}_1 = \sum_{k=1}^{K} \boldsymbol{D}_{1,k}$. Recall that $\boldsymbol{D}_{1,k} = \mathrm{diag}(q_k\lambda, q'_k\lambda')$. The problem under consideration is to determine the values of the probabilities q_k and q'_k that uniquely define the $\boldsymbol{D}_{1,k}$ matrices, given that λ and λ' are known from the $\boldsymbol{D}_1$ matrix fitted in Step 1.

Exact matching. We now give formulas to fit the probabilities q_k and q'_k from μ_k and $g_k(t)$. Note that the arbitrary timescale t used in the statement does not need to be equal to the timescales used for fitting the embedded MMPP.

Proposition 4. *Given ($\boldsymbol{D}_0$, $\boldsymbol{D}_1$) and, for each class k, μ_k and $g_k(t)$ at an arbitrary timescale t, the parameters of the M3PP can be computed as follows:*

$$q_k = f_{1,1}\mu_k + f_{1,2}g_k(t), \tag{11}$$

$$q'_k = f_{2,1}\mu_k + f_{2,2}g_k(t), \tag{12}$$

where $f_{1,1} = -F_2 x/F$, $f_{1,2} = \lambda' r/F$, $f_{2,1} = F_1 x/F$, $f_{2,2} = -\lambda r'/F$, in which $F = (F_1\lambda' r - F_2\lambda r')$,

$$F_1 = \lambda r' x^{-4}(2(\lambda - \lambda')re^{-xt} + tx(x^2 - 2(\lambda' - \lambda)r) + 2(\lambda' - \lambda)r),$$

$$F_2 = \lambda' r x^{-4}(2(\lambda' - \lambda)r' e^{-xt} + tx(x^2 - 2(\lambda - \lambda')r') + 2(\lambda - \lambda')r'),$$

and $x = r + r'$ is the solution of (6).

Proof. The expressions of μ_k and $g_k(t)$ for a two-state M3PP[K] process are found from the definitions to be

$$\mu_k = \frac{\lambda q_k r' + \lambda' q'_k r}{x}, \tag{13}$$

$$g_k(t) = F_1 q_k + F_2 q'_k, \tag{14}$$

where F_1 and F_2 are constant coefficients given $\boldsymbol{D}_0$, $\boldsymbol{D}_1$ and the reference timescale t. Solving (13) for q_k we obtain

$$q_k = \frac{\mu_k x - \lambda' q'_k r}{\lambda r'}. \tag{15}$$

Substituting (15) into (14), we obtain the fitting formulas (11) and (12). □

We are now ready to determine the contributions to the index of dispersion. Note that Proposition 4 may also be used to fit the contribution of class k to $I(t)$ in the embedded MMPP, i.e., $G_k(t) = g_k(t)/E[n(t)]$. This is because we can rewrite the fitting formulas as

$$q_k = c_{1,1}\mu_k + c_{1,2}G_k(t), \tag{16}$$

$$q'_k = c_{2,1}\mu_k + c_{2,2}G_k(t), \tag{17}$$

where $c_{1,1} = f_{1,1}$, $c_{1,2} = f_{1,2}\mu t$, $c_{2,1} = f_{2,1}$, $c_{1,1} = f_{2,2}\mu t$. The asymptotic expressions of the coefficients as $t \to \infty$ are then readily obtained as

$$c_{1,1} = -\frac{2(\lambda' - \lambda)r' + x^2}{2\lambda r'(\lambda - \lambda')},\ c_{1,2} = \frac{\mu x^2}{2\lambda r'(\lambda - \lambda')},$$
$$c_{2,1} = \frac{2(\lambda - \lambda')r + x^2}{2\lambda' r(\lambda - \lambda')},\ c_{2,2} = -\frac{\mu x^2}{2\lambda' r(\lambda - \lambda')}.$$

Combining these expressions with the requirements $q_k \geq 0$ and $q'_k \geq 0$, we find after simple passages this lower bound required to hold for the feasibility of the per-class contributions to the asymptotic index of dispersion

$$G_k \geq \frac{\mu_k}{\mu}\left(1 + 2\frac{\max((\lambda' - \lambda)r', (\lambda - \lambda')r)}{x^2}\right). \tag{18}$$

For two-state M3PPs, the minimum value of the right-hand side is achieved for Poisson processes, where $G_k = \mu_k/\mu$ and $\sum_k G_k = I = 1$.

Approximate matching. Some values of the descriptors may lead to unfeasible values of the parameters (e.g., negative probabilities). In this case, we choose to fit the per-class arrival rates μ_k exactly and find the feasible values $\tilde{G}_k$ that minimize the following L^2-norm

$$\sum_{k=1}^{K}\left(\frac{\tilde{G}_k - G_k}{G_k}\right)^2 = \frac{1}{2}\boldsymbol{x}^T H \boldsymbol{x} + f^T \boldsymbol{x} + K, \tag{19}$$

with $\boldsymbol{x} = (\tilde{G}_1, \ldots, \tilde{G}_K)$, $H = \mathrm{diag}(2/G_1^2, \ldots, 2/G_K^2)$, $f^T = (-2/G_1, \ldots, -2/G_K)$, K being the number of classes of the M3PP, and subject to the constraints

$$-c_{i,2}\tilde{G}_k \leq c_{i,1}\mu_k \qquad \forall i = 1, 2;\ k = 1, \ldots, K \tag{20}$$

$$c_{i,2}\sum_{k=1}^{K}\tilde{G}_k = 1 - c_{i,1}\mu \qquad \forall i = 1, 2. \tag{21}$$

These constraints are derived using equations (16) and (17) for q_k and q'_k. In particular, (20) stems directly from the constraints $q_k \geq 0$ and $q'_k \geq 0$, whereas (21) follows from the conditions $\sum_{k=1}^{K} q_k = 1$ and $\sum_{k=1}^{K} q'_k = 1$.

The above optimization program may also be used for fitting mean class rates μ_k and the contributions $g_k(t)$ to the index of dispersion $I(t)$. In order to do so, it is sufficient to replace the coefficients $c_{i,j}$ with the coefficients $f_{i,j}$, G_k with $g_k(t)$, and $\tilde{G}_k$ with $\tilde{g}_k(t)$.

In either case, the program has a convex quadratic objective function, thus its minimizer can be efficiently obtained using standard quadratic programming solvers. Once the problem is solved and the feasible values $\tilde{G}_k$ or $\tilde{g}_k(t)$ are obtained, the parameters q_k and q'_k are readily fitted using (16) and (17).

4. Compositional fitting of M3PP[*K*]s

In the previous section, we have defined a general purpose fitting method for two-state M3PPs. We now consider the problem of exploiting the additional degrees of freedom of a M3PP[*K*] to increase the flexibility of the fitting. In this case, exact fitting methods capable of exploiting all available degrees of freedom do not exist even for unmarked processes. Therefore we focus on compositional fitting, where one builds a large process by composition of smaller processes that are simpler to fit to count data. The drawback of compositional approaches of this kind is that they use just a few degrees of freedom of a general MMAP in return for ease of fitting.

We first review and generalize superposition methods for unmarked MMPPs. Afterwards, we introduce a novel form of composition, called *interposition*, which offers a different trade-off between accuracy and compactness of the representation. Note that we do not consider methods that are specific to inter-arrival processes, such as the Kronecker product composition (Casale et al., 2010).

4.1. Superposition

Unmarked case. Consider a set of J MMPPs $(\boldsymbol{D}_0^j, \boldsymbol{D}_1^j)$, $1 \leq j \leq J$, their superposition is the MMPP process $(\boldsymbol{D}_0^+, \boldsymbol{D}_1^+)$ where $\boldsymbol{D}_0^+ = \bigoplus_{j=1}^J \boldsymbol{D}_0^j$, $\boldsymbol{D}_1^+ = \bigoplus_{j=1}^J \boldsymbol{D}_1^j$, in which $\oplus$ denotes the Kronecker sum operator (Brewer, 1978). Superposition naturally arises in networking to describe the traffic process obtained by merging separate traffic flows, each described by a MMPP. The superposed process defines the multiplexing of J channels, each with inter-arrival times described by the j-th MMPP. This process has mean arrival rate and variance-time curve equal to the sum of mean arrival rates and variance-time curves of the individual MMPPs. The index of dispersion is a weighted sum of the IDCs of the individual MMPPs, i.e., $I(t) = \sum_j (\mu_j/\mu) I_j(t)$, where $\mu = \sum_j \mu_j$ is the mean arrival rate of the superposition. Also, if MMPP j has m_j states, the superposed process has $\prod_{j=1}^J m_j$ states, thus its size grows exponentially with J.

Marked case. Let $\mathcal{K} = \{1, \ldots, K\}$ be a set of classes. We consider J M3PPs and assume that M3PP j generates arrival of classes $\mathcal{K}_j \subset \mathcal{K}$, with $\mathcal{K}_1 \cup \mathcal{K}_2 \cup \cdots \cup \mathcal{K}_J = \mathcal{K}$. In this case, some M3PPs contribute to arrivals of class k, thus $\boldsymbol{D}_0^+ = \bigoplus_{j=1}^J \boldsymbol{D}_0^j$, $\boldsymbol{D}_{1,k}^+ = \bigoplus_{j=1}^J \boldsymbol{D}_{1,k}^j$, $\forall k \in \mathcal{K}$, where we define $\boldsymbol{D}_{1,k}^j = \boldsymbol{0}$ if $k \notin \mathcal{K}_j$, where $\boldsymbol{0}$ is here a matrix of all zeros of order m_j. The statistical properties of the embedded unmarked process are obtained as in an unmarked superposition. Lastly, note that if each M3PP has two states, the resulting process is a M3PP with 2^J states and K classes. Thus, the main drawback of the superposition method is the state-space explosion, which limits the composition to a small number of processes.

4.2. Interposition

We now propose a new form of composition for Markov-modulated processes that tackles the state-space explosion problem of superposition. Informally, our idea is to define an operator by which several M3PPs can by defined upon the same state space, without interfering with their respective marginal counting processes. This allows us to preserve in the interposition the counting process properties fitted in isolation for each composed M3PP. We characterize the main feature of this new composition operator in Theorem 1, given later. Note that the results in this section are also applicable to unmarked MMPPs, and thus represent advances also for unmarked processes.

Consider a set of J two-state M3PPs where the ith process has K_i classes. Assume without loss of generality that classes are labelled such that each class appears in one and only one M3PP. The M3PPs have representation

$$\boldsymbol{D}_0^i = \begin{bmatrix} -r_i - \lambda_i & r_i \\ r'_i & -r'_i - \lambda'_i \end{bmatrix}, \quad \boldsymbol{D}_{1,k}^i = \mathrm{diag}(q_{i,k}\lambda_i, q'_{i,k}\lambda'_i),$$
$$k = 1 \ldots, K_i,$$

where we define the probabilities $\sum_{k=1}^{K_i} q_{i,k} = 1$, $q_{i,k} \geq 0$, and $\sum_{k=1}^{K_i} q'_{i,k} = 1$, $q'_{i,k} \geq 0$, and the rates $r_i = \sum_{j=i}^{J} \alpha_j$ and $r'_i = \sum_{j=1}^{i} \beta_j$, for given constants $\alpha_j \geq 0$ and $\beta_j \geq 0$. Equivalently, given the values of the r_i and r'_i constants, we can define the rate differences $\alpha_i = r_i - r_{i+1}$ and $\beta_i = r'_i - r'_{i-1}$, with boundary values $\beta_1 = r'_1$ and $\alpha_J = r_J$.

We now define the interposition operator for M3PPs. Given the set of J two-state M3PPs $(\boldsymbol{D}_0, \boldsymbol{D}_{1,1}^i, \ldots, \boldsymbol{D}_{1,K_i}^i)$, the interposed process $(\boldsymbol{D}_0^*, \boldsymbol{D}_{1,1}^*, \ldots, \boldsymbol{D}_{1,K}^*)$ is the M3PP with $J+1$ states, $K = \sum_{i=1}^J K_i$

classes, and representation

$$\boldsymbol{D}_0^* = \begin{bmatrix} -\Sigma & \alpha_1 & \alpha_2 & \dots & \alpha_J \\ \beta_1 & -\Sigma & \alpha_2 & \dots & \alpha_J \\ \vdots & \ddots & -\Sigma & \ddots & \vdots \\ \beta_1 & \dots & \beta_{J-1} & -\Sigma & \alpha_J \\ \beta_1 & \dots & \beta_{J-1} & \beta_J & -\Sigma \end{bmatrix},$$

$$\boldsymbol{D}_{1,k}^* = \begin{bmatrix} q_{i,c}\lambda_i \boldsymbol{I}_i & \boldsymbol{0} \\ \boldsymbol{0} & q'_{i,c}\lambda'_i \boldsymbol{I}_{J+1-i} \end{bmatrix},$$

where $\boldsymbol{I}_n$ is the identity matrix of order n, class $k = \sum_{j=1}^{i-1} K_j + c$ is the class in the interposed process associated to class c of the i-th composed M3PP, and the diagonal elements of $\boldsymbol{D}_0^*$ are such that $(\boldsymbol{D}_0^* + \sum_k \boldsymbol{D}_{1,k}^*)\boldsymbol{1} = \boldsymbol{0}$. Throughout this section, we denote by $\mathcal{K}_i$ the set of class indexes in the interposed process associated to the i-th composed M3PP, and by $\bar{\mathcal{K}}_i = \{1, \dots, K\} \setminus \mathcal{K}_i$ its complement.

The interposed process may be seen as a M3PP modulated by a CTMC with an infinitesimal generator $\boldsymbol{Q}^* = \boldsymbol{D}_0^* + \sum_{k=1}^{K} \boldsymbol{D}_{1,k}^*$ that allows *exact* CTMC aggregation (Bolch, Greiner, de Meer, and Trivedi, 1998, Chap 4). Specifically, for each of the initial M3PPs, we can define a partition of the states of the interposed process such that, by exact aggregation of the CTMC $\boldsymbol{Q}^*$ one recovers the corresponding CTMC of the initial M3PP. For example, we may consider the aggregation

$$\boldsymbol{Q}^* = \left[\begin{array}{c|cccc} -r_1 & \alpha_1 & \alpha_2 & \dots & \alpha_J \\ \hline \beta_1 & -r_2 - r'_1 & \alpha_2 & \dots & \alpha_J \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \beta_1 & \dots & \beta_{J-1} & -r_J - r'_{J-1} & \alpha_J \\ \beta_1 & \dots & \beta_{J-1} & \beta_J & -r'_J \end{array}\right]$$

$$\Rightarrow \begin{bmatrix} -r_1 & \sum_{i=1}^{J} \alpha_i \\ \beta_1 & -\beta_1 \end{bmatrix} = \begin{bmatrix} -r_1 & r_1 \\ r'_1 & -r'_1 \end{bmatrix} = \boldsymbol{Q}_1,$$

where $\boldsymbol{Q}_1 = \boldsymbol{D}_0^1 + \sum_k \boldsymbol{D}_{1,k}^1$ is the infinitesimal generator of the ith composed M3PP. Similarly, for each partition, the definition of $\boldsymbol{D}_{1,k}^*$ ensures that the departure rates are identical to the ones in the composed M3PP.

In order to prove that the marginal counting process of each composed M3PP, for all classes $k \in \mathcal{K}_i$, is preserved by the interposition operator, we require additional notation. Let $\boldsymbol{P}(\boldsymbol{n}_i, t)$, $\boldsymbol{n}_i = (n_1, \dots, n_{K_i})$, be the 2×2 counting process matrix for the i-th M3PP at time t. Similarly, let $\boldsymbol{P}^*(\boldsymbol{n}, t)$ be the $(J+1) \times (J+1)$ counting process matrix for the composed M3PP, where $\boldsymbol{n} = (\boldsymbol{n}_1, \dots, \boldsymbol{n}_J)$. Let $\boldsymbol{0}_n$ and $\boldsymbol{1}_n$ be row vectors of size n of all zeros and all ones, respectively, and define the aggregation matrix (Buchholz, 1994)

$$\boldsymbol{V}_i = \begin{bmatrix} \boldsymbol{1}_i & \boldsymbol{0}_{J+1-i} \\ \boldsymbol{0}_i & \boldsymbol{1}_{J+1-i} \end{bmatrix}.$$

For an arbitrary $(J+1) \times (J+1)$ matrix $\boldsymbol{X}$, the 2×2 matrix $\boldsymbol{X}_i = \boldsymbol{W}_i \boldsymbol{X} \boldsymbol{V}_i^T$, where $\boldsymbol{W}_i = \mathrm{diag}(\boldsymbol{1}_{J+1}^T \boldsymbol{V}_i^T)^{-1} \boldsymbol{V}_i$, gives the aggregation of $\boldsymbol{X}$ into the two states associated to the i-th composed M3PP. Therefore

$$\boldsymbol{P}_i^*(\boldsymbol{n}, t) = \boldsymbol{W}_i \boldsymbol{P}^*(\boldsymbol{n}, t) \boldsymbol{V}_i^T,$$

is the counting process matrix aggregated onto the states of the i-th composed M3PP. From this matrix, we can readily compute the marginal counting process matrix of the i-th M3PP as

$$\boldsymbol{P}_i^*(\boldsymbol{n}_i, t) = \sum_{n_h : h \in \bar{\mathcal{K}}_i} \boldsymbol{P}_i^*(\boldsymbol{n}, t),$$

where the summation is over all classes not associated to the i-th M3PP. Using this notation, we prove the main characterization result for the interposed process, which states that the interposition operator does not affect the marginal counting process for the i-th M3PP.

Theorem 1. *Assume all processes to be time-stationary, then*

$$\boldsymbol{P}(\boldsymbol{n}_i, t) = \boldsymbol{P}_i^*(\boldsymbol{n}_i, t), \qquad \forall t \geq 0, \boldsymbol{n}_i \geq \boldsymbol{0}. \tag{22}$$

Proof. For the composed M3PP it is simple to verify that $\boldsymbol{V}_i$ has the following properties

$$\boldsymbol{Q}^* \boldsymbol{V}_i^T = \boldsymbol{V}_i^T \boldsymbol{Q}_i, \qquad \boldsymbol{D}_{1,k}^* \boldsymbol{V}_i^T = \boldsymbol{V}_i^T \boldsymbol{D}_{1,k}^i, \qquad \forall k \in \mathcal{K}_i. \tag{23}$$

Note that conditions of this kind naturally arise in the study of minimal representations of Markovian and Rational Arrival Processes (Buchholz & Telek, 2013).

Let us then consider the Kolmogorov forward equation for $\boldsymbol{P}^*(\boldsymbol{n}, t)$. Pre-multiplying (1) by $\boldsymbol{W}_i$ and post-multiplying by $\boldsymbol{V}_i^T$, we obtain

$$\begin{aligned} \frac{\partial \boldsymbol{P}_i^*(\boldsymbol{n}, t)}{\partial t} =& \boldsymbol{W}_i \boldsymbol{P}^*(\boldsymbol{n}, t) \boldsymbol{D}_0^* \boldsymbol{V}_i^T + \sum_{k \in \mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n} - \boldsymbol{e}_k, t) \boldsymbol{D}_{1,k}^i \\ &+ \sum_{k \in \bar{\mathcal{K}}_i} \boldsymbol{W}_i \boldsymbol{P}^*(\boldsymbol{n} - \boldsymbol{e}_k, t) \boldsymbol{D}_{1,k}^* \boldsymbol{V}_i^T, \end{aligned}$$

where we have used (23) to simplify the expression and the notation $\boldsymbol{n} - \boldsymbol{e}_k$ indicates the removal of a job of class k from $\boldsymbol{n}$. Now plugging the identity $\boldsymbol{D}_0^* = \boldsymbol{Q}^* - \sum_{k=1}^{K} \boldsymbol{D}_{1,k}^*$ and using again (23) we get

$$\begin{aligned} \frac{\partial \boldsymbol{P}_i^*(\boldsymbol{n}, t)}{\partial t} &= \boldsymbol{P}_i^*(\boldsymbol{n}, t)(\boldsymbol{Q}_i - \sum_{k \in \mathcal{K}_i} \boldsymbol{D}_{1,k}^i) + \sum_{k \in \mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n} - \boldsymbol{e}_k, t) \boldsymbol{D}_{1,k}^i \\ &\quad + \sum_{k \in \bar{\mathcal{K}}_i} \boldsymbol{W}_i (\boldsymbol{P}^*(\boldsymbol{n} - \boldsymbol{e}_k, t) - \boldsymbol{P}^*(\boldsymbol{n}, t)) \boldsymbol{D}_{1,k}^* \boldsymbol{V}_i^T \\ &= \boldsymbol{P}_i^*(\boldsymbol{n}, t) \boldsymbol{D}_0^i + \sum_{k \in \mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n} - \boldsymbol{e}_k, t) \boldsymbol{D}_{1,k}^i \\ &\quad + \sum_{k \in \bar{\mathcal{K}}_i} \boldsymbol{W}_i (\boldsymbol{P}^*(\boldsymbol{n} - \boldsymbol{e}_k, t) - \boldsymbol{P}^*(\boldsymbol{n}, t)) \boldsymbol{D}_{1,k}^* \boldsymbol{V}_i^T. \end{aligned}$$

We now consider the marginal probability $\boldsymbol{P}_i^*(\boldsymbol{n}_i, t)$, for which the last expression implies

$$\begin{aligned} \frac{\partial \boldsymbol{P}_i^*(\boldsymbol{n}_i, t)}{\partial t} =& \sum\nolimits_{n_h : h \in \bar{\mathcal{K}}_i} \boldsymbol{P}_i^*(\boldsymbol{n}, t) \boldsymbol{D}_0^i + \sum_{n_h : h \in \bar{\mathcal{K}}_i} \sum_{k \in \mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n} - \boldsymbol{e}_k, t) \boldsymbol{D}_{1,k}^i \\ &+ \sum_{n_h : h \in \bar{\mathcal{K}}_i} \sum_{k \in \bar{\mathcal{K}}_i} \boldsymbol{W}_i (\boldsymbol{P}^*(\boldsymbol{n} - \boldsymbol{e}_k, t) - \boldsymbol{P}^*(\boldsymbol{n}, t)) \boldsymbol{D}_{1,k}^* \boldsymbol{V}_i^T. \end{aligned} \tag{24}$$

Since the infinite summations in the last term of (24) are on the same class indexes ($\bar{\mathcal{K}}_i$ sets), and $\boldsymbol{P}^*(\boldsymbol{n} - \boldsymbol{e}_k, t) = \boldsymbol{0}$ when $n_k = 0$, by symmetry the double summation vanishes. Thus (24) reduces to

$$\frac{\partial \boldsymbol{P}_i^*(\boldsymbol{n}_i, t)}{\partial t} = \boldsymbol{P}_i^*(\boldsymbol{n}_i, t) \boldsymbol{D}_0^i + \sum_{k \in \mathcal{K}_i} \boldsymbol{P}_i^*(\boldsymbol{n}_i - \boldsymbol{e}_k, t) \boldsymbol{D}_{1,k}^i,$$

which is identical to the Kolmogorov forward equation for the counting process of the i-th composed M3PP. In order to prove this statement, we just need to show that the initial conditions of the Kolmogorov forward equations are the same, i.e., $\boldsymbol{P}(\boldsymbol{n}_i, 0) = \boldsymbol{P}_i^*(\boldsymbol{n}_i, 0)$, $\forall \boldsymbol{n}_i$. Since we are considering time-stationary processes, the initial state of the interposed process is determined by the equilibrium distribution of the CTMC with generator $\boldsymbol{Q}^*$. Conversely, for the i-th M3PP this is given by the equilibrium distribution of the CTMC with generator $\boldsymbol{Q}_i$. By (23) we see that for an arbitrary vector $\boldsymbol{\pi}$

$$\boldsymbol{W}_i \boldsymbol{\pi} \boldsymbol{Q}^* \boldsymbol{V}_i^T = \boldsymbol{W}_i \boldsymbol{\pi} \boldsymbol{V}_i^T \boldsymbol{Q}_i = \boldsymbol{\pi}_i \boldsymbol{Q}_i.$$

Therefore, if we choose the initial distribution to be the time-stationary distribution $\boldsymbol{\pi}\boldsymbol{Q}^* = \boldsymbol{0}$, $\boldsymbol{\pi}\boldsymbol{1} = 1$, we readily find that its aggregation corresponds to the time-stationary distribution of the i-th M3PP, i.e., $\boldsymbol{\pi}_i\boldsymbol{Q}_i = \boldsymbol{0}$ and $\boldsymbol{\pi}_i\boldsymbol{1} = 1$, where the last condition holds since we use exact aggregation. This concludes the proof. □

We remark that (23) provides a condition for a *general* MMAP to define a valid interposition. It is simple to see that the diagonal structure of the $\boldsymbol{D}_{1,k}$ matrices in a M3PP satisfies these assumptions. While our analysis does not exclude that other kinds of MMAPs may be used for interposition, the conditions required to satisfy (23) do not seem to readily suggest alternatives other than the M3PP. This motivates the use of M3PPs as a building block for interposition.

4.2.1. Feasibility of an interposition

We now give examples of valid and invalid interposed processes. Consider the following two-state M3PP[2]s:

$$\boldsymbol{A}_0 = \begin{bmatrix} -6 & 3 \\ 1 & -5 \end{bmatrix}, \quad \boldsymbol{B}_0 = \begin{bmatrix} -3 & 1 \\ 4 & -7 \end{bmatrix}, \quad \boldsymbol{C}_0 = \begin{bmatrix} -9 & 7 \\ 2 & -5 \end{bmatrix},$$

$$\boldsymbol{E}_0 = \begin{bmatrix} -7 & 4 \\ 1 & -3 \end{bmatrix},$$

where $\boldsymbol{A}_{1,1} = \mathrm{diag}(1,3)$, $\boldsymbol{A}_{1,2} = \boldsymbol{E}_{1,1} = \mathrm{diag}(2,1)$, $\boldsymbol{B}_{1,1} = \boldsymbol{C}_{1,1} = \mathrm{diag}(1,2)$, and $\boldsymbol{B}_{1,2} = \boldsymbol{C}_{1,2} = \boldsymbol{E}_{1,2} = \mathrm{diag}(1,1)$. It is easy to see that the interposition of $\boldsymbol{A} = (\boldsymbol{A}_0, \boldsymbol{A}_{1,1}, \boldsymbol{A}_{1,2})$ and $\boldsymbol{B} = (\boldsymbol{B}_0, \boldsymbol{B}_{1,1}, \boldsymbol{B}_{1,2})$ satisfies the assumptions on the non-negativity of the rates α_j and β_j and yields the composed 3-state M3PP[4]

$$\boldsymbol{D}_0^* = \begin{bmatrix} -8 & 2 & 1 \\ 1 & -8 & 1 \\ 1 & 3 & -11 \end{bmatrix},$$

with $\boldsymbol{D}_{1,1}^* = \mathrm{diag}(1,3,3)$, $\boldsymbol{D}_{1,2}^* = \mathrm{diag}(2,1,1)$, $\boldsymbol{D}_{1,3}^* = \mathrm{diag}(1,1,2)$, $\boldsymbol{D}_{1,4}^* = \mathrm{diag}(1,1,1)$. The per-class arrival rates and variances of counts in the interposed process match the corresponding statistics of the two classes of $\boldsymbol{A}$ and the two classes of $\boldsymbol{B}$.

Conversely, the interposition of processes $\boldsymbol{A}$ and $\boldsymbol{C} = (\boldsymbol{C}_0, \boldsymbol{C}_{1,1}, \boldsymbol{C}_{1,2})$ is invalid because the non-diagonal rates of $\boldsymbol{C}_0$ are both greater than the corresponding rates in $\boldsymbol{A}_0$. Similarly, the interposition of $\boldsymbol{A}$ and $\boldsymbol{E} = (\boldsymbol{E}_0, \boldsymbol{E}_{1,1}, \boldsymbol{E}_{1,2})$ is invalid, even though $\boldsymbol{B}$ and $\boldsymbol{E}$ are the same process, because their states are ordered in a different way and this affects the computation of the α_j and β_j coefficients. This last example provides intuition on the fact that, to obtain a feasible interposed process, one may need to re-order the states of the M3PPs. In the next section, we describe an algorithm to find a feasible interposition of a given set of M3PPs, if one exists.

4.2.2. Class covariances

While the interposed process preserves the marginal counting properties of the composed M3PPs, when compared to superposition this comes at the expense of introducing spurious covariances between arrivals of different M3PPs. This is because, by definition of the interposed process, a transition in the state space of a composed M3PP also changes the current state of the other composed M3PPs. In order to quantify the magnitude of these covariances, we look at the asymptotic covariances, which contribute to the index of dispersion. Observe that, by plugging (3) into (4), we find after some simplifications

$$\sigma_{k,h} = \lim_{t\to\infty} \mathrm{Cov}[n_k(t), n_h(t)] = -2\mu_k\mu_h + \boldsymbol{\pi}\boldsymbol{D}_{1,h}(\boldsymbol{1\pi} - \boldsymbol{Q})^{-1}\boldsymbol{D}_{1,k}\boldsymbol{1} + \boldsymbol{\pi}\boldsymbol{D}_{1,k}(\boldsymbol{1\pi} - \boldsymbol{Q})^{-1}\boldsymbol{D}_{1,h}\boldsymbol{1}.$$

Let $\boldsymbol{P} = (\boldsymbol{1\pi} - \boldsymbol{Q})^{-1}$ and observe that this by construction is a stochastic matrix with equilibrium vector $\boldsymbol{\pi}$. Using the fact that $\boldsymbol{\pi}\boldsymbol{D}_{1,k} = \mu_k$, for any class k, and after determining the structure of the $\boldsymbol{P}$ matrices for the interposed process, it is possible to compute their Jordan canonical form, which after algebraic simplifications yields the formula

$$\sigma_{k,h} = \frac{r_i' r_j (q_{j,h}\lambda_j - q_{j,h}'\lambda_j')(q_{i,k}\lambda_i - q_{i,k}'\lambda_i')(x_j + x_i)}{x_j^2 x_i^2}, \tag{25}$$

where it is assumed that $k \in \mathcal{K}_i$ and $h \in \mathcal{K}_j$, and $i < j$.

Some remarks on the formulas are as follows:

- When any of the two M3PPs tends to a Poisson process, a pair of departure rates at the numerator of (25) annihilates and $\sigma_{k,h}$ goes to zero.
- The order of the denominator suggests that a way to reduce the covariance introduced by interposition may consist in spending degrees of freedom to maximize x_i and x_j in the embedded MMPPs, for fixed arrival rates. When applying such a scheme, one should however take into account the bounds in Proposition 3, since an increase of the value of x_i also reduces fitting flexibility in the embedded MMPP.
- Since $x_i > 0$ for any i, we note that the sign of the covariance is determined only by the differences between the per-class arrival rates within each of the composed M3PPs.

5. Fitting the interposed process

In this section we consider two issues that arise in compositional fitting based on interposition. First, we consider the decision problem involving the mapping of a marked trace into a set of two-states M3PPs. Then we show that the problem of finding a feasible interposition of a set of J second-order M3PPs, where the i-th M3PP models any subset $\mathcal{K}_i$ of the K classes, can be formulated as a MILP and hence solved using an integer programming solver.

5.1. Fitting a marked trace into a set of M3PPs

Given a trace with arrivals of K classes, if the class arrivals are independent it is possible to fit the trace using a superposition of K independent MMPPs, one for each class, and then reduce the size of the resulting process using interposition. However, if the classes have a significantly large covariance, this method may not produce good results. Therefore, we propose two heuristic fitting methods to address these two cases. We assess the effectiveness of these methods later in Section 6.

5.1.1. Independent Method

In this method, we ignore class covariances and fit each class into a separate two-state MMPP. The method is similar to superposition, but returns a much smaller M3PP[K] with $K+1$ states, instead of 2^K states. Interposition uses the order of composition obtained from the MILP method that we present in Section 5.2.

5.1.2. Covariance-based method

We initially build the asymptotic co-variance matrix $\boldsymbol{\Sigma} = [\sigma_{k,h}] = \lim_{t\to\infty} \mathrm{Cov}[n_k(t), n_h(t)]$ between each pair of classes. We approximate the asymptotic timescale with the largest finite t for which we can average counts over at least 100 samples. The user is then requested to specify a covariance threshold δ. We then iterate on the classes, in decreasing order of variance $\sigma_{k,k}$. For each class k, we first determine all the classes $h \neq k$ with $\sigma_{h,k} \geq \delta$ and record the decision to fit these classes into the same M3PP[K]. The algorithm then continues by analyzing in a similar way the other classes not already planned for fitting in any M3PP. After this stage, each class k is mapped to a unique M3PP i. In order to obtain a representation for each M3PP i, we run the algorithm given in Section 5.2, which returns the parameters of the embedded MMPPs and their order of composition in the interposition.

With these, we can conclude by fitting the M3PPs with the method in Section 3.2.1 and applying the interposition operator.

5.2. Finding a feasible interposition of M3PPs

The definition of the interposed process implies that the $\boldsymbol{D}_{1,k}$ matrices are always feasible by construction. Thus, infeasible M3PPs may arise only if the $\boldsymbol{D}_0$ matrix has some negative elements. Since the entries of the $\boldsymbol{D}_0$ matrix depend only on the MMPPs embedded in the two-state M3PPs, a method to determine a feasible interposition should be run prior to fitting the MMPPs.

To find a feasible interposition, we consider permutations of the mapping of the rates to the states in the embedded MMPP and the order in which the J M3PPs are composed. Furthermore, when deciding the structure of the MMPPs, we exploit the degree of freedom given by ignoring the fitting of the third moment of counts. Let x_i and u_i, be the values of x and u for the i-th embedded MMPP. Sacrificing the fitting of the third moment allows us to decide the value for r_i, given x_i, which then implies $r_i' = x_i - r_i$. Recall from Section 3 that when ignoring the third moment of counts, infinite feasible MMPPs exist as long as $u_i \le r_i < x_i$ and $r_i' = x_i - r_i$, where x_i is found by (6). These bounds have been obtained in Section 3 under the assumption that $\lambda_i > \lambda_i'$. While this assumption does not have any impact on the feasibility of individual M3PPs, it has an impact on the feasibility of the interposition. Making the opposite assumption $\lambda_i \le \lambda_i'$, we obtain the specular constraints $u_i \ge r_i' > x_i$ and $r_i = x_i - r_i'$. Thus, the above bounds will need to be simultaneously considered and only one will be active depending on the relative value of λ_i to λ_i', which we decide using a binary variable b_i that is responsible for state order. Summarizing, the above bounds allow us to obtain a feasible interposition by choosing from a continuous set of feasible MMPPs.

Before solving the MILP formulation that decides the order of states and the order of composition, we assume that the fixed point Eq. (6) has been solved for each M3PP i for the values of x_i and u_i. For the MILP formulation, we then consider the following decision variables:

- the rates $r_i > 0$ and $r_i' > 0$ of the i-th M3PP, $\forall i \in \{1, ..., J\}$;
- the integer variables $p_i \in \mathbb{N}$, $\forall i \in \{1, ..., J\}$, deciding the position of the i-th M3PP in the interposed process, i.e., the order of composition.
- the auxiliary binary variables $z_{i,j} \in \{0, 1\}$, $\forall (i, j) \in \{1, ..., J\}^2$, $i \neq j$, such that $z_{i,j} = 1$ if and only if $p_i \le p_j$;
- the binary variables $b_i \in \{0, 1\}$, $\forall i \in \{1, ..., J\}$, encoding the choice of the state order for the i-M3PP, with $b_i = 0$ if $\lambda_i > \lambda_i'$ and $b_i = 1$ if $\lambda_i < \lambda_i'$.

Overall, the number of variables is $\mathcal{O}(J^2)$, i.e., quadratic in the number of composed M3PPs.

Let M be a large constant. We consider the following MILP formulation:

$$\text{minimize} \quad 0 \text{ (feasibility problem)} \tag{26}$$

$$\text{s.t.:} \quad u_i(1-b_i) \le r_i < x_i - u_i b_i, \qquad \forall i \in \{1, \dots, J\}, \tag{27}$$

$$u_i b_i \le r_i' < x_i - u_i(1-b_i), \qquad \forall i \in \{1, \dots, J\}, \tag{28}$$

$$z_{i,j} + z_{j,i} = 1, \qquad \forall (i,j) \in \{1, \dots, J\}^2 : i \neq j, \tag{29}$$

$$z_{i,j} \ge p_j - p_i, \qquad \forall (i,j) \in \{1, \dots, J\}^2, \tag{30}$$

$$r_i \ge r_j - (1 - z_{i,j})M, \qquad \forall (i,j) \in \{1, \dots, J\}^2, i \neq j, \tag{31}$$

$$r_i' \le r_j' + (1 - z_{i,j})M, \qquad \forall (i,j) \in \{1, \dots, J\}^2, i \neq j, \tag{32}$$

$$p_i \neq p_j, \qquad \forall (i,j) \in \{1, \dots, J\}^2 : i \neq j. \tag{33}$$

Strict bounds are obtained by adding small tolerance to the corresponding formulation with $\le$ inequalities. Observe that the objective function is a constant, since we just need a feasible solution. Constraint (27) imposes the bounds on r_i, with different bounds depending on whether the state order of the i-M3PP is inverted or not. Constraint (28) imposes specular bounds on r_i'. Antisymmetry constraints on the auxiliary variables $z_{i,j}$ are set in (29). Constraint (30) ensures that $z_{i,j} = 1$ if and only if $p_i \le p_j$. This constraint together with the uniqueness constraint on p_i ensures the transitivity property for the binary variables $z_{i,j}$. Ordering constraints on r_i and r_i' are expressed by (31) and (32), respectively. Two M3PP processes cannot be in the same position due to (33); note that inequality constraints can be handled for integer variables in MILP.

After solving the MILP, we have the parameters r_i and r_i' of each M3PP. Applying the fitting formulas of Section 3 we obtain the remaining parameters of the embedded MMPP, λ_i and λ_i', taking care to swap the two values when states are in inverted order, i.e., $b_i = 1$. Then we compute the class parameters $q_{i,k}$ and $q_{i,k}'$ for each class of the i-th M3PP and build the interposition of the J M3PPs as in Section 4.

6. Fitting results

6.1. Fitting random MMAPs

We begin by examining the applicability of two-state M3PPs and the interposition process in fitting the characteristics of randomly generated processes. We consider random M3PP[K]s, with $m \in \{4, 8, 16\}$ states and $K \in \{2, 3, 4\}$ classes. For each choice of m and K, we generate 100 random M3PP[K] as follows. First, we generate a random infinitesimal generator $\boldsymbol{Q} = \boldsymbol{D}_0 + \boldsymbol{D}_1$ using uniform random numbers. Then, for each class k, we first compute a vector of m uniform random numbers $\boldsymbol{u}$ and set $\boldsymbol{D}_{1,k} = \exp(\text{diag}(5\boldsymbol{u}))$. Lastly, we set $\boldsymbol{D}_1 = \sum_k \boldsymbol{D}_{1,k}$ and $\boldsymbol{D}_0 = \boldsymbol{Q} - \boldsymbol{D}_1$. The expression used to compute $\boldsymbol{D}_{1,k}$ provides a set of processes with index of dispersion I that is 3–6 times larger than the squared coefficient of variation (SCV), as shown in Table 1, which gives statistics for the randomly generated M3PPs. The statistics in each row are averaged across the 100 random instances and we report mean and standard deviation. The third columns gives the average ratio

Table 1
Statistics for randomly generated M3PPs. Each row refers to 100 random instances. Mean μ and standard deviation σ are reported as (μ, σ).

m	K	Cov percent	SCV	$I(\infty)$
4	2	0.29	(5.25, 4.17)	(32.06, 35.47)
8	2	0.31	(5.47, 2.54)	(21.57, 11.99)
16	2	0.30	(4.55, 1.15)	(11.43, 36.50)
4	4	0.57	(5.25, 3.90)	(27.27, 22.12)
8	4	0.57	(5.28, 2.22)	(19.93, 9.13)
16	4	0.56	(4.40, 1.07)	(11.94, 3.42)

$$\text{Cov percent} = \lim_{t\to\infty} \frac{\sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K} |\text{Cov}[n_k(t), n_{k'}(t)]|}{\sum_{c=1}^{K}\sum_{c'=1}^{K} |\text{Cov}[n_c(t), n_{c'}(t)]|}, \tag{34}$$

which quantifies the relative magnitude of cross-covariances.

M3PPs are fitted as follows. We first compute the statistical descriptors of the random process using theoretical expressions. Two-state M3PPs are then obtained using the method described in Section 3.2.1. The interposition process is fitted using the two methods described in Section 5.1 and the feasibility method in Section 5.2. The latter is implemented in MATLAB using

YALMIP (Löfberg, 2004) and the CBC branch-and-cut solver (Coin-or branch & cut project). Superposition is implemented as by the definition after fitting an independent MMPP for each class. Here and in the following sections, counts used to fit a class are obtained from inter-arrival times between successive arrivals of that same class.

Let $n_k(t)$ be the number of arrivals of class k in t time units for the randomly-generated M3PP and let $\tilde{n}_k(t)$ be the same descriptor in the fitted model. The metrics used to assess the quality of the fitting are the ability of the model to capture the asymptotic class variances and covariances, as quantified by the absolute relative errors

$$\epsilon_{var} = \lim_{t\to\infty} \frac{\sum_{k=1}^{K} |\mathrm{Var}[n_k(t)] - \mathrm{Var}[\tilde{n}_k(t)]|}{\sum_{c=1}^{K} \mathrm{Var}[n_c(t)]}, \tag{35}$$

$$\epsilon_{cov} = \lim_{t\to\infty} \frac{\sum_{k=1}^{K} \sum_{k'=1,k'\neq k}^{K} |\mathrm{Cov}[n_k(t), n_{k'}(t)] - \mathrm{Cov}[\tilde{n}_k(t), \tilde{n}_{k'}(t)]|}{\sum_{c=1}^{K} \sum_{c'=1,c'\neq c}^{K} |\mathrm{Cov}[n_c(t), n_{c'}(t)]|}. \tag{36}$$

We focus on second order class descriptors of the counting process since mean arrival rates can always be fitted exactly. We also quantify the absolute relative error on the variance of the underlying unmarked process

$$\epsilon_{unmark} = \lim_{t\to\infty} \frac{\sum_{k=1}^{K} \sum_{k'=1}^{K} |\mathrm{Cov}[n_k(t), n_{k'}(t)] - \mathrm{Cov}[\tilde{n}_k(t), \tilde{n}_{k'}(t)]|}{\sum_{c=1}^{K} \sum_{c'=1}^{K} |\mathrm{Cov}[n_c(t), n_{c'}(t)]|}. \tag{37}$$

For randomly-generated models, the values of ϵ_{var}, ϵ_{cov} and ϵ_{unmark} are averaged across all the models. The corresponding standard deviations are indicated with σ_{var}, σ_{cov} and σ_{unmark}, respectively. Lastly, we assess the total number of states m^{fit} of the fitted model and the time T required for the fitting algorithm to return a valid M3PP. In applying the M3PP fitting method of Section 3.2.1, we choose $t = 10E[X]$ as arbitrary timescale, being $E[X]$ the mean of the random M3PP[K], and we approximate asymptotic values using the timescale $t_\infty = 10^4 E[X]$. We have also repeated the experiments with $t = E[X]$ and $t = 100E[X]$, but the results closely resemble the ones reported in this section.

6.1.1. Results

Results of the validation against random M3PPs are given in Table 2. Remarks are as follows:

- The results indicate that interposition has the same efficiency of superposition in capturing class variances, whereas two-state M3PPs are better at capturing class covariances. This aligns with expectations, given that interposition matches marginal counting processes, whereas the method presented in Section 3.2.1 for two-state M3PPs matches asymptotic covariances.
- It is fairly surprising to note the good performance of two-state M3PPs in fitting even large processes, with several states and classes. However, we conjecture this to be because the arrivals have the same inter-arrival time distribution of the embedded MMPP. In fact, we will show that when this assumption is removed in the fitting of real traces, two-state M3PPs perform worse.
- Interposition is on most cases superior to superposition. The number of states is significantly lower, without appreciable loss of accuracy in matching variances. It should be noted that, while in some cases interposition performs worse than superposition in matching covariances, superposition always return zero covariance, thus it is uninformative. Conversely, the error on covariance of interposition appears to decrease with growing number of states and classes.
- The covariance-based fitting method used for interposition typically fares better than the interposition of independent MMPPs. The approach has similar accuracy for variance matching, it is generally more accurate in describing covariances, and requires less states.
- Computational times are small for all methods. However, there is an appreciable difference in the time to compute an interposition for large processes, due to the cost of solving the integer program introduced in Section 5.2. Still, it should be noted that when the number of classes grows beyond $K = 4$, superposition requires tens or even hundreds of states, thus it becomes far less tractable than interposition.

6.2. Fitting real traces

The analysis for random MMAPs is now repeated for fitting the BC-pAug89 and LBL-TCP-3 traces from the Internet Traffic Archive[1]. While these traces are commonly used in the literature of Markov-modulated processes, they are unmarked and thus cannot be readily used for validation of marked processes. In order to do so, we introduce a labelling that associates each arrival to a different bin of the histogram of packet sizes. We first consider the case where arrivals belong only to one of $K = 2$ classes, putting the separator at the p-th percentile of the inter-arrival time distribution, where $p \in \{25, 50, 75, 90\}$. In addition, we consider a separation of the histogram into five bins, defined by the same set of percentiles, which leads to a model with $K = 5$ classes.

Results are given in Tables 3 and 4. The trends in the two tables are qualitatively similar and indicate that interposition can in most cases fit better the traces than a 2-state model or superposition. There are also several instances on which the composed M3PPs have better ϵ_{cov} than the 2-state M3PP. This suggests that the good results on random instances for 2-state models may be biased by the fact that inter-arrival times of different classes are identically distributed in the random instances. This does not happen with the real traces, where the histogram bins do not overlap with each other.

It can also be noted that, as the number of classes grows to $K = 5$, the performance of the different methods get closer to each other, presumably due to the increased difficulty in matching class covariances.

6.2.1. Queuing analysis

Lastly, we consider a queueing analysis application. We use again the LBL-TCP-3 trace, but with the class marking defined in Buchholz et al. (2010). This marking defines 4 classes, each associated to a different bin of the histogram of packet sizes, and having service rates $\mu_1 = 300$, $\mu_2 = 250$, $\mu_3 = 200$, $\mu_4 = 100$. We simulate a queue with exponentially distributed service times, infinite buffer capacity, first-come first-served scheduling, and arrivals fed by the LBL-TCP-3 trace. In the simulations, we use 10^8 samples and record mean queue-lengths for each class. The results are compared with the predictions obtained from a MMAP/M/1 first-come first-served queue fed by a MMAP obtained by the following fitting methods:

- 2-state: the fitting of a M3PP[4] (2 states).
- superpos: superposition of 4 MMPPs, one per class (16 states).
- interpos-indep: interposition of the independent MMPPs used for the superposition, one per class (5 states).
- interpos-cov: interposed process obtained by the covariance-based method, which interposes 3 M3PP[2], the first for classes 1 and 3, which have the largest covariances, the second for class 2, the third for class 4 (4 states).

[1] http://ita.ee.lbl.gov/

Table 2
Fitting results for randomly generated M3PPs ($t = 10E[X]$).

m	K	*Method*	ϵ_{unmark}	ϵ_{var}	ϵ_{cov}	σ_{unmark}	σ_{var}	σ_{cov}	m^{fit}	Time [seconds]
4	2	2-state	0.184	0.118	0.972	0.251	0.145	0.000	2.00	0.181
4	2	superpos	0.296	0.002	1.000	0.146	0.001	0.000	4.00	0.171
4	2	interpos-independent	0.218	0.002	2.103	0.220	0.001	0.000	3.00	1.722
4	2	interpos-covariance	0.197	0.048	1.870	0.212	0.084	0.500	2.55	1.533
8	2	2-state	0.197	0.129	0.805	0.204	0.116	0.000	2.00	0.183
8	2	superpos	0.310	0.001	1.000	0.134	0.000	0.000	4.00	0.177
8	2	interpos-independent	0.212	0.001	1.464	0.231	0.000	0.000	3.00	1.825
8	2	interpos-covariance	0.188	0.022	1.321	0.197	0.047	0.453	2.72	1.681
16	2	2-state	0.227	0.153	0.600	0.168	0.094	0.000	2.00	0.175
16	2	superpos	0.300	0.000	1.000	0.094	0.000	0.000	4.00	0.184
16	2	interpos-independent	0.202	0.000	1.024	0.160	0.000	0.000	3.00	1.814
16	2	interpos-covariance	0.197	0.018	0.858	0.135	0.062	0.327	2.88	1.705
4	4	2-state	0.160	0.117	0.216	0.204	0.133	0.000	2.00	0.187
4	4	superpos	0.573	0.002	1.000	0.105	0.001	0.000	16.00	0.333
4	4	interpos-independent	0.409	0.002	0.755	0.261	0.001	0.000	5.00	4.439
4	4	interpos-covariance	0.280	0.074	0.491	0.237	0.098	1.067	3.64	2.788
8	4	2-state	0.180	0.147	0.223	0.160	0.098	0.000	2.00	0.200
8	4	superpos	0.574	0.001	1.000	0.075	0.000	0.000	16.00	0.361
8	4	interpos-independent	0.306	0.001	0.569	0.210	0.000	0.000	5.00	5.157
8	4	interpos-covariance	0.275	0.042	0.477	0.179	0.056	0.756	4.02	3.700
16	4	2-state	0.222	0.201	0.246	0.114	0.081	0.000	2.00	0.175
16	4	superpos	0.560	0.000	1.000	0.059	0.000	0.000	16.00	0.331
16	4	interpos-independent	0.320	0.000	0.594	0.158	0.000	0.000	5.00	4.978
16	4	interpos-covariance	0.307	0.019	0.557	0.158	0.044	0.514	4.72	4.570

Table 3
Fitting results for the BC-pAug89 trace for different percentile cut-points of the packet size histogram.

Trace	K	*Cut-points p*	*Method*	ϵ_{unmark}	ϵ_{var}	ϵ_{cov}	m	Time [seconds]
BC-pAug89	2	25	2-state	0.718	0.635	0.984	2	1.159
BC-pAug89	2	25	superpos	0.277	0.052	1.000	4	0.922
BC-pAug89	2	25	interpos-independent	0.278	0.052	1.005	3	1.436
BC-pAug89	2	25	interpos-covariance	0.163	0.052	0.519	3	1.489
BC-pAug89	2	50	2-state	0.358	0.357	0.360	2	1.263
BC-pAug89	2	50	superpos	0.377	0.048	1.000	4	1.156
BC-pAug89	2	50	interpos-independent	0.070	0.048	0.113	3	2.564
BC-pAug89	2	50	interpos-covariance	0.070	0.047	0.112	3	2.064
BC-pAug89	2	75	2-state	0.257	0.186	0.646	2	0.880
BC-pAug89	2	75	superpos	0.195	0.047	1.000	4	0.699
BC-pAug89	2	75	interpos-independent	0.083	0.047	0.278	3	1.435
BC-pAug89	2	75	interpos-covariance	0.082	0.046	0.277	3	1.642
BC-pAug89	2	90	2-state	0.257	0.186	0.646	2	0.963
BC-pAug89	2	90	superpos	0.195	0.047	1.000	4	0.777
BC-pAug89	2	90	interpos-independent	0.195	0.047	1.001	3	1.550
BC-pAug89	2	90	interpos-covariance	0.195	0.046	1.001	3	1.425
BC-pAug89	5	25, 50, 75, 90	2-state	0.382	0.399	0.363	2	1.342
BC-pAug89	5	25, 50, 75, 90	superpos	0.487	0.045	1.000	16	1.223
BC-pAug89	5	25, 50, 75, 90	interpos-independent	0.313	0.045	0.624	5	2.363
BC-pAug89	5	25, 50, 75, 90	interpos-covariance	0.680	0.045	1.416	5	2.380

We have observed that different choices of the timescales used for fitting have a quite visible effect on the rate at which mean queue-lengths build up. Therefore, for complex traces such as LBL-TCP-3, we recommend to perform some calibration tests in order to find the best assignment of the arbitrary timescale $t = t_1 = t_2$ used in M3PP fitting. For this trace we perform calibration by varying t in {0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100} seconds and approximating the asymptotic timescale in each experiment as $t_\infty = 10t$. Then we select the optimal timescale as the value for which superposition predicts the most accurate aggregated mean queue-length at 90 percent utilization. Using this calibration, we settle on $t = 50$ seconds and $t_\infty = 500$ seconds.

The resulting MMAP/M/1 queueing systems are analyzed using Q-MAM (Bini et al., 2012). Results are illustrated in Fig. 1. The simulation results indicate that mean queue-lengths grow exponentially, with classes 1–3 growing up to a few thousand jobs, whereas class 4 grows up to 86 jobs. The trends indicate that are all methods are accurate in heavy-load, above 60 percent utilization. Interposed processes are the most effective in capturing the job growths for the dominating classes 1–3. In heavy load interpos-cov is more accurate than interpos-indep, e.g., the mean class-2 queue length at 90 percent utilization predicted by interpos-cov is 3442 jobs, against a simulated value of 3881 jobs, whereas interpos-indep predicts 2196 jobs and superposition predicts 2024 jobs; the 2-state M3PP is also accurate in heavy load, with a prediction of 3416 jobs at 90 percent utilization, but this method is significantly worse than the other methods at lower loads.

Summarizing, by comparing against superposition as a baseline, the interposition method performs best in fitting real-world traces, delivering a more accurate estimate in loads that exceed 60 percent utilization. Combined with the previous validations, this outcome suggests that the interposition methods introduced in this paper can be helpful for real-world fitting and queueing analyses.

Table 4
Fitting results for the LBL-TCP-3 trace for different percentile cut-points of the packet size histogram.

Trace	*K*	*Cut-points p*	*Method*	ϵ_{unmark}	ϵ_{var}	ϵ_{cov}	*m*	Time [seconds]
LBL-TCP-3	2	25	2-state	0.468	0.661	0.185	2	1.768
LBL-TCP-3	2	25	superpos	0.426	0.034	1.000	4	1.600
LBL-TCP-3	2	25	interpos-independent	0.709	0.034	1.699	3	1.826
LBL-TCP-3	2	25	interpos-covariance	0.064	0.085	0.034	2	2.502
LBL-TCP-3	2	50	2-state	0.454	0.357	0.700	2	1.647
LBL-TCP-3	2	50	superpos	0.313	0.043	1.000	4	1.516
LBL-TCP-3	2	50	interpos-independent	0.032	0.042	0.005	3	1.765
LBL-TCP-3	2	50	interpos-covariance	0.324	0.060	0.995	3	1.845
LBL-TCP-3	2	75	2-state	0.563	0.383	1.207	2	1.661
LBL-TCP-3	2	75	superpos	0.252	0.043	1.000	4	1.428
LBL-TCP-3	2	75	interpos-independent	0.061	0.043	0.125	3	1.925
LBL-TCP-3	2	75	interpos-covariance	0.308	0.118	0.989	3	2.053
LBL-TCP-3	2	90	2-state	0.175	0.107	8.916	2	2.090
LBL-TCP-3	2	90	superpos	0.045	0.037	1.000	4	1.528
LBL-TCP-3	2	90	interpos-independent	0.045	0.037	1.009	3	2.200
LBL-TCP-3	2	90	interpos-covariance	0.045	0.037	1.009	3	2.650
LBL-TCP-3	5	25, 50, 75, 90	2-state	0.660	0.493	0.858	2	2.646
LBL-TCP-3	5	25, 50, 75, 90	superpos	0.482	0.047	1.000	32	2.484
LBL-TCP-3	5	25, 50, 75, 90	interpos-independent	0.663	0.047	1.397	6	3.690
LBL-TCP-3	5	25, 50, 75, 90	interpos-covariance	0.611	0.047	1.284	6	4.571

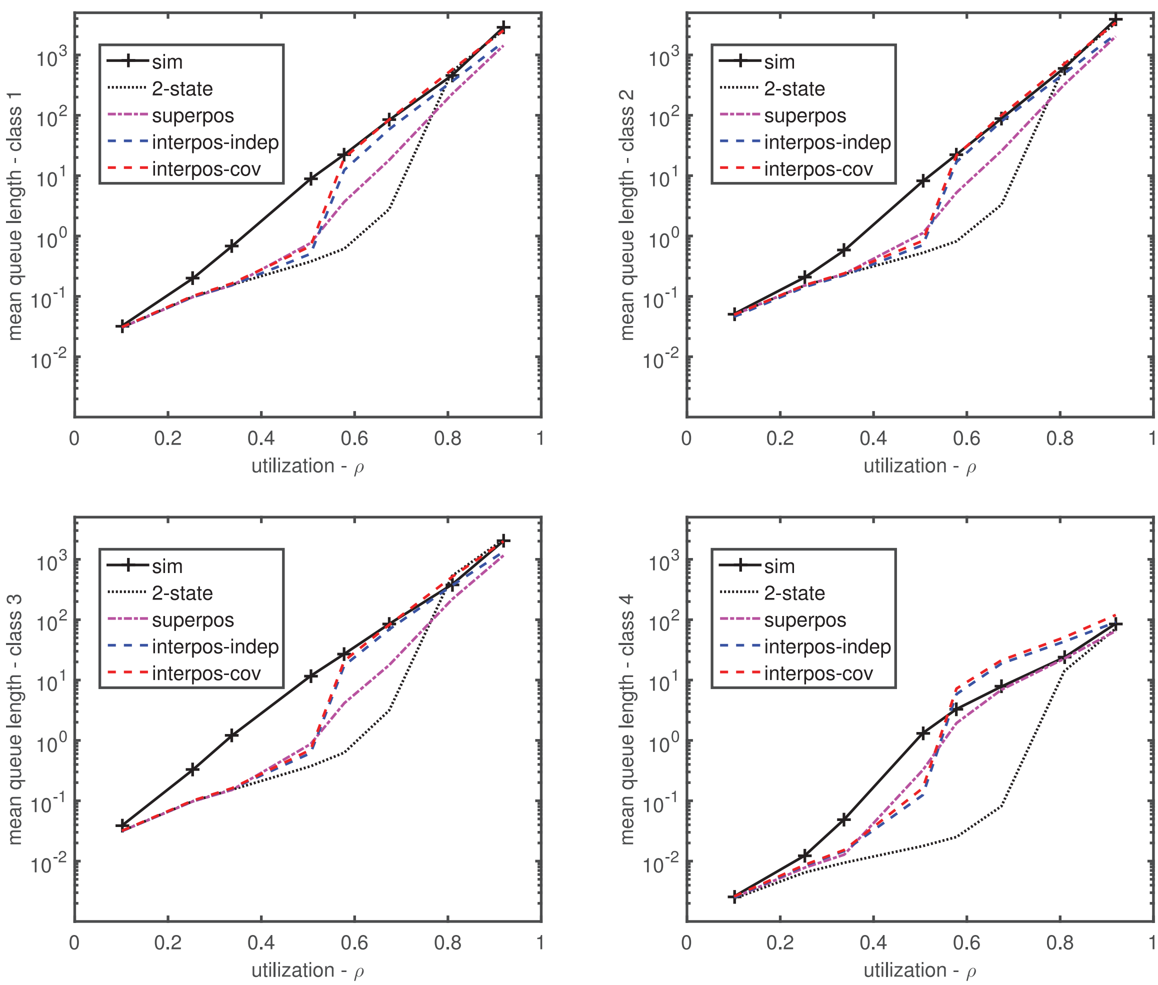

Fig. 1. LBL-TCP-3 trace – queueing analysis results.

7. Conclusion

In this paper, we have presented novel methods to fit multiclass arrival processes using marked Markov-modulated Poisson processes (M3PPs). We have defined exact and approximate algorithms to fit two-state M3PPs with arbitrary number of classes and introduced a new composition operator, called interposition, which enables composing several M3PPs while preserving their marginal counting processes. The state space of the interposed process grows linearly in the total number of composed M3PPs, instead than exponentially as in ordinary superposition. Experiments reveal that the interposed process can be effective in fitting real-world traces, in matching the descriptors of more complex marked processes, and in queueing analysis.

Future work should also investigate how interposition and exact aggregations may be fruitfully applied to MMPPs and the general class of MMAPs. Also, it would be interesting to apply M3PPs to decomposition analysis of multiclass queueing networks, for example by iterative matching of the departure flows of the marked requests from queues.

Acknowledgment

The authors wish to thank the anonymous referees for their comments that greatly helped in improving this manuscript. The work of G. Casale has received funding from the Engineering and Physical Sciences Research Council (EPSRC) under Grant agreement no. EP/M009211/1 (OptiMAM) and from the European Union under grant agreement H2020-644869 (DICE). The dataset associated to this paper is available at https://zenodo.org/record/55833.

References

Andersen, A. T., & Nielsen, B. F. (1998). A Markovian approach for modeling packet traffic with long-range dependence. *IEEE Journal on Selected Areas in Communications, 16*(5), 719–732.

Bini, D., Meini, B., Steffé, S., Pérez, J. F., & Houdt, B. V. (2012). SMCSolver and Q–MAM: Tools for matrix-analytic methods. *ACM Performance Evaluation Review, 39*(4), 46.

Bodrog, L., Heindl, A., Horváth, G., & Telek, M. (2008). A Markovian canonical form of second-order matrix-exponential processes. *European Journal of Operations Research, 190*(2), 459–477.

Bolch, G., Greiner, S., de Meer, H., & Trivedi, K. S. (1998). *Queueing networks and Markov chains*. John Wiley and Sons.

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.

Brewer, J. W. (1978). Kronecker products and matrix calculus in system theory. *IEEE Transactions on Circuits and Systems, 25*(9), 772–781.

Buchholz, P. (1994). Exact and ordinary lumpability in finite Markov chains. *Journal of Applied Probability, 31*(1), 59–75.

Buchholz, P., Kemper, P., & Kriege, J. (2010). Multi-class Markovian arrival processes and their parameter fitting. *Performance Evaluation, 67*(11), 1092–1106. Elsevier

Buchholz, P., & Telek, M. (2013). On minimal representations of rational arrival processes. *Annals of Operations Research, 202*(1), 35–58.

Casale, G., Zhang, E. Z., & Smirni, E. (2010). Trace data characterization and fitting for Markov modeling. *Performance Evaluation, 67*(2), 61–79.

Coin-or branch and cut project:. https://projects.coin-or.org/Cbc. Accessed 17.06.16.

Fischer, W., & Meier-Hellstern, K. S. (1993). The Markov-modulated poisson process (MMPP) cookbook. *Performance Evaluation, 18*(2), 149–171.

He, Q.-M., & Neuts, M. F. (1998). Markov chains with marked transitions. *Stochastic Processes and their Applications, 74*(1), 37–52.

He, Q.-M. (2001). The versatility of MMAP[K] and the MMAP[K]/G[K]/1 queue. *Queueing Systems: Theory and Applications, 38*(4), 397–418.

Heffes, H., & Lucantoni, D. (1986). A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications, 4*(6), 856–868.

Heindl, A., Horváth, G., & Gross, K. (2006). Explicit inverse characterizations of acyclic MAPs of second order. In *Proceedings of European performance engineering workshop, EPEW: 4054* (pp. 108–122). Lecture notes in computer science, Springer.

Horváth, A., Horváth, G., & Telek, M. (2009). A traffic based decomposition of two-class queueing networks with priority service. *Computer Networks, 53*(8), 1235–1248.

Horváth, A., & Telek, M. (2002). Markovian modeling of real data traffic: Heuristic phase type and MAP fitting of heavy tailed and fractal like samples*: 2459*. *Performance evaluation of complex systems: Techniques and tools* (pp. 405–434). Lecture notes in computer science, Springer.

Horváth, G., & Okamura, H. (2013). A fast EM algorithm for fitting marked Markovian arrival processes with a new special structure. In *Proceedings of European performance engineering workshop, EPEW: 8168* (pp. 119–133). Lecture notes in computer science, Springer.

Horváth, G., & Telek, M. (2009). On the canonical representation of phase type distributions. *Performance Evaluation, 66*(8), 396–409.

Horváth, G. (2012). Efficient analysis of the queue length moments of the MMAP/MAP/1 preemptive priority queue. *Performance Evaluation, 69*(12), 684–700.

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

Klemm, A., Lindemann, C., & Lohmann, M. (2003). Modeling IP traffic using the batch Markovian arrival process. *Performance Evaluation, 54*(2), 149–173.

Löfberg, J. (2004). Yalmip : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD conference, Taipei, Taiwan*.

Li, H., Muskulus, M., & Wolters, L. (2006). Modeling job arrivals in a data-intensive grid. In *Proceedings of Job Scheduling Strategies for Parallel Processing, JSSPP* (pp. 210–231).

Mi, N., Zhang, Q., Riska, A., Smirni, E., & Riedel, E. (2007). Performance impacts of autocorrelated flows in multi-tiered systems. *Performance Evaluation, 64*(9–12), 1082–1101.

Neuts, M. F. (1979). A versatile markovian point process. *Journal of Applied Probability, 16*(4), 764–779.

Okamura, H., Dohi, T., & Trivedi, K. S. (2009). Markovian arrival process parameter estimation with group data. *IEEE/ACM Transactions on Networking, 17*(4), 1326–1339.

Perez-Palacin, D., Merseguer, J., & Mirandola, R. (2012). Analysis of bursty workload-aware self-adaptive systems. In *Proceedings of International conference on performance engineering, ICPE* (pp. 75–84). ACM.

Pérez, J. F., Velthoven, J. V., & Houdt, B. V. (2008). Q-MAM: A tool for solving infinite queues using matrix-analytic methods. In *Proceedings of ICST conference on valuetools* (pp. 16:1–16:9).

Sriram, K., & Whitt, W. (1986). Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications, 4*(6), 833–846.

Verma, A., & Anand, A. (2007). General store placement for response time minimization in parallel disks. *Journal of Parallel and Distributed Computing, 67*(12), 1286–1300.