Bayesian Analysis (2022) **17**, Number 3, pp. 765–790

On a Dirichlet Process Mixture Representation of Phase-Type Distributions*

Daniel Ayala†, Leonardo Jofré‡, Luis Gutiérrez§, and Ramsés H. Mena¶

Abstract. An explicit representation of phase-type distributions as an infinite mixture of Erlang distributions is introduced. The representation unveils a novel and useful connection between a class of Bayesian nonparametric mixture models and phase-type distributions. In particular, this sheds some light on two hot topics, estimation techniques for phase-type distributions, and the availability of closed-form expressions for some functionals related to Dirichlet process mixture models. The power of this connection is illustrated via a posterior inference algorithm to estimate phase-type distributions, avoiding some difficulties with the simulation of latent Markov jump processes, commonly encountered in phase-type Bayesian inference. On the other hand, closed-form expressions for functionals of Dirichlet process mixture models are illustrated with density and renewal function estimation, related to the optimal salmon weight distribution of an aquaculture study.

Keywords: Bayesian nonparametrics, Erlang distribution, mixture model, renewal function.

MSC2020 subject classifications: Primary 62G05, 62M05; secondary 60J28.

1 Introduction

Mixture models are ubiquitous in statistics. Their study can be traced back to Pearson (1894) with excellent, up to date, reviews by Titterington et al. (1985), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006). A mixture model can be written as

$$f_Y(\cdot) = \sum_{h=1}^{N} w_h K(\cdot|\theta_h), \tag{1}$$

with $N = 1, \ldots, \infty$, where $K(\cdot|\theta)$ is a kernel density, here supported in $\mathbb{R}_+$, and $\{\theta_h, w_h\}_{h\geq 1}$ are N-dimensional parameters satisfying $\sum_{h=1}^{N} w_h = 1$. Depending on the kernel and weights specification, this model might induce a dense class of densities,

*The work of the first author was supported by "Becas Doctorado Nacional CONICYT 2017 Folio No. 21171601". The work of the third author was supported by "Proyecto REDES ETAPA INICIAL Convocatoria 2017 REDI170094" and by ANID–Millennium Science Initiative Program–NCN17_059. The fourth author acknowledges the support of CONTEX project 2018-9B.

†Departamento de Estadística, Pontificia Universidad Católica de Chile.

‡Departamento de Estadística, Pontificia Universidad Católica de Chile.

§Departamento de Estadística, Pontificia Universidad Católica de Chile. ANID–Millennium Science Initiative Program–Millennium Nucleus Center for the Discovery of Structures in Complex Data, llgutier@mat.uc.cl

¶IIMAS-UNAM, México

 https://doi.org/10.1214/21-BA1272

namely it could capture any density on $\mathbb{R}_+$. When $N = \infty$ and the parameters are random, mixture models (1) are widely studied in Bayesian nonparametrics (see, e.g., Ghosh and Ramamoorthi, 2003; Müller and Quintana, 2004; Dunson, 2010; Müller and Mitra, 2013), with the benchmark model being the celebrated Dirichlet process mixture (DPM) model (Ferguson, 1973, 1974; Lo, 1984; Escobar and West, 1995). The DPM model can be defined as the random density model

$$f_P(y) = \int_{\mathbb{R}_+} K(y \mid \xi)\, P(d\xi), \tag{2}$$

driven by a Dirichlet Process (DP), $P \sim \mathrm{DP}(\alpha, P_0)$, *i.e.* a random probability measure $P = \sum_{h\geq 0} w_h \delta_{\theta_h}$, where weights, $\{w_h\}_{h\geq 1}$, and locations, $\{\theta_h\}_{h\geq 1}$, are random and independent, given by

$$w_h = v_h \prod_{\ell<h}(1 - v_\ell), \quad v_h \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha), \quad \alpha > 0, \tag{3}$$

and $\theta_h \overset{\text{iid}}{\sim} P_0$, respectively. Here, P_0, sometimes referred as the baseline distribution, is assumed to be a non-atomic distribution on $\mathbb{R}_+$. The Bayesian nonparametric literature offers a wide choice of other models for P, being of practical interest those falling in the general class of species sampling models (Ghosal and van der Vaart, 2017). Among other aspects, the Dirichlet process stands out within this latter general class as being the only one tractable for atomic P_0. Though this discreteness of P_0 could be potentially relevant for our purposes below, letting P_0 to be atomic is not always adequate, as it prevents posterior distributions to smoothly deviate from the prior. Indeed, this has encouraged other proposals in the literature (Canale and Dunson, 2011; Canale and Prünster, 2017) to overcome the difficulty of modeling random mass probability functions. Hence, we keep the assumption of non-atomic P_0. Notice that random density (2) can be also simplified as

$$f_P(y) = \sum_{h=1}^{\infty} w_h\, K(y \mid \theta_h), \tag{4}$$

and, when describing a set of iid observations $\{y_1, \ldots, y_n\}$ from it, sometimes written in the hierarchical representation form

$$\begin{aligned} y_k \mid \theta_k &\overset{\text{ind}}{\sim} K(y \mid \theta_k)\,, \qquad k = 1, 2, \ldots, n, \\ \theta_k \mid P &\overset{\text{iid}}{\sim} P, \\ P &\sim \mathrm{DP}(\alpha, P_0)\,. \end{aligned} \tag{5}$$

On an unseemly connected direction, phase-type distributions (Neuts, 1975, 1978) (sometimes abbreviated as PH-distributions) have been mainly studied in the applied probability literature, see, e.g., Bladt and Nielsen (2017) for extensive treatment. The basic idea of phase-type distributions starts by considering a Markov jump process $\{X_t\}_{t\geq 0}$ with state space $E = \{1, 2, \ldots, p, p+1\}$, where states $1, 2, \ldots, p$ are transient, and state $p+1$ is absorbing. This process is driven by an intensity matrix of the form

$$\boldsymbol{\Lambda} = \begin{pmatrix} \mathbf{T} & \boldsymbol{t} \\ \mathbf{0} & 0 \end{pmatrix},$$

where $\mathbf{T} \in \mathbb{T}$, with $\mathbb{T}$ denoting the space of subintensity matrices of dimension $p \times p$. Given the row elements in an intensity matrix add up to 0 (i.e. $\mathbf{\Lambda 1} = \mathbf{0}$), the space $\mathbb{T}$ contains all the square matrices whose row sums are non-positive, and contains negative elements in the main diagonal, that is,

$$T(i,j) = \begin{cases} -\lambda_i & \text{if } i = j \\ \lambda_{ij} & \text{if } i \neq j \end{cases},$$

where $\lambda_i > 0$ corresponds to the parameter of an exponential distribution. This amounts to say that the process remains in state i, an exponential time with rate λ_i, and then jumps to state j with transition probability $p(i,j) = P(X_{n+1} = j \mid X_n = i) = \lambda_{ij}/\lambda_i$. Notice that λ_{ij} is the rate at which transition from i to j occurs. Given the finite nature of this Markov chain, transition probabilities are typically represented with a matrix $\mathbf{P}$ with elements $p(i,j)$, for $i \neq j$, and $p(i,i) = 0$. The vector $\boldsymbol{t} = -\mathbf{T1}$ is the exit rate, as it contains the jump rates to the absorbing state (Bladt and Nielsen, 2017). Here, $\mathbf{1}$ is a p-dimensional column vector of ones. Now let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$, with $\pi_i = \mathbb{P}(X_0 = i)$, be a row vector in the p-dimensional simplex space $\mathbb{S}^p$. A phase-type distribution with dimension p is defined as the time until absorption of the Markov jump process $\{X_t\}_{t \geq 0}$, with initial distribution $\boldsymbol{\pi}$ and subintensity matrix $\mathbf{T}$, namely the distribution of the random variable $Y := \inf\{t > 0 \mid X_t = p + 1\}$. Accordingly, we will use the notation $Y \sim \text{PH}_p(\boldsymbol{\pi}, \mathbf{T})$ for a phase type distribution with the embedded Markov process $\{X_t\}_{t \geq 0}$ characterized by $\boldsymbol{\pi}$ and $\mathbf{T}$. It is worth noting that, to keep the standard notation in phase-type distributions literature, we are keeping $\boldsymbol{\pi}$ as a row vector.

If $Y \sim \text{PH}_p(\boldsymbol{\pi}, \mathbf{T})$, the corresponding density and cumulative distribution function (cdf) are given by

$$f(y) = \boldsymbol{\pi} e^{\mathbf{T}y} \boldsymbol{t} \quad \text{and} \quad F(y) = 1 - \boldsymbol{\pi} e^{\mathbf{T}y} \mathbf{1}\,, \tag{6}$$

respectively, where $e^{\mathbf{T}y} = \sum_{\ell=0}^{\infty} \frac{1}{\ell!} (\mathbf{T}y)^{\ell}$ and the parameter space is given by $\Theta = \mathbb{S}^p \times \mathbb{T}$.

Just as for DPM models, this class can be dense in the space of distributions on the positive real line when the number of phases p tends to infinity (Asmussen, 2000a). Phase-type distributions are appealing in a variety of complex statistical problems. They are closed under convolutions and mixing, they have closed-form expressions to solve problems in Survival Analysis (Aalen, 1995), Renewal theory (Asmussen and Bladt, 1996), computation of Ruin probabilities (Asmussen, 2000b), the estimation of the Lorenz curve and Gini index (Bladt and Nielsen, 2011), and various other estimation problems (see, e.g., Bladt et al., 2016). As it will become clear later, these results unveil various novel applications of DPM models.

Although connections between phase-type distributions and mixture models have been described in the literature (see, e.g., Cumani, 1982; O'Cinneide, 1989; Mocanu and Commault, 1999; Lee and Lin, 2010, 2012; Zadeh and Stanford, 2016), they have not been exploited in depth. Inference for Bayesian nonparametric mixture models, is nowadays relatively standard (see, e.g., Escobar and West, 1995; Neal, 2000; Ishwaran and James, 2001; Walker, 2007; Kalli et al., 2011; Miller and Harrison, 2018), whereas inference for phase-type distributions is still challenging task (Bladt et al., 2003; Aslett, 2012). Both classical and Bayesian approaches to phase-type distribution inference available in the literature, resort to the underlying Markov jump processes

$\{X_t\}_{t\geq 0}^k$, $k = 1, \ldots, n$ to write the complete likelihood, and thus to estimate the parameters using the Expectation-Maximization (EM) (Asmussen et al., 1996) or Gibbs Sampling algorithms (Bladt et al., 2003). In both approaches, it is difficult scaling up the corresponding algorithms to consider relatively large sample sizes n. In the EM algorithm, the computation of the Expectation step, based on latent trajectory of embedded jump process, is necessary for each data in the sample. On the other hand, in the Gibbs algorithm, the simulation of the associated Markov jump process for each data point is required. Furthermore, a question that arises in both approaches is: How to determine the number of phases p? As we will show below, our proposal solves all these difficulties.

The main purpose of this work is to reveal an appealing connection between DPM mixtures and phase-type distributions, thus mutually benefiting both research areas. With this in mind, we present an explicit representation of phase-type distributions as an infinite mixture of Erlang distributions. This new representation is derived using the corresponding Laplace transform, which admits loops of the process to the same state, and generalizes an existing result by Zadeh and Stanford (2016).

The manuscript is organized as follows. In Section 2, after presenting some background material related to infinite-dimensional phase-type distributions also known as SPH-distributions, we present results that connect phase-type distributions with infinite mixtures of Erlang distributions. This section includes the connection with Bayesian nonparametrics, which then allows to adapt known posterior inference techniques, shown in Section 3. An extensive Monte Carlo simulation study is included in Section 4. Section 5 illustrates the availability of a closed expression for the renewal function and density estimation with aquaculture data. Some concluding remarks are deferred to Section 6.

2 A SPH-distribution representation via Erlang kernels

When the number of transient states is infinity, $p = \infty$, PH-distributions are not necessarily proper, defining the class of infinite-dimensional phase-type (IPH) distributions. Shi et al. (1996) give conditions under which a subset of this class, identified with the acronym SPH, contains only proper distribution functions. Let us recast their result:

Corollary 1 (Shi et al., 2005)**.** Let $(w_1, w_2, \ldots)$ be a probability vector and for each $h = 1, 2, \ldots$, let $F_h(t)$ be the cdf of a $\mathrm{PH}_{p_h}(\boldsymbol{\pi}_h, \mathbf{T}(p_h))$ distribution, with initial distribution $\boldsymbol{\pi}_h$ and subintensity matrix $\mathbf{T}(p_h)$, where $(p_h)_{h=1}^{\infty}$ is a sequence of finite dimension values. Now assume there exists $\lambda := \sup_{h,j} |T_{jj}(p_h)| < \infty$, where T_{jj} are the diagonal elements of $\mathbf{T}$. Hence, the mixture model $\sum_{h=1}^{\infty} w_h F_h(t)$ can be represented as a $\mathrm{PH}_{\infty}(\boldsymbol{\phi}, \mathbf{W})$ distribution, with

$$\boldsymbol{\phi} = (w_1\boldsymbol{\pi}_1, w_2\boldsymbol{\pi}_2, \cdots), \quad \text{and} \quad \mathbf{W} = \begin{bmatrix} \mathbf{T}(p_1) & 0 & \cdots \\ 0 & \mathbf{T}(p_2) & \ddots \\ \vdots & \ddots & \ddots \end{bmatrix}.$$

Such distribution is referred to as a SPH-distribution.

Theorem 2.1 in Shi et al. (1996) establishes that SPH-distributions are proper if and only if the infinite matrix $\mathbf{W}$ is invertible. SPH-distributions share similar closure properties with PH-distributions; in fact, the class of finite PH-distributions is contained in the SPH class. In particular, if all diagonal elements of the matrix $\mathbf{W}$ are bounded and $\mathbf{W}$ is invertible, then $\sum_{h=1}^{\infty} w_h F_h(t)$ defines a proper density of the form in (6) with parameters $(\boldsymbol{\phi}, \mathbf{W})$, which is established in Shi et al. (Theorem 2.2, 1996). The following sections are valid for p arbitrarily large but finite. The conditions in Shi et al. (1996) justify using infinite mixtures of Erlang distributions, which satisfy such conditions, have a phase-type representation, and can be seen as an infinite-dimensional PH-distribution.

In general, PH-distributions are not identifiable (Telek and Horváth, 2007), meaning two different sets of parameters, $(\boldsymbol{\pi}, \mathbf{T})$ and $(\boldsymbol{\pi}^*, \mathbf{T}^*)$, might account for the same probability mass. This lack of identifiability can be seen as consequence of the observation process: there are different possible trajectories, of the embedded Markov jump process, $\{X_t\}_{t\geq 0}^1, \ldots, \{X_t\}_{t\geq 0}^n$, that lead to the same observed absorption time. Additionally, even identifiable cases such as the Exponential distribution of rate one, can be represented as a non-identifiable PH-distribution with a higher dimension p. Notice that there is a difference between the dimension and the order of a PH-distribution, the latter being the smallest dimension among all its representations. All this complicates the learning process in estimation procedures, as noted by Asmussen et al. (1996) when fitting of the Old Faithful geyser data.

As mentioned above, the literature offers some instances of connections between phase-type distributions and mixture models. However, to the best of our knowledge, such connections have not been used for Bayesian inference. Here, we use a novel characterization of PH-distributions as SPH-distributions, with Erlang kernels.

Proposition 1. Let $Y \sim \text{PH}_p(\boldsymbol{\pi}, \mathbf{T})$, where $\boldsymbol{\pi} \in \mathbb{S}^p$, $\mathbf{T} \in \mathbb{T}$, $p < \infty$, and denote by $\mathbf{P}$ the transition matrix of the embedded Markov process. Then the Laplace transform of Y can be represented as $\mathcal{L}(s) = \boldsymbol{\pi}\left(\sum_{k=0}^{\infty} (\mathbf{DP})^k\right)\mathbf{D}(\mathbf{I}-\mathbf{P})\mathbf{1}$, for $s > 0$, where $\mathbf{D}$ denotes the diagonal matrix with elements $\left(\frac{\lambda_i}{\lambda_i + s}\right)$, corresponding to Laplace transforms of exponential distributions with rates $\lambda_i > 0$ for all $i = 1, \ldots, p$.

Proof. Using the same notation as in Section 1, the Laplace transform of Y, $\mathcal{L}(s) := \mathbb{E}[e^{-sY}]$, can be represented as a non-homogeneous linear system of equations, as there exists a positive probability that the embedded Markov process $\{X_t\}_{t\geq 0}$ jumps to the absorbing state directly from any initial state. That is $\mathcal{L}(s) = \sum_{i=1}^{p} \pi_i \mathcal{L}_i(s)$, where $\mathcal{L}_i(s) := \mathbb{E}[e^{-sY} \mid X_0 = i]$, with

$$\mathcal{L}_i(s) = \frac{\lambda_i}{\lambda_i + s} \sum_{j=1, j\neq i}^{p} p_{ij}\mathcal{L}_j(s) + \frac{\lambda_i}{\lambda_i + s}\left(1 - \sum_{j=1, j\neq i}^{p} p_{ij}\right).$$

Here, p_{ij} denotes the ij-element of $\mathbf{P}$. The first term of $\mathcal{L}_i$ corresponds to the case where the embedded Markov process $\{X_t\}_{t\geq 0}$ jumps to state j, after an exponential time in state i. The second term corresponds to the case where the process jumps to the absorbing state directly from any initial state i.

Using the notation $\mathbf{L} := (\mathcal{L}_1(s), \ldots, \mathcal{L}_p(s))^t$, we have $\mathbf{L} = \mathbf{DPL} + \mathbf{D}(\mathbf{1} - \mathbf{P1})$, which solving with respect to $\mathbf{L}$, simplifies as $\mathbf{L} = (\mathbf{I} - \mathbf{DP})^{-1}\mathbf{D}(\mathbf{1} - \mathbf{P1})$.

Hence, the Laplace transform of Y can be expressed as $\mathcal{L}(s) = \boldsymbol{\pi}\mathbf{L} = \boldsymbol{\pi}(\mathbf{I} - \mathbf{DP})^{-1}\mathbf{D}(\mathbf{1} - \mathbf{P1})$ and, for $s > 0$, represented as $\mathcal{L}(s) = \boldsymbol{\pi}(\sum_{k=0}^{\infty}(\mathbf{DP})^k)\mathbf{D}(\mathbf{I} - \mathbf{P})\mathbf{1}$. This latter representation relies on the series expansion

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k,$$

which exist when the norm $\|\mathbf{A}\|_\infty < 1$. Here $\|\mathbf{A}\|_\infty := \max_i \sum_j |a_{ij}|$.

We have to show that $\|\mathbf{DP}\|_\infty < 1$. Using the sub-multiplicative property of the matrix norm, and the fact that the norm of a sub-stochastic matrix $\mathbf{P}$ is less than or equal to one, we have

$$\begin{aligned}
\|\mathbf{DP}\|_\infty &\leq \|\mathbf{D}\|_\infty \|\mathbf{P}\|_\infty \\
&= \max_i \sum_j |d_{ij}| \max_i \sum_j |p_{ij}| \\
&\leq \max\left\{\frac{\lambda_1}{\lambda_1 + s}, \ldots, \frac{\lambda_p}{\lambda_p + s}\right\} \\
&= \frac{\lambda_{\max}}{\lambda_{\max} + s},
\end{aligned}$$

where the second inequality follows from $\max_i \sum_j |p_{ij}| \leq 1$, and $\lambda_{\max} = \max\{\lambda_1, \ldots, \lambda_p\}$. Hence, $\mathbf{DP}\|_\infty < 1$ implying that $(\mathbf{I} - \mathbf{DP})^{-1} = \sum_{k=0}^{\infty}(\mathbf{DP})^k$. □

Note that the Laplace transform of Phase-type distributions can be well defined for negative values of its argument. However, having $s > 0$ is sufficient for the celebrated Laplace uniqueness Theorem (e.g. Feller, 1971, Section XIII) to follow, and thus to ensure the coincidence in distribution of positive random variables. This will be used in the results below.

Proposition 2. Let $Y \sim \mathrm{PH}_p(\boldsymbol{\pi}, \mathbf{T})$, where $\boldsymbol{\pi} \in \mathbb{S}^p$, $\mathbf{T} = \lambda(\mathbf{P} - \mathbf{I})$, that is, $\mathbf{T}$ is a subintensity matrix with all the elements on the main diagonal equal to $-\lambda$, with $\mathbf{P}$ the associated transition matrix of the embedded Markov process and $\lambda > 0$. Denote by $\mathrm{Er}(a, b)$, the Erlang distribution with mean a/b. Hence, the density of Y can be represented as

$$f_Y(\cdot) = \sum_{k=0}^{\infty} \alpha_k \mathrm{Er}(\cdot \mid k + 1, \lambda),$$

where $\alpha_k = \boldsymbol{\pi}\mathbf{P}^k(\mathbf{I} - \mathbf{P})\mathbf{1}$. Additionally, if $\mathbf{P}$ is a nilpotent matrix, that is, $\mathbf{P}^k = \mathbf{0}$ for some positive integer k, then

$$f_Y(\cdot) = \sum_{k=0}^{p-1} \alpha_k \mathrm{Er}(\cdot \mid k + 1, \lambda),$$

which is an identifiable statistical model.

Proof. The proof follows directly from Proposition 1. In fact, as $\mathbf{T} = \lambda(\mathbf{P} - \mathbf{I})$, then $\mathbf{D} = \frac{\lambda}{\lambda+s}\mathbf{I}$, where $\frac{\lambda}{\lambda+s}$ is the Laplace transform of an exponential distribution with rate $\lambda > 0$, and

$$\begin{aligned}\mathcal{L}(s) &= \boldsymbol{\pi}\left(\sum_{k=0}^{\infty}(\mathbf{DP})^k\right)\mathbf{D}(\mathbf{I}-\mathbf{P})\mathbf{1}\\ &= \boldsymbol{\pi}\left(\sum_{k=0}^{\infty}\left(\left(\frac{\lambda}{\lambda+s}\right)\mathbf{IP}\right)^k\right)\left(\frac{\lambda}{\lambda+s}\right)\mathbf{I}(\mathbf{I}-\mathbf{P})\\ &= \sum_{k=0}^{\infty}\boldsymbol{\pi}\mathbf{P}^k(\mathbf{I}-\mathbf{P})\mathbf{1}\left(\frac{\lambda}{\lambda+s}\right)^{k+1}.\end{aligned}$$

Also note that,

$$\begin{aligned}\sum_{k=0}^{\infty}\alpha_k &= \sum_{k=0}^{\infty}\boldsymbol{\pi}\mathbf{P}^k(\mathbf{I}-\mathbf{P})\mathbf{1}\\ &= \boldsymbol{\pi}\left(\sum_{k=0}^{\infty}(\mathbf{P}^k-\mathbf{P}^{k+1})\right)\mathbf{1}\\ &= \lim_{k\to\infty}\boldsymbol{\pi}(\mathbf{P}^0-\mathbf{P}^k)\mathbf{1}\\ &= \boldsymbol{\pi}\mathbf{P}^0\mathbf{1}\\ &= 1.\end{aligned}$$

Then, when the matrix $\mathbf{T}$ has all the elements of its diagonal equal to λ, the phase-type distribution has an equivalent representation as an infinite mixture of Erlang distributions. The mixture is finite when $\mathbf{P}$ is nilpotent because $\mathbf{P}^q = \mathbf{0}$ for all $q \geq p$. Note that the family of Erlang kernels in the mixture has a lexicographical order, then using Theorem 2 in (Teicher, 1963), one can deduce that a phase-type distribution with $\mathbf{T} = \lambda(\mathbf{P} - \mathbf{I})$ and $\mathbf{P}$ nilpotent is an identifiable statistical model. □

Proposition 3. Let $Y \sim \mathrm{PH}_p(\boldsymbol{\pi}, \mathbf{T})$, where $\boldsymbol{\pi} \in \mathbb{S}^p$, $\mathbf{T} \in \mathbb{T}$, $p < \infty$. Denote by $\mathbf{P}^* := \frac{1}{\lambda}\mathbf{T} + \mathbf{I}$ a transition matrix that admits loops to the same state, where λ is any arbitrary value such that $\lambda > \max\{\text{-diag}(\mathbf{T})\}$. Then, the density of Y can be represented as

$$f_Y(\cdot) = \sum_{k=0}^{\infty}\alpha_k^*\mathrm{Er}(k+1,\lambda),$$

where $\alpha_k^* := \boldsymbol{\pi}\mathbf{P}^{*k}(\mathbf{I}-\mathbf{P}^*)\mathbf{1}$.

Proof. First, note that $\mathbf{P}^*$ is a sub-stochastic matrix. In fact, the diagonal elements are of the form $p_{ii} = 1 - \lambda_i/\lambda$ and the off-diagonal elements $p_{ij} = \lambda_{ij}/\lambda$. Each row adds up to a number less or equal than one. On the other hand, let $\mathbf{D}^* = \frac{\lambda}{\lambda+s}\mathbf{I}$, $s > 0$. Then,

we have

$$\begin{aligned}\|\mathbf{D}^*\mathbf{P}^*\|_\infty &\leq \|\mathbf{D}^*\|_\infty \|\mathbf{P}^*\|_\infty \\ &= \max_i \sum_j |d^*_{ij}| \underbrace{\max_i \sum_j |p^*_{ij}|}_{\leq 1} \leq \frac{\lambda}{\lambda+s} < 1.\end{aligned}$$

Using the result in Proposition 1, we have

$$\begin{aligned}\mathcal{L}(s) &= \boldsymbol{\pi}\left(\mathbf{I}-\mathbf{DP}\right)^{-1}\mathbf{D}\left(\mathbf{1}-\mathbf{P1}\right) \\ &= \boldsymbol{\pi}\left(\sum_{k=0}^{\infty}\left(\mathbf{D}^*\mathbf{P}^*\right)^k\right)\mathbf{D}^*(\mathbf{I}-\mathbf{P}^*)\mathbf{1} \\ &= \boldsymbol{\pi}\sum_{k=0}^{\infty}\left(\left(\frac{\lambda}{\lambda+s}\right)\mathbf{P}^*\right)^k\left(\frac{\lambda}{\lambda+s}\right)(\mathbf{I}-\mathbf{P}^*)\mathbf{1} \\ &= \sum_{k=0}^{\infty}\alpha^*_k\left(\frac{\lambda}{\lambda+s}\right)^{k+1},\end{aligned}$$

where $\alpha^*_k := \boldsymbol{\pi}\mathbf{P}^{*k}(\mathbf{I}-\mathbf{P}^*)\mathbf{1}$. □

Example 1. Let $Y \sim \mathrm{PH}_2(\boldsymbol{\pi}, \mathbf{T})$, where

$$\boldsymbol{\pi} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} -1 & 1 \\ 0 & -2 \end{pmatrix}.$$

The Laplace transform is given by $L_Y(s) = \boldsymbol{\pi}(s\mathbf{I}-\mathbf{T})^{-1}\boldsymbol{t} = \frac{2}{(s+1)(s+2)}$. The equivalent representation is given by the transition matrix

$$\mathbf{P}^* = \begin{pmatrix} 1-\frac{1}{\lambda} & \frac{1}{\lambda} \\ 0 & 1-\frac{2}{\lambda} \end{pmatrix} \text{ and } \mathbf{D}^* = \frac{\lambda}{\lambda+s}\mathbf{I}.$$

With the above parameters we have that $\alpha^*_k = \frac{2}{\lambda}\left[\left(1-\frac{1}{\lambda}\right)^k - \left(1-\frac{2}{\lambda}\right)^k\right]$ for $\lambda > 2$. In such a case, the density of Y has an equivalent representation as an infinite mixture of Erlang distributions

$$f_Y(\cdot) = \sum_{k=0}^{\infty}\frac{2}{\lambda}\left[\left(1-\frac{1}{\lambda}\right)^k - \left(1-\frac{2}{\lambda}\right)^k\right]\mathrm{Er}(k+1,\lambda). \tag{7}$$

Mixture distribution (7) can be equivalently rewritten as a convolution. This can be derived using the Laplace transform of Proposition 1, that is

$$L_Y(s) = \sum_{k=0}^{\infty}\frac{2}{\lambda}\left[\left(1-\frac{1}{\lambda}\right)^k - \left(1-\frac{2}{\lambda}\right)^k\right]\left(\frac{\lambda}{\lambda+s}\right)^{k+1} = \frac{2}{(s+1)(s+2)},$$

namely the Laplace transform of a convolution o two Erlang distributions, $\mathrm{Er}(1,1)$ and $\mathrm{Er}(1,2)$. The representation of $\mathbf{P}^*$, $\mathbf{D}^*$ and $\boldsymbol{\pi}$ can be equivalently expressed without loops to the same state, but with a double number of states, such as, $Y \sim \mathrm{PH}_4(\boldsymbol{\nu}, \mathbf{Q})$,

where

$$\boldsymbol{\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}, \text{ and } \mathbf{Q} = \begin{pmatrix} 0 & \frac{1}{\lambda} & 1-\frac{1}{\lambda} & 0 \\ 0 & 0 & 0 & 1-\frac{2}{\lambda} \\ 1-\frac{1}{\lambda} & \frac{1}{\lambda} & 0 & 0 \\ 0 & 1-\frac{2}{\lambda} & 0 & 0 \end{pmatrix}.$$

Computing the eigenvalues and eigenvectors of $\mathbf{Q}$ it is immediate to show that the Laplace transform of such representation is also given by $L_Y(s) = \frac{2}{(s+1)(s+2)}$. This example illustrates how a phase-type distribution with representation $(\boldsymbol{\pi}, \mathbf{P}^*)$ can be represented without loops to the same state, but with a double number of states.

Proposition 3 states that any phase-type distribution has an equivalent representation as an infinite mixture of Erlang kernels. Hence, a natural way to represent a genuine infinite mixture distribution, without resorting to truncation, is to use an infinite-dimensional prior distribution like the Dirichlet process. Note that, given a value of λ and the number of phases p, the infinite sequence of weights $\alpha_k^*, k = 1, 2, \ldots$ can always be recovered with a stick-breaking construction as the one used to construct the Dirichlet process, see e.g. Bissiri and Ongaro (2014). The parameter λ, somehow mimics a similar effect of the total mass parameter in the DP. Indeed, fitting mixture model of Proposition 3 can be achieved by fitting a DPM model with the following hierarchical representation

$$\begin{aligned} y_k \mid \phi_k &\overset{\text{iid}}{\sim} \text{Er}\big(y \mid \lceil \phi_k \rceil, \lambda\big), \\ \phi_k \mid P &\overset{\text{ind}}{\sim} P\,, \\ P \mid \alpha, P_0 &\sim \text{DP}(\alpha, P_0)\,, \\ \lambda &\sim \text{Ga}(a_1, b_1)\,, \\ \alpha &\sim \text{Ga}(a_\alpha, b_\alpha)\,, \end{aligned} \tag{8}$$

with $P_0 = \text{Ga}(a_0, b_0)$, the Gamma distribution with mean a_0/b_0, and where $\lceil x \rceil$ denotes the least integer greater than or equal to x. Though one could think of modeling the shape parameter of the Erlang distribution above with a Dirichlet Process with atomic baseline, P_0, this results in a poor posterior performance, as noted by Canale and Prünster (2017). Therefore, to avoid this, as well as to keep mixtures of Erlangs and benefit of the closed-form expressions for phase-type distributions, we resorted to above truncated mechanism. Indeed, this resembles the rounding function approach suggested by Canale and Dunson (2011).

For the sake of simplicity, model (8) has common rate parameter λ, however this is not a limitation, as mixtures of Erlang distributions with common rate parameter λ are dense in the space of distributions with support on the positive real numbers (Tijms, 1994; Lee and Lin, 2010).

3 Posterior inference

Assume that we have i.i.d. observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ from model (8). The main difficulty in the estimation process is the infinite-dimensional nature of the parameter

space. In Bayesian nonparametric literature, dealing with the infinite-dimensional nature of a model is virtually a routine problem. We begin by examining the likelihood function of the observed data:

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\phi}, \lambda \mid \boldsymbol{y}) = \prod_{k=1}^{n} \sum_{h=1}^{\infty} w_h \operatorname{Er}\big(y_k \mid \lceil \phi_h \rceil, \lambda\big) . \tag{9}$$

Now, let d_k be a latent variable such that

$$(y_k \mid d_k = h) \quad \sim \quad \operatorname{Er}\big(y_k \mid \lceil \phi_h \rceil, \lambda\big) ,$$

for $k = 1, \ldots, n,\ h = 1, 2, \ldots$, and $\mathbb{P}[d_k = h] = w_h$. This inclusion leads to a simplification of (9), which can be rewritten as

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\phi}, \lambda \mid \boldsymbol{y}, \boldsymbol{d}) = \prod_{h=1}^{\infty} w_h^{n_h} \left[\prod_{\{k : d_k = h\}} \operatorname{Er}\big(y_k \mid \lceil \phi_h \rceil, \lambda\big) \right] , \tag{10}$$

where $\boldsymbol{d} = (d_1, \ldots, d_n)$ and $n_h = \sum_{k=1}^{n} \mathbb{1}\{d_k = h\},\ h = 1, 2, \ldots$.

To bypass the computation of an infinite number of terms in (10), Walker (2007) introduced a set of latent variables $\{u_k\}_{k=1}^{n}$, such that:

$$f(y_k, u_k \mid \boldsymbol{w}, \boldsymbol{\phi}, \lambda) = \sum_{h=1}^{\infty} \mathbb{1}\{u_k < w_h\} \operatorname{Er}\big(y_k \mid \lceil \phi_h \rceil, \lambda\big) .$$

The advantage of this approach is that only a finite subset of w_h's will satisfy the condition $(w_h > u_k, k = 1, \ldots, n)$. Hence, in a sampling scenario it is only necessary to sample N parameter sets, where $N = \max_k\{N_k\}$ and N_k is the smallest integer such that $\sum_{h=1}^{N_k} w_h > 1 - u_k$ (Walker, 2007; Kalli et al., 2011). Namely, in a particular iteration, the infinite-dimensional parameters $\boldsymbol{w}, \boldsymbol{\phi}$ reduce to a finite value. The number of components, N, is random and its magnitude depends on the complexity of the data, influenced by aspects like the number of modes, skewness levels or outlier observations. Using the stick-breaking representation of (3), with stick lengths denoted by $\boldsymbol{v}_k s$, and adding the latent variable d_k previously defined, the joint distribution is given by

$$f(y_k, d_k, u_k \mid \boldsymbol{w}, \boldsymbol{\phi}, \lambda) = \mathbb{1}\{u_k < w_{d_k}\} \operatorname{Er}(y_k \mid (\lceil \phi_{d_k} \rceil), \lambda) ,$$

and the complete data likelihood for n observations ends up being

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\phi}, \lambda \mid \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{u}) = \prod_{k=1}^{n} \mathbb{1}\{u_k < w_{d_k}\} \operatorname{Er}(y_k \mid \lceil \phi_{d_k} \rceil, \lambda) . \tag{11}$$

The following algorithm implements a slice Gibbs sampler based on the above likelihood. Details of the corresponding full conditionals can be found in the Supplementary Material (Ayala et al., 2021).

To learn about the precision parameter α, one can further implement the step described in Escobar and West (1995). See details in the Supplementary Material. This algorithm was implemented in an `R` function, `mcmcErlangMix`. The code is included in the Supplementary Material. It is worth emphasizing that any other valid algorithm for DPM models could be alternatively used.

Algorithm 1: Slice sampler.

1 Initialize $N, \boldsymbol{\phi}_N^{(0)}, \lambda^{(0)}, \boldsymbol{v}_N^{(0)}, \boldsymbol{u}^{(0)}$ and $\boldsymbol{d}^{(0)}$

2 Sample ϕ_h from $p(\phi_h \mid \ldots) \propto \left\{ \prod_{\{k:d_k=h\}} \frac{\lambda^{\lceil \phi_h \rceil} y_k^{\lceil \phi_h \rceil - 1}}{\left(\lceil \phi_h \rceil - 1\right)!} \right\} \times \phi_h^{a_0-1} e^{-b_0\,\phi_h}$

3 Sample λ from $\text{Gamma}\big(a_1 + \sum_{k=1}^{n} \lceil \phi_{d_k} \rceil\,,\, b_1 + \sum_{k=1}^{n} y_k\big)$

4 Sample v_h from $\text{Beta}\big(1 + \sum_{k=1}^{n} \mathbb{1}\{d_k = h\}\,, \alpha + \sum_{k=1}^{n} \mathbb{1}\{d_k > h\}\big)$, set $w_h = v_h \prod_{\ell<h}(1 - v_\ell)$

5 Sample u_k from $(u_k \mid \ldots) \sim \text{Unif}(0, w_{d_k}),\ \ k = 1, \ldots, n$. Then set N as the smallest integer for which $\sum_{h=1}^{N} w_h > 1 - u^*$, where $u^* = \min_k\{u_k\}$

6 With probability $\mathbb{P}[d_k = h \mid \ldots] \propto \mathbb{1}\{h : w_h > u_k\}\, \text{Er}\big(y_k \mid \lceil \phi_{d_k} \rceil, \lambda\big)$, set $d_k = h;\ k = 1, \ldots, n$

7 Repeat steps 2 through 6 until reaching stationarity.

Remark 1. Note that the mixture model induced in each iteration of Algorithm 1 is given by

$$f_Y(y) = \sum_{h=1}^{N^*} w_h^{(r)}\, \text{Er}(y \mid \lceil \phi_h^{(r)} \rceil, \lambda^{(r)})\,,$$

where the superscript r denotes the iteration, $N^* \leq N$ is the effective number of mixture components, and it comes from the number of different $\lceil \phi_h^{(r)} \rceil$ values of $\boldsymbol{\phi}^{(r)}$. Consequently, it is necessary to factorize weights $w_h^{(r)}$ in $\boldsymbol{w}^{(r)}$ whose corresponding $\phi_h^{(r)}$ produce the same $\lceil \phi_h^{(r)} \rceil$ value.

3.1 Phase-type representation

Here, we develop an algorithm to recover the parameters of the phase-type representation from model (8). The algorithm is based on the phase-type representation of the mixture

$$f(y) = \sum_{h=1}^{N} w_h Er(y \mid h, \lambda), \tag{12}$$

which corresponds to parameters $\boldsymbol{\pi} = (w_N, \ldots, w_1)$,

$$\mathbf{T} = \begin{bmatrix} -\lambda & \lambda & \cdots & 0 \\ 0 & -\lambda & \lambda & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & -\lambda \end{bmatrix}.$$

Note that the maximum value of h determines the number of phases. The initial probabilities are given by the weights in reverse order. For instance, the first component in the

mixture is an Erlang distribution with parameters 1 and λ, which is represented by the initial probability $\pi_N = w_1$, that is, a process that starts on state N where it remains an exponential time, and then jumps to the absorbing state. On the other hand, the last component in the mixture is an Erlang distribution with parameters N and λ. In this case, the initial probability is $\pi_1 = w_N$; here, the process starts in the state 1 and then jumps to the following states until absorption from state N.

In our case, the parameters for r-th posterior sample are $\boldsymbol{\phi}_{N^*}^{(r)} = \left(\phi_1^{(r)}, \ldots, \phi_{N^*}^{(r)}\right)$, $\boldsymbol{w}_{N^*}^{(r)} = \left(w_1^{(r)}, \ldots, w_{N^*}^{(r)}\right)$, and $\lambda^{(r)}$, for $r = 1, \ldots, R$. The number of phases is computed using the same logic of mixture (12), that is $p = \max\left\{\lceil\phi_1^{(r)}\rceil, \ldots, \lceil\phi_{N^*}^{(r)}\rceil\right\}$. Note that the number of phases depend on the magnitude of parameters $\lceil\phi\rceil$ and not on the number of components in the mixture. However, the values $\lceil\phi_1^{(r)}\rceil, \ldots, \lceil\phi_{N^*}^{(r)}\rceil$ are not ordered nor consecutive as it is the case for the parameters $h = 1, \ldots, N$ in mixture (12). Then the values of $\boldsymbol{w}_{N^*}^{(r)}$ are sorted according to the values of $\boldsymbol{\phi}_{N^*}^{(r)}$ which are in increasing order: set $\boldsymbol{\pi} = \mathbf{0}$, and replace the elements of $\boldsymbol{\pi}$ at positions $p - \lceil\phi_h^{(r)}\rceil + 1$ with $\pi_{p-\lceil\phi_h^{(r)}\rceil+1} = w_h^{(r)}$, for $h = 1, \ldots, N^*$.

Finally, the subintensity matrix $\mathbf{T}$ is constructed as a $p \times p$ bi-diagonal matrix, with $t_{jj} = -\lambda^{(r)}$ for $j = 1 \ldots, p$, and $t_{j,j+1} = \lambda^{(r)}$ for $j = 1, \ldots, p-1$, using the same logic of mixture (12). The estimated $\boldsymbol{\pi}$ has a sparse structure, which makes sense as we are representing a phase-type distribution with $p < \infty$, through an infinite-dimensional mixture distribution.

It is important to emphasize that even though the parameter space is of infinite dimension by definition, the resulting mixture estimate and corresponding phase-type representation can only be expressed as finite-dimensional object, due to the random truncating nature of Algorithm 1.

4 Monte Carlo study

To assess the behavior of Algorithm 1 for model (8), a Monte Carlo (MC) study was designed to estimate six different density functions on $\mathbb{R}^+$. Distributions, their corresponding density functions and parameter values are shown in Table 1. It is worth noting that these distributions do not belong to the matrix-exponential distributions family, except obviously for the phase-type case.

The simulation study was structured as follows: first, we generated 100 random samples for every distribution, each with three sample sizes ($n = 125, 250, 500$). Subsequently, Markov chain density estimates were obtained for all samples via Algorithm 1. The hyper-parameters for the base measure were fixed at $a_0 = 2$ and $b_0 = 0.1$, reflecting our lack of knowledge about their values, resulting in a prior mean for ϕ equal to 20 and a variance of 200. Similarly, the hyper-parameters for the parameter λ were fixed at $a_1 = a_2 = 0.1$, resulting in a prior mean for λ equal to 1 and a variance of 10, supporting the parametric space. Finally, for the hyper-prior distribution of the precision parameter α, we assumed $a_\alpha = b_\alpha = 1$.

Distribution name	**Density function**	**Parameter values**
Log-normal LN(μ, σ)	$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right\}$	$\mu = 0.0$ $\sigma = 0.25$
Log-normal mixture	$f(y) = \omega_1 \,\mathrm{LN}(\mu_1, \sigma_1) + \omega_2 \,\mathrm{LN}(\mu_2, \sigma_2)$	$\omega_1 = 0.6; \mu_1 = 0.0; \sigma_1 = 1.0$ $\omega_2 = 0.4; \mu_2 = 1.0; \sigma_2 = 0.25$
Generalized Inverse Gaussian GIG(λ, χ, ψ)	$f(y) = \frac{(\psi/\chi)^{\lambda/2}}{2\boldsymbol{K}_\lambda(\sqrt{\chi\psi})} y^{\lambda-1} \exp\{-(\chi y^{-1} + \psi y)/2\}$ where $\boldsymbol{K}_\lambda(z) = 2^{-\lambda-1} z^\lambda \int_0^\infty u^{-\lambda-1} e^{-u-\frac{z^2}{4u}}\, du$	$\lambda = 2$ $\chi = 2$ $\psi = 1$
GIG mixture	$f(y) = \omega_1 \,\mathrm{GIG}(\lambda_1, \chi_1, \psi_1) + \omega_2 \,\mathrm{GIG}(\lambda_2, \chi_2, \psi_2)$	$\lambda_1 = 12; \chi_1 = 1; \psi_1 = 2$ $\omega_1 = 0.65$ $\lambda_2 = 30; \chi_2 = 1; \psi_2 = 2$ $\omega_2 = 0.35$
Three-parameter Weibull Weibull(α, β, θ)	$f(y) = \frac{\alpha}{\beta}\left(\frac{y-\theta}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{y-\theta}{\beta}\right)^{\alpha}\right\}$	$\alpha = 5$ $\beta = 2$ $\theta = 1$
Phase-type PH$_p(\boldsymbol{\pi}, \mathbf{T})$	$f(y) = \boldsymbol{\pi} e^{\mathbf{T}y} \boldsymbol{t}$ where $\boldsymbol{t} = -\mathbf{T}\mathbf{1}$	$\boldsymbol{\pi} = (0.6, 0, 0, 0, 0, 0, 0, 0.4, 0, 0)$ $\mathbf{T}$ is a bidiagonal matrix with $p = 10$ and $\lambda = 1.5$

Table 1: Monte Carlo study density functions.

Each run comprised 10,000 iterations, with a burn-in period of 2,000 iterations and a thinning value of 8. Therefore, all density estimates were constructed by averaging over 1,000 posterior density draws. Figure 1 shows the posterior mean of the density for each of the 100 MC replicates. In particular, this figure presents the results for $n = 500$ (see the Supplementary Material for posterior mean estimates when $n = 125$ and $n = 250$).

Subsequently, to gauge the estimation process's overall performance quantitatively, the Mean Integrated Squared Error (MISE), $\mathbb{E}\|f_n - f\|_2^2 = \mathbb{E}\int (f_n(x) - f(x))^2 dx$, was estimated for every density function and sample size, over a reasonable grid which covers up to two times the maximum of sampled deviates. Here, f and f_n denote the true and estimated densities, respectively. The results for the MISE are reported in Table 2, and they show an improvement in the precision of density estimates as sample size increases for each distribution.

Distribution	**MISE**		
	$n = 125$	$n = 250$	$n = 500$
Log-normal	0.02597	0.01073	0.00438
Log-normal mixture	0.01018	0.00462	0.00265
GIG	0.00190	0.00094	0.00064
GIG mixture	0.00121	0.00056	0.00027
3-parameter Weibull	0.01675	0.00689	0.00344
Phase-type	0.00336	0.00194	0.00088

Table 2: Mean Integrated Squared Error estimates for the Monte Carlo study.

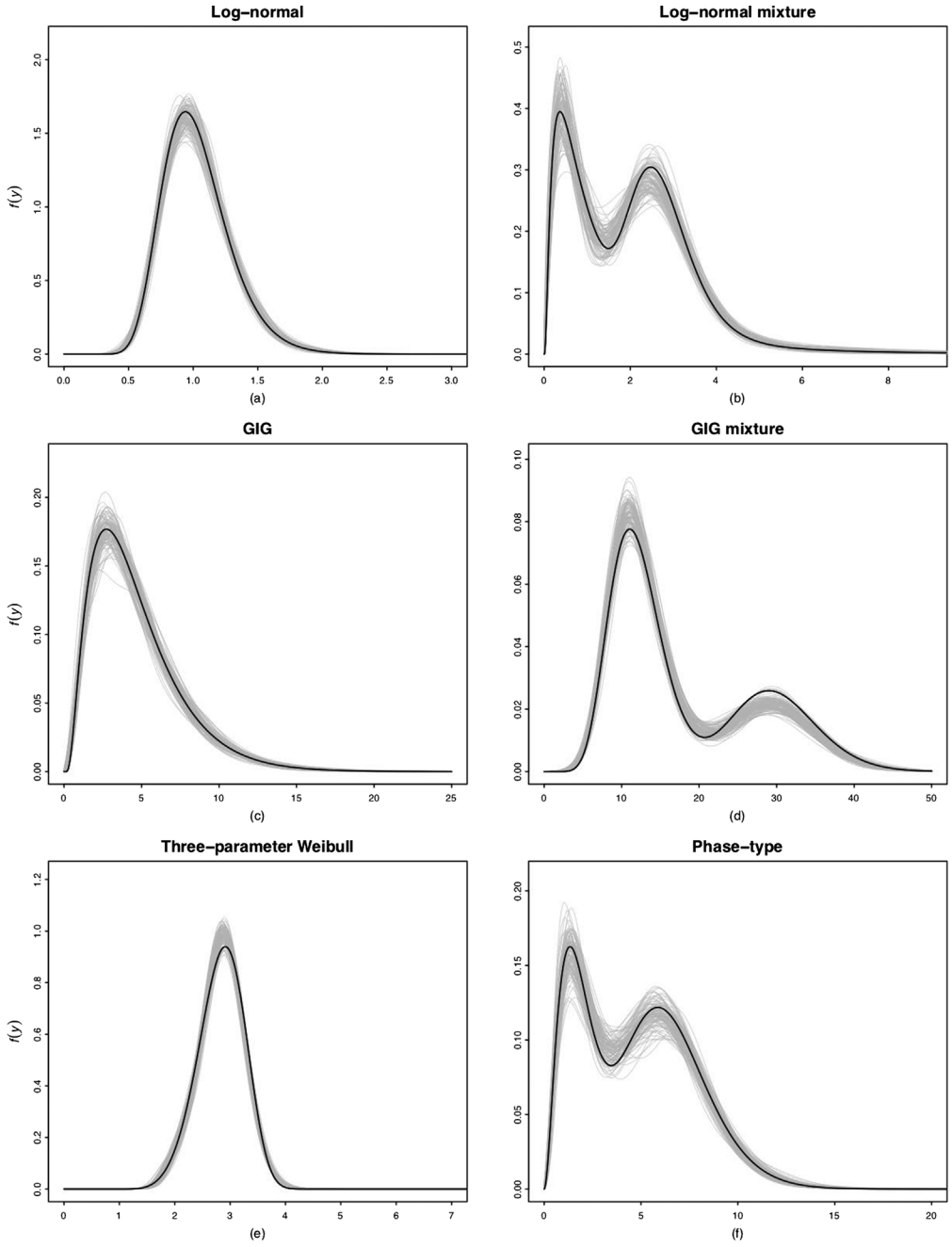

Figure 1: True density function (solid lines) and posterior mean density estimates (dotted lines) of the 100 Monte Carlo runs for the selected distributions and sample size $n = 500$.

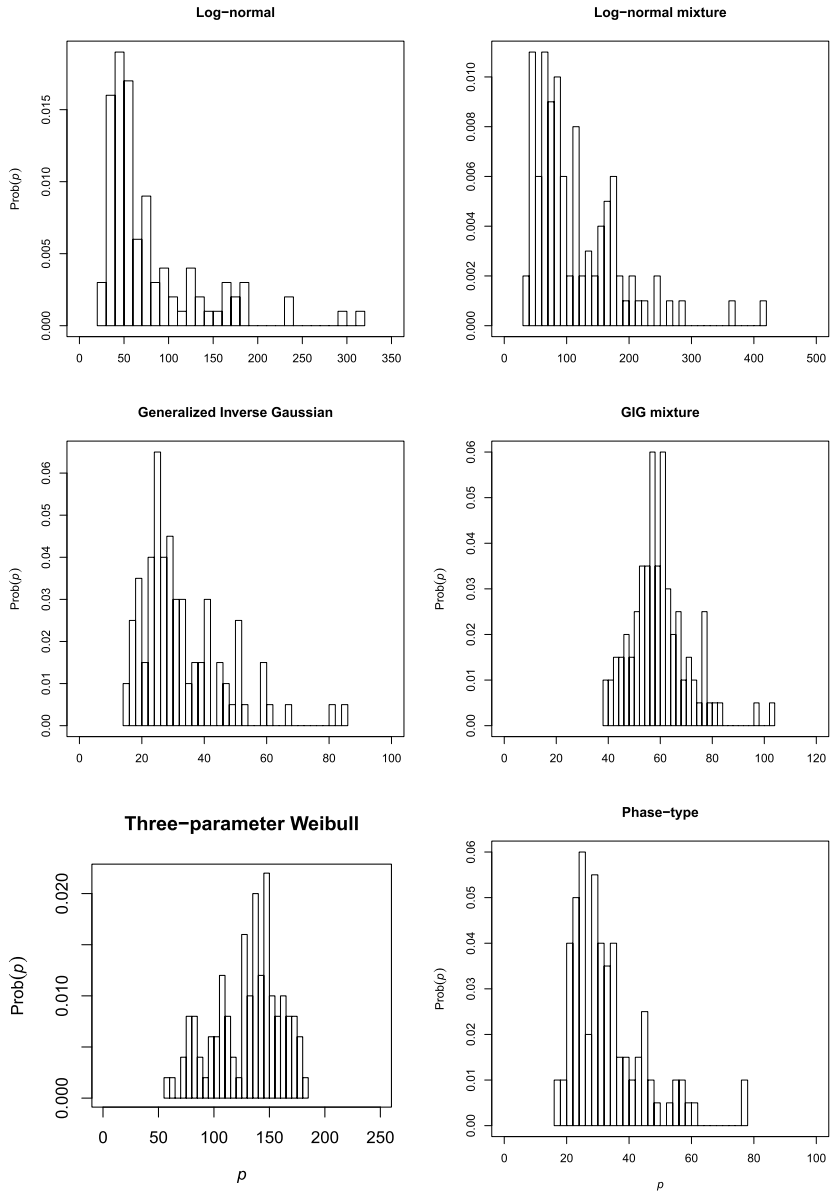

Figure 2: Mode of the number of phases for each replication in the Monte Carlo study: $n = 500$.

Lastly, we analyzed the resulting dimension p of the equivalent phase-type representation for every replication of the Monte Carlo study. Then, for each estimation, we computed the mode for p in the posterior samples and reported the mode for each MC sample in Figure 2. We observe that for distributions with lowest MISE values (GIG,

GIG mixture, and phase-type), the estimated dimension needed to accommodate for the given data complexity is also low. In contrast, the required dimension for the phase-type representation is noticeable more spread out for the remaining distributions.

Table 3 includes the mode for the estimated dimension p for each scenario, that is, the mode in the 100 replications. It also shows the sample average skewness, $\overline{\gamma_1}$, and sample kurtosis of all generated samples of size 500. We observe that for unimodal distributions, the magnitude of the mode for p is related to negative values of $\overline{\gamma_1}$; specifically, the three-parameter Weibull distributed data produced the largest estimated value of p, with 109 transient states.

Distribution	**Mode**(p)	$\overline{\gamma_1}$	**Kurtosis**
Log-normal	40	0.70788	0.81165
Log-normal mixture	80	2.11441	10.23081
GIG	26	1.24060	2.05129
GIG mixture	58	0.91080	−0.26986
3-parameter Weibull	109	−0.24396	−0.13597
Phase-type	30	0.38288	−0.56179

Table 3: Monte Carlo study for the dimension mode, p, average empirical skewness and kurtosis. Sample size $n = 500$.

4.1 Comparison with other inference approaches

Here we provide two simulated examples to compare our approach with the results from the `R` packages `mapfit` and `PhaseType`. The `mapfit` package was developed by Okamura and Dohi (2015) based on a variation of the Expectation-Maximization algorithm proposed by Asmussen et al. (1996). The `PhaseType` package (Aslett, 2012), follows the Bayesian approach proposed by (Bladt et al., 2003). This latter approach resorts to the Metropolis-Hasting algorithm to simulate the underlying Markov jump processes and then performs the inference, defining a gamma distribution as a prior on the elements of $\mathbf{T}$ and a Dirichlet prior for the initial probabilities $\boldsymbol{\pi}$. We simulate $n = 1{,}000$ realizations from the phase-type distributions $\mathrm{PH}_5(\boldsymbol{\pi}_1, \mathbf{T}_1)$ and $\mathrm{PH}_{10}(\boldsymbol{\pi}_2, \mathbf{T}_2)$ where $\boldsymbol{\pi}_1 = (0, 0, 0, 1, 0)$, $\boldsymbol{\pi}_2 = (0.6, 0, 0, 0, 0, 0, 0, 0.4, 0, 0)$,

$$\mathbf{T}_1 = \begin{pmatrix} -3.96 & 0 & 0 & 0 & 3.96 \\ 0 & -0.64 & 0 & 0.64 & 0 \\ 1.10 & 0.47 & -1.58 & 0 & 0 \\ 0.78 & 0 & 0 & -0.78 & 0 \\ 0 & 0 & 2.01 & 0 & -3.95 \end{pmatrix},$$

and $\mathbf{T}_2$ a bidiagonal matrix with elements $t_{ii} = -1.5$ and $t_{i,i+1} = 1.5$, $i = 1, \ldots, 10$. Figure 3 shows the estimated densities. For the distribution $\mathrm{PH}_5(\boldsymbol{\pi}_1, \mathbf{T}_1)$, which is unimodal, the `mapfit` package and our proposal got very good fits. The performance of the `PhaseType` package is low. For the distribution $\mathrm{PH}_{10}(\boldsymbol{\pi}_2, \mathbf{T}_2)$ our proposal gives good results, the `mapfit` and `PhaseType` were unable to detect the two modes. The number of phases in the `mapfit` and `PhaseType` packages requires to be fixed by the

user. We have tried different values, e.g., 10, 20, 30. Figure 3 shows the best results we observed, corresponding to $p = 30$. On contrast, our inference strategy adjusts the number of phases automatically, which is an advantage. O'Cinneide (1990, 1999) demonstrates that a phase-type distribution of order p is determined by at most $2p-1$ independent parameters. The results of the `mapfit` and `PhaseType` packages are based on $p^2 + p - 1$ parameters, which is clearly larger than $2p - 1$. In general, for highly redundant parametrizations, the performance of phase-type distributions is well-known to have problems, see, e.g., Asmussen et al. (1996). In our inference strategy reduces to estimate p parameters. Overall, our method avoids the simulation of latent processes, used by most approaches available in the literature, which results in a more efficient technique.

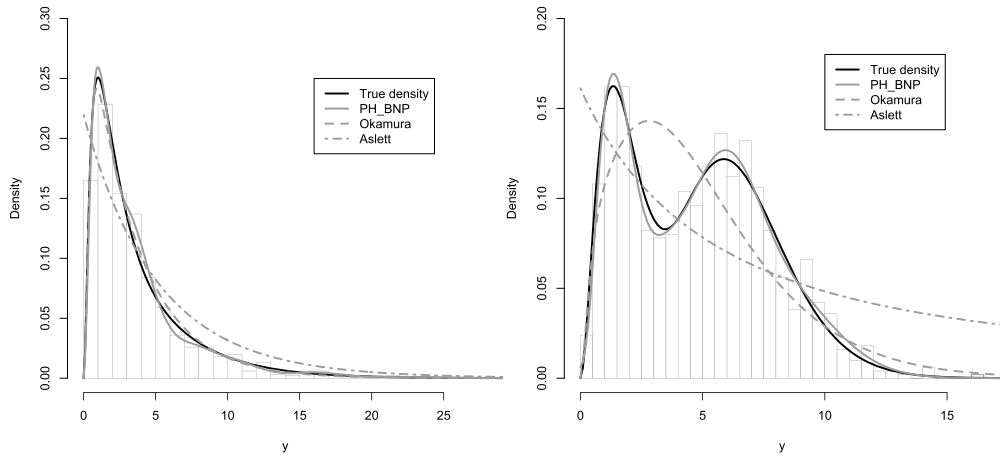

Figure 3: Density estimation for two phase-type distributions using the `mapfit` (Okamura, with $p = 30$ phases), `PhaseType` (Aslett, with $p = 30$ phases) packages, and our proposal (PH_BNP).

4.2 Renewal function estimation

Within the context of Renewal Theory, a quantity of great interest is the expected number of renewals at time τ, $U(\tau) := \mathbb{E}[\eta(\tau)]$, of the associated counting process η (Feller, 1971). In the case of phase-type distributions, $U(\tau)$ has an analytical expression (Bladt and Nielsen, 2017) given by

$$U(\tau) = \frac{\tau}{\boldsymbol{\pi}\mathbf{T}^{-2}\boldsymbol{t}} - \boldsymbol{\pi}\left(\mathbf{I} - e^{(\mathbf{T}+\boldsymbol{t}\boldsymbol{\pi})\tau}\right)(\mathbf{T} + \boldsymbol{t}\boldsymbol{\pi} - \boldsymbol{s}\boldsymbol{\vartheta})^{-1}\boldsymbol{t} \tag{13}$$

with $\boldsymbol{\vartheta} = \boldsymbol{\pi}\mathbf{T}^{-1}/\boldsymbol{\pi}\mathbf{T}^{-2}\boldsymbol{t}$ and $\boldsymbol{s} = (-\mathbf{T})^{-1}\boldsymbol{t}$.

Accordingly, we exemplify the computation of the renewal function for the phase-type simulated data sets described in Table 1. To that effect, we constructed the set of

parameters $(\boldsymbol{\pi}, \mathbf{T})$ from the Erlang mixture parameters estimates, as specified in sub-Section 3.1, and then continued with the calculation of $U(\tau)$. Resulting $U(\tau)$ estimates are shown in Figure 4.

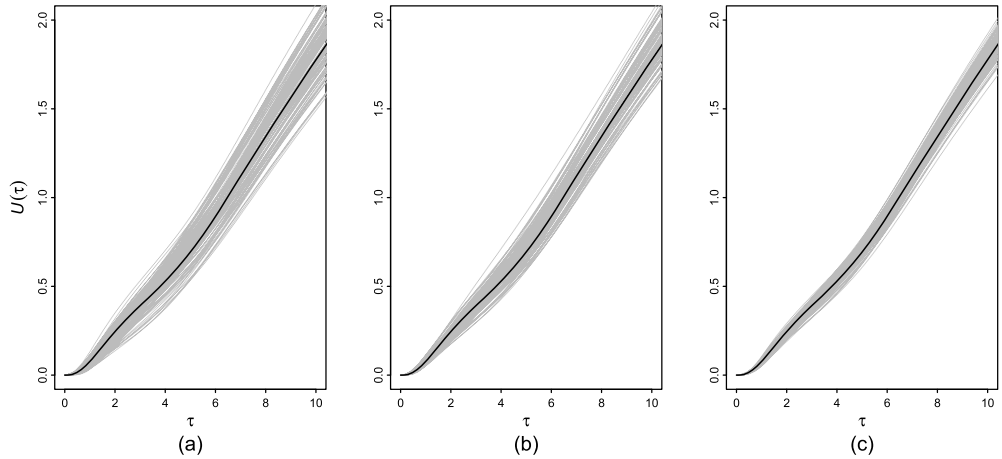

Figure 4: True renewal function (solid lines) and corresponding posterior mean estimates (dotted lines) of the 100 Monte Carlo runs for phase-type simulated data: (a) $n = 125$; (b) $n = 250$; (c) $n = 500$.

Here we can see that computed renewal functions do not differ significantly from the true renewal function. Furthermore, we point out the ability to capture the non-linear behavior of the true renewal function, a feature not captured by simple Poisson counting processes (Winkelmann, 1995).

5 Application to real data sets

Here we compute density function and renewal function estimation for three datasets: the first two have been studied in the Renewal processes literature, and a third one belongs to an aquaculture study. The first dataset consists of the Old Faithful geyser eruption data explored by Asmussen et al. (1996). However, we use the more comprehensive data set, which consists of 299 eruption observations (duration and waiting time) from August 1$^{\text{st}}$ to August 15$^{\text{th}}$, 1985 (Azzalini and Bowman, 1990).

The estimated density clearly captures the bi-modal shape of the observed data (Figure 5, panel (a)), as well as the *delay* for times lower than 40 minutes. The renewal function $U(\tau)$ in this case presents a distinctly nonlinear pattern, reflecting the multi-modal density function (Figure 5, panel (b)). The second dataset comprises coal-mining disasters presented in Jarrett (1979) and studied by Xiao (2015). This reports the number of days between 191 successive explosions on a coal-mining site and involving ten or more men killed. The dataset includes information of a period from March 15$^{\text{th}}$, 1851 to March 22$^{\text{nd}}$, 1962. The observations exhibit an exponential-like shape with a heavy tail,

which is known to be challenging to model when using finite-dimensional phase-type distributions (Bladt and Rojas-Nandayapa, 2018). Nevertheless, our proposed model captures correctly the shape, as shown in Figure 6.

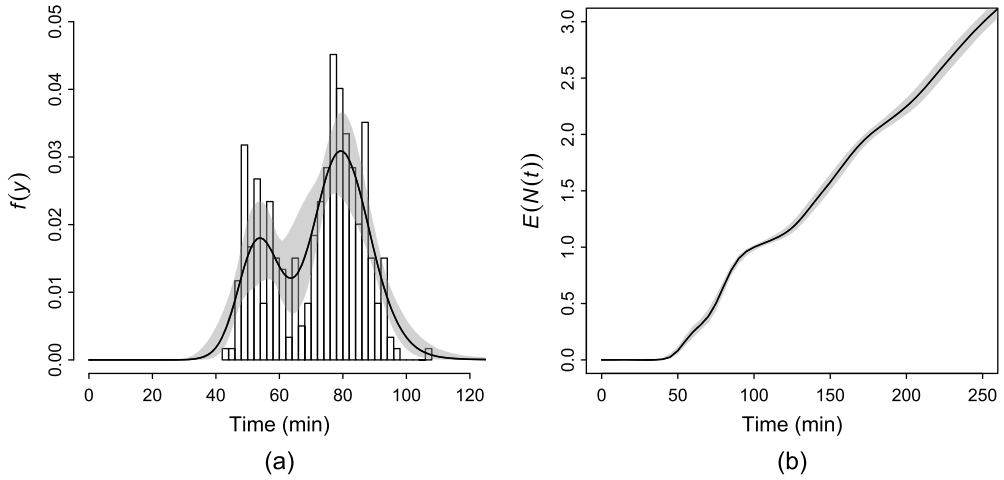

Figure 5: Posterior mean estimates (solid lines) and 95% credible intervals (shaded areas) for the geyser waiting times between eruptions: (a) density function; (b) renewal function.

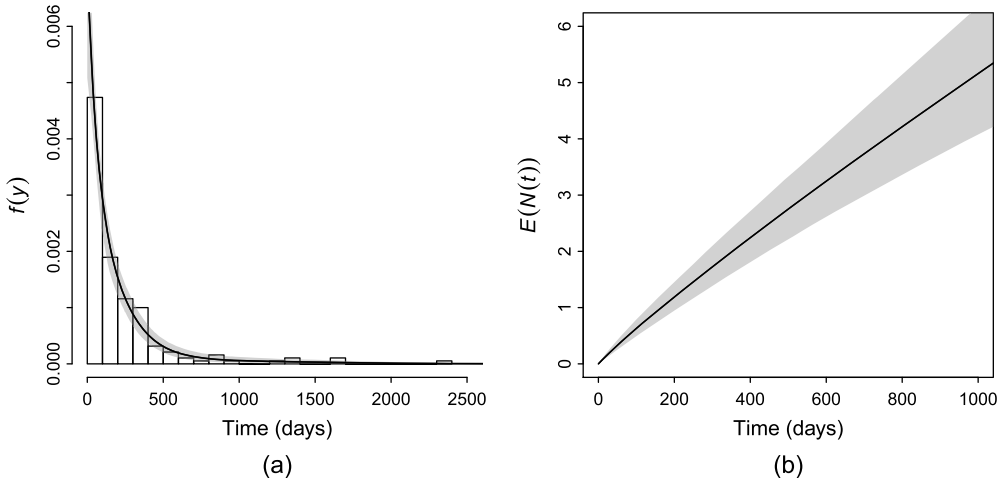

Figure 6: Posterior mean estimates (solid lines) and 95% credible intervals (shaded areas) for the coal mine waiting times between explosions: (a) density function; (b) renewal function.

Regarding the renewal function estimation, it is also non-linear. Having expression (13) at hand allows us to compute the number of expected events for different time windows. For example, by day 365, one can expect the occurrence of two (≈ 2.06) mine disasters with the characteristics of interest. By day 700, the predicted number consists of almost four events (≈ 3.88), which is of paramount importance for risk assessment.

Finally, we tackled an estimation problem within a fish farming set up. The main goal was to estimate the salmon weight population's density function in a cage throughout time. Although the underlying Markov jump process in phase-type distributions represents the time until absorption, phase-type distributions are suitable for any random variable with support on the positive real numbers. The dataset contains weight measurements of sampled fish in a culture tank at day 15 ($n = 243$), day 34 ($n = 256$), day 74 ($n = 195$) and day 154 ($n = 251$). The relevance of density estimation in this context lies in the necessity to know the proportion of fish that is in a given weight range due to the different commercial value according to the fish's size.

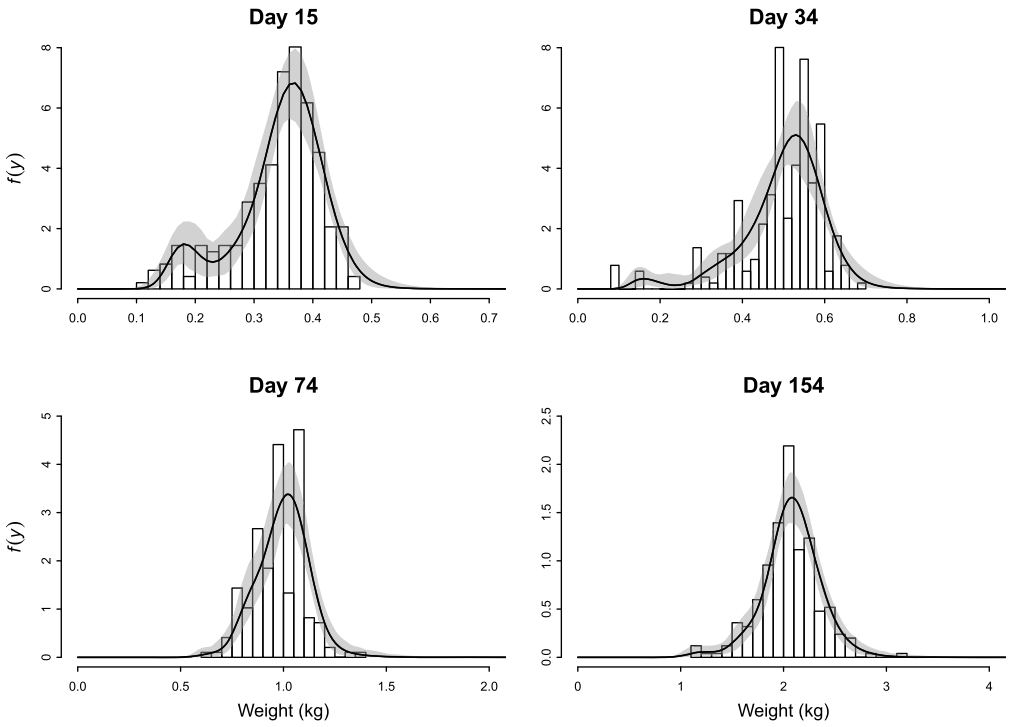

Figure 7: Posterior mean estimates (solid lines) and 95% credible intervals (shaded areas) for the salmon weight density by day.

Figure 7 shows the posterior mean estimate of the density for the cage's weight distribution. In the first stages of development, we can observe more variability and even two modes. The weight distribution can be explained, in part, by the vaccination

effect. It is well known that a proportion of the individuals in a cage do not receive the corresponding doses by difficulties in the capture.

6 Discussion and concluding remarks

We demonstrated a clear connection between phase-type distributions, mixtures of Erlang distributions, and Bayesian nonparametric inference in this work. As established in Propositions 1 to 3, any phase-type distribution has an equivalent infinite mixture of Erlang distributions representation, and if its associated transition matrix $\mathbf{P}$ is nilpotent, then the corresponding Erlang mixture is finite-dimensional. In addition, if $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$, the finite mixture is an identifiable statistical model.

Although some links between phase-type distributions and Erlang mixtures have been explored in the past, none of them have exploited this relationship for inference. Bayesian nonparametric methods allow us to treat infinite-dimensional statistical models and thus fit phase-type distributions using their infinite mixture model representations. Under our framework, one avoids the simulation of a latent Markov jump process for each observation when implementing a Gibbs sampler, overcoming serious identifiability and numerical problems inherent to other methods available in the literature, e.g. Bladt et al. (2003) and Aslett (2012). The significant reduction of the computational burden translates into faster and more efficient algorithms.

As a byproduct of the established connection, we are able to use well-known closed-form expressions obtainable for phase-type distributions' density functionals, not readily available from a DPM models standpoint. See, e.g., heavy-tailed data modeling in queueing theory (Greiner et al., 1999); the numerical approximations to estimate the Lorenz curve and Gini index in Hasegawa and Kozumi (2003); or the numerical inversion of the Laplace transform to compute the renewal function, when inter-arrival times follow an infinite mixture of Erlang distributions in Xiao (2015). In particular, the renewal function is no longer an approximation but an exact analytic quantity. Therefore, we are now capable of performing inference on a counting process by analyzing its waiting times, even though their distribution does not belong to the exponential distributions family, which has been the usual approach. This is an exceptional result as we can study non-regular counting processes, as long as their inter-arrival times are assumed i.i.d. phase-type random variables. The connection makes feasible model over and under dispersion, something not possible in the Poisson-Exponential scenario (Cox, 1962).

Supplementary Material

Supplementary material for: On a Dirichlet process mixture representation of phase-type distributions (DOI: 10.1214/21-BA1272SUPP; .pdf). The online Supplementary Material contains the posterior inference derivations of the algorithm of Section 3, some results of the Monte Carlo simulation study of Section 4, and an `R` function for posterior sampling.

References

Aalen, O. O. (1995). "Phase type distributions in survival analysis." *Scandinavian Journal of Statistics*, 22(4): 447–463. URL http://www.jstor.org/stable/4616373. 767

Aslett, L. J. M. (2012). "MCMC for Inference on Phase-type and Masked System Lifetime Models." Ph.D. thesis, Trinity College Dublin. 767, 780, 785

Asmussen, S. (2000a). "Matrix-analytic Models and their Analysis." *Scandinavian Journal of Statistics*, 27(2): 193–226. MR1777501. doi: https://doi.org/10.1111/1467-9469.00186. 767

Asmussen, S. (2000b). *Ruin Probabilities*, volume 2 of *Advanced Series on Statistical Science & Applied Probability*. World Scientific Publishing Co. Pte. Ltd. URL https://www.worldscientific.com/doi/abs/10.1142/2779. MR1794582. doi: https://doi.org/10.1142/9789812779311. 767

Asmussen, S. and Bladt, M. (1996). "Renewal theory and queueing algorithms for matrix-exponential distributions." In Chakravarthy, S. R. and Alfa, A. S. (eds.), *Matrix-analytic methods in stochastic models*, Lecture notes in pure and applied Mathematics, 313–341. CRC Press. MR1427279. 767

Asmussen, S., Nerman, O., and Olsson, M. (1996). "Fitting Phase-type distributions via the EM algorithm." *Scandinavian Journal of Statistics*, 23(4): 419–441. MR1439706. 768, 769, 780, 781, 782

Ayala, D., Jofré, L., Gutiérrez, L., and Mena, R. H. (2021). "Supplementary material for: On a Dirichlet process mixture representation of phase-type distributions." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1272SUPP. 774

Azzalini, A. and Bowman, A. W. (1990). "A look at some data on the Old Faithful geyser." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(3): 357–365. 782

Bissiri, P. G. and Ongaro, A. (2014). "On the topological support of species sampling priors." *Electronic Journal of Statistics*, 8(1): 861–882. MR3229100. doi: https://doi.org/10.1214/14-EJS912. 773

Bladt, M., Campillo Navarro, A., and Nielsen, B. (2016). "On the use of functional calculus for phase-type and related distributions." *Stochastic Models*, 32(1): 1–19. MR3457119. doi: https://doi.org/10.1080/15326349.2015.1064773. 767

Bladt, M., González, A., and Lauritzen, S. L. (2003). "The estimation of phase-type related functionals using Markov chain Monte Carlo methods." *Scandinavian Actuarial Journal*, 2003(4): 280–300. MR2025352. doi: https://doi.org/10.1080/03461230110106435. 767, 768, 780, 785

Bladt, M. and Nielsen, B. (2011). "Moment distributions of phase type." *Stochastic Models*, 27(4): 651–663. MR2854237. doi: https://doi.org/10.1080/15326349.2011.614192. 767

Bladt, M. and Nielsen, B. (2017). *Matrix-Exponential Distributions in Applied Probability*, volume 81 of *Probability Theory and Stochastic Modelling*. Springer-Verlag. MR3616926. doi: https://doi.org/10.1007/978-1-4939-7049-0. 766, 767, 781

Bladt, M. and Rojas-Nandayapa, L. (2018). "Fitting phase–type scale mixtures to heavy–tailed data and distributions." *Extremes*, 21(2): 285–313. MR3800300. doi: https://doi.org/10.1007/s10687-017-0306-4. 783

Canale, A. and Dunson, D. B. (2011). "Bayesian Kernel Mixtures for Counts." *Journal of the American Statistical Association*, 106(496): 1528–1539. MR2896854. doi: https://doi.org/10.1198/jasa.2011.tm10552. 766, 773

Canale, A. and Prünster, I. (2017). "Robustifying Bayesian nonparametric mixtures for count data." *Biometrics*, 73(1): 174–184. MR3632363. doi: https://doi.org/10.1111/biom.12538. 766, 773

Cox, D. R. (1962). *Renewal Theory*. Methuen's monographs on applied probability and statistics. Methuen & Co., 1 edition. MR0153061. 785

Cumani, A. (1982). "On the canonical representation of homogeneous Markov processes modelling failure – time distributions." *Microelectronics Reliability*, 22(3): 583–602. 767

Dunson, D. B. (2010). "Nonparametric Bayes applications to biostatistics." In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds.), *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, 223–273. Cambridge University Press. MR2730665. 766

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 766, 767, 774

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume 2 of *Wiley Series in Probability and Statistics*. John Wiley & Sons Inc., 2 edition. MR0270403. 770, 781

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. MR0350949. 766

Ferguson, T. S. (1974). "Prior Distributions on Spaces of Probability Measures." *The Annals of Statistics*, 2(4): 615–629. MR0438568. 766

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, 1 edition. MR2265601. 765

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. MR3587782. doi: https://doi.org/10.1017/9781139029834. 766

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. Springer New York. MR1992245. 766

Greiner, M., Jobmann, M., and Lipsky, L. (1999). "The importance of power-tail distributions for modeling queueing systems." *Operations Research*, 47(2): 313–326. MR2455624. doi: https://doi.org/10.1007/978-0-387-49706-8. 785

Hasegawa, H. and Kozumi, H. (2003). "Estimation of Lorenz curves: a Bayesian Nonparametric approach." *Journal of Econometrics*, 115(2): 277–291. URL http://www.sciencedirect.com/science/article/pii/S0304407603000988. MR1984778. doi: https://doi.org/10.1016/S0304-4076(03)00098-8. 785

Ishwaran, H. and James, L. F. (2001). "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 767

Jarrett, R. G. (1979). "A note on the intervals between coal-mining disasters." *Biometrika*, 66(1): 191–193. 782

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21(1): 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 767, 774

Lee, S. C. K. and Lin, X. S. (2010). "Modeling and evaluating insurance losses via mixtures of Erlang distributions." *North American Actuarial Journal*, 14(1): 107–130. MR2720423. doi: https://doi.org/10.1080/10920277.2010.10597580. 767, 773

Lee, S. C. K. and Lin, X. S. (2012). "Modeling dependent risks with multivariate Erlang mixtures." *ASTIN Bulletin*, 42(1): 153–180. MR2963333. 767

Lo, A. Y. (1984). "On a class of Bayesian Nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12(1): 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 766

McLachlan, G. and Peel, D. (2000). *Finte Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 1 edition. MR1789474. doi: https://doi.org/10.1002/0471721182. 765

Miller, J. W. and Harrison, M. T. (2018). "Mixture Models With a Prior on the Number of Components." *Journal of the American Statistical Association*, 113(521): 340–356. PMID: 29983475. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 767

Mocanu, Ş. and Commault, C. (1999). "Sparse representations of phase-type distributions." *Communications in Statistics – Stochastic Models*, 15(4): 759–778. MR1708450. doi: https://doi.org/10.1080/15326349908807561. 767

Müller, P. and Mitra, R. (2013). "Bayesian Nonparametric Inference – Why and How." *Bayesian Analysis*, 8(2): 269–302. MR3066939. doi: https://doi.org/10.1214/13-BA811. 766

Müller, P. and Quintana, F. A. (2004). "Nonparametric Bayesian data analysis." *Statistical Science*, 19(1): 95–110. MR2082149. doi: https://doi.org/10.1214/088342304000000017. 766

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. URL http://www.jstor.org/stable/1390653. MR1823804. doi: https://doi.org/10.2307/1390653. 767

Neuts, M. F. (1975). "Probability distribution of Phase Type." In *Liber amicorum professor emeritus H. Florin*, 173–206. Department of Mathematics, University of Louvain. 766

Neuts, M. F. (1978). "Renewal processes of phase type." *Naval Research Logistics Quarterly*, 25(3): 445–454. MR0518067. doi: https://doi.org/10.1002/nav.3800250307. 766

O'Cinneide, C. A. (1989). "On non-uniqueness of representations of phase-type distributions." *Communications in Statistics – Stochastic Models*, 5(2): 247–259. MR1000633. doi: https://doi.org/10.1080/15326348908807108. 767

O'Cinneide, C. A. (1990). "Characterization of phase-type distributions." *Communications in Statistics – Stochastic Models*, 6(1): 1–57. MR1047102. doi: https://doi.org/10.1080/15326349908807134. 781

O'Cinneide, C. A. (1999). "Phase-type distributions: open problems and a few properties." *Communications in Statistics – Stochastic Models*, 15(4): 731–757. MR1708454. doi: https://doi.org/10.1080/15326349908807560. 781

Okamura, H. and Dohi, T. (2015). "mapfit: An R-Based Tool for PH/MAP Parameter Estimation." In Campos, J. and Haverkort, B. R. (eds.), *Quantitative Evaluation of Systems*, 105–112. Cham: Springer International Publishing. 780

Pearson, K. (1894). "Contributions to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London A*, 185: 71–110. 765

Shi, D., Guo, J., and Liu, L. (2005). "On the SPH-distribution class." *Acta Mathematica Scientia*, 25(2): 201–214. MR2133060. doi: https://doi.org/10.1016/S0252-9602(17)30277-1. 768

Shi, D. H., Guo, J., and Liu, L. (1996). "SPH-Distributions and the Rectangle-Iterative algorithm." In Chakravarthy, S. R. and Alfa, A. S. (eds.), *Matrix-analytic methods in stochastic models*, Lecture notes in pure and applied Mathematics, 207–224. CRC Press. MR1427274. 768, 769

Teicher, H. (1963). "Identifiability of finite mixtures." *The Annals of Mathematical Statistics*, 34(4): 1265–1269. MR0155376. doi: https://doi.org/10.1214/aoms/1177703862. 771

Telek, M. and Horváth, G. (2007). "A minimal representation of Markov arrival processes and a moments matching method." *Performance Evaluation*, 64(9-12): 1153–1168. 769

Tijms, H. C. (1994). *Stochastic models: an algorithmic approach*. Wiley series in probability and mathematical statistics. Chichester; New York: Wiley, 1 edition. MR1314821. 773

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester: Wiley., 1 edition. MR0838090. 765

Walker, S. G. (2007). "Sampling the Dirichlet Mixture Model with Slices." *Communications in Statistics – Simulation and Computation*, 36(1): 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 767, 774

Winkelmann, R. (1995). "Duration dependence and dispersion in count-data models." *Journal of Business & Economic Statistics*, 13(4): 467–474. MR1354532. doi: https://doi.org/10.2307/1392392. 782

Xiao, S. (2015). "Bayesian Nonparametric modeling for some classes of temporal point processes." PhD. thesis, University of California, Santa Cruz. MR3347098. 782, 785

Zadeh, A. H. and Stanford, D. A. (2016). "Bayesian and Bühlmann credibility for phase-type distributions with a univariate risk parameter." *Scandinavian Actuarial Journal*, 4: 338–355. MR3435187. doi: https://doi.org/10.1080/03461238.2014.926977. 767, 768

Acknowledgments

The authors are grateful for the valuable comments made by two anonymous referees, the Associate Editor and the Editor. The authors thank professor Ricardo Olea Ortega by providing the salmon weights data set.