FITTING PHASE–TYPE SCALE MIXTURES TO HEAVY–TAILED DATA AND DISTRIBUTIONS

MOGENS BLADT AND LEONARDO ROJAS-NANDAYAPA

ABSTRACT. We consider the fitting of heavy tailed data and distributions with a special attention to distributions with a non–standard shape in the "body" of the distribution. To this end we consider a dense class of heavy tailed distributions introduced in [6], employing an EM algorithm for the maximum likelihood estimation of its parameters. We present methods for fitting to observed data, histograms, censored data, as well as to theoretical distributions. Numerical examples are provided with simulated data and a benchmark reinsurance dataset. Empirical examples show that the methods will in most cases adequately fit both body and tail simultaneously.

Keywords. Statistical inference, heavy–tailed, phase–type, scale mixtures, approximating distributions, EM algorithm.

1. INTRODUCTION

In this paper we consider the statistical inference problem for univariate and independent heavy–tailed data. We propose to fit heavy–tailed data with distributions in the class of *scale mixtures of phase–type distributions* (NPH) as introduced in [6]. The NPH class consists of distributions which can be expressed as the distribution of a product $N \cdot X$, where N is a discrete random variable with some distribution π and X has a phase–type distribution (hence the abbreviation NPH). We refer to π as the scaling distribution.

The NPH class is appropriate for modelling purposes because it is a class of absolutely continuous distributions which is dense in the nonnegative distributions. It is particularly well suited for heavy–tailed data because it contains heavy–tailed distributions (i.e. nonnegative distributions having Laplace transforms that fail to converge for any positive value of the argument); in fact, an NPH distribution is heavy–tailed if and only if the scaling distribution is unbounded [18]. We concentrate on scaling distributions which can be parametrized by some vector θ of finite dimension, so the number of parameters of the NPH distribution remains finite as well.

The NPH class allows for the simultaneous modelling of the "body" and the "tail" of general heavy–tailed distributions. The idea behind this approach is that N shall provide an accurate fitting of the tail of the distribution while X will take care of its body. This flexibility will allow to fit general heavy–tailed data more accurately, including those data sets which may look distinctively different from distributions usually found in catalogues of heavy–tailed distributions. The most crucial decision a modeller must take is regarding the appropriate selection of a scaling distribution, which will commonly be obtained from discretizing a parametric heavy tailed distribution of a certain class (like e.g. Pareto or Weibull).

The problem of fitting heavy–tailed data is closely related to the statistics of univariate extreme value theory (EVT), where limiting laws for sample maxima [8, 9], upper order statistics [7, 11] and excesses over high thresholds [3, 16] are the foundational theoretical results for an extensive list of statistical methods (see [10] for a recent survey). However, the EVT approach has drawn some criticism; for instance, the focus is on the upper order statistics (thus implying an imminent waste of data), while the resulting estimators depend significantly on thresholds and/or block data sizes selected by the modeller. The alternative approach of selecting a specific parametric distribution is often too rigid for modelling the tail behaviour of general distributions.

With an appropriate selection of the scaling distribution, our method can be used not only to derive estimates for the unique EVT shape parameter, but will also provide an accurate description of the data in its whole range. Our method makes use of the full data set and it is not necessary to make subjective decisions concerning threshold levels or the block sizes as required by EVT methods. Finally it should be remarked that our approach allows for fitting NPH distributions having asymptotically equivalent tails with respect to a fixed distribution. For instance, the modeller can fit the tail using another competing (classical) approach, then fix the appropriate parameters of the scaling distribution and thereafter run the NPH method, thus obtain a good fit in the body of the data while maintaining its tail.

Apart from providing an adequate description of heavy–tailed data, an NPH distribution having an unbounded and discrete scaling distribution can also be seen as infinite–dimensional phase–type distribution, which for all intents and purposes is as tractable as its finite–dimensional counterpart. Much of the machinery available for finite–dimensional phase–type distributions is also applicable to the whole NPH class. For instance, algorithms are available for the exact calculations of quantities related to renewal theory, random walks (ladder processes) and ruin probabilities (see [6] for details).

The interpretation of an NPH distribution as an infinite–dimensional phase–type distribution will allow for the approximation of the maximum likelihood estimators via the EM algorithm and the estimation is hence considered classical. The implementation of the EM algorithm is carried out in a similar way as for finite–dimensional phase–type distributions [2]; in addition, it will also be possible to adapt the EM algorithm for approximating theoretical distributions as well as for fitting (right-, left- or interval-) censored data [15] . However, new challenges arise from the presence of the scaling component. Firstly, the estimation cannot be dealt as a simple extension of the EM algorithm for finite–dimensional PH distributions to infinite dimensions. Neither a cut–off in the number of dimensions would be a satisfactory solution because this would need to be fixed in advance and would be equivalent to light–tailed estimation. We shall see, however, that because the scaling distribution is selected from a parametric family, it is then possible to reduce the expressions for the estimators to simple (one–dimensional) infinite series which can be numerically evaluated up to any specified precision. Secondly, the estimation of the vector of parameters θ of the scaling distribution will depend on the particular choice made by the modeller, which in many cases requires a (non-trivial) numerical optimization in each iteration of the EM algorithm. The selection of an adequate scaling distribution can result in a simpler optimization of the log-likelihood function, as well as in reducing the required number of terms of the infinite series.

Further improvements in numerical performance of the algorithm are also described in this paper. We show that the implementation of the EM algorithm essentially requires the empirical cumulative distribution function associated to the heavy–tailed data, so a significant increase in speed may be obtained by implementing certain data–reduction techniques (data binning/bucketing) with minimal loss of information. We also simplify various expressions which have their analogues in the original algorithm [2] and employ alternative methods to compute matrix exponentials in a more efficient way (see Remark 3.1).

The rest of the paper is organized as follows. In Section 2 we provide an overview of the PH and NPH classes of distributions making some emphasis on the characteristics of the scaling distribution. In Section 3 we develop the main algorithm for estimating independent and identically distributed data sampled from an NPH distribution. In Section 4 we adapt the proposed method to cope with the presence of left-, right- or interval censored data. Section 5 provides an EM–algorithm for adjusting an NPH distribution to a theoretical distribution function F, which in turn is equivalent to finding the distribution in the NPH class with a given order which minimizes the Kullback–Leibler distance to F. In Section 6 we provide some numerical examples from both simulated data, real data and fits to a theoretical distribution. In there, we highlight that with an

adequate selection of the scaling distribution we can accurately and efficiently fit any general Regularly Varying distribution as well as any Weibullian distribution.

2. PHASE–TYPE DISTRIBUTIONS AND THE NPH CLASS

Before proceeding with a more detailed account of the statistical approach for heavy–tailed data proposed here, we first establish some notation and provide some background on PH (phase–type) distributions and the extended class NPH of scale mixtures of phase–type distributions.

We follow the convention that matrices are written in bold capital letters from the Roman alphabet and their elements are represented with the corresponding minuscules (i.e. $\boldsymbol{A} = \{a_{ij}\}$). In this paper, bold minuscule letters from the Greek alphabet $(\boldsymbol{\alpha}, \boldsymbol{\pi})$ will represent row vectors, while bold minuscule letters from the Roman alphabet $(\boldsymbol{t}, \boldsymbol{e})$ will represent column vectors. Elements of a vector are denoted with the same letter in plain style (i.e. $\boldsymbol{t} = \{t_i\}$). In that way, the dimensions of both matrices and vectors will become clear from the context and left unspecified unless stated otherwise.

The class PH of phase–type distributions consists of distributions of (random) times until a finite state Markov jump process exits a set of transient states. This can be made precise by letting $\mathscr{E} = \{1, 2, ..., p, p+1\}$ denote the state–space of a Markov jump process $\{X_t\}_{t \geq 0}$, where states $1, 2, ..., p$ are transient and $p+1$ is absorbing. The intensity matrix for $\{X_t\}_{t \geq 0}$ can then be written on the form

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

where $\boldsymbol{T}$ is a $p \times p$ sub–intensity matrix, $\boldsymbol{t}$ is a p–dimensional column vector and $\boldsymbol{0}$ the p–dimensional row vector of zeroes. We set $\alpha_i = \mathbb{P}(X_0 = i)$, $\sum_{i=1}^{p} \alpha_i = 1$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)$. Then we say that

$$\tau = \inf\{s > 0 : X_s = p+1\}$$

has a PH (phase–type) distribution with representation $\mathrm{PH}_p(\boldsymbol{\alpha}, \boldsymbol{T})$. Since $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$, where $\boldsymbol{e}$ denotes the column vector of ones, the distribution of τ is fully specified in terms of $\boldsymbol{\alpha}$ and $\boldsymbol{T}$. For further background on PH distributions we refer e.g. to [4], [5], [12] or [14].

The PH class of distributions is widely used in the area of Applied Probability, where they often provide exact (or even explicit) solutions in complex stochastic models. This is for example the case for ruin probabilities in risk theory or steady–state waiting time distributions for queues. Any distribution with support on the positive real numbers may be approximated arbitrarily close by a PH distribution. In spite of this denseness property, PH distributions are all light tailed and consequently any approximating (finite–dimensional) PH distribution will not be able to capture a possible heavy tailed behaviour.

In [6] the dense class, NPH, of genuinely heavy tailed distributions is proposed in terms of infinite–dimensional phase–type distributions. The idea is very simple and goes as follows. Let N be a discrete random variable supported over $\{s_i : i \in \mathbb{N}, s_i > 0\}$ with distribution $\pi_i = \mathbb{P}(N = s_i)$, and $\tau \sim \mathrm{PH}_p(\boldsymbol{\alpha}, \boldsymbol{T})$ be independent of N. We say that the random variable $Y := N \cdot \tau$ has an NPH distribution with parameters $\boldsymbol{\alpha}$, $\boldsymbol{T}$ and $\boldsymbol{\pi} = \{\pi_i\}_{i \geq 1}$ (the vector of probabilities characterising the distribution of N), and we write

$$Y \sim \mathrm{NPH}_p(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{T}).$$

While the support of $\boldsymbol{\pi}$ is important, we will not denote it explicitly in the parametrisation of Y. The cdf of an $\mathrm{NPH}_p(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{T})$ distribution can be written as the Mellin-Stieltjes convolution of a cdf $G_\tau \sim \mathrm{PH}_p(\boldsymbol{\alpha}, \boldsymbol{T})$ and the distribution of N, i.e.,

$$F_Y(y) = \sum_{i=1}^{\infty} \pi_i \cdot G_\tau(y/s_i) = \sum_{i=1}^{\infty} \pi_i \cdot \boldsymbol{\alpha} e^{\boldsymbol{T} y/s_i} \boldsymbol{e}, \qquad y > 0.$$

Observe from the above that $\boldsymbol{T}$ is a scale parameter of the finite–dimensional PH distribution, so the distribution $\text{NPH}_p(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{T})$ can be seen as a scale mixture of phase–type distributions, and for this reason we often refer to N as the scaling random variable and $\boldsymbol{\pi}$ as the scaling distribution. An NPH distribution is absolutely continuous with density

$$f_Y(y) = \sum_{i=1}^{\infty} \pi_i \cdot \boldsymbol{\alpha} e^{\boldsymbol{T} y/s_i} \boldsymbol{t}/s_i = \sum_{i=1}^{\infty} \pi_i \cdot \boldsymbol{\alpha} e^{\boldsymbol{T}_i y/s_i} \boldsymbol{t}_i, \qquad y > 0,$$

where $\boldsymbol{T}_i = \boldsymbol{T}/s_i$ and $\boldsymbol{t}_i = -\boldsymbol{T}_i \boldsymbol{e}$, the latter corresponding to the exit rate vector at level i. Sampling from $\text{NPH}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{T})$ is also simple as we first draw an index i (referred to as the *level*) from $\boldsymbol{\pi}$ and then a τ which will be simulated as a PH random variable from $\text{PH}_p(\boldsymbol{\alpha}, \boldsymbol{T}/s_i)$.

The distribution of Y can also be interpreted as an infinite dimensional phase–type distribution since we can write

$$f_Y(y) = (\boldsymbol{\pi} \otimes \boldsymbol{\alpha})\, e^{\boldsymbol{\Gamma} y} \boldsymbol{\gamma},$$

where

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{T}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{T}_2 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{T}_3 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{T}_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} = \begin{pmatrix} \boldsymbol{T}/s_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{T}/s_2 & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{T}/s_3 & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{T}/s_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and $\boldsymbol{\gamma} = -\boldsymbol{\Gamma}\boldsymbol{e}$. Here $\boldsymbol{e}$ is now the infinite–dimensional column vector of ones; we notice that the exponential of the infinite dimensional matrix $\boldsymbol{\Gamma}$ is well defined since it is a *bounded operator* (because the sequence $\{s_i : s_i > 0\}$ is bounded away from zero).

In what follows, we will concentrate on NPH distributions with scaling distributions that belong to some parametric family of nonnegative discrete distributions. We will write $\pi_i := \pi_i(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters of finite dimension. In that way, a distribution $\text{NPH}(\boldsymbol{\pi}(\boldsymbol{\theta}), \boldsymbol{\alpha}, \boldsymbol{T})$ will have a finite number of parameters.

The quality of the estimation method to be proposed below (in terms of accuracy and efficiency) will depend on the adequate choice of the scaling distribution. The scaling distribution not only determines if the NPH distribution is light- or heavy–tailed (recall that if N is unbounded, then Y has a heavy–tailed distribution [cf. 18]), but will also determine the type of tail behaviour that the NPH distribution will have. It is because of this feature that the NPH class contains a very rich variety of heavy–tailed distributions. Below we consider two important general cases:

Example 2.1 (Regularly Varying distributions)**.** In the regularly varying case, Breiman's lemma implies that if the scaling random variable N has a regularly varying distribution with tail index $-\xi < 0$, then the distribution of $Y := N \cdot X$ will also be regularly varying with the same index. Moreover, the exact asymptotic tail behaviour is then given by

$$\mathbb{P}(Y > t) = \mathbb{E}[\tau^{\xi}]\mathbb{P}(N > t)(1 + o(1)), \qquad t \to \infty.$$

Therefore, if data are assumed to have been drawn from some Regularly Varying distribution, then we should select a discrete distribution which is regularly varying. This strategy will also be effective if we choose a scaling distribution which is tail-equivalent. Recall that Regularly Varying distributions with tail index $-\xi$ together with their tail equivalent distributions correspond to the Frechét maximum domain of attraction with shape parameter ξ. Hence an estimator for the parameter index ξ in our model can also be taken as an estimator for the extreme value shape parameter.

Example 2.2 (Weibullian Distributions)**.** A similar feature holds true for the the class of Weibullian distributions [1] since it is a closed class under Mellin-Stieltjes convolution (multiplication of

independent random variable). Such a class is defined as the collection of nonnegative distributions having survival function

$$\overline{F}(x) = x^{\delta} e^{-(\lambda x)^p} (C + \mathrm{o}(1)), \qquad x > 0, \lambda, p > 0, \delta \in \mathbb{R}. \tag{1}$$

A Weibullian distribution is heavy–tailed iff its parameter $p \in (0,1)$ (and light–tailed otherwise). Since PH distributions are Weibullian with parameter $p = 1$, then Lemma 2.1 of [1] implies that if the scaling distribution $\boldsymbol{\pi}$ is Weibullian with parameter p then $Y \sim \mathrm{NPH}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{T})$ is Weibullian with parameter $p/(1+p)$. The exact tail asymptotics is also known and given in [1] but the expressions are somewhat complicated and thus omitted here.

In practice, we will frequently use scaling distributions constructed as discretizations of some continuous distribution H over a countable set of nonnegative numbers. For instance, a discretization of H over an arithmetic progression with step length Δ is such that $s_i = i\Delta$ and $\pi_i = H(i\Delta) - H((i-1)\Delta)$, $i = 1, 2, \ldots$. A discretization over a geometric progression with initial value $s_1 > 0$ and ratio r is such that $s_i = s_1 r^{i-1}$, so $\pi_1 = H(s_1)$ and $\pi_{i+1} = H(s_1 r^i) - H(s_1 r^{i-1})$. The latter can be seen as an exponential transformation of a distribution supported over a regular lattice with step length $\Delta = \log r$. Such a selection is quite natural because heavy–tailed distributions often arise as exponential transformation of well-known light tailed distributions. Furthermore, if we use a geometric progression the estimation will be more efficient because in this way an accurate description of the tail is obtained through a smaller number of terms π_i's.

3. Estimation

In this section we address the problem of fitting an NPH distribution to a data set. We assume that $y_1, y_2, ..., y_M$ form an i.i.d. data set sampled from $\mathrm{NPH}_p(\boldsymbol{\pi}(\boldsymbol{\theta}), \boldsymbol{\alpha}, \boldsymbol{T})$ for some fixed p and where $\boldsymbol{\pi}(\boldsymbol{\theta})$ is some parametric nonnegative discrete distribution of N. We assume that the support for $\boldsymbol{\pi}(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. We shall estimate the parameters $\boldsymbol{\theta}, \boldsymbol{\alpha}$ and $\boldsymbol{T}$.

Hence, with probability $\pi_i(\boldsymbol{\theta})$, y_n is the realization of the i–th level phase–type distribution $\mathrm{PH}_p(\boldsymbol{\alpha}, \boldsymbol{T}/s_i)$, but both the level i and the actual sample path of the underlying Markov jump process are unobservable. In this context we may consider the data as being incomplete, and a standard method for maximizing the (incomplete data) likelihood is via the EM algorithm. To this end we must first attend the calculation of the likelihood function for complete data.

Assume that apart from $y_1, ..., y_M$ we have also observed all sample paths of the underlying Markov jump process and their levels. Let I_n denote the level of the phase–type distribution of the n'th sample path, so that

$$L^i = \sum_{n=1}^{M} 1\{I_n = i\}$$

equals the number of i–level sample paths in the data. Next consider the Markov jump process $\{J_u^{(n)}\}_{u \geq 0}$ underlying the n'th phase–type distribution (which generates the data y_n) and let

$$B_k^i = \sum_{n=1}^{M} 1\{J_0^{(n)} = k, I_n = i\}$$

be the total number of i–level sample paths initiated in state k. Define

$$Z_k^i = \sum_{n=1}^{M} \int_0^{y_n} 1\{J_u^{(n)} = k, I_n = i\} du,$$

which is the total time all underlying i–level sample paths spend in state k and let $N_{k\ell}^i$ denote the total count of jumps from state k to ℓ within all i–level sample paths. Finally, let N_k^i be the total number of i–level sample paths that exit to the absorbing state from state k.

Then the complete data likelihood is easily seen to be (see e.g. [2] or [5] for further comments on this)

$$L_c(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T}) = \prod_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})^{L^i}\prod_{k=1}^{p}\alpha_k^{B_k^i}\prod_{\substack{k,\ell=1\\ \ell\neq k}}^{p}\left(\frac{t_{k\ell}}{s_i}\right)^{N_{k\ell}^i}\exp\left(-\frac{t_{k\ell}}{s_i}Z_k^i\right)\prod_{k=1}^{p}\left(\frac{t_k}{s_i}\right)^{N_k^i}\exp\left(-\frac{t_k}{s_i}Z_k^i\right),$$

where L^i, B_k^i, $N_{k\ell}^i$, N_k^i and Z_k^i are the (unobserved) sufficient statistics, and $\boldsymbol{\alpha}=(\alpha_1,...,\alpha_p)$, $\boldsymbol{T}=\{t_{ij}\}_{i,j=1,...,p}$ are the parameters of a representation for $\tau\sim \mathrm{PH}_p(\boldsymbol{\alpha},\boldsymbol{T})$, and $\boldsymbol{t}=(t_1,...,t_p)'=-\boldsymbol{T}\boldsymbol{e}$.

The corresponding log–likelihood is thus given by

$$\begin{aligned}\ell_c(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T}) &= \sum_{i=1}^{\infty}L^i\log\pi_i(\boldsymbol{\theta})+\sum_{i=1}^{\infty}\sum_{k=1}^{p}B_k^i\log\alpha_k+\sum_{i=1}^{\infty}\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{p}N_{k\ell}^i\log\left(\frac{t_{k\ell}}{s_i}\right)\\ &\quad -\sum_{i=1}^{\infty}\sum_{\substack{k,\ell=1\\ \ell\neq k}}^{p}\frac{t_{k\ell}}{s_i}Z_k^i+\sum_{i=1}^{\infty}\sum_{k=1}^{p}N_k^i\log\left(\frac{t_k}{s_i}\right)-\sum_{i=1}^{\infty}\sum_{k=1}^{p}\frac{t_k}{s_i}Z_k^i \qquad (2)\end{aligned}$$

with the convention that $0\cdot\log(0)=0$ (or, equivalently, a reduction of the state–space if $t_{k\ell}$ is assumed to be zero, which will happen in sub–models such as generalized Erlang or Coxian).

We now turn to the EM–algorithm for maximizing the incomplete data likelihood. Let $f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})$ denote the density function of $Y\sim \mathrm{NPH}_p(\boldsymbol{\pi}(\boldsymbol{\theta}),\boldsymbol{\alpha},\boldsymbol{T})$. The EM–algorithm then maximizes the incomplete data likelihood

$$L(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T};\boldsymbol{y})=\prod_{k=1}^{M}f_Y(y_k;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T}),$$

by using the complete data likelihood L_c (or equivalently the log-likelihood ℓ_c) in the following way.

0: Initialize with some "arbitrary" $(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\boldsymbol{T}_0)$ and let $n=0$.

1: (E–step) Calculate the function

$$h:(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})\to\mathbb{E}_{(\boldsymbol{\theta}_n,\boldsymbol{\alpha}_n,\boldsymbol{T}_n)}\left(\ell_c(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})|\boldsymbol{Y}=\boldsymbol{y}\right).$$

2: (M–step) Let $(\boldsymbol{\theta}_{n+1},\boldsymbol{\alpha}_{n+1},\boldsymbol{T}_{n+1}):=\mathrm{argmax}_{(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}h(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})$.

3: n=n+1; GOTO 1.

In each iteration, the incomplete likelihood is increased (i.e. $L(\boldsymbol{\theta}_n,\boldsymbol{\alpha}_n,\boldsymbol{T}_n)\leq L(\boldsymbol{\theta}_{n+1},\boldsymbol{\alpha}_{n+1},\boldsymbol{T}_{n+1})$), which hence implies that the procedure converges (possibly to a local maximum or saddlepoint though).

From the actual form of the log–likelihood (2) we see that it is a linear function of the sufficient statistics, and the conditional expected value of the log–likelihood given the data will then be the log–likelihood function with the sufficient statistics replaced by their conditional expectations given the data. Hence the calculation of the h function is straightforward since we only have to replace the unobservable sufficient statistics in the complete log–likelihood and plug instead their respective conditional expectations given the data.

E–step: First we consider one (generic) data point ($M=1$) and let $y=y_1$. We need to calculate the conditional expected values of the sufficient statistics given $Y=y$. All distributions and expectations are under $\mathbb{E}_{\boldsymbol{\theta}_n,\boldsymbol{\alpha}_n,\boldsymbol{T}_n}$, but we will omit the index in order to ease the exposition.

Concerning the total number of i-level paths L^i, this is equal to one if i is the chosen level and zero otherwise. Let I denote the random variable indicating the chosen level. Then

$$\begin{aligned}\mathbb{E}(L^i|Y\in dy) &= \mathbb{P}(I=i|Y\in dy)\\ &= \frac{\mathbb{P}(Y\in dy|I=i)\mathbb{P}(I=i)}{\mathbb{P}(Y\in dy)}\end{aligned}$$

$$= \frac{\pi_i(\boldsymbol{\theta})\boldsymbol{\alpha}\exp(y\boldsymbol{T}_i)\boldsymbol{t}_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} = \pi_i(\boldsymbol{\theta})\frac{\boldsymbol{\alpha}\exp(y\boldsymbol{T}s_i)\boldsymbol{t}/s_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}. \tag{3}$$

Similarly, for total number of i-level paths started in state k, i.e. B_k^i we have

$$\begin{aligned}
\mathbb{E}(B_k^i|Y\in dy) &= \mathbb{E}(1\{J_0=k,I=i\}|Y\in dy)\\
&= \frac{\mathbb{P}(Y\in dy|J_0=k,I=i)\mathbb{P}(I=i)\mathbb{P}(J_0=k)}{\mathbb{P}(Y\in dy)}\\
&= \pi_i(\boldsymbol{\theta})\frac{\alpha_k\boldsymbol{e}_k'\exp(y\boldsymbol{T}_i)\boldsymbol{t}_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} = \pi_i(\boldsymbol{\theta})\frac{\alpha_k\,\boldsymbol{e}_k'\exp(y\boldsymbol{T}/s_i)\boldsymbol{t}/s_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}.
\end{aligned}$$

Regarding Z_k^i,

$$\begin{aligned}
\mathbb{E}\left(Z_k^i|Y\in dy\right) &= \mathbb{E}\left(1\{I=i\}\int_0^Y 1\{J_u=k\}du\middle|Y\in y\right)\\
&= \frac{\mathbb{E}\left(1\{Y\in dy\}\int_0^\infty 1\{J_u=k,u<Y\}du\middle|I=i\right)\mathbb{P}(I=i)}{\mathbb{P}(Y\in dy)}\\
&= \mathbb{P}(I=i)\frac{\int_0^\infty \mathbb{P}(Y\in dy,J_u=k,Y<u|I=i)du}{\mathbb{P}(Y\in dy)}\\
&= \pi_i(\boldsymbol{\theta})\frac{\int_0^\infty \mathbb{P}(Y\in dy,u<Y|J_u=k,I=i)\mathbb{P}(J_u=k|I=i)du}{\mathbb{P}(Y\in dy)}\\
&= \pi_i(\boldsymbol{\theta})\frac{\int_0^\infty 1\{u<y\}\boldsymbol{e}_k'e^{(y-u)\boldsymbol{T}_i}\boldsymbol{t}_i\boldsymbol{\alpha}e^{u\boldsymbol{T}_i}\boldsymbol{e}_k du}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}\\
&= \pi_i(\boldsymbol{\theta})\frac{\int_0^y \boldsymbol{e}_k'e^{(y-u)\boldsymbol{T}/s_i}(\boldsymbol{t}/s_i)\boldsymbol{\alpha}e^{u\boldsymbol{T}/s_i}\boldsymbol{e}_k du}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}.
\end{aligned}$$

Similar calculations yield that

$$\mathbb{E}(N_{k\ell}^i|Y=y) = \pi_i(\boldsymbol{\theta})\frac{t_{k\ell}}{s_i}\frac{\int_0^y \boldsymbol{e}_\ell'e^{(y-u)\boldsymbol{T}/s_i}(\boldsymbol{t}/s_i)\boldsymbol{\alpha}e^{u\boldsymbol{T}/s_i}\boldsymbol{e}_k du}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}$$

and

$$\mathbb{E}(N_k^i|Y=y) = \pi_i(\boldsymbol{\theta})\frac{t_k}{s_i}\frac{\boldsymbol{\alpha}e^{y\boldsymbol{T}/s_i}\boldsymbol{e}_k}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}.$$

In general, for $M>1$ datapoints we simply sum the previous formulas with arguments y_j, $j=1,...,M$. Indeed, by independence of the data, the log–likelihood function will be a sum of log–likelihoods for single data points and hence the expected values of the sufficient statistics given data will be the sum of the sufficient statistics as given above for individual data points.

Remark 3.1. We see that the formulas in the E–step involve both matrix–exponentials and integrals thereof, and by defining the generic integral

$$\boldsymbol{J}(y;\boldsymbol{\alpha},\boldsymbol{T}) = \int_0^y e^{(y-u)\boldsymbol{T}}\boldsymbol{t}\boldsymbol{\alpha}e^{u\boldsymbol{T}}du, \tag{4}$$

we have that (see [19])

$$\exp\left(\begin{pmatrix} \boldsymbol{T} & \boldsymbol{t\alpha} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix} y\right) = \begin{pmatrix} e^{\boldsymbol{T}y} & \boldsymbol{J}(y;\boldsymbol{\alpha},\boldsymbol{T}) \\ \boldsymbol{0} & e^{\boldsymbol{T}y} \end{pmatrix}. \tag{5}$$

Thus a simple (and numerically efficient) way of obtaining $\boldsymbol{J}(y;\boldsymbol{\alpha},\boldsymbol{T})$ is by calculating the matrix exponential on the left hand side.

M-step. Next we determine the point $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{T}})$ that maximizes the h–function (the conditional expectation of the complete log–likelihood given the data). The estimation of $\boldsymbol{\theta}$ should be treated on a case-by-case basis, while for the parameters α_i we shall use Lagrange multipliers due to the constraints on them summing up to one. In the case of the non–diagonal elements $t_{k\ell}$ and the absorption rates t_k we will calculate these in a straightforward way by obtaining the first order derivatives and equating them to zero.

For the parameter of the scaling distribution we write

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{\infty} \mathbb{E}[L^i|\boldsymbol{Y}=\boldsymbol{y}] \log \pi_i(\boldsymbol{\theta}),$$

so it depends on the particular form of the discrete distribution $\boldsymbol{\pi}(\boldsymbol{\theta})$ whether it can be estimated explicitly or numerically. We shall consider some particular examples later on.

As for the parameters of the phase–type component we first address the estimator for the parameter $\boldsymbol{\alpha}$. Consider the Lagrange function

$$M(\boldsymbol{\alpha}) = \sum_{i=1}^{\infty} \mathbb{E}[B_k^i|\boldsymbol{Y}=\boldsymbol{y}] \log \alpha_k + \mu(1-\sum_k \alpha_k),$$

where μ is a Lagrange multiplier. Then

$$\frac{\partial M}{\partial \alpha_k} = \sum_{i=1}^{\infty} \frac{\mathbb{E}[B_k^i|\boldsymbol{Y}=\boldsymbol{y}]}{\alpha_k} - \mu = 0,$$

which result in

$$\alpha_k \mu = \sum_{i=1}^{\infty} \mathbb{E}[B_k^i|\boldsymbol{Y}=\boldsymbol{y}].$$

Summing over k yields

$$\mu = \sum_{i=1}^{\infty}\sum_{k=1}^{p} \mathbb{E}[B_k^i|\boldsymbol{Y}=\boldsymbol{y}] = M$$

so

$$\hat{\alpha}_k = \frac{1}{M}\sum_{i=1}^{\infty} \mathbb{E}[B_k^i|\boldsymbol{Y}=\boldsymbol{y}].$$

Next we consider the non-elements of $t_{k\ell}$, $k \neq \ell$. For each entry we take partial derivatives and equate to 0

$$\frac{\partial h}{\partial t_{k\ell}} = \sum_{i=1}^{\infty} \mathbb{E}[N_{k\ell}^i|\boldsymbol{Y}=\boldsymbol{y}]\frac{1}{t_{k\ell}} - \sum_{i=1}^{\infty} \frac{\mathbb{E}[Z_k^i|\boldsymbol{Y}=\boldsymbol{y}]}{s_i} = 0$$

implying

$$\hat{t}_{k\ell} = \frac{\sum_{i=1}^{\infty} \mathbb{E}[N_{k\ell}^i|\boldsymbol{Y}=\boldsymbol{y}]}{\sum_{i=1}^{\infty} \mathbb{E}[Z_k^i|\boldsymbol{Y}=\boldsymbol{y}]/s_i}.$$

Similarly,

$$\hat{t}_{k} = \frac{\sum_{i=1}^{\infty} \mathbb{E}[N_{k}^i|\boldsymbol{Y}=\boldsymbol{y}]}{\sum_{i=1}^{\infty} \mathbb{E}[Z_k^i|\boldsymbol{Y}=\boldsymbol{y}]/s_i}.$$

We then obtain the estimator for the diagonal elements of the matrix $\boldsymbol{T}$ by

$$\hat{t}_{kk} = -\sum_{\ell \neq k} \hat{t}_{k\ell} - \hat{t}_k.$$

The EM–algorithm can now be stated as follows.

Algorithm 3.2 (EM–algorithm)**.**

0: *Initialize with some "arbitrary"* $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{T})$.

1: *The E-step and the M-step together yield the following estimators.*

$$\begin{aligned}
\widehat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}^*} \sum_{i=1}^{\infty} \sum_{j=1}^{M} \pi_i(\boldsymbol{\theta}) \frac{\boldsymbol{\alpha} \exp(y_j \boldsymbol{T}/s_i)\boldsymbol{t}/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} \log \pi_i(\boldsymbol{\theta}^*) \\
\hat{\alpha}_k &= \frac{1}{M} \sum_{i=1}^{\infty} \sum_{j=1}^{M} \pi_i(\boldsymbol{\theta}) \frac{\alpha_k \boldsymbol{e}_k' \exp(y_j \boldsymbol{T}/s_i)\boldsymbol{t}/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} \\
\hat{t}_{k\ell} &= \frac{\sum_{i=1}^{\infty} \sum_{j=1}^{M} \pi_i(\boldsymbol{\theta}) \frac{\boldsymbol{J}_{\ell k}(y_j;\boldsymbol{\alpha},\boldsymbol{T}/s_i) \cdot t_{k\ell}/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}}{\sum_{i=1}^{\infty} \sum_{j=1}^{M} \pi_i(\boldsymbol{\theta}) \frac{\boldsymbol{J}_{kk}(y_j;\boldsymbol{\alpha},\boldsymbol{T}/s_i)/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}}, \quad k \neq \ell \\
\hat{t}_k &= \frac{\sum_{i=1}^{\infty} \sum_{j=1}^{M} \pi_i(\boldsymbol{\theta}) \frac{\boldsymbol{\alpha} e^{\boldsymbol{T} y_j/s_i} \boldsymbol{e}_k t_k/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}}{\sum_{i=1}^{\infty} \sum_{j=1}^{M} \pi_i(\boldsymbol{\theta}) \frac{\boldsymbol{J}_{kk}(y_j;\boldsymbol{\alpha},\boldsymbol{T}/s_i)/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}}.
\end{aligned}$$

Assign the diagonal values

$$\hat{t}_{kk} = -\sum_{l \neq k} \hat{t}_{k\ell} - \hat{t}_k.$$

2: *Reassign values to initial parameters*

$$\begin{aligned}
\boldsymbol{\theta} &:= \widehat{\boldsymbol{\theta}} \\
\boldsymbol{\alpha} &:= \widehat{\boldsymbol{\alpha}} \\
\boldsymbol{T} &:= \widehat{\boldsymbol{T}}
\end{aligned}$$

3: *GOTO 1.*

The EM–estimators for the NPH class (only) differ from their analogues for the PH class in the presence of the scaling component. The estimators above are written as infinite series which can be interpreted in terms of expected values associated to the EM estimators for scaled PH distributions (seen as functions of the random scaling).

In practice we can only compute these infinite series up to a finite number of terms. It is then desirable to compute enough terms so the probability $\mathbb{P}(N > i)$ is smaller than the computational numerical precision, hence implying that the error becomes negligible. A natural trade–off between precision and speed then ensues because computing a larger number of terms will come at a high computational cost. Hence it is convenient to choose a scaling distribution $\pi_i(\boldsymbol{\theta})$ that converges to 0 rather quickly as $i \to \infty$ but still provides a good description of the tail behaviour. A solution proposed here is to consider distributions supported over geometric progressions as these will converge fast, thus achieving precision with a few terms while still preserving the shape of the tail. Below is an important example for the Regularly Varying case.

Example 3.3. Here we present an important special case for the choice of the scaling distribution $\{\pi_i(\boldsymbol{\theta})\}$ of N for the Regularly Varying case. Consider a Pareto distribution with parameter θ and supported over $(1,\infty)$, so its cdf distribution is given by $H(t) = 1 - t^{-\theta}$, $t > 1$. We consider a scaling distribution as the discretization of H over a geometric progression with initial value $s_1 = 1$ and ratio e^c, $c > 0$. That is

$$\begin{aligned}\pi_i(\boldsymbol{\theta}) = \mathbb{P}(N = s_i) = H(s_{i+1}) - H(s_i) &= \frac{1}{s_i^{\theta}} - \frac{1}{s_{i+1}^{\theta}} \\ &= (e^{-\theta c})^{(i-1)}(1 - e^{-\theta c}), \qquad i = 1,2,\ldots\end{aligned}$$

The distribution above also corresponds to the exponential transformation e^{cW} of a geometric random variable W with parameter $e^{-\theta c}$. Notice that the probability $\pi_i(\boldsymbol{\theta})$ will decay exponentially fast to 0 as $i \to \infty$, so we will require to compute fewer terms to obtain estimators with negligible numerical errors while the cdf of $\boldsymbol{\pi}(\boldsymbol{\theta})$ will remain close to H specially in the tail region. A further advantage of selecting a scaling distribution of this form is that it features an explicit solution to the M–step maximization problem of the function

$$\boldsymbol{\theta} \to \sum_{i=1}^{\infty} \log \pi_i(\boldsymbol{\theta}) w_i,$$

where $w_i = \mathbb{E}(L_i | \boldsymbol{Y} = \boldsymbol{y})$. It is not difficult to see that such an explicit solution is given by

$$\widehat{\theta} = -\frac{1}{c} \log \left(1 - \frac{\sum_{i=1}^{\infty} w_i}{\sum_{i=1}^{\infty} i w_i} \right) = -\frac{1}{c} \log \left(1 - \frac{M}{\sum_{i=1}^{\infty} i w_i} \right),$$

where M is the sample size. We shall remark however, that since the distances between consecutive points in the support are unbounded, then the scaling distribution suggested above is no longer regularly varying in the strict sense and Breiman's lemma cannot be applied. Nevertheless, the tail probability of $\boldsymbol{\pi}(\boldsymbol{\theta})$ will oscillate between two regularly varying distributions with the same index $\boldsymbol{\theta}$, so this model still provides an accurate approximation to any regularly varying distribution (see [18]).

The computational effort necessary to obtain the EM estimators for the NPH class will also depend on the sample size. Below we discuss data-reduction techniques that can be used to speed–up the algorithm without loosing important information in the tail.

Remark 3.4. Suppose that there are repeated values among the data points such that there are $\tilde{y}_1, \ldots, \tilde{y}_D$ different values and that $\tilde{y}_i$ appears $M_i \geq 1$ times in the original data. Then $M_1 + \cdots M_D = M$ and the sums over $j = 1, \ldots, M$ in expected values can then be reduced to weighted sums of fewer terms instead. For example,

$$\mathbb{E}\left(Z_k^i | \boldsymbol{Y} = \boldsymbol{y}\right) = \pi_i(\boldsymbol{\theta}) \sum_{j=1}^{D} M_j \frac{\boldsymbol{J}(\tilde{y}_j; \boldsymbol{\alpha}, \boldsymbol{T}/s_i)_{kk}}{f_Y(\tilde{y}_j; \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{T})}.$$

Hence, Algorithm 3.2 can be used to estimate the parameters of an NPH distribution when data are represented by a histogram rather than the raw data.

Remark 3.5. The computational burden can also be reduced if we use *data binning* (*bucketing*) on those regions of the support where data gather rather densely. This strategy is particularly effective for heavy–tailed data, which typically concentrates on some interval of the distribution while it is more scarce in the "tail".

We split the support into D disjoint sub–intervals $[T_j, T_{j+1})$, $j = 0, 1, \ldots, D-1$ where $0 = T_0 < T_1 < \cdots < T_D = \infty$ and let M_j denote the number of data points falling into $[T_j, T_{j+1})$. Then we use the repeated data reduction of Remark 3.4 by treating the average of the data points falling in the j-th bin, $j = 0, ..., D-1$ as data points with counts M_j. In regions where scattering of the data is too diffuse (for instance in the tail-regions), bins will typically contain at most one data point so the average will correspond to the raw data, so there will be minimal loss of information. Alternatively, we might choose any point in $[T_j, T_{j+1})$ as a representative, including any data point falling in this interval. Notice however that if we choose the left points of the intervals T_j as representatives, we may need to make special arrangements regarding the first interval in order not to provoke an atom at zero.

4. CENSORING

In certain situations some data may be censored. We say a data point is right–censored at y_* if it takes a value above y_* but unknown; it is left–censored at y^* if it is below y^* but the exact value is unknown, and it is interval-censored if it is contained in the interval $(y_*, y^*]$ but its exact value is unknown. Left–censoring is a special case of interval–censoring with $y_* = 0$ while right–censoring can be obtained by fixing y_* and letting $y^* \to \infty$.

The EM algorithm works entirely in the same way as for uncensored data with the only difference that we are no longer observing a data point $Y = y$ but $Y \in (y_*, y^*]$. Formulas for right censoring will, however, appear as a part of the derivation of interval censoring. This will only change the E–step where we now have to calculate the conditional expectations:

$$\mathbb{E}(L^i | Y \in (y_*, y^*]),\ \mathbb{E}(B_k^i | Y \in (y_*, y^*]),\ \mathbb{E}(Z_k^i | Y \in (y_*, y^*]),\ \mathbb{E}(N_{k\ell}^i | Y \in (y_*, y^*]),\ \mathbb{E}(N_k^i | Y \in (y_*, y^*]).$$

We consider first the case of a single data point y taken from a realization of $Y \sim \mathrm{NPH}_p(\pi(\boldsymbol{\theta}), \boldsymbol{\alpha}, \boldsymbol{T})$. Concerning L^i, notice that

$$\begin{aligned} \mathbb{E}(L^i 1\{Y > y^*\}) &= \mathbb{P}(Y > y^* | I = i)\mathbb{P}(I = i) \\ &= \pi_i(\boldsymbol{\theta})\boldsymbol{\alpha} e^{y^* \boldsymbol{T}/s_i}\boldsymbol{e}. \end{aligned}$$

Thus for interval–censored data we obtain that

$$\begin{aligned} \mathbb{E}(L^i | Y \in (y_*, y^*]) &= \frac{\mathbb{E}(L^i 1\{Y \in (y_*, y^*]\})}{\mathbb{P}(Y \in (y_*, y^*])} \\ &= \frac{\mathbb{E}(L^i 1\{Y > y_*\}) - \mathbb{E}(L^i 1\{Y > y^*\})}{\mathbb{P}(Y \in (y_*, y^*])} \\ &= \frac{\pi_i(\boldsymbol{\theta})(\boldsymbol{\alpha} e^{y_* \boldsymbol{T}/s_i}\boldsymbol{e} - \boldsymbol{\alpha} e^{y^* \boldsymbol{T}/s_i}\boldsymbol{e})}{\sum_{i=1}^{\infty} \pi_i(\boldsymbol{\alpha})(\boldsymbol{\alpha} e^{y_* \boldsymbol{T}/s_i}\boldsymbol{e} - \boldsymbol{\alpha} e^{y^* \boldsymbol{T}/s_i}\boldsymbol{e})}. \end{aligned}$$

For right–censored data we get

$$\mathbb{E}(L^i | Y > y_*) = \frac{\pi_i(\boldsymbol{\theta})\boldsymbol{\alpha} e^{y_* \boldsymbol{T}/s_i}\boldsymbol{e}}{\mathbb{P}(Y > t)} = \frac{\pi_i(\boldsymbol{\theta})\boldsymbol{\alpha} e^{y_* \boldsymbol{T}/s_i}\boldsymbol{e}}{\sum_{i=1}^{\infty} \pi_i(\boldsymbol{\theta})\boldsymbol{\alpha} e^{y_* \boldsymbol{T}/s_i}\boldsymbol{e}}.$$

The rest of the formulas are derived as for censored phase–type distributions (see [15]) conditionally on the level L^i, with parameters $\boldsymbol{\alpha}$, $\boldsymbol{T}_i = \boldsymbol{T}/s_i$, which happens with probability $\pi_i(\boldsymbol{\theta})$. Thus we get that

$$\mathbb{E}(B_k^i | Y \in (y_*, y^*]) = \frac{\pi_i(\boldsymbol{\theta})\left(\alpha_k \boldsymbol{e}_k' e^{y_* \boldsymbol{T}/s_i}\boldsymbol{e} - \alpha_k \boldsymbol{e}_k' e^{y^* \boldsymbol{T}/s_i}\boldsymbol{e}\right)}{\mathbb{P}(Y \in (y_*, y^*])}$$

$$\mathbb{E}(Z_k^i|Y\in(y_*,y^*]) = \frac{\pi_i(\boldsymbol{\theta})\left(\int_{y_*}^{y^*}\boldsymbol{\alpha}e^{u\boldsymbol{T}/s_i}\boldsymbol{e}_k du + \boldsymbol{J}_{kk}(y_*;\boldsymbol{\alpha},\boldsymbol{T}/s_i) - \boldsymbol{J}_{kk}(y^*;\boldsymbol{\alpha},\boldsymbol{T}/s_i)\right)}{\mathbb{P}(Y\in(y_*,y^*])}$$

$$\mathbb{E}(N_{k\ell}^i|Y\in(y_*,y^*]) = \frac{\pi_i(\boldsymbol{\theta})\frac{t_{k\ell}}{s_i}\left(\int_{y_*}^{y^*}\boldsymbol{\alpha}e^{u\boldsymbol{T}/s_i}\boldsymbol{e}_k du + \boldsymbol{J}_{k\ell}(y_*;\boldsymbol{\alpha},\boldsymbol{T}/s_i) - \boldsymbol{J}_{k\ell}(y^*;\boldsymbol{\alpha},\boldsymbol{T}/s_i)\right)}{\mathbb{P}(Y\in(y_*,y^*])}$$

$$\mathbb{E}(N_k^i|Y\in(y_*,y^*]) = \frac{\pi_i(\boldsymbol{\theta})\frac{t_k}{s_i}\int_{y_*}^{y^*}\boldsymbol{\alpha}e^{u\boldsymbol{T}/s_i}\boldsymbol{e}_k du}{\mathbb{P}(Y\in(y_*,y^*])},$$

where $\boldsymbol{J}(\cdot;\boldsymbol{\alpha},\boldsymbol{T})$ is as defined in (4). The integral is calculated similarly as in the uncensored case, namely

$$\int_{y_*}^{y^*}\boldsymbol{\alpha}e^{u\boldsymbol{T}/s_i}\boldsymbol{e_k}du \;=\; s_i\boldsymbol{\alpha}\boldsymbol{T}^{-1}(e^{y_*\boldsymbol{T}/s_i}-e^{y^*\boldsymbol{T}/s_i})\boldsymbol{e}_k.$$

For more than one data point, the data are split into a group of uncensored data and into other groups of different types of censored data. The conditional expectations are then calculated for all data points subject to their group classification, and all conditional expectations of the same kind (jumps, occupation times etc.) are then summed over all data. This amounts to the E–step in an EM algorithm, the rest of which is identical to Theorem 3.2.

Remark 4.1. In Remark 3.5 we suggested the use of data–binning if the amount of data is large. Interval censoring provides a feasible alternative in this same direction.

5. Fitting to a known distribution

It may be of interest to approximate a given (heavy tailed) distribution H with a distribution in the NPH class if for example methods for calculation the ruin probability based on the claim size distribution H are not known. In [2] it was shown how the EM algorithm can be modified in order to approximate phase–type distributions to a given distribution with non–negative support. The EM algorithm then converges to a limit which minimizes the Kullback–Leibler distance between phase–type distributions of a specified order and the target distribution.

The idea is to consider sequences of empirical distributions with increasing sample sizes. These empirical distributions converge almost surely to the target distribution H as $M\to\infty$. Using Algorithm 3.2 and dominated convergence we obtain that

$$\widehat{\alpha}_k = \sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})\frac{1}{M}\sum_{j=1}^{M}\frac{\alpha_k\boldsymbol{e}_k'\exp(\boldsymbol{T}y_j/s_i)\boldsymbol{t}/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} \to \sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})\int_0^{\infty}\frac{\alpha_k\boldsymbol{e}_k'\exp(\boldsymbol{T}y/s_i)\boldsymbol{t}/s_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}dH(y).$$

Similarly it is not difficult to see that for $k\neq\ell$ we have that

$$\widehat{t}_{k\ell} = \frac{\sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})\frac{1}{M}\sum_{j=1}^{M}\frac{\boldsymbol{J}_{\ell k}(y_j;\boldsymbol{\alpha},\boldsymbol{T}/s_i)\cdot t_{k\ell}/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}}{\sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})\frac{1}{M}\sum_{j=1}^{M}\frac{\boldsymbol{J}_{kk}(y_j;\boldsymbol{\alpha},\boldsymbol{T}/s_i)/s_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}} \to \frac{\sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})\frac{t_{k\ell}}{s_i}\int_0^{\infty}\frac{\boldsymbol{J}_{\ell k}(y;\boldsymbol{\alpha},\boldsymbol{T}/s_i)}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}dH(y)}{\sum_{i=1}^{\infty}\frac{\pi_i(\boldsymbol{\theta})}{s_i}\int_0^{\infty}\frac{\boldsymbol{J}_{kk}(y;\boldsymbol{\alpha},\boldsymbol{T}/s_i)}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}dH(y)}$$

and

$$\widehat{t}_k \to \frac{\sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})\frac{t_k}{s_i}\int_0^{\infty}\frac{\boldsymbol{\alpha}e^{\boldsymbol{T}_i y}\boldsymbol{e}_k}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}dH(y)}{\sum_{i=1}^{\infty}\frac{\pi_i(\boldsymbol{\theta})}{s_i}\int_0^{\infty}\frac{\boldsymbol{J}_{kk}(y;\boldsymbol{\alpha},\boldsymbol{T}/s_i)}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}dH(y)}.$$

Concerning $\boldsymbol{\theta}$, maximizing the following function with respect to $\boldsymbol{\theta}^*$

$$\boldsymbol{\theta}^* \to \sum_{i=1}^{\infty} \mathbb{E}(L^i|\boldsymbol{Y}=\boldsymbol{y}) \log \pi_i(\boldsymbol{\theta}^*)$$

is equivalent to maximizing

$$\boldsymbol{\theta}^* \to \sum_{i=1}^{\infty} \frac{1}{M} \mathbb{E}(L^i|\boldsymbol{Y}=\boldsymbol{y}) \log \pi_i(\boldsymbol{\theta}^*).$$

As $M \to \infty$ the latter converges to

$$\boldsymbol{\theta}^* \to \sum_{i=1}^{\infty} \pi_i(\boldsymbol{\theta}) \left(\int_0^{\infty} \frac{\boldsymbol{\alpha} \exp(y\boldsymbol{T}/s_i)\boldsymbol{t}/s_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} dH(y) \right) \log \pi_i(\boldsymbol{\theta}^*).$$

Let $\widehat{\boldsymbol{\theta}}$ denote the argument which maximizes this function.

In general none of the integrals will have explicit solutions, and approximations (e.g. numerical integration, Quasi Monte Carlo methods or more sophisticated variants) will have to be employed.

6. EXAMPLES

In this section we consider various numerical examples. In the first example we consider the simplest case of simulated data from a scale mixture of exponential distributions, where the scaling distribution has a Pareto type of tail. The exponential distribution is a particular case of an Erlang distribution; so we will treat the more general case instead. We will see that the canonical parametrization of the Erlang as a PH distribution will impose certain restrictions on the canonical parameters $\boldsymbol{\alpha}$ and $\boldsymbol{T}$ that will help to simplify the EM estimators.

In the second example we fit NPH distributions to a real data set (Danish reinsurance data of fire claims), while in the last three examples we consider the fitting of NPH distributions to the theoretical distributions Loggamma, Weibull and Lognormal respectively.

Example 6.1 (Erlang distributions)**.** In the first example we will estimate the parameters of data simulated from an NPH model where the PH component is Erlang distributed. Recall that a q dimensional Erlang distribution, $\mathrm{ER}_q(\lambda)$, is a phase–type distribution with canonical representation

$$\boldsymbol{\alpha} = (1,0,0,...,0), \quad \boldsymbol{T} = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 \\ 0 & 0 & -\lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda \end{pmatrix} \quad \text{and} \quad \boldsymbol{t} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ \lambda \end{pmatrix}.$$

Hence, it is not difficult to see that the complete likelihood for λ and $\boldsymbol{\theta}$ satisfies

$$L_c(\lambda,\boldsymbol{\theta}) \propto \prod_{i=1}^{\infty} [\pi_i(\boldsymbol{\theta})\lambda^q]^{L_i} e^{-Z_i\lambda/s_i}, \tag{6}$$

where $L^i = \sum_{j=1}^{M} 1(I_j = i)$ is the total number of paths started at level i, while $Z^i = \sum_{j=1}^{M} y_j \cdot 1(I_j = i)$ is the aggregation of the absorption times of all i-level paths. The complete loglikelihood is then given by

$$\ell_c(\lambda,\boldsymbol{\theta}) = \sum_{i=1}^{\infty} L_i \log \pi_i(\boldsymbol{\theta}) + Mq \log \lambda - \lambda \sum_{i=1}^{\infty} \frac{Z_i}{s_i} + k, \tag{7}$$

where k is some constant not depending on λ or $\boldsymbol{\theta}$. The E–step is similar as for the general model. As for the sufficient statistic L^i, its expected value is analogue and found to be equal to

$$\mathbb{E}\left(L^i|\boldsymbol{Y}=\boldsymbol{y}\right)=\sum_{j=1}^{M}\mathbb{P}(I_j=i|Y=y_j)=\sum_{j=1}^{M}\frac{\pi_i(\boldsymbol{\theta})g(y_j/s_i;q,\lambda)}{f(y_j;\boldsymbol{\theta},\lambda)},$$

where $g(\cdot;q,\lambda)$ is the density of an Erlang(q,λ) distribution. In the case of the sufficient statistic Z^i we have

$$\mathbb{E}(Z^i|\boldsymbol{Y}=\boldsymbol{y})=\sum_{j=1}^{M}y_j\mathbb{P}(I_j=i|Y=y_j)=\sum_{j=1}^{M}y_j\frac{\pi_i(\boldsymbol{\theta})g(y_j;q,\lambda/s_i)}{f(y_j;\boldsymbol{\theta},\lambda)}.$$

The EM estimators for λ and $\boldsymbol{\theta}$ are then easily calculated to be

$$\widehat{\lambda}^{-1}=\frac{1}{qM}\sum_{i=1}^{\infty}\sum_{j=1}^{M}y_j\frac{\pi_i(\boldsymbol{\theta})g(y_j;q,\lambda/s_i)/s_i}{f(y_j;\boldsymbol{\theta},\lambda)}$$

$$\widehat{\boldsymbol{\theta}}=\arg\max_{\theta^*}\sum_{i=1}^{\infty}\sum_{j=1}^{M}y_j\frac{\pi_i(\boldsymbol{\theta})g(y_j;q,\lambda/s_i)}{f(y_j;\boldsymbol{\theta},\lambda)}\log\pi_i(\theta^*).$$

We present a simulation study for the case of $q=1$ (corresponding to an infinite–dimensional hyperexponential distribution), $s_i=i$ and scaling distribution

$$\pi_i(\theta)=\frac{i^{-\theta}}{\zeta(\theta)},\qquad i=1,2,...,$$

where

$$\zeta(\theta)=\sum_{i=1}^{\infty}i^{-\theta}$$

is the Riemann Zeta function with parameter θ. This distribution is known as the Zipf distribution, Riemann Zeta distribution or discrete Pareto distribution (since its tail resemble that of a Pareto distribution).

In order to find the EM–estimate of $\theta^{(n+1)}$, we have to maximize the function

$$h(\theta)=-\sum_{i=1}^{\infty}\mathbb{E}(L^i|\boldsymbol{Y}=\boldsymbol{y})\theta\log(i)-\sum_{i=1}^{\infty}\mathbb{E}(L^i|\boldsymbol{Y}=\boldsymbol{y})\log(\zeta(\theta)).$$

Differentiating with respect to θ and rearranging implies that we have to solve the following equation w.r.t. θ,

$$\frac{\zeta'(\theta)}{\zeta(\theta)}=-\frac{1}{M}\sum_{i=1}^{\infty}\mathbb{E}\left(L^i|\boldsymbol{Y}=\boldsymbol{y}\right)\log(i), \tag{8}$$

which is done numerically using a simple Newton–Raphson procedure.

Results of the simulation study, which is shown in Table 1, reveal that the EM algorithm is able to recover the underlying structure of the data, and that the estimation, as expected, improves with larger sample sizes. The EM algorithm was in in all cases initiated with a randomly generated seed. For these examples we choose a Gamma$(1,1)$ distribution for both λ and θ to allow for all possible values in the parametric space. We remark this selection is arbitrary and similar estimates were obtained when using some other fixed starting points.

Example 6.2. We consider 2167 reinsurance data for Danish fire insurance claims above 1 million DKR for the period 1980–1993. These data have been widely studied in Extreme Value Theory [13, 17]. The data corresponds to claims in millions of Danish Kroner for the period 1980–1993, the amounts being adjusted for inflation to prices of 1985. We subtracted 1 (million) from all data in order to shift the support to $[0,\infty)$ which is the natural support for a phase–type distributions. Of

parameters		sample size									
λ	θ	100		500		1000		5000		10000	
1.0	2.0	1.04	1.88	0.95	2.09	1.14	1.93	0.99	2.00	1.01	2.01
1.0	2.5	0.85	2.67	0.90	2.52	1.01	2.63	0.94	2.64	1.00	2.52
1.0	5.0	0.92	7.1	1.09	7.26	1.05	4.65	1.00	4.85	1.01	5.03

Table 1: EM–estimates of $(\widehat{\lambda}, \widehat{\theta})$ infinite dimensional hyper–exponential distributions with varying parameters and sample size

the 2167 data, 519 are repeated values so only 1648 have different values. Just above 90% of the data is below 5, while less than 10% falls between 5 and the maximum observation of 262.2504.

We propose an NPH model having a scaling distribution that is assumed to follow the discretized Pareto distribution of Example 3.3. We conducted various numerical experiments to analyse the effect of the various components of the model. For each model we ran the EM starting from various initial points chosen at random; more precisely, the values of $\boldsymbol{\alpha}$, the non-diagonal elements of $\boldsymbol{T}$ and the elements of the vector $\boldsymbol{t}$ were chosen from a uniform distribution, and then normalized in the case of $\boldsymbol{\alpha}$. The parameter θ was drawn from a Gamma$(2,1)$ distribution. The iterations of the EM were stopped when the relative change in the log-likelihood was smaller than 10^{-10}. The EM algorithm was restarted several times and the model with the largest value of the incomplete loglikelihood was selected as the output.

First we tested the difference in performance between the EM–algorithm using raw data and one using the hybrid method proposed in Remark 3.5. We considered bins of the same length 0.05, thus treating losses within 50,000 DKR as all being the same. For this example we choose the average of all data points falling in the j-th bin as the data points. This amounts to a total of 265 distinct data points. The actual binning of the data does not appear to have any effect at all on the estimation, as compared with the EM–algorithm of the raw data. The speed was increased by a factor 6.12 which is more or less consistent with the execution times depending linearly on the number of data points, in which case we should have expected an increase in speed by a factor $1648/265 = 6.21$. Moreover, we observed a significant faster convergence of the EM algorithm of a model with full data in cases where the initial parameters are taken as the output of a data–reduced model. In all the EM algorithms which follow we have used similar data–reduction techniques.

By fitting models with $p = 1,...,6$, we selected a model with $p = 5$ phases based on visual inspection of its fit to the histogram. Model selection based on e.g. the AIC index for phase–type distributions is not well established since the number of free parameters in a phase–type distribution is not known in general. Also, a p–dimensional phase–type distributions may be parametrized in terms of at most $2p-1$ parameters which are the coefficients of the polynomials in its rational Laplace transform. Based on this parametrization, the model with $p = 4$ would have been the choice.

We now investigate the effect of the choice of the parameter c (see Example 3.3) will have on the estimation. Intuitively, the smaller we choose the ratio of the geometric progression, the better the estimation. Fixing the dimension of the phase–type distribution to $p = 5$, we select three different values for $c = 1, 1/2, 1/4$. The estimates are as follows

$c = 1$**:**

$$\begin{aligned} \widehat{\theta} &= 1.2743 \\ \widehat{\boldsymbol{\alpha}} &= (0.6415, 0.0099, 0.0055, 0.2115, 0.1316) \end{aligned}$$

$$\widehat{\boldsymbol{T}} = \begin{pmatrix} -2.7430 & 1.3565 & 0 & 0 & 0 \\ 0.0003 & -3.0398 & 0 & 0 & 0 \\ 0 & 0 & -2.6313 & 1.1226 & 0.2167 \\ 0 & 0 & 0.7223 & -1.3953 & 0.4089 \\ 0 & 0 & 1.5129 & 0.8388 & -2.4779 \end{pmatrix}$$

$c = 1/2$**:**

$$\widehat{\theta} = 1.3136$$
$$\widehat{\boldsymbol{\alpha}} = (0.7692, 0.0149, 0.1154, 0.0148, 0.0857)$$
$$\widehat{\boldsymbol{T}} = \begin{pmatrix} -3.0713 & 0.1234 & 0.2319 & 0.0655 & 1.6457 \\ 0.2239 & -1.9576 & 0.3787 & 0.9651 & 0.3899 \\ 0.1824 & 0.7410 & -3.0705 & 0.7978 & 0.4776 \\ 0.3099 & 0.6150 & 0.4751 & -1.5588 & 0.1588 \\ 0.0034 & 0.0446 & 0.2263 & 0.0446 & -3.4154 \end{pmatrix}$$

$c = 1/4$**:**

$$\widehat{\theta} = 1.3230$$
$$\widehat{\boldsymbol{\alpha}} = (0.0267, 0.4563, 0.0010, 0.2603, 0.2556)$$
$$\widehat{\boldsymbol{T}} = \begin{pmatrix} -0.7896 & 0.2531 & 0.0385 & 0.2652 & 0.0265 \\ 0.0810 & -3.7022 & 1.7356 & 1.1626 & 0.3934 \\ 0.5838 & 0.0013 & -3.7890 & 0.0576 & 0.0302 \\ 0.3386 & 0.0606 & 1.0320 & -3.7735 & 0.1957 \\ 0.6694 & 0.0292 & 0.2376 & 0.3920 & -3.4129 \end{pmatrix}$$

That the estimates for $(\boldsymbol{\alpha}, \boldsymbol{T})$ look distinct is because phase–type representations are not unique. In Figure 1, the estimated densities are plotted against the almost identical histogram of the data (actually indistinguishable in the plotted range of $[0,6]$). It is clear that the different phases do not have a physical interpretation but are merely dummy (or black box) states used for the only purpose of obtaining a proper adjustment of the distribution.

In order to inspect the tail behaviour we plotted the log–survival function in Figure 2. We have provided two plots different ranges in order to be able to inspect the tail behaviour well beyond the largest data value (which is 262.2504). The different values for θ (obtained as a consequence of the different choices of c) does result in a slightly different tail behaviour for very large claim sizes. We compare these estimates of θ against the ones obtained by [13] and [17]. The first reference fits a Generalized Pareto Distribution via maximum likelihood estimation while the second employs the Hill estimator. The implementation of the methods described above involve the selection of an appropriate threshold which should be chosen by the modeler. According to [17] the values in the interval [1.40, 1.46] produce excellent fits, the value recommend by the same author being 1.45. In our case, smaller choices of c will typically result in values of the parameter θ closer to these ideal values, hence suggesting that smaller values of c will provide better fits.

Finally, we compare the models above against an NPH model having a fixed scaling distribution. The scaling distribution is chosen as in Example 3.3, but this time we fix $\theta = 1.45$ as this is the recommended by [17]. This can be done by running the EM algorithm in the usual way but avoiding any adjustments in θ. The results of the estimation are given next

$\theta = 1.45$**:** Fixed tail.

$$\widehat{\boldsymbol{\alpha}} = (0.1114, 0.0080, 0.3366, 0.3600, 0.1840)$$

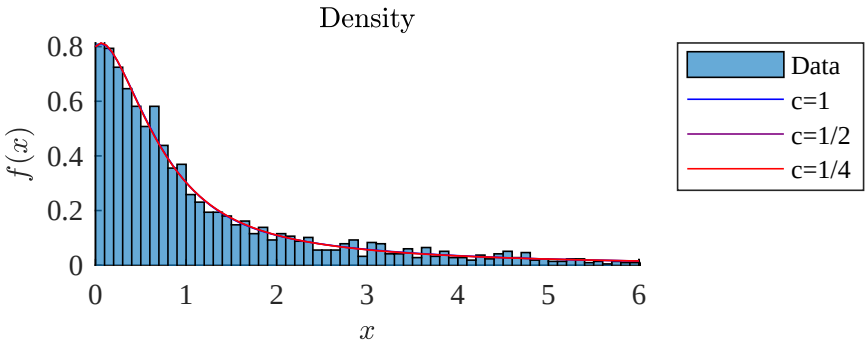

FIGURE 1. Estimated densities against histogram of the NPH models for Danish reinsurance data.

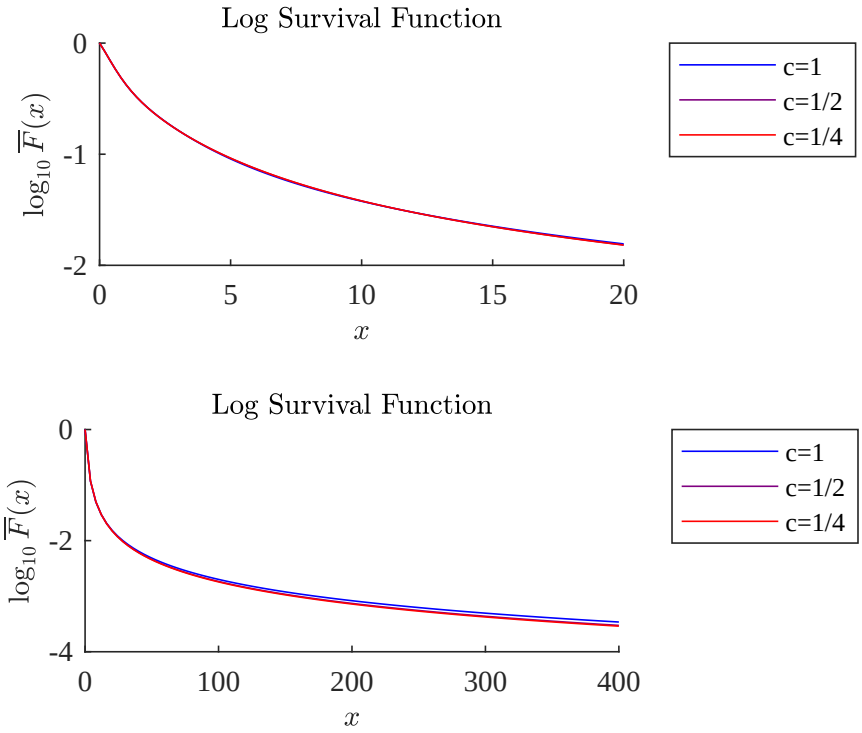

FIGURE 2. Estimated survival functions of the NPH models for Danish reinsurance data corresponding to $\widehat{\theta}$ equal to $1.2748, 1.3115, 1.3387$.

$$\widehat{\boldsymbol{T}} = \begin{pmatrix} -3.0292 & 0.8810 & 0.0655 & 0.1600 & 0.1538 \\ 0.1204 & -0.6765 & 0.3295 & 0.0829 & 0.0827 \\ 1.4308 & 0.2854 & -4.2611 & 1.1133 & 1.4316 \\ 0.4463 & 0.2717 & 0.3082 & -3.6472 & 1.3508 \\ 0.0958 & 0.0739 & 0.0230 & 0.0265 & -3.2740 \end{pmatrix}.$$

This representation does not resemble any of the previous estimations (due to non–uniqueness of representations) but from figures 3 and 4 it is clear that there is no distinguishable difference between densities for the models with fixed and EM adjusted tails in the range $[0, 6]$ where the

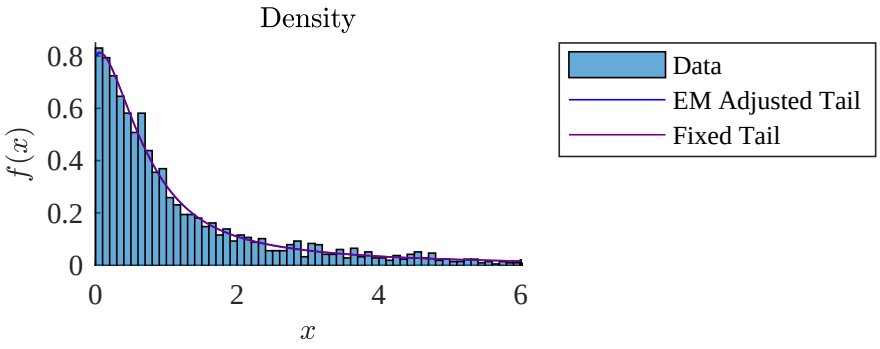

FIGURE 3. Estimated densities of the NPH models with EM adjusted ($\widehat{\theta} = 1.3387$) and fixed ($\widehat{\theta} = 1.45$) tails against histogram of the Danish reinsurance data.

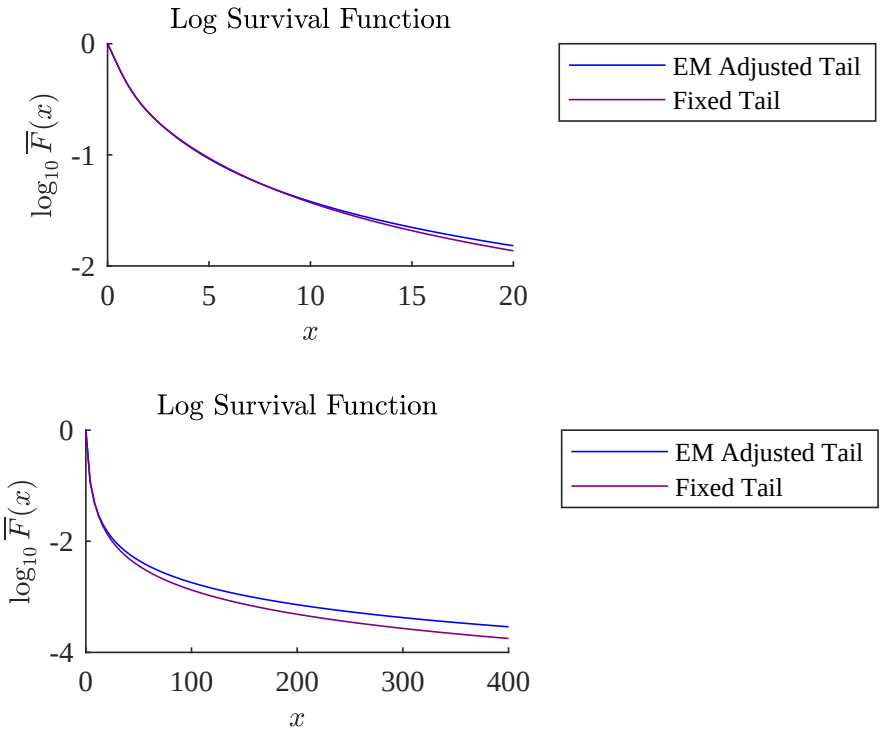

FIGURE 4. Estimated survival functions of the NPH models with EM adjusted ($\widehat{\theta} = 1.3387$) and fixed ($\widehat{\theta} = 1.45$) tails.

main body of the distribution is situated. The tail behaviour also looks quite similar though the EM fitted tail seems to be slight heavier than the fixed tail.

Example 6.3 (Fitting to a theoretical distribution: Loggamma)**.** Next we consider the problem of approximating a theoretical distribution via an NPH model. In our first example we take as target a Loggamma distribution with shape parameter α, scale parameter β and shifted one unit to the left so its support is $[0,\infty)$, which is the natural support for the class NPH. Its density function is then given by

$$f(x) = \frac{\beta^{\alpha} \log^{\alpha-1}(x+1)}{\Gamma(\alpha)(x+1)^{\beta+1}}, \qquad x \geq 0.$$

This distribution is regularly varying with parameter β. For this example we choose the parameters $\alpha = \beta = 2$, for the target distribution with the purpose of analysing a distribution with a mode away from 0 and a moderately heavy–tail. We consider an NPH model where the underlying phase–type distribution has five phases and the scaling distribution is that of Example 3.3 with $c = 1$. With this example, we want to test if general Regularly Varying distributions can be correctly fitted with the general model suggested in Example 3.3.

We employed Quasi Monte Carlo ideas to approximate the integrals in the EM steps and iterated the algorithm until the relative error was smaller than 10^{-9}. The results are given below

$$\begin{aligned}
\widehat{\theta} &= 1.6031 \\
\widehat{\boldsymbol{\alpha}} &= (0.5717, 0.0330, 0.0000, 0.3954, 0.0000) \\
\widehat{\boldsymbol{T}} &= \begin{pmatrix} -1.9634 & 0.0609 & 0.5025 & 0.1249 & 1.2751 \\ 0.0616 & -0.3372 & 0.0775 & 0.0382 & 0.1428 \\ 0.7529 & 0.1178 & -2.2723 & 0.4797 & 0.0068 \\ 0.7278 & 0.3060 & 1.1458 & -4.8966 & 2.7170 \\ 0.8923 & 0.0317 & 0.0482 & 0.2021 & -3.4321 \end{pmatrix}
\end{aligned}$$

The densities of the Loggamma distribution and its NPH approximation are plotted in Figure 5. The densities of the Loggamma and its NPH approximation are almost indistinguishable from each other in the body region. The tail behavior is correctly captured by the NPH model as seen in Figure 6. The shape of the tail of the NPH is very close to that of the Loggamma distribution although the NPH estimate has a heavier tail (the estimated regularly varying parameter was $\widehat{\theta} = 1.6031$ as compared to $\theta = \beta = 2$).

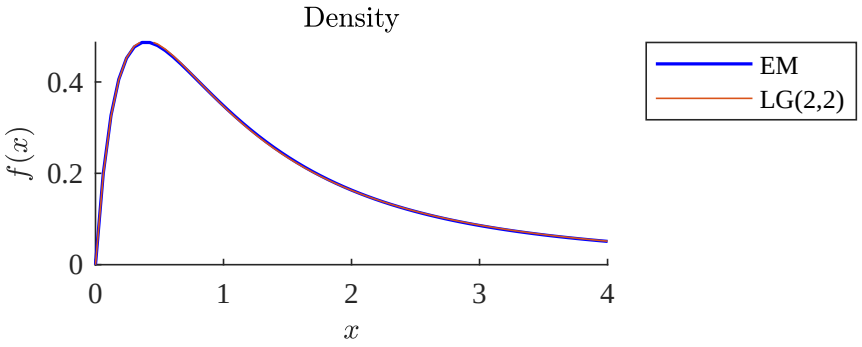

FIGURE 5. Density of the Loggamma distribution with parameters $(2,2)$ and a fit of an NPH model using the EM algorithm.

Example 6.4 (Fitting to a theoretical distribution: Weibull)**.** Next we move away from the Regularly Varying case and consider instead the Weibullian case. As a target distribution we consider a classical two–parameter Weibull with $\lambda = 1$ and $p = 1/2$ so its density is $e^{-\sqrt{x}}/2\sqrt{x}$, $x \geq 0$. For adjusting this model we consider an NPH family of distributions where the phase–type part has five phases and the scaling distribution is supported over a geometric progression $e^c, e^{2c}, e^{3c}, \ldots$ with $c = 1$ and taken as a discretization of a classical two–parameter Weibull distribution with $\delta = p - 1$. More precisely, its density is given by

$$f(x) = p\lambda^p x^{p-1} e^{-(\lambda x)^p}, \qquad x > 0.$$

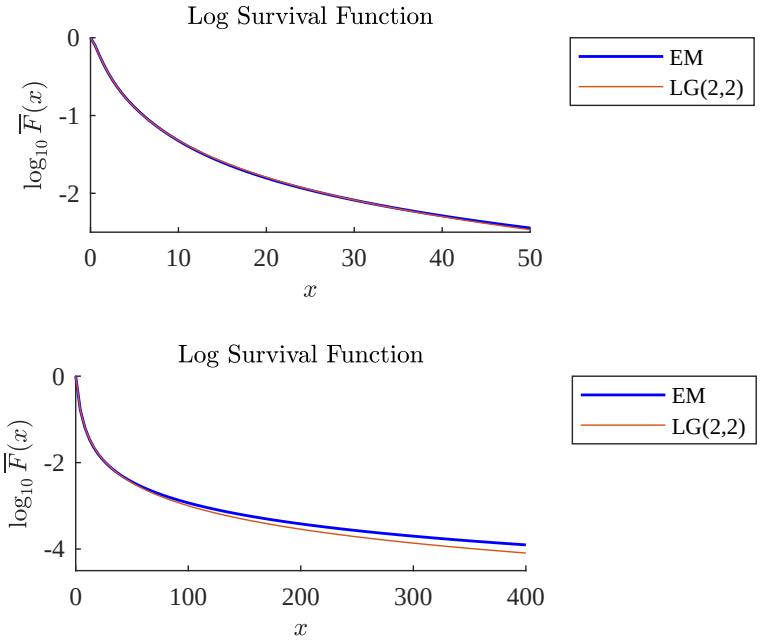

FIGURE 6. Survival functions of the Loggamma distribution LG$(2,2)$ and a fit of an NPH model using the EM algorithm.

Notice, that the target distribution is in the same two–parameter family of Weibull distributions. The results are given below.

$$\begin{aligned}
\widehat{\lambda} &= 0.6181,\ \ \widehat{p} = 1.1673 \\
\widehat{\boldsymbol{\alpha}} &= (0.2370, 0.3349, 0.0901, 0.3380, 0) \\
\widehat{\boldsymbol{T}} &= \begin{pmatrix} -3.0858 & 0.6964 & 0.2188 & 0.3355 & 0.5572 \\ 51.0253 & -207.2799 & 18.3961 & 18.5851 & 46.0724 \\ 0.4369 & 0.2511 & -0.9839 & 0.0122 & 0.0487 \\ 1.1297 & 0.2448 & 0.9388 & -11.0548 & 0.7637 \\ 0.7583 & 1.0994 & 0.9844 & 0.4882 & -3.3303 \end{pmatrix}
\end{aligned}$$

The agreement between the Weibull distribution and its NPH approximation is very good in the body of the distribution. The approximation of the tail is also particularly good for values going up to 100 which correspond to probabilities of order 10^{-4}. The NPH adjusted model is Weibullian with parameter $\widehat{p}(1+\widehat{p})^{-1} \approx 1.1673/2.1673 \approx 0.5386$. Recall that the parameter of the target distribution is 0.5.

Example 6.5 (Fitting to a theoretical distribution: Lognormal)**.** Finally we consider a Lognormal distribution with location parameter $\mu = 0$ and dispersion parameter $\sigma^2 = 1$. The lognormal distribution is heavy–tailed but its survival function ultimately decays faster than a Regularly Varying function (but slower than a Weibullian function). We consider the lognormal case more difficult because a scaling distribution $\pi(\theta)$ for which the NPH model has the same tail behavior as the lognormal distribution is unknown.

We consider two alternative NPH models for fitting the data. For both we have selected a phase–type part with eight phases and the scaling distribution is chosen as a discretization of a Lognormal distribution LN(μ,σ^2) and supported over the set $\{s_i = e^i : i = 0,1,\dots\}$. The difference between the two models is that for the first one we let the EM algorithm to estimate the values of the

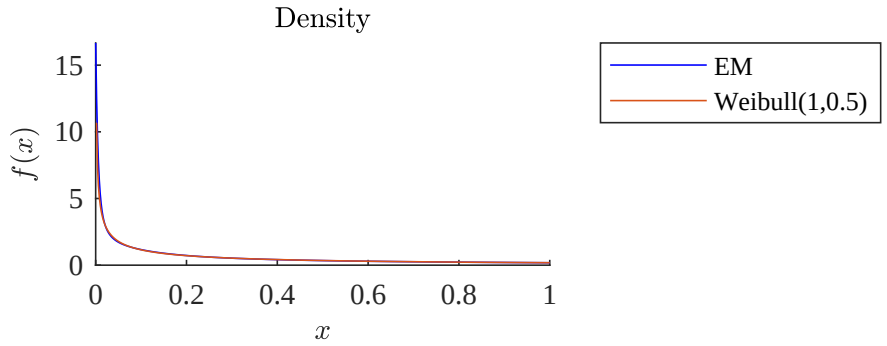

FIGURE 7. Density of the Weibull distribution with parameters $(1,0.5)$ and a fit of an NPH model using the EM algorithm.

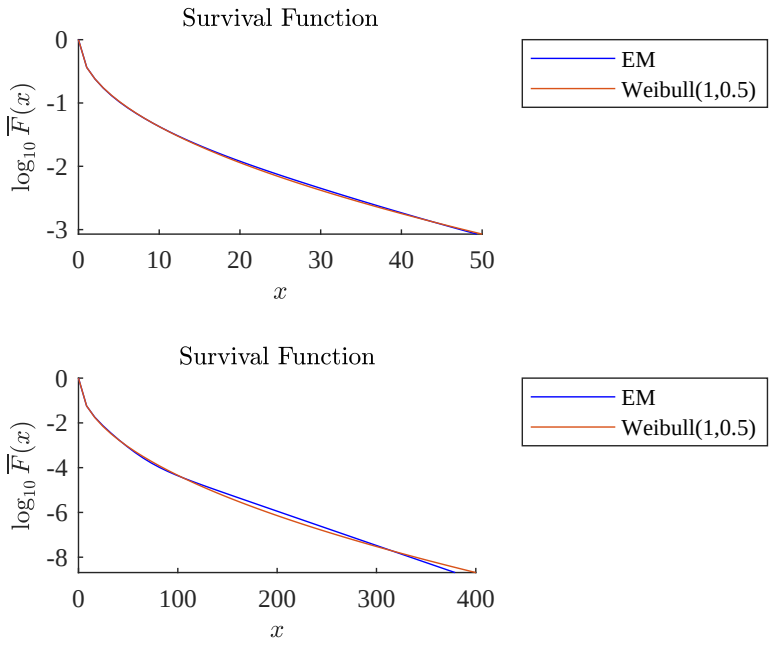

FIGURE 8. Survival functions of the Weibull$(1,0.5)$ and a fit of an NPH model using the EM algorithm. The EM model is Weibullian with shape parameter $p = \widehat{p}(1+\widehat{p}) \approx 0.5386$.

parameters μ and σ, while for the second one we take these values to be fixed and equal to 0 and 1 respectively. The results for the first model are given below

From Figure 9 we can observe that the fit to the body of the distribution of the first model is appropriate, but the tail probabilities differ. In the lognormal case, the *heaviness* of the distribution is mostly determined by the parameter σ. The larger the parameter σ the more heavier its tail. In our estimations we obtained an estimate $\widehat{\sigma} = 0.5979$ which suggest a lighter tail, and this is confirmed in the last panel of Figure 10.

For our second model we obtained a very poor fit, but this is somewhat expected since the tail distribution of the NPH model is very different from the tail behavior of its scaling distribution (as in the case of Regularly Varying or Weibullian cases). In fact, [18] demonstrate that the tail probability of the NPH model is significantly heavier and different to the scaling distribution. More

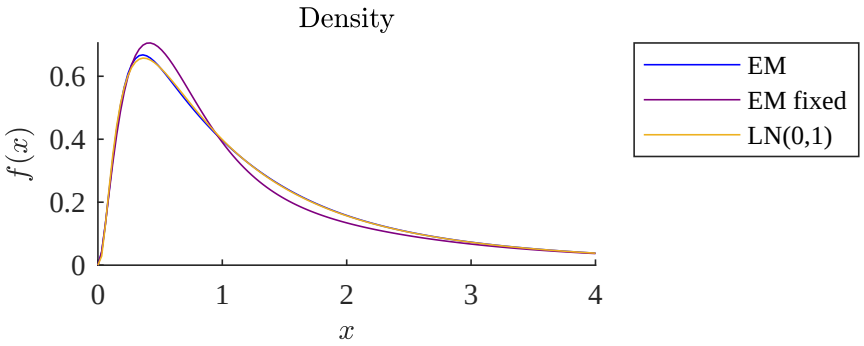

FIGURE 9. Density of the Lognormal distribution with parameters $(0,1)$ and a fit of an NPH model using the EM algorithm.

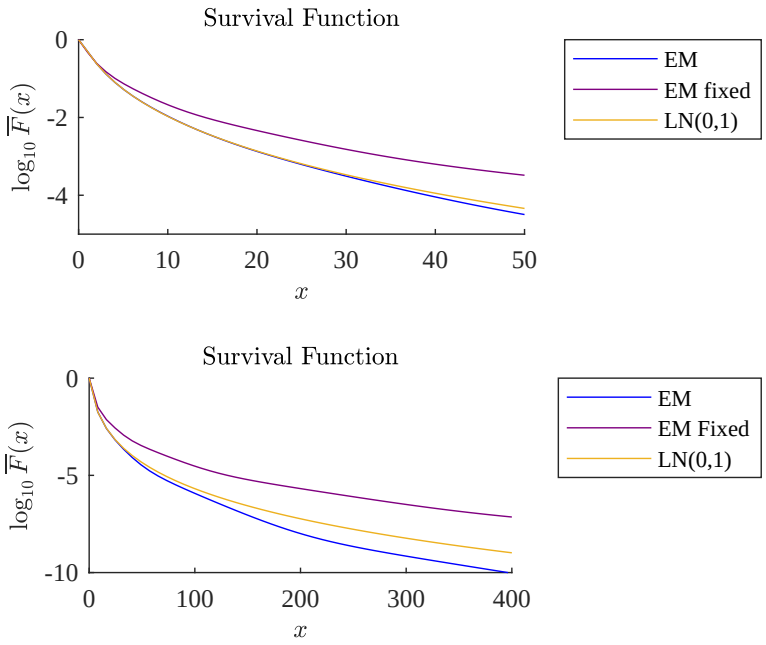

FIGURE 10. Survival functions of the Lognormal distribution $\mathrm{LN}(0,1)$ and a fit of an NPH model using the EM algorithm.

precisely, it is shown that if $N \sim \mathrm{LN}(0,1)$, then

$$\lim_{t\to\infty} \frac{\mathbb{P}(\tau N > t)}{\mathbb{P}(N > t)} = \infty.$$

7. CONCLUSIONS

In this paper we have adapted a known method for estimating phase–type distributions to the extended class NPH of phase–type scale mixture distributions with the purpose of fitting heavy–tailed data and distributions. While the model only requires the specification of the dimension of the underlying phase–type distribution and a parametric family of discrete scaling distributions, the suggested algorithm simultaneously fit both the body and the tail of general distributions.

Since the class of NPH distributions are generally genuinely heavy tailed (if N has unbounded support) and dense in the class of heavy tailed distributions with support on $\mathbb{R}_+$, we may in principle approximate any heavy tailed distribution (data) arbitrarily close with an NPH distribution.

In particular, for the cases of Regularly Varying and Weibullian distributions the aforementioned approximation is not only in the limit (denseness) but can be effectively carried out in praxis.

REFERENCES

[1] M. Arendarczyk and K. Dębicki. Asymptotics of supremum distribution of a Gaussian process over a Weibullian time. *Bernoulli*, 17(1):194–210, 2011.

[2] S. Asmussen, O. Nerman, and M. Olsson. Fitting Phase-Type Distributions via the EM Algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[3] A. Balkema and L. de Haan. Residual life time at great age. *Ann. Probab.*, 2:792–804, 1974.

[4] M. Bladt. A review on phase-type distributions and their use in risk theory. *Astin Bulletin*, 35(1):145–161, 2005.

[5] M. Bladt and B. F. Nielsen. *Matrix-Exponential Distributions in Applied Probability*, volume 81 of *Probability Theory and Stochastic Modelling*. Springer US, 2017.

[6] M. Bladt, B. F. Nielsen, and G. Samorodnitsky. Calculation of ruin probabilities for a dense class of heavy tailed distributions. *Scandinavian Actuarial Journal*, 2015(7):573–591, 2015.

[7] M. Dwass. Extremal Processes. *Ann. Math. Stat.*, 35:1718–1725, 1964.

[8] R. Fisher and L. Tippett. Limiting forms of the frequency of the largest or smalles member of a sample. *Math. Proc. Cambridge Philos. Soc*, 24:180–190, 1928.

[9] B. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatorie. *Ann. Math.*, 44:423–453, 1943.

[10] M. Gomes and A. Guillou. Extreme Value Theory and Statistics of Univariate Extremes: A Review. *International Statistical Review*, 83:263–292, 2015.

[11] J. Lamperti. On extreme order statistics. *Ann. Math. Stat.*, 35:1726–1737, 1964.

[12] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*, volume 5. Society for Industrial and Applied Mathematics, 1987.

[13] A. J. McNeil. Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *ASTIN Bulletin*, 27(1):117–137, 1997.

[14] M. F. Neuts. *Matrix-geometric Solutions in Stochastic Models*. An Algorithmic Approach. Courier Dover Publications, 1981.

[15] M. Olsson. Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, pages 443–460, 1996.

[16] I. Pickands, J. Statistical inference using extreme order statistics. *Ann. Stat.*, 3:119–131, 1975.

[17] S. I. Resnick. Discussion of the Danish Data on Large Fire Insurance Losses. *ASTIN Bulletin*, 27(1):139–151, 1997.

[18] L. Rojas-Nandayapa and W. Xie. Asymptotic tail behavior of phase-type scale mixture distributions. *Annals of Actuarial Science*, pages 1–21, 2017.

[19] C. Van Loan. Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control*, 23(3):395–404, 1978.

INSTITUTE FOR APPLIED MATHEMATICS AND SYSTEMS, NATIONAL UNIVERSITY OF MEXICO, A.P. 20-726, 01000 MEXICO, D.F., MEXICO

E-mail address: bladt@sigma.iimas.unam.mx

INSTITUTE FOR FINANCIAL AND ACTUARIAL MATHEMATICS, UNIVERSITY OF LIVERPOOL, MATHEMATICAL SCIENCES BUILDING, L69 7ZL, LIVERPOOL, UNITED KINGDOM

E-mail address: leorojas@liverpool.ac.uk