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 distribution 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 [3], employing an EM algorithm for the the maximum likelihood estimates 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. We empirically demonstrate that our model can provide excellent fits to heavy–tailed data/distributions with minimal assumptions.

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

1. INTRODUCTION

In this paper we consider the maximum likelihood estimation for a dense class of nonnegative heavy–tailed tailed distributions, referred to as scale mixtures of phase–type distribuitions (NPH), which was defined in [3]. Distributions in the NPH class allow for the simultaneous modelling of the "body" and the "tail" of general distributions which are assumed to be absolutely continuous and nonnegative while their "tails" are assumed to belong to some general class of heavy tailed distributions, like for instance Regularly Varying or Weibullian. These very general assumptions allows us to fit heavy–tailed distributions which may look distinctively different from distributions usually found in catalogues.

Apart from providing an adequate description of data, distributions from NPH can also be seen as infinite–dimensional phase–type distributions, which to all intents and purposes are as tractable as their finite–dimensional counterparts. Much of the machinery available for finite–dimensional phase–type distributions is also applicable to the extended class. Algorithms are for example available for the exact calculations of properties related to renewal theory, random walks (ladder processes) and ruin probabilities (see [3] for details).

The maximum likelihood estimation will be carried out employing an EM algorithm similarly as for finite–dimensional phase–type distributions [2]. The main challenge we face is the algorithmic implementation resulting from the extension to infinite dimensions since we cannot make a pre–fixed cut–off in the number of dimensions as this would be equivalent to a light–tailed estimation. We shall see, however, that it is possible to reduce the formulas to simple (one–dimensional) infinite series which can be numerically evaluated to a specified degree of precision. We are also able to simplify certain expressions of the original paper ([2]), obtaining explicit expressions which allow for an improved numerical performance.

The rest of the paper is organized as follows. In Section 3 we develop the main algorithm for estimating independent and identically distributed data sampled from an NPH distribution. We show that the algorithm essentially uses the empirical cumulative distribution function, and that a significant increase in speed may be obtained by grouping the data into intervals on parts of the support and considering the corresponding histograms. In Section 4 we adjust the algorithm 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 a fits to theoretical distribution. In there, we highlight some minimal assumptions made for efficiently fitting general Regularly Varying and Weibullian distributions.

2. Phase–type distributions and the NPH class

Before proceeding with a more detailed account, we first provide some background on phase–type distributions and the extended class NPH of scale mixtures of phase–type distributions.

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 $E = \{1,2,...,p,p+1\}$ 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 follow the convention that matrices are written in capital bold and their elements with the corresponding minuscule (like $\boldsymbol{A} = \{a_{ij}\}$), bold minuscule Greek letters (like $\boldsymbol{\alpha}$) are row vectors while bold minuscule Roman letters (like $\boldsymbol{t}$) are column vectors. Their dimensions are usually clear from the context and left unspecified unless necessary. Let $\alpha_i = \mathbb{P}(X_0 = i)$, $\sum_{i=1}^{p}\alpha_i = 1$ and $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_p)$. Then we say that

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

has a 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 phase–type distributions we refer e.g. to [**?**], [4], [7] or [8].

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

In [3] the dense class, NPH, of genuinely heavy tailed distributions is proposed in terms of infinite–dimensional phase–type distributions with a finite number of parameters. The idea is very simple and goes a follows. Let N be a discrete random variable with support on $\{s_i : i\in\mathbb{N}, s_i > 0\}$ and distribution $\boldsymbol{\pi} = \{\pi_i\}_{i\geq 1}$ where $\pi_i = \mathbb{P}(N = s_i)$. If N, for example, is a discretization of a continuous distribution G at step length Δ, then $s_i = i\Delta$ and $\pi_i = G(i\Delta) - G((i-1)\Delta)$, $i = 1,2,\ldots$. Let $\tau\sim \mathrm{PH}_p(\boldsymbol{\alpha},\boldsymbol{T})$ be independent of N. When sampling from Y, we first draw an index i (referred to as the *level*) from $\boldsymbol{\pi}$ and then a phase–type random variable from $\mathrm{PH}_p(\boldsymbol{\alpha},\boldsymbol{T}/s_i)$. Hence, the random variable $Y = N\tau$ may be seen as a scale mixture of phase–type distributions, so we often refer to N as the scaling random variable and its distribution $\boldsymbol{\pi}$ as the scaling distribution of the NPH. The density of Y can written as

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

The distribution of Y can also be seen 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 and we notice that the exponential of the infinite dimensional matrix $\boldsymbol{\Gamma}$ is well defined since it is a *bounded operator* (as the sequence $\{s_i : s_i > 0\}$ is bounded away from zero). We also let $\boldsymbol{t}_i = -\boldsymbol{T}_i\boldsymbol{e}$. We shall 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.

A key feature of the NPH class is that it contains a rich variety of heavy–tailed distributions. For instance, if $\boldsymbol{\pi}$ has unbounded support, then Y has a heavy–tailed distribution [**?**, cf.]]RojasXie. In the regularly varying case, the tail of Y greatly follows that of N. Breiman's lemma implies that if N has a regularly varying with tail index $-\alpha < 0$, then the tail of Y will also be regularly varying with the same index. More precisely,

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

A similar feature occurs for the the class of Weibullian distributions [1], which is defined as the collection of nonnegative distributions having survival function

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

If the parameter $p \in (0,1)$, then the distribution is heavy–tailed (and light–tailed otherwise). Since a PH distribution is Weibullian with parameter $p = 1$, then Lemma 2.1 of [1] implies that if the scaling distribution $\boldsymbol{\pi}$ is Weibullian with parameter p_1 then $Y \sim \mathrm{NPH}(\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{T})$ is Weibullian with parameter $p_1/(1 + p_1)$.

3. Estimation

In this section we address the problem of fitting an NPH distribution to a data set. We assume that $y_1, y_2, ..., y_M$ forms an i.i.d. data set sampled from $\mathrm{NPH}_p(\pi(\boldsymbol{\theta}), \boldsymbol{\alpha}, \boldsymbol{T})$, where $\boldsymbol{\pi}(\boldsymbol{\theta})$ is some parametric distribution describing the 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 $\boldsymbol{\pi}(\boldsymbol{\theta})_i$, 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 realization of the underlying Markov jump process are not observable. Thus, we resort to the EM algorithm for approximating the maximum likelihood estimators. To this end we first attend the corresponding estimation problem for complete data.

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

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

denote the number of i–level processes 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 number of i–level processes that are 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 Markov jump processes spend in state k and let $N_{k\ell}^i$ denote the total number of jumps from state k to ℓ within all i–level Markov jump processes. Finally, let N_k^i be the number of i–level processes that exit to the absorbing state from state k.

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

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

with corresponding log–likelihood

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

In order to calculate the maximum likelihood estimator $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}})$, which is the point that maximizes the likelihood (or log–likelihood) function, we calculate first order partial derivatives of $t_{k\ell}$, t_k, while we use the Lagrange multiplier methods for α_i and $\pi_i(\boldsymbol{\theta})$ due to the constraints on them summing up to one.

Consider the parameter $t_{k\ell}$, $k \neq \ell$. Then

$$\frac{\partial \ell_c}{\partial t_{k\ell}} = \sum_{i=1}^{\infty} N_{k\ell}^i \frac{1}{t_{k\ell}} - \sum_{i=1}^{\infty} \frac{Z_k^i}{s_i} = 0$$

implying

$$\hat{t}_{k\ell} = \frac{\sum_{i=1}^{\infty} N_{k\ell}^i}{\sum_{i=1}^{\infty} Z_k^i / s_i}.$$

Similarly,

$$\hat{t}_k = \frac{\sum_{i=1}^{\infty} N_k^i}{\sum_{i=1}^{\infty} Z_k^i / s_i}.$$

The diagonal terms $\hat{t}_{kk} = -\sum_{\ell \neq k} \hat{t}_{k\ell} - \hat{t}_k$. Regarding $\boldsymbol{\alpha}$, consider the Lagrange function

$$M(\boldsymbol{\alpha}) = \sum_{i=1}^{\infty} B_k^i \log \alpha_k + \mu(1 - \sum_k \alpha_k),$$

where μ is a Lagrange multiplier. Then

$$\frac{\partial M}{\partial \alpha_k} = \sum_{i=1}^{\infty} \frac{B_k^i}{\alpha_k} - \mu = 0,$$

which result in

$$\alpha_k \mu = \sum_{i=1}^{\infty} B_k^i.$$

Summing over k yields

$$\mu = \sum_{i=1}^{\infty}\sum_{k=1}^{p} B_k^i = M$$

so

$$\hat{\alpha}_k = \frac{1}{M}\sum_{i=1}^{\infty} B_k^i.$$

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

We now consider the case of incomplete data. We shall employ the EM–algorithm, which optimizes the incomplete data likelihood L,

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

using the complete likelihood L_c (or ℓ_c) in the following way. Here $f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})$ denotes the density function of $Y \sim \mathrm{NPH}_p(\pi(\boldsymbol{\theta}),\boldsymbol{\alpha},\boldsymbol{T})$.

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. $L(\boldsymbol{\theta}_n,\boldsymbol{\alpha}_n,\boldsymbol{T}_n) \le L(\boldsymbol{\theta}_{n+1},\boldsymbol{\alpha}_{n+1},\boldsymbol{T}_{n+1})$) and the procedure hence converges (possibly to a local maximum or saddlepoint though).

We notice that the actual calculations in the EM–algorithm only involve the complete data likelihood. 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 M–step is trivial since we only have to plug in the conditional expectations given the data instead of the sufficient statistics which are not available

Concerning the E–step we proceed as follows. 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 $\mathbb{P} = \mathbb{P}_{\boldsymbol{\Psi}}$, $\boldsymbol{\Psi} = (\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})$, but we will omit the index in order to ease the exposition.

Concerning L^i, it equals one if i is the chosen level and zero otherwise. Let I denote the random variable indicating the chosen level. Then

$$\begin{aligned}
\mathbb{E}(L^i|Y=y) &= \mathbb{P}(I=i|Y=y) \\
&= \frac{\mathbb{P}(Y\in dy|I=i)\mathbb{P}(I=i)}{\mathbb{P}(Y\in dy)} \\
&= \frac{\pi_i(\boldsymbol{\theta})\boldsymbol{\alpha}\exp(\boldsymbol{T}_i y)\boldsymbol{t}_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}. \qquad (3)
\end{aligned}$$

Similarly, for B_k^i we have

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

so

$$\mathbb{E}\left(\sum_{i=1}^{\infty} B_k^i \middle| Y=y\right) = \frac{\sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})\alpha_k \boldsymbol{e}_k' \exp(\boldsymbol{T}_i y)\boldsymbol{t}_i}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}.$$

Regarding Z_k^i,

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

The formula for $\mathbb{E}(N_{k\ell}^i|Y=y)$ is derived in a similar way, resulting in

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

Finally,

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

The $\boldsymbol{\theta}$ needs to be treated separately on a case–by–case basis and it generally involves a numerical procedure for finding the next iteration $\boldsymbol{\theta}^{(n+1)}$. We need to maximize

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

subject to $\sum_{i=1}^{\infty}\pi_i(\boldsymbol{\theta})=1$, where (see (3))

$$\mathbb{E}(L^i|Y=y) = \frac{\pi_i(\boldsymbol{\theta})}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}\boldsymbol{\alpha}\exp(\boldsymbol{T}_i y)\boldsymbol{t}_i.$$

In general, for $M>1$ datapoints we simply sum the previous formulas with arguments y_j, $j=1,...,M$.

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

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

we have that (see [12])

$$\exp\left(\begin{pmatrix}\boldsymbol{T} & \boldsymbol{t}\boldsymbol{\alpha}\\ \boldsymbol{0} & \boldsymbol{T}\end{pmatrix}y\right) = \begin{pmatrix} e^{\boldsymbol{T}y} & \boldsymbol{J}(y)\\ \boldsymbol{0} & e^{\boldsymbol{T}y}\end{pmatrix}. \tag{4}$$

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

The EM–algorithm can now be stated as follows.

Theorem 3.2 (EM–algorithm)**.**

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

1: *(E–step) Compute*

$$\begin{aligned}
\mathbb{E}(L^i|\boldsymbol{Y}=\boldsymbol{y}) &= \pi_i(\boldsymbol{\theta}) \sum_{j=1}^{M} \frac{\boldsymbol{\alpha}\exp(\boldsymbol{T}_i y_j)\boldsymbol{t}_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} \\
\mathbb{E}(B_k^i|\boldsymbol{Y}=\boldsymbol{y}) &= \pi_i(\boldsymbol{\theta}) \sum_{j=1}^{M} \frac{\alpha_k \boldsymbol{e}_k' \exp(\boldsymbol{T}_i y_j)\boldsymbol{t}_i}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} \\
\mathbb{E}\left(Z_k^i|\boldsymbol{Y}=\boldsymbol{y}\right) &= \pi_i(\boldsymbol{\theta}) \sum_{j=1}^{M} \frac{\boldsymbol{J}(y_j/s_i;\boldsymbol{\alpha},\boldsymbol{T})_{kk}}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} \\
\mathbb{E}(N_{k\ell}^i|\boldsymbol{Y}=\boldsymbol{y}) &= \frac{\pi_i(\boldsymbol{\theta})}{s_i} \sum_{j=1}^{M} \frac{\boldsymbol{J}(y_j/s_i;\boldsymbol{\alpha},\boldsymbol{T})_{\ell k} t_{k\ell}}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})} \\
\mathbb{E}(N_k^i|\boldsymbol{Y}=\boldsymbol{y}) &= \frac{\pi_i(\boldsymbol{\theta})}{s_i} \sum_{j=1}^{M} \frac{\boldsymbol{\alpha} e^{\boldsymbol{T}_i y_j} \boldsymbol{e}_k t_k}{f_Y(y_j;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}
\end{aligned}$$

2: *(M–step) Maximize*

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

subject to $\sum_i \pi_i(\boldsymbol{\theta}) = 1$ *and let* $\hat{\boldsymbol{\theta}}$ *be the argument at which the maximum is attained. Let*

$$\begin{aligned}
\hat{\alpha}_k &= \frac{1}{M} \sum_{i=1}^{\infty} \mathbb{E}(B_k^i|\boldsymbol{Y}=\boldsymbol{y}) \\
\hat{t}_{k\ell} &= \frac{\sum_{i=1}^{\infty} \mathbb{E}(N_{k\ell}^i|\boldsymbol{Y}=\boldsymbol{y})}{\sum_{i=1}^{\infty} \frac{1}{s_i}\mathbb{E}(Z_k^i|\boldsymbol{Y}=\boldsymbol{y})} \\
\hat{t}_k &= \frac{\sum_{i=1}^{\infty} \mathbb{E}(N_k^i|\boldsymbol{Y}=\boldsymbol{y})}{\sum_{i=1}^{\infty} \frac{1}{s_i}\mathbb{E}(Z_k^i|\boldsymbol{Y}=\boldsymbol{y})}
\end{aligned}$$

and assign diagonal values $\hat{t}_{kk} = -\sum_{l\neq k} \hat{t}_{k\ell} - \hat{t}_k$ *so* $\hat{\boldsymbol{T}} = \{\hat{t}_{k\ell}\}_{k,\ell=1,\ldots,p}$.

3: *Reassign values to initial parameters*

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

4: *GOTO 1.*

Remark 3.3. Consider the E–step of Theorem 3.2. Suppose that there are repeated values among the data points such that among the data $y_1, \ldots, y_M$ there are $\tilde{y}_1, \ldots, \tilde{y}_D$ different values and that $\tilde{y}_i$ appears $k_i \geq 1$ times in the original data. Then $k_1 + \cdots k_D = M$ and the sums over $j = 1, \ldots, M$ in expected values can then be reduced to weighted sums of fewer terms instead. For example,

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

This shows that the EM–algorithm of Theorem 3.2 can also be used to estimate an NPH distribution when data are represented by a histogram rather than the raw data. This is particularly suitable when the amount of data is large since the histogram may here adequately represent the target distribution.

Remark 3.4. In order to reduce the computational burden, we may replace the raw data by a histogram in certain regions of the support where data gather rather densely. Heavy–tailed data with support on $[0,\infty)$ will typically concentrate in an interval $[0,T)$ (the "body" of the distribution) while it will be more scarce in $[T,\infty)$ (the "tail").

Now assume that M is large and that the support for the data may be split into $[0,T)$ and $[T,\infty)$ for some $T>0$ such that the concentration of data points in $[0,T]$ is large. Then divide $[0,T]$ into K disjoint sub–intervals $[T_i,T_{i+1})$, $i=0,1,\ldots,K-1$ where $0=T_0<T_1<\cdots<T_K=T$ and k_i denote the number of data points falling into $[T_i,T_{i+1})$. Then we may use the repeated data reduction of Remark 3.3 for the interval $[0,T)$ by treating $\frac{T_i+T_{i+1}}{2}$, $i=0,...,K-1$ as data points with counts k_i. In fact we might choose any point in $[T_i,T_{i+1})$ as a representative, including any data point falling in this interval, however, if we choose the left points of the intervals, T_i, as representatives, we have to make special arrangements regarding the first interval in order not to provoke an atom at zero.

Concerning the tail, $[T,\infty)$, we continue to use the raw data since the scattering in the tail is typically too diffuse in order to be represented by a histogram without too much loss of information.

Example 3.5. 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. It features an explicit solution to the M–step maximization problem of the function

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

where $w_i=\mathbb{E}(L_i|\boldsymbol{Y}=\boldsymbol{y})$. Define

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

This distribution corresponds to a the discretization of a Pareto distribution F and the resulting discrete distribution is supported by $\{e^{ic}:i=\in\mathbb{N}\}$. For $c>0$, N with parameter $e^{-\theta c}$. The advantage of representing the scaling Pareto distribution on this form is that its argument which maximizes

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

where $w_i=\mathbb{E}(L_i|\boldsymbol{Y}=\boldsymbol{y})$, and it is not difficult to see it has an explicit solution given by

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

The selected support of N is also convenient for implementation purposes. In practice, the infinite series defining the estimators can only be computed up to a finite number of terms. The sequence $\pi(\theta)$ converges faster to 0 so a significant reduction of terms to be computed is required in order to attain a certain specified precision.

Since the distance between consecutive points in the support of N is unbounded, then the distribution of N is not regularly varying and Breiman's lemma no longer applies. Nevertheless, the tail probability of Y oscillates between two regularly varying functions, so this model can provide an accurate approximation to any regularly varying distribution [11].

4. CENSORING

In certain situations some data may be censored. Again we consider first the case of a single data point y taken from a realization of $Y \sim \mathrm{NPH}_p(\pi(\boldsymbol{\theta}), \boldsymbol{\alpha}, \boldsymbol{T})$. The data point is right censored at t if the only knowledge about Y is that $Y > t$, left censored at t if $Y \leq t$ and interval censored at $(s,t]$ if $Y \in (s,t]$. Left censoring is a special case of interval censoring with $s = 0$ while right censoring can be obtained by fixing s and letting $t \to \infty$. Formulas for right censoring will, however, appear as a part of the derivation of interval censoring.

The EM algorithm works entirely the same way as for uncensored data with the only difference that we are no longer observing a data point $Y = y$ but $Y \in (s,t]$. This will only change the E–step where we now have to calculate the following conditional expectations:

$$\mathbb{E}(L^i|Y \in (s,t]),\ \mathbb{E}(B_k^i|Y \in (s,t]),\ \mathbb{E}(Z_k^i|Y \in (s,t]),\ \mathbb{E}(N_{k\ell}^i|Y \in (s,t])\ \text{ and }\ \mathbb{E}(N_k^i|Y \in (s,t]).$$

Concerning L^i, notice that

$$\begin{aligned}\mathbb{E}(L^i 1\{Y \leq t\}) &= \mathbb{P}(Y \leq t|I = i)\mathbb{P}(I = i)\\ &= \pi_i(\boldsymbol{\theta})(1 - \boldsymbol{\alpha} e^{\boldsymbol{T}_i t}\boldsymbol{e})\end{aligned}$$

and

$$\mathbb{E}(L^i 1\{Y > t\}) = \pi_i(\boldsymbol{\theta})\boldsymbol{\alpha} e^{\boldsymbol{T}_i t}\boldsymbol{e}.$$

Thus

$$\begin{aligned}\mathbb{E}(L^i|Y \in (s,t]) &= \frac{\mathbb{E}(L^i 1\{Y \in (s,t]\})}{\mathbb{P}(Y \in (s,t])}\\ &= \frac{\mathbb{E}(L^i) - \mathbb{E}(L^i 1\{Y \leq s\}) - \mathbb{E}(L^i 1\{Y > t\})}{\mathbb{P}(Y \in (s,t])}\\ &= \frac{\pi_i(\boldsymbol{\theta}) - \pi_i(\boldsymbol{\theta})(1 - \boldsymbol{\alpha} e^{\boldsymbol{T}_i s}\boldsymbol{e}) - \pi_i(\boldsymbol{\theta})\boldsymbol{\alpha} e^{\boldsymbol{T}_i t}\boldsymbol{e}}{\mathbb{P}(Y \in (s,t])}\\ &= \frac{\pi_i(\boldsymbol{\theta})(\boldsymbol{\alpha} e^{\boldsymbol{T}_i s}\boldsymbol{e} - \boldsymbol{\alpha} e^{\boldsymbol{T}_i t}\boldsymbol{e})}{\sum_{i=1}^{\infty}\pi_i(\boldsymbol{\alpha})(\boldsymbol{\alpha} e^{\boldsymbol{T}_i s}\boldsymbol{e} - \boldsymbol{\alpha} e^{\boldsymbol{T}_i t}\boldsymbol{e})}.\end{aligned}$$

For right censored data we get

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

The rest of the formulas are derived as for censored phase–type distributions (see [9]) conditionally on the level L^i, with parameters $\boldsymbol{\alpha}, \boldsymbol{T}_i$, which happens with probability $\pi_i(\boldsymbol{\theta})$. Thus we get that

$$\mathbb{E}(B_k^i|Y \in (s,t]) = \frac{\pi_i(\boldsymbol{\theta})\left(\alpha_k \boldsymbol{e}_k' e^{\boldsymbol{T}_i s}\boldsymbol{e} - \alpha_k \boldsymbol{e}_k' e^{\boldsymbol{T}_i e}\boldsymbol{e}\right)}{\mathbb{P}(Y \in (s,t])}$$

$$\mathbb{E}(N_{k\ell}^i|Y \in (s,t]) = \frac{\pi_i(\boldsymbol{\theta}) t_{k\ell}/s_i\left(\int_s^t \boldsymbol{\alpha} e^{\boldsymbol{T}_i u}\boldsymbol{e}_k du - \int_s^t \boldsymbol{e}_\ell' e^{\boldsymbol{T}_i (t-u)}\boldsymbol{e}\boldsymbol{\alpha} e^{\boldsymbol{T}_i u}\boldsymbol{e}_k du\right)}{\mathbb{P}(Y \in (s,t])}$$

$$\mathbb{E}(Z_k^i|Y \in (s,t]) = \frac{\pi_i(\boldsymbol{\theta})\left(\int_s^t \boldsymbol{\alpha} e^{\boldsymbol{T}_i u}\boldsymbol{e}_k du - \int_s^t \boldsymbol{e}_k' e^{\boldsymbol{T}_i (t-u)}\boldsymbol{e}\boldsymbol{\alpha} e^{\boldsymbol{T}_i u}\boldsymbol{e}_k du\right)}{\mathbb{P}(Y \in (s,t])}$$

$$\mathbb{E}(N_k^i|Y \in (s,t]) = \frac{\pi_i(\boldsymbol{\theta}) t_k/s_i \int_s^t \boldsymbol{\alpha} e^{\boldsymbol{T}_i u}\boldsymbol{e}_k du}{\mathbb{P}(Y \in (s,t])}$$

with similar and obvious formulas for the right censored case. The integrals are calculated similarly as in the uncensored case, namely

$$\begin{aligned}\int_s^t \boldsymbol{\alpha} e^{\boldsymbol{T}_i u}\boldsymbol{e_k} du &= \boldsymbol{\alpha}\boldsymbol{T}_i^{-1}(e^{\boldsymbol{T}_i s}-e^{\boldsymbol{T}_i t})\boldsymbol{e}_k \\ &= s_i\boldsymbol{\alpha}\boldsymbol{T}^{-1}(e^{\boldsymbol{T}_i s}-e^{\boldsymbol{T}_i t})\boldsymbol{e}_k \\ \int_s^t \boldsymbol{e}'_\ell e^{\boldsymbol{T}_i(t-u)}\boldsymbol{e}\boldsymbol{\alpha}e^{\boldsymbol{T}_i u}\boldsymbol{e}_k du &= \boldsymbol{J}^i(t)_{\ell k}-\boldsymbol{J}^i(s)_{\ell k},\end{aligned}$$

where

$$\boldsymbol{J}^i(y)=\exp\left[\begin{pmatrix}\boldsymbol{T}_i & \boldsymbol{e}\boldsymbol{\alpha}\\ \boldsymbol{0} & \boldsymbol{T}_i\end{pmatrix}y\right].$$

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 3.4 we suggested the use of a histogram for representing the "body" of the distribution 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 by a distribution $G\in$ NPH if for example methods for calculation the ruin probability based on the claim size distribution H is 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 F 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 distribution F.

The idea is to let the number of data points $M\to\infty$ such that the distribution H is seen as an empirical distribution function for a data set with $M=+\infty$. If we let $M\to\infty$ then (see Theorem 3.2)

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

Similarly it is not difficult to see that

$$\begin{aligned}\hat{t}_{k\ell}&=\frac{\sum_{i=1}^{\infty}\mathbb{E}(N^i_{k\ell}|\boldsymbol{Y}=\boldsymbol{y})}{\sum_{i=1}^{\infty}\frac{1}{s_i}\mathbb{E}(Z^i_k|\boldsymbol{Y}=\boldsymbol{y})}=\frac{\sum_{i=1}^{\infty}\frac{1}{M}\mathbb{E}(N^i_{k\ell}|\boldsymbol{Y}=\boldsymbol{y})}{\sum_{i=1}^{\infty}\frac{1}{s_i}\frac{1}{M}\mathbb{E}(Z^i_k|\boldsymbol{Y}=\boldsymbol{y})}\\ &\to\frac{\displaystyle\sum_{i=1}^{\infty}\frac{\pi_i(\boldsymbol{\theta})}{s_i}\int_0^{\infty}\frac{\boldsymbol{J}(y/s_i;\boldsymbol{\alpha},\boldsymbol{T})_{\ell k}t_{k\ell}}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}dH(y)}{\displaystyle\sum_{i=1}^{\infty}\frac{\pi_i(\boldsymbol{\theta})}{s_i}\int_0^{\infty}\frac{\boldsymbol{J}(y/s_i;\boldsymbol{\alpha},\boldsymbol{T})_{kk}}{f_Y(y;\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{T})}dH(y)}\end{aligned}$$

and

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

Concerning $\boldsymbol{\theta}$, maximizing

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

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

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

which is then the function to be maximized. Let $\hat{\boldsymbol{\theta}}$ 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. The EM algorithm now works as follows:

0: initiate with some parameters $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{T})$.
1: calculate $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{T}}$ and $\hat{\boldsymbol{\theta}}$.
2: assign $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{T}) = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}})$.
3: GOTO 1.

6. EXAMPLES

In this section we provide five 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. 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 log–gamma, Weibull and log–normal respectively.

Example 6.1 (Erlang distributions)**.** A q dimensional Erlang distribution, $\mathrm{ER}_q(\lambda)$, is a phase–type distribution with canonical representation

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

Hence, the correspoding NPH distribution has density

$$f_Y(y;\boldsymbol{\theta},\lambda) = \sum_{i=1}^{\infty} \pi_i(\boldsymbol{\theta}) h(y/s_i;q,\lambda) = \sum_{i=1}^{\infty} \pi_i(\boldsymbol{\theta}) \left(\frac{\lambda}{s_i}\right)^q \frac{y^{q-1}}{(q-1)!} e^{-y\lambda/s_i}, \tag{5}$$

and the maximum likelihood estimator for λ is easily calculated to be

$$\widehat{\lambda}^{-1} = \frac{1}{qM} \sum_{i=1}^{\infty} \frac{Z^i}{s_i}, \qquad Z^i = \sum_{j=1}^{N} y_j \cdot 1(I_j = i),$$

while the maximum likelihood estimate for $\boldsymbol{\theta}$ has the general form

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\theta} \sum_{i=1}^{\infty} L^i \log \pi_i(\theta), \qquad L^i = \sum_{j=1}^{N} 1(I_j = i).$$

The E–step is similar as for the general model. In the case of the sufficient statistic Z^i we have

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

As for the sufficient statistic L^i, its expected value is analogue for the general model in the previous section and found to be equal to

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

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

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

where

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

is the Riemann Zeta function with parameter θ. This distribution is also known as the Riemann Zeta 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,

$$\frac{\zeta'(\theta)}{\zeta(\theta)} = -\frac{1}{M} \sum_{i=1}^{\infty} \mathbb{E}\left(L^i|\boldsymbol{Y}=\boldsymbol{y}\right) \log(i), \tag{6}$$

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

Results of the simulation study, which is shown in Table 1, reveals 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.

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

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 [6, 10]. 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 the 2167 data, 519 are repeated values so only 1648 have different values.

We propose an NPH model for the data by mixing a five dimensional phase–type distribution with an N that is assumed to follow the discretized Pareto distribution of Example 3.5.

Just above 90% of the data is below 5, while less than 10% falls between 5 and the maximum observation of 262.2504. Thus we select the "body part" of the distribution as the densely populated region $[0,5]$ while the much more diffuse region $[5,\infty)$ is then considered the "tail".

First we test the difference in performance between the EM–algorithm using raw data and one using the hybrid method proposed in Remark 3.4.

We divide $[0,5]$ into 100 subintervals of the same size, thus treating losses within 50,000 DKR as all being the same. We choose the centre of intervals as data points. In the tail we have 186 data points of which 8 are repeated values so a total of 178 different values. This amounts to a total of 278 distinct data points. The result of the experiment is following. While the actual discretization of $[0,5]$ 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 5.75 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 $1648/278 = 5.92$.

In all the EM algorithms which follows we have used the hybrid method of Remark 3.4 with $[0,5]$ divided into 100 equally sized subintervals.

We now investigate the effect the choice of the parameter c (see Example 3.5) will have on the estimation. Intuitively, the smaller we choose the discretization steps of the geometric progression, the better the estimation. Fixing the dimension of the phase–type distribution to five, we select three different values for c, 1, $1/2$ and $1/4$. To obtain the EM estimates we randomly selected various sets of initial and iterated the EM until the relative change in the loglikelihood was smaller than 10^{-8}. We kept the model with the largest likelihood. The estimates are as follows

$c = 1$**:**

$$\begin{aligned}
\hat{\theta} &= 1.2743 \\
\hat{\boldsymbol{\alpha}} &= (0.6415, 0.0099, 0.0055, 0.2115, 0.1316) \\
\hat{\boldsymbol{T}} &= \begin{pmatrix} -2.7430 & 1.3565 & 0 & 0 & 0 \\ 0.0003 & -3.0398 & 0 & 0 & 0 \\ 0 & 0 & -2.6313 & 1.1226 & 0.2167 \\ 0 & 0 & 0.7223 & -1.3953 & 0.4089 \\ 0 & 0 & 1.5129 & 0.8388 & -2.4779 \end{pmatrix}
\end{aligned}$$

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

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

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

$$\begin{aligned}
\hat{\theta} &= 1.3230 \\
\hat{\boldsymbol{\alpha}} &= (0.0267, 0.4563, 0.0010, 0.2603, 0.2556)
\end{aligned}$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -0.7896 & 0.2531 & 0.0385 & 0.2652 & 0.0265 \\ 0.0810 & -3.7022 & 1.7356 & 1.1626 & 0.3934 \\ 0.5838 & 0.0013 & -3.7890 & 0.0576 & 0.0302 \\ 0.3386 & 0.0606 & 1.0320 & -3.7735 & 0.1957 \\ 0.6694 & 0.0292 & 0.2376 & 0.3920 & -3.4129 \end{pmatrix}$$

That the estimates for $(\boldsymbol{\alpha}, \boldsymbol{T})$ look distinct is because phase–type representations are by no means unique, however, the estimated densities which are plotted in Figure 1 against the histogram of the data are almost identical (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 of the survival function in 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 only result in a slightly different tail behaviour for very large claim sizes. We conclude that the estimation does not depend significantly on the choice of c within the possible choices of $c = 1, 1/2$ or $1/4$, and can probably be extrapolated to concluding robustness against choices in c within any "reasonable" range for c.

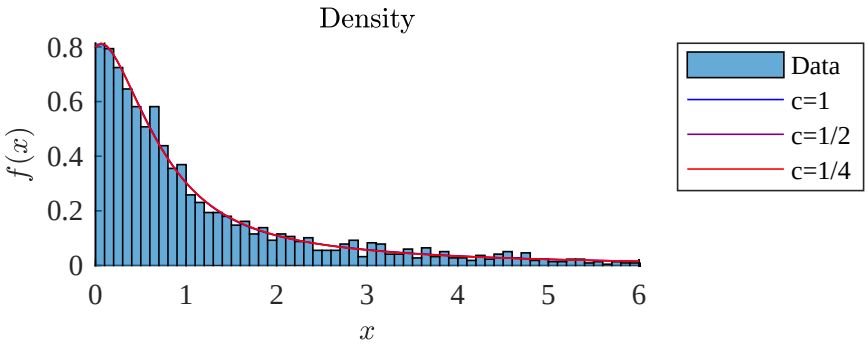

FIGURE 1. Estimated densities against histogram of the NPH models for Danish reinsurance data.

Finally we compare the fit for $c = 1/4$ to one obtained by [6] and [10]. 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 [10] the values in the interval [1.40, 1.46] produce excellent fits, the value recommend by the same author being 1.45.

We finally compare our estimated model with $c = 1/4$ to a NPH model where we fix $\theta = 1.45$ as recommended by [10]. 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.

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

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

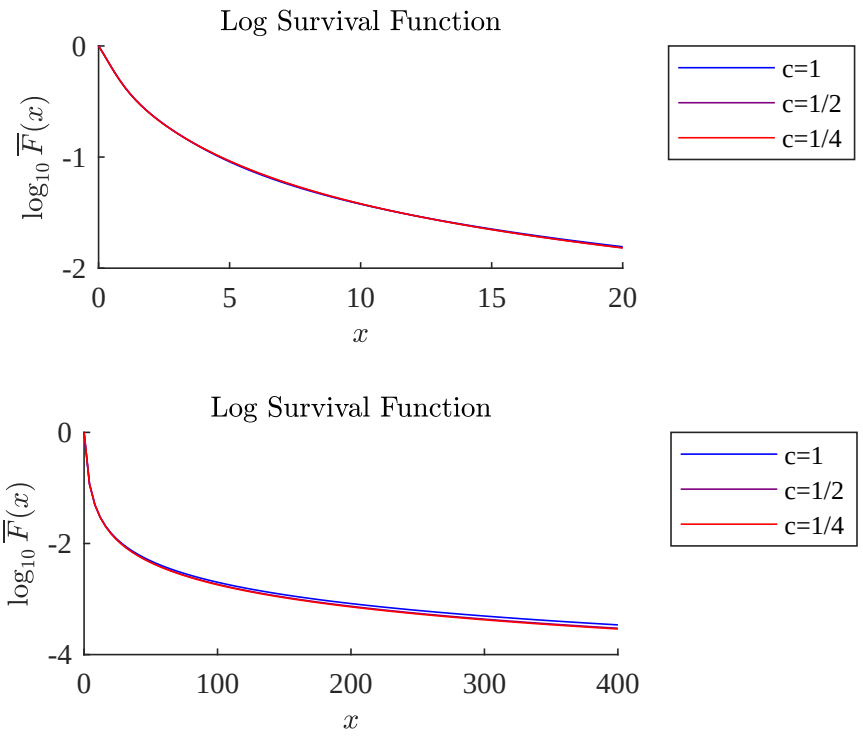

FIGURE 2. Estimated survival functions of the NPH models for Danish reinsurance data corresponding to θ equal to $1.2748, 1.3115, 1.3387$.

Again 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 fixed and EM adjusted tails in the range $[0,6]$ where the 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.

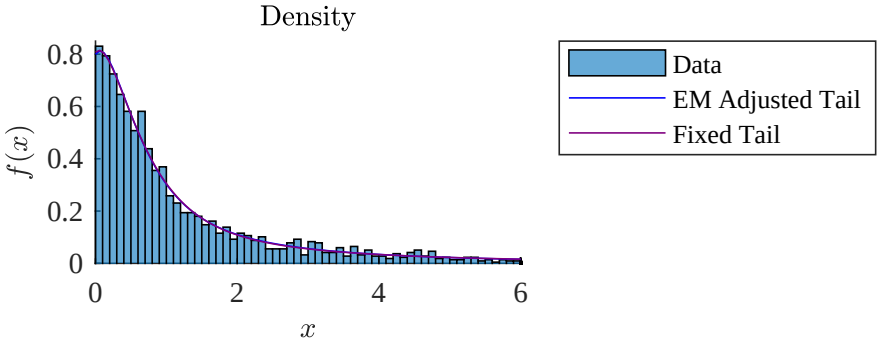

FIGURE 3. Estimated densities of the NPH models with EM adjusted ($\hat{\theta} = 1.3387$) and fixed ($\hat{\theta} = 1.45$) tails against histogram of the Danish reinsurance data.

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 log–gamma distribution with shape parameter α, scale parameter β and shifted one unit to the left

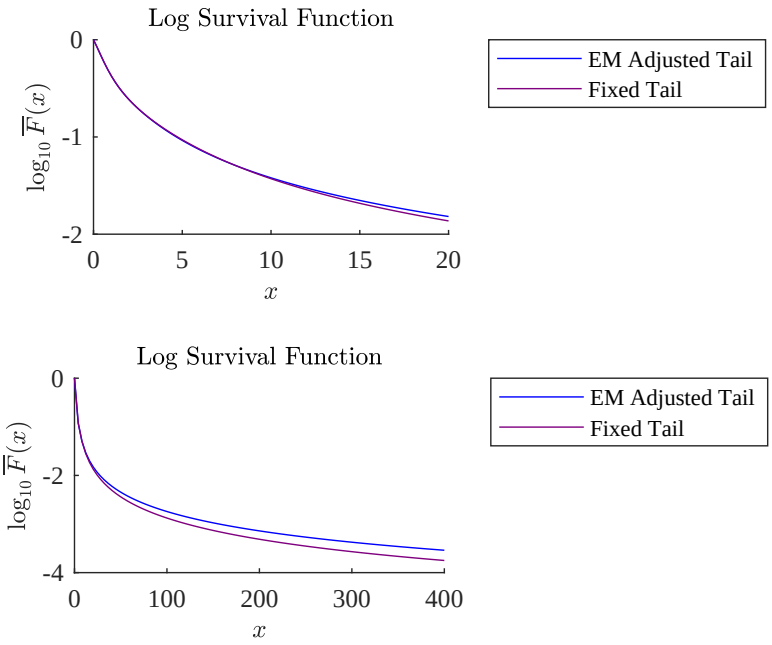

FIGURE 4. Estimated survival functions of the NPH models with EM adjusted ($\hat{\theta} = 1.3387$) and fixed ($\hat{\theta} = 1.45$) tails.

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 purposes 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 3.5 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 3.5.

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

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

The densities of the log-gamma distribution and its NPH approximation are plotted in Figure 5. The densities of the log-gamma 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 $\hat{\theta} = 1.6031$).

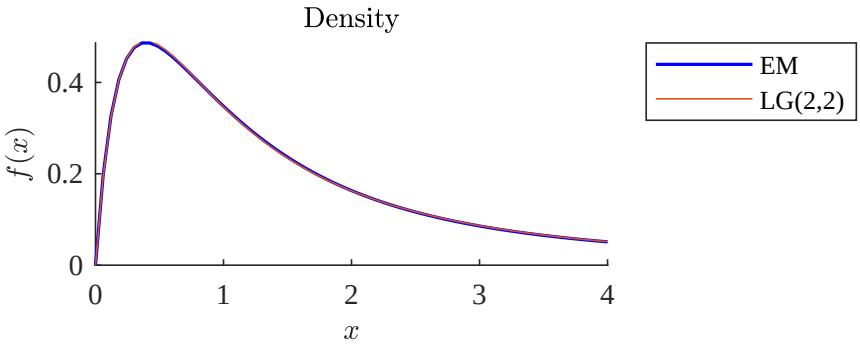

FIGURE 5. Density of the Loggamma distribution with parameters $(2,2)$ and an approximation via an NPH model with EM adjustment.

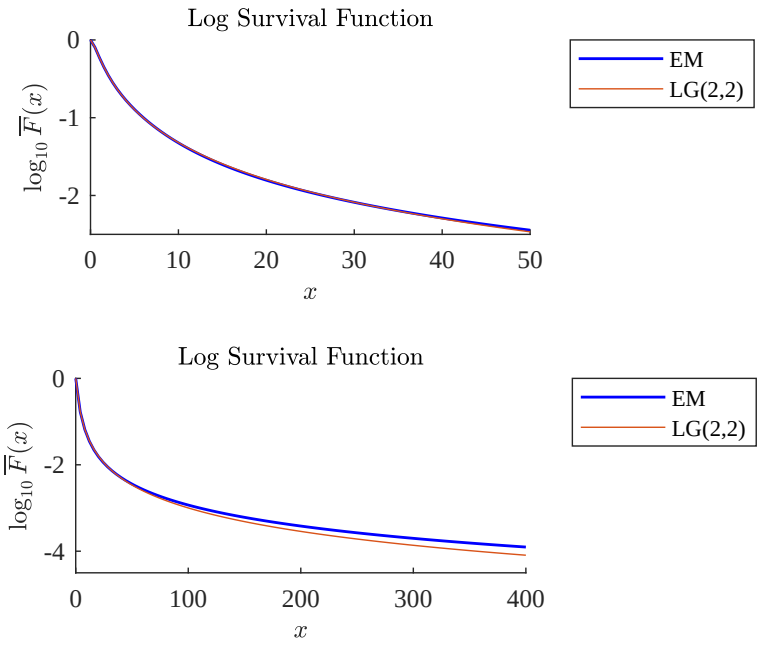

FIGURE 6. Survival functions of the Loggamma distribution LG$(2,2)$ and an approximation via an NPH model with EM adjustment.

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

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

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

$$\begin{aligned} \hat{\lambda} &= 0.6181, \ \hat{p} = 1.1673 \\ \hat{\boldsymbol{\alpha}} &= (0.2370, 0.3349, 0.0901, 0.3380, 0) \end{aligned}$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -3.0858 & 0.6964 & 0.2188 & 0.3355 & 0.5572 \\ 51.0253 & -207.2799 & 18.3961 & 18.5851 & 46.0724 \\ 0.4369 & 0.2511 & -0.9839 & 0.0122 & 0.0487 \\ 1.1297 & 0.2448 & 0.9388 & -11.0548 & 0.7637 \\ 0.7583 & 1.0994 & 0.9844 & 0.4882 & -3.3303 \end{pmatrix}$$

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.

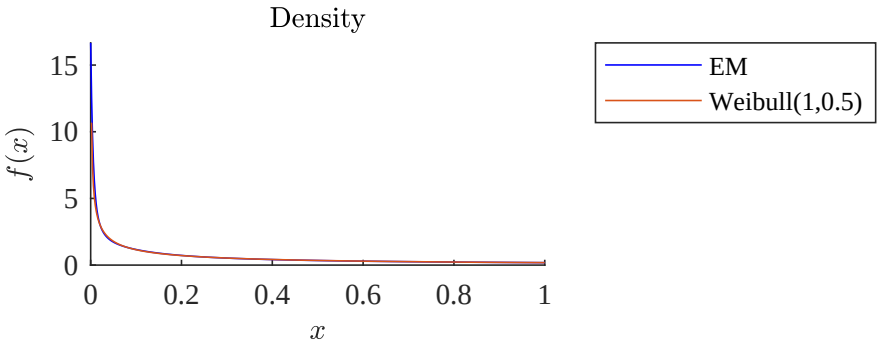

FIGURE 7. Density of the Weibull distribution with parameters $(1,0.5)$ and an approximation via NPH models with EM adjustment.

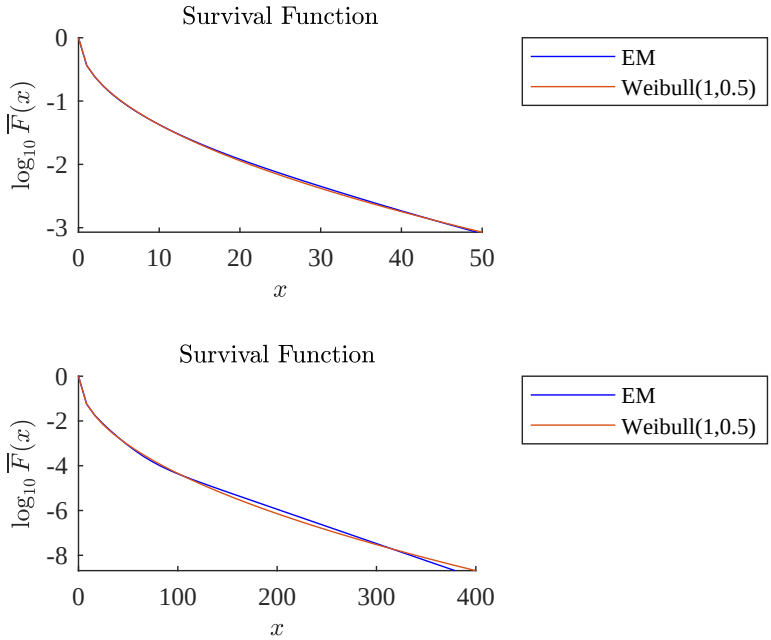

FIGURE 8. Survival functions of the Weibull$(1,0.5)$ and an approximation via an NDPH model with EM adjustment. The EM model is Weibullian with shape parameter $p = \widehat{p}(1+\widehat{p}) \approx 0.5386$.

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 $\boldsymbol{\pi}(\boldsymbol{\theta})$ for which the NPH model has the same tail behavior as the lognormal distribution is unknown.

Therefore, we consider two alternative NPH models to adjust the data. For both we have selected a phase–type part has eight phases and the scaling distribution is chosen as a discretization of a Lognormal distribution $\mathrm{LN}(\mu, \sigma^2)$ 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 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

$$\hat{\mu} = 0.1909, \quad \hat{\sigma} = 0.5979$$

$$\hat{\boldsymbol{\alpha}} = (0.0000, 0.0000, 0.0000, 0, 0.0000, 0.0439, 0.1360, 0.8201)$$

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -7.1741 & 0.1557 & 0.3338 & 6.2239 & 0.0195 & 0.1985 & 0.1061 & 0.1366 \\ 0.2056 & -1.1622 & 0.0528 & 0.2434 & 0.1133 & 0.0519 & 0.0328 & 0.0782 \\ 0.7528 & 0.0891 & -6.2221 & 4.7969 & 0.1873 & 0.2546 & 0.1238 & 0.0176 \\ 0.3840 & 0.1133 & 0.1303 & -9.4035 & 0.1845 & 0.0654 & 0.2161 & 0.1597 \\ 0.9550 & 0.4274 & 1.5430 & 0.9113 & -5.2223 & 0.2926 & 1.0875 & 0.0056 \\ 0.2341 & 0.0842 & 2.3467 & 0.0000 & 1.3606 & -4.6381 & 0.0621 & 0.5504 \\ 1.8359 & 0.0083 & 1.0178 & 0.0000 & 0.0211 & 0.2123 & -3.3825 & 0.2870 \\ 4.9757 & 0.0000 & 0.9689 & 0.0000 & 3.4702 & 0.2689 & 0.3086 & -9.9923 \end{pmatrix}$$

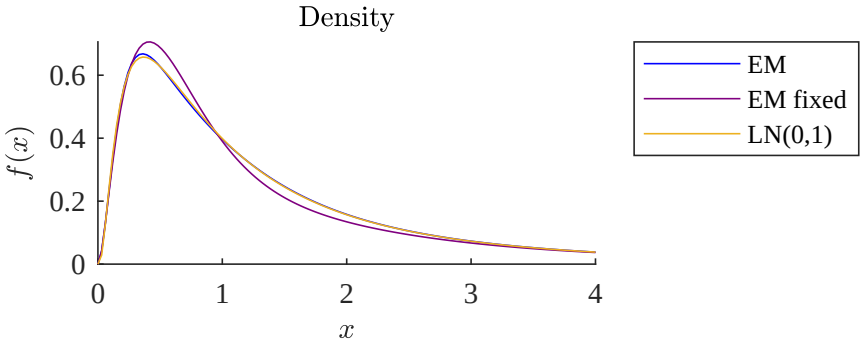

FIGURE 9. Density of the Lognormal distribution with parameters $(0, 1)$ and an approximation via NPH models with EM adjustment.

From Figure 9 we can observe that the adjustment of the body of the distributions of the first model is excellent, 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, [11] demonstrate that the tail probability of the NPH model is significantly heavier and different to the scaling distribution. More

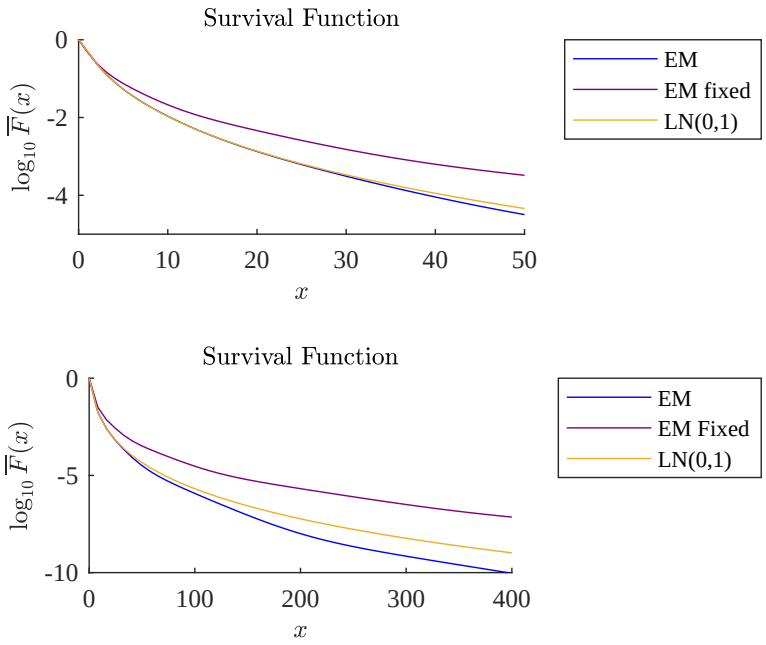

FIGURE 10. Survival functions of the Lognormal distribution $\mathrm{LN}(0,1)$ and an approximation via an NPH model with EM adjustment.

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 introduced a new methodology for adjusting phase–type scale mixture distributions to heavy–tailed data. 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 by a NPH distribution. In particular, for the case of regular varying and Weibullian distributions the aforementioned approximation is not only in the limit (denseness) but can be effectively carried out in praxis.

REFERENCES

[1] Marek Arendarczyk and Krzysztof 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] M Bladt, B. F. Nielsen, and G. Samorodnitsky. Calculation of ruin probabilities for a dense class of heavy tailed distributions. *Scandinavian Actuarial Journal*, pages 573–591, 2015.

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

[5] Arnoldo Frigessi, Ola Haug, and Håvard Rue. A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3):219–235, 2002.

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

[7] Guy Latouche and Vaidyanathan Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*, volume 5. Society for Industrial Mathematics, 1987.
[8] Marcel F Neuts. *Matrix-geometric Solutions in Stochastic Models*. An Algorithmic Approach. Courier Dover Publications, 1981.
[9] M Olsson. Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics*, pages 443–460, 1996.
[10] Sidney I. Resnick. Discussion of the Danish Data on Large Fire Insurance Losses. *ASTIN Bulletin*, 27(1):139–151, 1997.
[11] Leonardo Rojas-Nandayapa and Wangyue Xiei. Asymptotic tail behavior of phase-type scale mixture distributions. *Submitted*.
[12] Charles 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

SCHOOL OF MATHEMATICS AND PHYSICS, THE UNIVERSITY OF QUEENSLAND, ST. LUCIA 4072, BRISBANE, AUSTRALIA
E-mail address: l.rojas@uq.edu.au