SCANDINAVIAN ACTUARIAL JOURNAL
https://doi.org/10.1080/03461238.2022.2097019

OPEN ACCESS

Phase-type mixture-of-experts regression for loss severities

Martin Bladt[a] and Jorge Yslas[b]

[a]Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland; [b]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

ABSTRACT
The task of modeling claim severities is addressed when data is not consistent with the classical regression assumptions. This framework is common in several lines of business within insurance and reinsurance, where catastrophic losses or heterogeneous sub-populations result in data difficult to model. Their correct analysis is required for pricing insurance products, and some of the most prevalent recent specifications in this direction are mixture-of-experts models. This paper proposes a regression model that generalizes the latter approach to the phase-type distribution setting. More specifically, the concept of mixing is extended to the case where an entire Markov jump process is unobserved and where states can communicate with each other. The covariates then act on the initial probabilities of such underlying chain, which play the role of expert weights. The basic properties of such a model are computed in terms of matrix functionals, and denseness properties are derived, demonstrating their flexibility. An effective estimation procedure is proposed, based on the EM algorithm and multinomial logistic regression, and subsequently illustrated using simulated and real-world datasets. The increased flexibility of the proposed models does not come at a high computational cost, and the motivation and interpretation are equally transparent to simpler MoE models.

ARTICLE HISTORY
Received 31 October 2021
Accepted 29 June 2022

KEYWORDS
Mixture-of-experts; regression modeling; phase-type distributions; multinomial regression

1. Introduction

The correct estimation of claim severities is a classical problem in actuarial science, and yet the task remains challenging and often only solvable by partially formal procedures. For instance, when dealing with data arising from reinsurance of natural catastrophes or from third-party liability insurance, very large claims are treated differently to the bulk of smaller – or attritional – claim sizes. Multimodality of the attritional claims can further exacerbate the problem. Such heterogeneity in the data is often still present after segmentation, possibly due to pooled and unlabeled sub-populations in the dataset. In addition, most common commercial software, which mostly uses generalized linear models (GLM), does not capture quantiles correctly. Consequently, risk managers and actuaries interested in understanding their fitted probabilistic models (and not only using them for prediction) keep returning to the drawing board to obtain more interpretable, flexible and effective statistical tools for their practice.

Several statistically coherent approaches have been proposed in recent years to overcome multimodality and heavy-tailedness. For instance, mixing Erlang distributions results in multimodal histograms, Lee & Lin (2010) being the first to consider such a model for insurance, and then later extended by Tzougas et al. (2014), Miljkovic & Grün (2016) for more general mixtures. More recent

CONTACT Martin Bladt martin.bladt@unil.ch Faculty of Business and Economics, University of Lausanne, Quartier de Chambronne, Lausanne 1015, Switzerland

approaches, such as Fung et al. (2019), adopt a mixture-of-experts approach, which consists of regressing the component probabilities of a finite mixture model. Regarding the heavy-tailed component, the formal way of dealing with the attritional and large claims jointly has been using splicing – also referred to as composite models – which have a different tail and body distribution (see Grün & Miljkovic 2019 for a comparison and a good literature review). Combining the two approaches is the state-of-the-art of probabilistic models for loss severity modeling, referred to as composite models. Reynkens et al. (2017) were the first to consider this global approach, and Fung et al. (2021b) suggested a feature-selection variant.

The main idea of this paper is to use phase-type (PH) distributions to capture the specificities of heterogeneous data more intuitively and effectively than mixing. The underlying multi-state model is easy to motivate and understand for practitioners and is mathematically convenient for developing their estimation. More specifically, we propose to use PH distributions to describe claim severities and build our regression framework with PH building blocks. PH distributions are defined as the absorption time of a time-homogeneous Markov pure-jump process on a finite state space. In life insurance, such a framework is familiar and understood as the traversing of healthy, disabled, and dead states, with time corresponding to calendar time. In non-life insurance, the states can be regarded as unobserved steps in legal cases or reparations of a building, and time now corresponds to the incurred monetary loss.

Many distributions such as the Erlang, generalized Coxian, and finite mixtures between them are all PH distributions (see Neuts (1975, 1981) for the first systematic approaches, and Bladt & Nielsen (2017) for a recent comprehensive treatment), and they are even known to be dense in weak convergence on the set of distributions of positive-valued risks (cf. Asmussen 2003). Most applications of PH were initially in the field of applied probability, but their estimation became widely used after (Asmussen et al. 1996) laid out the EM-algorithm for statistical fitting. To correct for non-exponential tail behavior, Albrecher & Bladt (2019), Albrecher et al. (2022) defined and provided estimation approaches for transformed PH distributions, also known as inhomogeneous phase-type (IPH) laws.

To incorporate rating factors into our model, we consider regressing the initial probabilities of the underlying stochastic process starting in a given state. This is in the same spirit as the mixture-of-experts approaches, cf. Yuksel et al. (2012) for a survey, which can roughly be described as machine learning methods where inhomogeneous data regions are divided into homogeneous ones, where simpler models can suffice for their description. A different approach to regression with PH distributions was considered in Albrecher et al. (2021a), Bladt (2021), where the proportional intensities (PI) model was proposed. The strength of PH regression models is that no threshold selection is required, and a tail behavior specification can be easily done by choosing an appropriate inhomogeneity function. Moreover, the interaction between the hidden states allows for complex density shapes, going beyond what simple mixing can account for, for a given number of experts. In essence, the latter property can have a parsimonious effect on the number of estimation parameters. However, the phase-type mixture-of-experts (PH-MoE) approach can obtain a wider range of variation for a fixed state-space size than the PI approach and can be faster to estimate for a small number of covariates.

The PH-MoE model is rather flexible, illustrated by two denseness results on: (a) multinomial experiments with arbitrary distributions assigned to each outcome; (b) more general regression models, subject to some technical conditions. The marginal and conditional distributions associated with the PH-MoE specification fall into the IPH class, for which many closed-form formulas exist, and their tail behavior is well understood. Furthermore, their estimation can be carried out using an ingenious decomposition of the fully observed likelihood into two components: one which may be maximized using a variant of the EM algorithm for PH distributions; and another component can be seen as a weighted one multinomial logistic regression problem.

The remainder of the paper is structured as follows. First, in Section 2, we provide a short reminder of IPH distributions, specify the main regression model, and derive its basic properties, along with the first denseness result. We then prove the denseness of PH-MoE on regression models in Section 3 and provide an effective estimation technique based on the EM algorithm and weighted multinomial regression in Section 4, along with a goodness of fit consideration. In Section 5, we review some common choices of inhomogeneity functions for global fitting and introduce a new alternative to composite splicing models based on piecewise-continuous inhomogeneity functions. Subsequently, we show in Section 6 the practical feasibility of our approach on synthetic and real insurance data. Finally, Section 7 concludes.

2. Phase-type mixture-of-experts regression model

2.1. Preliminaries

Let $(J_t)_{t\geq 0}$ be a time-inhomogeneous Markov pure-jump process on the finite state space $\{1,\dots,p,p+1\}$, where states $1,\dots,p$ are transient and $p+1$ is absorbing. Then, the transition probabilities

$$p_{kl}(s,t) = \mathbb{P}(J_t = l \,|\, J_s = k), \quad 0 \leq k,\ l \leq p+1,$$

can be written in matrix form as

$$\boldsymbol{P}(s,t) = \prod_{s}^{t}(\boldsymbol{I} + \boldsymbol{\Lambda}(u)\,\mathrm{d}u) := \boldsymbol{I} + \sum_{i=1}^{\infty}\int_s^t \int_s^{u_i} \cdots \int_s^{u_2} \Lambda(u_1)\cdots\Lambda(u_i)\,\mathrm{d}u_1\cdots\mathrm{d}u_i,$$

for $s < t$, where $\boldsymbol{\Lambda}(t)$ is called the intensity matrix, that is, a matrix with negative diagonal elements and non-negative off-diagonal elements such that the rows sum to zero. If we further require that the matrices $\boldsymbol{\Lambda}(s)$ and $\boldsymbol{\Lambda}(t)$ commute for every $s < t$, this can be done by assuming the following structure of the intensity matrix

$$\boldsymbol{\Lambda}(t) = \lambda(t)\begin{pmatrix} \boldsymbol{T} & \boldsymbol{t} \\ \boldsymbol{0} & 0 \end{pmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}, \quad t \geq 0,$$

where $\boldsymbol{T}$ is a $p \times p$ sub-intensity matrix, $\boldsymbol{t}$ is a p-dimensional column vector providing the exit rates to the absorbing state, $\boldsymbol{0}$ is a p-dimensional row vector of zeroes, and $\lambda(\cdot)$ is some known positive real function. Since the rows of the intensity matrix sum to zero, the relationship $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$ holds, where $\boldsymbol{e}$ denotes the p-dimensional column vector of ones. In what follows, we always assume this structure of $\boldsymbol{\Lambda}(t)$ and that the function $\lambda(\cdot) > 0$ satisfies for $y > 0$

$$(0,\infty) \ni \int_0^y \lambda(t)\,\mathrm{d}t \overset{y\to\infty}{\to} \infty. \tag{1}$$

For future reference, we write $\boldsymbol{e}_k$ for the kth canonical basis vector in $\mathbb{R}^p$. Concerning the sub-intensity matrix and the vector of exit rates, we introduce the following notation for their entries

$$\boldsymbol{T} = (t_{kl})_{k,l=1,\dots,p}, \quad \boldsymbol{t} = (t_1,\dots,t_p)^{\mathsf{T}}.$$

We will make use of functions of matrices in the sequel. The standard unambiguous way of defining them is in terms of the Cauchy formula as follows. Let h be any analytic function and $\boldsymbol{A}$ a square matrix. Then we define

$$h(\boldsymbol{A}) = \frac{1}{2\pi i}\oint_{\Gamma} h(w)(w\boldsymbol{I} - \boldsymbol{A})^{-1}\,\mathrm{d}w, \quad \boldsymbol{A} \in \mathbb{R}^{p\times p},$$

with Γ a simple path enclosing the eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{I}$ is the identity matrix of the same dimension.

Definition 2.1: Let $\boldsymbol{\pi}_0$ be an initial distribution on $\{1,\dots,p\}$. Then, if $J_0 \sim \boldsymbol{\pi}_0$, we say that

$$Y_0 = \inf\{t > 0 : J_t = p+1\},$$

follows an inhomogeneous phase-type (IPH) distribution and we write $Y_0 \sim \mathrm{IPH}(\boldsymbol{\pi}_0, \boldsymbol{T}, \lambda)$.

The density f and distribution function F of $Y_0 \sim \mathrm{IPH}(\boldsymbol{\pi}_0, \boldsymbol{T}, \lambda)$ are explicit in terms of functions of matrices and given by

$$f(y) = \lambda(y)\boldsymbol{\pi}_0 \exp\left(\int_0^y \lambda(s)\,\mathrm{d}s\boldsymbol{T}\right)\boldsymbol{t}, \quad y \geq 0,$$
$$F(y) = 1 - \boldsymbol{\pi}_0 \exp\left(\int_0^y \lambda(s)\,\mathrm{d}s\boldsymbol{T}\right)\boldsymbol{e}, \quad y \geq 0.$$

Another attractive property of IPH distributions is that a random variable following this specification can be expressed as the transformation of a phase-type (PH) distributed random variable, that is, the homogenous case corresponding to $\lambda \equiv 1$. More specifically, if $Y_0 \sim \mathrm{IPH}(\boldsymbol{\pi}_0, \boldsymbol{T}, \lambda)$, then

$$Y_0 \overset{d}{=} g(Z_0), \tag{2}$$

where $Z_0 \sim \mathrm{PH}(\boldsymbol{\pi}_0, \boldsymbol{T})$ and g is defined through its inverse in terms of λ by

$$g^{-1}(y) = \int_0^y \lambda(s)\,\mathrm{d}s, \quad y \geq 0.$$

This representation is particularly useful to derive further properties of IPH distributions by exploiting the known PH machinery. For instance, the following explicit asymptotic behavior for the tails $\bar{F} = 1 - F$ of IPH distributions can be deduced using this representation in conjunction with the corresponding asymptotic result for PH distribution:

$$\bar{F}(y) \sim c[g^{-1}(y)]^{m-1}\exp(-\eta g^{-1}(y)), \quad y \to \infty, \tag{3}$$

where c is a positive constant depending on $\boldsymbol{\pi}$ and $\boldsymbol{T}$, $-\eta$ is the largest real eigenvalue of $\boldsymbol{T}$, and m is the size of the Jordan block associated with η.

2.2. The regression model

Define the mapping

$$\boldsymbol{\pi} : D \subset \mathbb{R}^d \to \Delta^{p-1},$$

where $\Delta^{p-1} = \{(\pi_1,\dots,\pi_p) \in \mathbb{R}^p \mid \sum_k \pi_k = 1 \mathrm{and} \pi_k \geq 0 \mathrm{for all} k\}$ is the standard $(p-1)$-simplex. Thus, for any given $\boldsymbol{x} \in \mathbb{R}^d$, we may endow the process with the initial probabilities

$$\mathbb{P}(J_0 = k) = \pi_k(\boldsymbol{x}) := (\boldsymbol{\pi}(\boldsymbol{x}))_k, \quad k = 1,\dots,p,$$

and $\mathbb{P}(J_0 = p+1) = 0$.

As a particular consequence, the following random variable

$$Y = \inf\{t > 0 : J_t = p+1\},$$

satisfies that

$$Y \sim \mathrm{IPH}(\boldsymbol{\pi}(\boldsymbol{x}), \boldsymbol{T}, \lambda) \;\Leftrightarrow\; J_0 \sim \boldsymbol{\pi}(\boldsymbol{x}).$$

Definition 2.2: Let $\boldsymbol{X}$ be a d-dimensional vector of covariates. Then we say that

$$Y \,|\, \boldsymbol{X} \sim \mathrm{IPH}(\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T}, \lambda)$$

is a phase-type mixture-of-experts (PH-MoE) model.

Remark 2.1: The above model obtains its name since we may write

$$\mathbb{P}(Y > y \,|\, \boldsymbol{X} = \boldsymbol{x}) = \sum_{k=1}^{p} \mathbb{P}(Y > y \,|\, J_0 = k)\pi_k(\boldsymbol{x}),$$

which is a mixture of p PH distributions with different initial distributions, each assigning all its mass to a given state. In particular, we have the simple identity

$$\mathbb{E}[Y \,|\, \boldsymbol{X} = \boldsymbol{x}] = \sum_{k=1}^{p} \mathbb{E}[Y \,|\, J_0 = k]\pi_k(\boldsymbol{x}).$$

For instance, in the homogeneous case we obtain

$$\mathbb{E}[Y \,|\, \boldsymbol{X} = \boldsymbol{x}] = \sum_{k=1}^{p} \pi_k(\boldsymbol{x})\boldsymbol{e}_k^{\mathsf{T}}[-\boldsymbol{T}]^{-1}\boldsymbol{e} = \boldsymbol{\pi}(\boldsymbol{x})^{\mathsf{T}}[-\boldsymbol{T}]^{-1}\boldsymbol{e}. \tag{4}$$

Example 2.3: If $D = \{\boldsymbol{x}_0\}$ is a singleton, then the PH-MoE model exactly spans the class of IPH distributions.

The following result shows that random covariates do not extend the marginal distribution beyond the above example.

Proposition 2.4: *Let $\boldsymbol{X}$ be a random vector in a convex $D \subset \mathbb{R}^d$. Then the PH-MoE model has marginal distribution given by*

$$\mathrm{IPH}(\boldsymbol{\pi}(\boldsymbol{x}^*), \boldsymbol{T}, \lambda),$$

for some $\boldsymbol{x}^* \in D$. In fact, $\boldsymbol{\pi}(\boldsymbol{x}^*) = \mathbb{E}(\boldsymbol{\pi}(\boldsymbol{X}))$.

Proof: Let $\boldsymbol{X}$ have density $f : D \to \mathbb{R}_+$. We simply observe that by disintegration we get

$$\mathbb{P}(J_0 = k) = \int_D \pi_k(\boldsymbol{x})f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \pi_k(\boldsymbol{x}^*), \quad k = 1, \ldots, p,$$

and since the process $(J_t)_{t\geq 0}$ otherwise has the same dynamics after any initiation, the other parameters are unchanged. It remains to notice that $\boldsymbol{x}^* \in D$ by convexity. ■

For the above reason, the PH-MoE model is most useful in its conditional form, and can be used for regression purposes. We now formulate a particularly advantageous parametrization for when $D = \mathbb{R}^d$.

Definition 2.5: We say that the PH-MoE model with initial probabilities $\boldsymbol{\pi}(\boldsymbol{X};\boldsymbol{\alpha}) = (\pi_k(\boldsymbol{X};\boldsymbol{\alpha}))_{k=1,\ldots,p}$ given by

$$\pi_k(\boldsymbol{X};\boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_k)}{\sum_{j=1}^{p}\exp(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_j)}, \quad k = 1, \ldots, p, \tag{5}$$

satisfies the softmax parametrization. Here, $\boldsymbol{\alpha}_k \in \overline{\mathbb{R}}^d$, $k = 1, \ldots, p$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathsf{T}}, \ldots, \boldsymbol{\alpha}_p^{\mathsf{T}})^{\mathsf{T}} \in \overline{\mathbb{R}}^{(p\times d)}$.

Remark 2.2: In essence, we consider the coefficients of $\boldsymbol{\alpha}$ as assigning 'expertly' each observation to an initial distribution $\boldsymbol{\pi}(\boldsymbol{X})$ according to their information $\boldsymbol{X}$.

For the above parametrization, we have that the logarithm of the ratio between any two probabilities is linear in that for any $k, j \in \{1, \dots, p\}$,

$$\log\left(\frac{\pi_k(\boldsymbol{X};\boldsymbol{\alpha})}{\pi_j(\boldsymbol{X};\boldsymbol{\alpha})}\right) = \boldsymbol{X}^{\mathsf{T}}(\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_j) = \sum_{i=1}^{d} X_i(\alpha_{ki} - \alpha_{ji}).$$

However, for two individuals with covariate information $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, a class-specific correction arises, as follows:

$$\log\left(\frac{\pi_k(\boldsymbol{X}_1;\boldsymbol{\alpha})}{\pi_j(\boldsymbol{X}_2;\boldsymbol{\alpha})}\right) = \boldsymbol{X}_1^{\mathsf{T}}\boldsymbol{\alpha}_k - \boldsymbol{X}_2^{\mathsf{T}}\boldsymbol{\alpha}_j + \log\left(\frac{\sum_{i=1}^{p}\exp(\boldsymbol{X}_2^{\mathsf{T}}\boldsymbol{\alpha}_i)}{\sum_{l=1}^{p}\exp(\boldsymbol{X}_1^{\mathsf{T}}\boldsymbol{\alpha}_l)}\right).$$

In regression analyses, in particular, in the analysis of variance (ANOVA) it can often be the case that there is a discrepancy in the mean between observations belonging to two or more categories. However, the conditional distributions may not be Gaussian, or may even be different between different groups. We make a technical definition for such common situations.

Definition 2.6: Let $W_1, \dots, W_n$ be positive and continuous random variables having otherwise arbitrary distributions, and let $\eta \in \{1, \dots, n\}$ be a multinomial random variable, such that

$$W_i \perp\!\!\!\perp W_j, \quad \forall i \neq j, \quad \text{and} \quad W_i \perp\!\!\!\perp_{\boldsymbol{X}} \eta, \quad \forall i,$$

and such that $\boldsymbol{X}$ contains at least an intercept. Then we say that $W_\eta \,|\, \boldsymbol{X}$ follows a multinomial mixture distribution.

Proposition 2.7: *Let $W \,|\, \boldsymbol{X}$ follow a multinomial mixture distribution. Then there exist a sequence of PH-MoE models $(Y_m \,|\, \boldsymbol{X})_{m \geq 0}$ such that*

$$Y_m \,|\, \boldsymbol{X} \xrightarrow{d} W \,|\, \boldsymbol{X}, \quad m \to \infty.$$

Moreover, the softmax parametrization may be chosen.

Proof: We have by definition that $W \,|\, \boldsymbol{X} \overset{d}{=} W_\eta \,|\, \boldsymbol{X}$ for $W_1, \dots, W_n$ some conditionally independent variables, and a conditionally independent multinomial variable $\eta \in \{1, \dots, n\}$. Given $\boldsymbol{X}$, and by the denseness of PH distributions, there exist sequences of PH distributed random variables Y_{im}, $i = 1, \dots, n$, $m = 1, 2, \dots$, such that, as $m \to \infty$,

$$Y_{im} \xrightarrow{d} W_i, \quad i = 1, \dots, n.$$

Since η takes finitely many values, it follows that even

$$Y_{\eta m} \,|\, \boldsymbol{X} \xrightarrow{d} W_\eta \,|\, \boldsymbol{X}, \quad m \to \infty.$$

It remains to note that for any given m, $Y_{\eta m} \,|\, \boldsymbol{X}$ is a finite mixture of independent PH variables, and thus PH distributed as well, with dimension p^* at most the sum of the individual mixture-component dimensions. Finally, since the covariates contain an intercept term, the softmax function (as a function of $\boldsymbol{\alpha}$)

$$\pi_k(\boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_k)}{\sum_{j=1}^{p^*}\exp(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_j)}, \quad k = 1, \dots, p^*,$$

is a surjective map from $\mathbb{R}^{(d \times p^*)}$ to Δ^{p^*-1}, and we may choose, for each m, $\boldsymbol{\alpha}$ to exactly match the required initial distribution of $Y_{\eta m} \,|\, \boldsymbol{X}$. ■

Remark 2.3: In the above result, any pre-specified tail behavior of $W \,|\, \boldsymbol{X}$ may be exactly matched by all the $Y_m \,|\, \boldsymbol{X}$, $m = 1, 2, \ldots$, by using the appropriate inhomogeneity function λ. The details are straightforward but technical and thus omitted.

3. Denseness on regression models

This section is devoted to showing a stronger version of Proposition 2.7, under some more restrictive conditions on the covariate space and the associated conditional distributions.

Definition 3.1: Let $\mathcal{A}$ be the set of possible values of the covariates $\boldsymbol{X}$. A severity regression model is the set of conditional distributions of claim severity, given the covariates, that is, the set of laws of

$$Y \,|\, \boldsymbol{X} = \boldsymbol{x}, \quad \boldsymbol{x} \in \mathcal{A}.$$

Given a severity regression model, we say that a sequence of severity regression models converges weakly (respectively, uniformly weakly) to it, if all the associated conditional distributions converge weakly for each $\boldsymbol{x} \in \mathcal{A}$ (respectively, uniformly weakly in $\boldsymbol{x} \in \mathcal{A}$).

Definition 3.2: A feature space $\mathcal{A}$ is said to be regular if it is of the form $\mathcal{A} = \{1\} \times [a, b]^{d-1}, a, b \in \mathbb{R}$, that is, the covariates contain an intercept and are otherwise contained in a hypercube.

Condition 3.3: A regression model is said to satisfy the tightness and Lipschitz conditions on $\mathcal{A}$ if

$$\{\mathbb{P}(Y \in \cdot \,|\, \boldsymbol{X} = \boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{A}}$$

is a tight family of distributions, and for each $y \geq 0$, the function

$$\boldsymbol{x} \mapsto \mathbb{P}(Y \leq y \,|\, \boldsymbol{X} = \boldsymbol{x})$$

is Lipschitz continuous in $\mathcal{A}$.

To allow zeroes in the vector of initial probabilities, we can assume without loss of generality that the vector of initial probabilities is of the form $\boldsymbol{\pi}(\boldsymbol{X}; \boldsymbol{\alpha}) = (\tilde{\boldsymbol{\pi}}^{\mathsf{T}}(\boldsymbol{X}; \boldsymbol{\alpha}), \boldsymbol{0})^{\mathsf{T}}$, where $\tilde{\boldsymbol{\pi}}(\boldsymbol{X}; \boldsymbol{\alpha}) = (\pi_k(\boldsymbol{X}; \boldsymbol{\alpha}))_{k=1,\ldots,q}$ is a q-dimensional column vector, $q \leq p$, with

$$\pi_k(\boldsymbol{X}; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{\alpha}_k)}{\sum_{j=1}^{q} \exp(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{\alpha}_j)}, \quad k = 1, \ldots, q.$$

Indeed, we can always reorder the states of the PH representation in such a way that the first $q \leq p$ entries of $\boldsymbol{\pi}(\boldsymbol{X}; \boldsymbol{\alpha})$ are the ones corresponding to the values different from zero.

Proposition 3.4 (Denseness): *Let a regression model satisfy the tightness and Lipschitz conditions on a regular $\mathcal{A}$. Then, there exists a sequence of PH-MoE regression models converging uniformly weakly to it.*

Proof: The claim follows from Theorem 3.3 in Fung et al. (2019) by noticing that LRMoE models with Erlang distributed severities (which satisfy Property 3 of Proposition 3.1 in Fung et al. 2019) are particular instances of the PH-MoE model. ■

Remark 3.1: Proposition 3.4 also implies that the PH-MoE models, with fixed inhomogeneity transformation g, form a dense class on the set of univariate severity regression distributions.

This is relevant since it allows us to obtain different tail behaviors for modeling claim severities. For instance, in Fung et al. (2019), it is shown that Pareto distributions fail to fulfill the denseness conditions in the LRMoE model. This implies that for a fixed splicing threshold, denseness and heavy-tails are not possible in that setting. In contrast, Pareto tail behavior is now possible using a PH-MoE model, while still preserving the denseness property.

It is worth mentioning that another alternative was recently introduced in Fung et al. (2021a), which was termed the TG-LRMoE model. The main idea of this model consists of transforming Gamma-distributed random variables to obtain heavy tails for the severity distributions (including Pareto tails). Note, however, that when considering Gamma random variables with integer shape parameters (i.e. Erlang) in the TG-LRMoE specification, we obtain a particular case of a PH-MoE model with intensity $\lambda(y) = (1+y)^{\gamma-1}$, $\gamma > 0$.

4. Estimation

4.1. The EM algorithm

Suppose that we have a PH-MoE specification

$$Y \mid \boldsymbol{X} \sim \mathrm{IPH}(\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T}, \lambda).$$

By a simple inhomogeneity transformation, we may momentarily concentrate on the homogeneous case as follows:

$$Z \mid \boldsymbol{X} := g^{-1}(Y \mid \boldsymbol{X}) \sim \mathrm{PH}(\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T}).$$

Now, let $B_k(\boldsymbol{X})$ be the number of times that the process $(J_t)_{t\geq 0}$ with initial distribution $\boldsymbol{\pi}(\boldsymbol{X})$ starts in state k, $N_{kl}(\boldsymbol{X})$ the total number of jumps from state k to l conditional on $\boldsymbol{\pi}(\boldsymbol{X})$, $N_k(\boldsymbol{X})$ the number of times that we reach the absorbing state $p+1$ from state k conditional on $\boldsymbol{\pi}(\boldsymbol{X})$, and let $V_k(\boldsymbol{X})$ be the total time that the underlying Markov jump process spends in state k prior to absorption given $\boldsymbol{\pi}(\boldsymbol{X})$. Then, given a sample of absorption times $\boldsymbol{z} = (z_1, \dots, z_N)^{\mathsf{T}}$ and the corresponding paired covariate information $\bar{\boldsymbol{x}} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$, the completely observed likelihood can be written in terms of the previously defined statistics as follows:

$$\begin{aligned}
\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T} \mid \boldsymbol{z}, \bar{\boldsymbol{x}}) &= \prod_{i=1}^{N} \mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T} \mid z_i, \boldsymbol{X} = \boldsymbol{x}_i) \\
&= \prod_{i=1}^{N}\prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{B_k(\boldsymbol{x}_i)} \prod_{k=1}^{p}\prod_{l\neq k} t_{kl}{}^{N_{kl}(\boldsymbol{x}_i)} \exp(-t_{kl}V_k(\boldsymbol{x}_i)) \prod_{k=1}^{p} t_k{}^{N_k(\boldsymbol{x}_i)} \exp(-t_k V_k(\boldsymbol{x}_i)) \\
&= \left(\prod_{i=1}^{N}\prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{B_k(\boldsymbol{x}_i)}\right) \prod_{k=1}^{p}\prod_{l\neq k} t_{kl}{}^{\sum_{i=1}^{N} N_{kl}(\boldsymbol{x}_i)} \exp\left(-t_{kl}\sum_{i=1}^{N} V_k(\boldsymbol{x}_i)\right) \\
&\quad \times \prod_{k=1}^{p} t_k{}^{\sum_{i=1}^{N} N_k(\boldsymbol{x}_i)} \exp\left(-t_k \sum_{i=1}^{N} V_k(\boldsymbol{x}_i)\right) \\
&= \left(\prod_{i=1}^{N}\prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{B_k(\boldsymbol{x}_i)}\right) \prod_{k=1}^{p}\prod_{l\neq k} t_{kl}{}^{N_{kl}} \exp(-t_{kl}V_k) \prod_{k=1}^{p} t_k{}^{N_k} \exp(-t_k V_k),
\end{aligned}$$

with

$$N_{kl} := \sum_{i=1}^{N} N_{kl}(\boldsymbol{x}_i), \quad V_k := \sum_{i=1}^{N} V_k(\boldsymbol{x}_i), \quad N_k := \sum_{i=1}^{N} N_k(\boldsymbol{x}_i).$$

The above specification partially belongs to the exponential family of distributions and thus has semi-explicit maximum likelihood estimators.

Since the full-trajectory data is not observed, we employ the expectation-maximization (EM) algorithm to estimate part of the MLE iteratively. This implies that at each iteration, the conditional expectations of the sufficient statistics $B_k(\boldsymbol{x}_i)$, N_{kl}, N_k, and V_k given the absorption times $\boldsymbol{z}$ are computed, corresponding to the E-step. Then $\mathcal{L}_c(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{z})$ is maximized by replacing the values of the statistics by their corresponding expected values from the previous step, obtaining in this way updated parameters $(\boldsymbol{\pi}, \boldsymbol{T})$, commonly referred to as the M-step.

Then the detailed formulas are given as follows:

(1) *E-step, conditional expectations:*

$$\mathbb{E}(B_k(\boldsymbol{x}_i) \mid Z = z_i, \boldsymbol{X} = \boldsymbol{x}_i) = \frac{\pi_k(\boldsymbol{x}_i)\boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T} z_i)\boldsymbol{t}}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i)\boldsymbol{t}}, \quad i = 1, \dots, N,$$

$$\mathbb{E}(V_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}}) = \sum_{i=1}^{N} \frac{\int_0^{z_i} \boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{t}\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} u)\boldsymbol{e}_k \, \mathrm{d}u}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i)\boldsymbol{t}},$$

$$\mathbb{E}(N_{kl} \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}}) = \sum_{i=1}^{N} t_{kl} \frac{\int_0^{z_i} \boldsymbol{e}_l{}^{\mathsf{T}} \exp(\boldsymbol{T}(z_i - u))\boldsymbol{t}\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} u)\boldsymbol{e}_k \, \mathrm{d}u}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i)\boldsymbol{t}},$$

$$\mathbb{E}(N_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}}) = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i)\boldsymbol{e}_k}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i)\boldsymbol{t}}.$$

(2) *M-step, explicit maximum likelihood estimators:*

$$\hat{t}_{kl} = \frac{\mathbb{E}(N_{kl} \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}{\mathbb{E}(V_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}, \quad \hat{t}_k = \frac{\mathbb{E}(N_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}{\mathbb{E}(V_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}, \quad \hat{t}_{kk} = -\sum_{l \neq k} \hat{t}_{kl} - \hat{t}_k.$$

(3) *R-step, weighted multinomial regression estimation:*

$$\hat{\pi}(\cdot) = \arg\max_{\pi(\cdot) \in \Delta^{p-1}} \left(\prod_{i=1}^{N} \prod_{k=1}^{p} \pi_k(x_i)^{\mathbb{E}(B_k(x_i) \mid Z = z_i, X = x_i)} \right),$$

where as before Δ^{p-1} is the standard $(p-1)$-simplex.

Finally, to incorporate the inhomogeneity transformation $g(\cdot)$, we assume that this is a parametric function depending on some vector $\boldsymbol{\theta}$, that is, we consider $g(\cdot; \boldsymbol{\theta})$. Then, $\boldsymbol{\theta}$ is updated in a subsequent step consisting of direct maximization of the incomplete likelihood function with respect to (solely) this parameter.

Remark 4.1: In general, the set of all functions in the simplex is too broad, and a parametric family is chosen – such as the softmax functions. Even then, no explicit solution for the R-step is available. We describe the entire procedure for the softmax case in Algorithm 1.

In view that the R-step is computed numerically even for the simplest logistic case, we see that Algorithm 1 easily extends to the case where an arbitrary regression model with a categorical response is used to predict the initial Markov probabilities, for instance, when specifying $\boldsymbol{x} \mapsto \boldsymbol{\pi}(\boldsymbol{x})$ as a neural network. In this framework, the multinomial logistic regression model can be seen as a 0-layer neural network.

Direct calculations, or general results from EM theory, yield the following result.

Algorithm 1 EM algorithm for PH-MoE (Softmax parametrization)

```
Input: Positive data points $\boldsymbol{y} = (y_1, \dots, y_N)^{\mathsf{T}}$, covariates $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, and initial parameters $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\theta})$.
(1) Mixture specification: Set
    $\pi_k(\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_k)}{\sum_{j=1}^{p} \exp(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_j)}, \quad i = 1, \dots, N, \ k = 1, \dots, p.$
(2) Inhomogeneity transformation: Transform the data into
    $z_i = g^{-1}(y_i; \boldsymbol{\theta}), \quad i = 1, \dots, N.$
(3) E-step: Compute the statistics
    $\mathbb{E}(B_k(\boldsymbol{x}_i) \mid Z = z_i, \boldsymbol{X} = \boldsymbol{x}_i) = \frac{\pi_k(\boldsymbol{x}_i) \boldsymbol{e}_k^{\mathsf{T}} \exp(\boldsymbol{T} z_i) \boldsymbol{t}}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i) \boldsymbol{t}}, \quad i = 1, \dots, N,$
    $\mathbb{E}(V_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}}) = \sum_{i=1}^{N} \frac{\int_0^{z_i} \boldsymbol{e}_k^{\mathsf{T}} \exp(\boldsymbol{T}(z_i - u)) \boldsymbol{t} \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} u) \boldsymbol{e}_k \, \mathrm{d}u}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i) \boldsymbol{t}},$
    $\mathbb{E}(N_{kl} \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}}) = \sum_{i=1}^{N} t_{kl} \frac{\int_0^{z_i} \boldsymbol{e}_l^{\mathsf{T}} \exp(\boldsymbol{T}(z_i - u)) \boldsymbol{t} \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} u) \boldsymbol{e}_k \, \mathrm{d}u}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i) \boldsymbol{t}},$
    $\mathbb{E}(N_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}}) = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i) \boldsymbol{e}_k}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}_i) \exp(\boldsymbol{T} z_i) \boldsymbol{t}}.$
(4) M-step: Let
    $\hat{t}_{kl} = \frac{\mathbb{E}(N_{kl} \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}{\mathbb{E}(V_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}, \quad \hat{t}_k = \frac{\mathbb{E}(N_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}{\mathbb{E}(V_k \mid \boldsymbol{Z} = \boldsymbol{z}, \bar{\boldsymbol{x}})}, \quad \hat{t}_{kk} = -\sum_{l \neq k} \hat{t}_{kl} - \hat{t}_k.$
(5) R-step: Maximize the weighted multinomial logistic regression
    $\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^{(p \times d)}} \sum_{i=1}^{N} \sum_{k=1}^{p} \mathbb{E}(B_k(\boldsymbol{x}_i) \mid Z = z_i, \boldsymbol{X} = \boldsymbol{x}_i) \log(\pi_k(\boldsymbol{x}_i; \boldsymbol{\alpha})),$
    and set
    $\hat{\pi}_k(\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) = \frac{\exp(\boldsymbol{x}_i^{\mathsf{T}} \hat{\boldsymbol{\alpha}}_k)}{\sum_{j=1}^{p} \exp(\boldsymbol{x}_i^{\mathsf{T}} \hat{\boldsymbol{\alpha}}_j)}, \quad i = 1, \dots, N, \ k = 1, \dots, p.$
(6) Inhomogeneity optimization: Maximize
    $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log \left( \lambda(y_i; \boldsymbol{\theta}) \hat{\boldsymbol{\pi}}^{\mathsf{T}}(\boldsymbol{x}_i) \exp\left( \int_0^{y_i} \lambda(s; \boldsymbol{\theta}) \, \mathrm{d}s \hat{\boldsymbol{T}} \right) \hat{\boldsymbol{t}} \right).$
(7) Update the current parameters to $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}}, \hat{\boldsymbol{\theta}})$. Return to step 1 unless a stopping rule is satisfied.
Output: Fitted representation $(\boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{\theta})$.
```

Proposition 4.1: *The likelihood function is increasing at each iteration of Algorithm 1. For a given p, the likelihood is also bounded, and we guarantee convergence to a (possibly local) maximum.*

Notice that although convergence occurs, even if the parameters are such that the MLE of the PH distribution is asymptotically consistent, convergence to such MLE is still not guaranteed.

4.2. Censoring

In applications, an observation may be partially observed, in that only upper and/or lower bounds may be determined, but not its actual size. This incurs in a large bias if the bounds are far apart, and thus a statistical correction is required. Below we outline such adaptation to the estimation technique for PH-MoE models.

In essence, the EM Algorithm 1 can be modified to work with censored observations, with just some adjustments on the formulas of the E-step being required. Recall that a data point is said to be right-censored at a if it takes an unknown value above a, left-censored at b if it takes an unknown value below b, and more generally interval-censored at $(a, b]$ if it takes an unknown value within the interval $(a, b]$. Moreover, note that for any censored observation of a PH-MoE model $Y \mid \boldsymbol{X} \sim \mathrm{IPH}(\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T}, \lambda)$, the inhomogeneity transformation $g^{-1}(\cdot)$ results on a censored observation (of the same type) in the homogeneous setting $Z \mid \boldsymbol{X} = g^{-1}(Y \mid \boldsymbol{X}) \sim \mathrm{PH}(\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T})$, meaning that we formally only need to deal with the latter case.

In the following, we provide the explicit formulas for the E-step in the interval-censoring setting. Results for left and right censoring then follow as special cases, given that left-censoring can be seen as interval-censoring with $a = 0$ and right-censoring is retrieved by fixing a and letting $b \to \infty$. Thus, for a single generic interval-censored observation $Z \in (a, b]$ with covariate information $\boldsymbol{X} = \boldsymbol{x}$, we have that

$$\mathbb{E}(B_k(\boldsymbol{x}) \mid Z \in (a,b], \boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_k(\boldsymbol{x})\boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T}a)\boldsymbol{e} - \pi_k(\boldsymbol{x})\boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T}b)\boldsymbol{e}}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}a)\boldsymbol{e} - \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}b)\boldsymbol{e}},$$

$$\begin{aligned}
&\mathbb{E}(V_k \mid Z \in (a,b], \boldsymbol{X} = \boldsymbol{x}) \\
&\quad = \frac{1}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}a)\boldsymbol{e} - \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}b)\boldsymbol{e}} \left[\boldsymbol{x} \int_a^b \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}u)\boldsymbol{e}_k \,\mathrm{d}u \right. \\
&\qquad - \int_0^b \boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T}(b-u))\boldsymbol{t}\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}u)\boldsymbol{e}_k \,\mathrm{d}u \\
&\qquad \left. + \int_0^a \boldsymbol{e}_k{}^{\mathsf{T}} \exp(\boldsymbol{T}(a-u))\boldsymbol{t}\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}u)\boldsymbol{e}_k \,\mathrm{d}u \right],
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}(N_{kl} \mid Z \in (a,b], \boldsymbol{X} = \boldsymbol{x}) \\
&\quad = \frac{t_{kl}}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}a)\boldsymbol{e} - \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}b)\boldsymbol{e}} \left[\boldsymbol{x} \int_a^b \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}u)\boldsymbol{e}_k \,\mathrm{d}u \right. \\
&\qquad - \int_0^b \boldsymbol{e}_l{}^{\mathsf{T}} \exp(\boldsymbol{T}(b-u))\boldsymbol{t}\,\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}u)\boldsymbol{e}_k \,\mathrm{d}u \\
&\qquad \left. + \int_0^a \boldsymbol{e}_l{}^{\mathsf{T}} \exp(\boldsymbol{T}(a-u))\boldsymbol{t}\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}u)\boldsymbol{e}_k \,\mathrm{d}u \right],
\end{aligned}$$

$$\mathbb{E}(N_k \mid Z \in (a,b], \boldsymbol{X} = \boldsymbol{x}) = t_k \frac{\int_a^b \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}u)\boldsymbol{e}_k \,\mathrm{d}u}{\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}a)\boldsymbol{e} - \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}b)\boldsymbol{e}}.$$

The other steps of the algorithm are unchanged.

4.3. Goodness of fit for phase-type regression models

We propose a common visual tool for assessing the goodness of fit of the overall model. The procedure is described for the case when right-censored observations are present since it is the most common scenario in applications. We define the residuals of a PH-MoE model by

$$r_i = -\log\left(\boldsymbol{\pi}(\boldsymbol{x}_i;\boldsymbol{\alpha})\exp\left(\int_0^{y_i}\lambda(s;\boldsymbol{\theta})\,\mathrm{d}s\boldsymbol{T}\right)\boldsymbol{e}\right),\quad i=1,\dots,N,$$

which under right-censoring completely at random and assuming (essentially, specifying the null-hypothesis) that the true distribution is indeed such PH-MoE, then by plugging in the estimated parameters from the EM algorithm, we obtain a dataset

$$\{(r_1\ \delta_1),\ (r_2,\ \delta_2),\dots,(r_N,\ \delta_N)\},$$

which follows a right-censored mean one exponential distribution. Here, δ_i, $i=1,\dots,N$, denote censoring indicators. In turn, we may construct a Kaplan-Meier survival curve for this dataset, $S(r)$, which should roughly resemble $S_0(r)=\exp(-r)$. To obtain a confidence band, we may use Greenwood's formula $\mathrm{Var}(S(r))=S(r)^2\sum_{i:r_i\le r}d_i/(n_i(n_i-d_i))$, where d_i is the number of tied values at r_i, and n_i all values yet to be observed (or at risk).

5. Transforms

The shape of the intensity function λ is a central assumption of the PH-MoE model, which in particular determines the tail behavior, as can be deduced from (3). This section introduces two useful global parametrizations for heavy-tailed distributions, and subsequently considers semi-composite models, which combine the conceptual approaches of splicing with our current setting.

Before introducing the parametric forms, we provide the exact tail behavior, which follows immediately from (3).

Proposition 5.1: *Let $Y\,|\,\boldsymbol{X}$ be a PH-MoE specification. Let $\boldsymbol{T}(\boldsymbol{X})$ be the sub-intensity matrix associated with the Markov jump-process $(J_t)_{t\ge0}$ restricted to the accessible states $A(\boldsymbol{X})\subset\{1,\dots,p\}$ when starting according to the distribution $\boldsymbol{\pi}(\boldsymbol{X})$. Then*

$$\bar{F}_{Y|\boldsymbol{X}}(y\,|\,\boldsymbol{x})\sim c(\boldsymbol{x})[g^{-1}(y)]^{m(\boldsymbol{x})-1}\exp(-\eta(\boldsymbol{x})g^{-1}(y)),\quad y\to\infty,$$

where $c(\boldsymbol{x})$ is a positive constant depending on $\boldsymbol{\pi}(\boldsymbol{x})$ and $\boldsymbol{T}(\boldsymbol{x})$, $-\eta(\boldsymbol{x})$ is the largest real eigenvalue of $\boldsymbol{T}(\boldsymbol{x})$, and $m(\boldsymbol{x})$ is the size of the Jordan block associated with $\eta(\boldsymbol{x})$.

In particular, if $\boldsymbol{\pi}(\boldsymbol{x})$ never has zeros or if all states of the Markov process communicate, then $\boldsymbol{T}(\boldsymbol{x})=\boldsymbol{T}$, for all $\boldsymbol{x}$, and all subgroups of the population have the same tail parameters.

5.1. Global models

Global models in the context of PH-MoE refer to parametrizations of λ with respect to the same function on all of $\mathbb{R}_+$, as opposed to piece-wise functions. Such specifications are natural when considering the interpretation of the g^{-1} function: it serves as a time transform that changes throughout time in a smooth way, that is, with continuous derivatives.

5.1.1. Pareto PH-MoE

Consider the transformation

$$Y\,|\,\boldsymbol{X}=\theta(\exp(Z\,|\,\boldsymbol{X})-1),$$

where $Z\,|\,\boldsymbol{X}\sim\mathrm{PH}(\boldsymbol{\pi}(\boldsymbol{X}),\boldsymbol{T})$ and $\theta>0$. Then, for $y\ge0$,

$$\bar{F}_{Y|\boldsymbol{X}}(y\,|\,\boldsymbol{x})=\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x})\left(\frac{y}{\theta}+1\right)^{\boldsymbol{T}}\boldsymbol{e},$$

$$f_{Y|\boldsymbol{X}}(y\,|\,\boldsymbol{x}) = \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \left(\frac{y}{\theta} + 1\right)^{\boldsymbol{T}-\boldsymbol{I}} \boldsymbol{t}\frac{1}{\theta}.$$

Here, $g(y) = \theta(\exp(y) - 1)$ and $g^{-1}(y) = \log(y/\theta + 1)$. Consequently, the intensity function is given by

$$\lambda(y) = \frac{1}{y+\theta}.$$

We refer to $Y|\boldsymbol{X}$ as a Pareto PH-MoE. It then follows from Proposition 5.1 that

$$\bar{F}_{Y|\boldsymbol{X}}(y\,|\,\boldsymbol{x}) \sim L(y,\boldsymbol{x})y^{-\eta(\boldsymbol{x})},$$

as $y \to \infty$, where $L(\cdot,\boldsymbol{x})$ is a slowly varying function, that is, it satisfies that $\lim_{y\to\infty} L(cy,\boldsymbol{x})/L(y,\boldsymbol{x}) = 1$ for all $c > 0$, and $-\eta(\boldsymbol{x})$ is the largest real eigenvalue of $\boldsymbol{T}(\boldsymbol{x})$. The Pareto MoE is designed to capture heavy-tailed (in the sense of regular variation) distributions with additional flexibility in the body of the distribution arising from the matrix parameters.

5.1.2. Weibull PH-MoE

If we now instead consider

$$Y\,|\,\boldsymbol{X} = (Z\,|\,\boldsymbol{X})^{1/\theta},$$

where $Z\,|\,\boldsymbol{X} \sim \mathrm{PH}(\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T})$ and $\theta > 0$, then for $y \geq 0$,

$$\bar{F}_{Y|\boldsymbol{X}}(y\,|\,\boldsymbol{x}) = \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}y^{\theta})\boldsymbol{e},$$
$$f_{Y|\boldsymbol{X}}(y\,|\,\boldsymbol{x}) = \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T}y^{\theta})\boldsymbol{t}\theta y^{\theta-1}.$$

Hence, $g(y) = y^{1/\theta}, g^{-1}(y) = y^{\theta}$, and

$$\lambda(y) = \theta y^{\theta-1}.$$

We refer to this model as a Weibull PH-MoE, where loosely speaking we obtain, for each observation, a Weibull tail behavior with a matrix in place of the usual scale parameter. From Proposition 5.1, it follows that

$$\bar{F}_{Y|\boldsymbol{X}}(y\,|\,\boldsymbol{x}) \sim c(\boldsymbol{x})y^{\gamma(\boldsymbol{x})} \exp(-\eta(\boldsymbol{x})y^{\theta}), \tag{6}$$

as $y \to \infty$, where $c(\boldsymbol{x}) > 0$, $\gamma(\boldsymbol{x}) \geq 0$, and $-\eta(\boldsymbol{x})$ is as above.

This model is suitable for a wider range of applications, since it falls into the Gumbel max-domain of attraction, which implies that it has strictly lighter tails than those of Pareto-type. However, for $\theta < 1$ (respectively, $\theta > 1$), we get that (6) specifies heavier (respectively, lighter) tails than exponentially decaying ones.

An interesting feature of this specification, and contrary to the Pareto case, is that conditional means are fully explicit and given by

$$\mathbb{E}(Y^{\zeta}\,|\boldsymbol{X}) = \Gamma(1 + \zeta/\theta)\boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{X})(-\boldsymbol{T})^{-\zeta/\theta}\boldsymbol{e} \quad \forall \zeta > 0.$$

Example 5.2 (Different tail behaviors): We illustrate the importance of Proposition 5.1 by providing a simple two-groups case where different tail behavior arises. Consider the matrix

$$\boldsymbol{T} = \begin{pmatrix} -1 & 0.5 & 0 \\ 1 & -2 & 0 \\ 0 & 0 & -3 \end{pmatrix},$$

and two groups with initial distributions

$$\boldsymbol{\pi}(\mathrm{Group1}) = (1, 0, 0),$$

$$\boldsymbol{\pi}(\text{Group2}) = (0, 0, 1),$$

respectively. Then the first group can only access the first two states, and thus its tail is of the order $\exp(-0.634y)$ (since -0.634 is the largest eigenvalue of the sub-matrix $(t_{kl})_{k,l=1,2}$), while the second group can only access the third state, and thus has a tail of order $\exp(-3y)$.

Similarly, if a Pareto inhomogeneity function is used, the tail index will vary between the two groups. Thus, after estimation, it is important to check which states are accessible by which sub-populations, in order to deduce their precise conditional tail asymptotics.

5.2. Semi-composite models as an alternative to splicing

When heavy tails are present, a standard approach to obtain a global model for claim severities is to model the body and tail separately, and then combine them through splicing or mixing. For describing the tail of the distribution, extreme value tools are typically employed. Although this two-step procedure is not fully satisfactory, the outcome can be more reliable when the parameter of interest is the tail coefficient, see, for instance, Embrechts et al. (2013).

On the other hand, the models presented in the previous section are attractive alternatives to obtain global models, due to their authentic heavy tails and denseness. Moreover, the fitting of these models does not require any form of threshold selection, as in traditional extreme value techniques. However, their estimation methods give the same weight to all data points, and hence the automatic modeling of the tails may not be as satisfactory as when targeting the tail via thresholding. Furthermore, in some situations, even if the tail is correctly specified via fitting a PH-MoE model using the EM algorithm, a risk manager might be interested in at least partially separating the analysis above and below a certain threshold.

Below we see how certain piecewise specifications for the inhomogeneity function λ can achieve a compromise between the two above approaches, while still formally falling into the class of standard PH-MoE models. Specifically, we consider inhomogeneity transformations which are defined differently below and above a certain threshold (and the idea can be extended to several layers).

Definition 5.3: We say that a PH-MoE model is semi-composite if its intensity function is of the form

$$\lambda(t) = \begin{cases} \lambda_1(t), & t \le y_0, \\ \lambda_2(t), & t > y_0, \end{cases}$$

for any two intensities λ_1, λ_2.

An immediate consequence is the following:

Proposition 5.4: *For a semi-composite PH-MoE model we have that*

$$g^{-1}(y) = \begin{cases} g_1^{-1}(y), & y \le y_0, \\ g_2^{-1}(y) + g_1^{-1}(y_0) - g_2^{-1}(y_0), & y > y_0, \end{cases}$$

and so in particular

$$\bar{F}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \begin{cases} \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T} g_1^{-1}(y))\boldsymbol{e}, & y \le y_0, \\ \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp((g_2^{-1}(y) + g_1^{-1}(y_0) - g_2^{-1}(y_0))\boldsymbol{T})\boldsymbol{e}, & y > y_0. \end{cases}$$

Hence, $Y \,|\, \boldsymbol{X}$ is tail-equivalent to a PH-MoE model with intensity $\lambda_2(t)$ for all $t \ge 0$.

Below we outline the details of two cases which give rise to tails which are commonly used for loss modeling.

Example 5.5 (PH body with Weibull tail): Specify

$$\lambda(t) = \begin{cases} 1, & t \leq y_0, \\ \theta(t - y_0)^{\theta - 1}, & t > y_0. \end{cases}$$

In this way

$$g^{-1}(y) = \begin{cases} y, & y \leq y_0, \\ y_0 + (y - y_0)^{\theta}, & y > y_0, \end{cases}$$

and

$$g(y) = \begin{cases} y, & y \leq y_0, \\ y_0 + (y - y_0)^{1/\theta}, & y > y_0. \end{cases}$$

Hence

$$\bar{F}_{Y|\boldsymbol{X}}(y \,|\, \boldsymbol{x}) = \begin{cases} \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T} y)\boldsymbol{e}, & y \leq y_0, \\ \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp((y_0 + (y - y_0)^{\theta})\boldsymbol{T})\boldsymbol{e}, & y > y_0. \end{cases}$$

Example 5.6 (PH body with Pareto tail): We may instead specify

$$\lambda(t) = \begin{cases} 1, & t \leq y_0, \\ (t - y_0 + \theta)^{-1}, & t > y_0. \end{cases}$$

Which now yields

$$g^{-1}(y) = \begin{cases} y, & y \leq y_0, \\ y_0 + \log((y - y_0)/\theta + 1), & y > y_0, \end{cases}$$

and

$$g(y) = \begin{cases} y, & y \leq y_0, \\ y_0 + \theta(\exp(y - y_0) - 1), & y > y_0. \end{cases}$$

Hence

$$\bar{F}_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \begin{cases} \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(\boldsymbol{T} y)\boldsymbol{e}, & y \leq y_0, \\ \boldsymbol{\pi}^{\mathsf{T}}(\boldsymbol{x}) \exp(y_0 \boldsymbol{T}) \left(\dfrac{y - y_0}{\theta} + 1\right)^{\boldsymbol{T}} \boldsymbol{e}, & y > y_0. \end{cases}$$

Note that for fix $\boldsymbol{x}$, the proposed models are dense in the class of distributions in the positive real line. This follows from the fact that they belong to the IPH class, which possesses the said property for any λ satisfying (1).

Given that the intensity function λ is a parametric function depending on the parameters θ and y_0, we can employ Algorithm 1 for the estimation of the above semi-composite specifications. However, the changepoint may alternatively be specified in advance and then fixed through the fitting procedure. The latter approach is preferable in almost all cases, and in particular when working with regularly-varying heavy tails, where the threshold may be determined by well-founded visual tools, such as the Hill estimator, cf. Hill (1975).

Table 1. Regression coefficients $\hat{\boldsymbol{\alpha}}$ for the categorical variable of the simulated data. Group A is the baseline level.

State	(Intercept)	Group B	Group C	Group D
2	−9.986 (6.673)	17.488 (24.118)	8.732 (6.675)	13.983 (30.537)
3	−4.32 (0.395)	−3.113 (0.044)	4.308 (0.41)	17.281 (29.532)
4	−12.642 (25.17)	25.136 (34.211)	12.002 (25.17)	8.02 (300.3)
5	−4.488 (0.429)	6 (25.6)	3.079 (0.464)	13.74 (29.534)

In parenthesis, the standard errors are displayed. The coefficient associated with state 1 can be deduced from the constraint $\sum_{k=1}^{5} \pi_k(\boldsymbol{X}) = 1$.

6. Numerical examples

This section illustrates the statistical feasibility of the methods developed above. We do not aim to be comprehensive in our treatment, but instead point out the general direction which seems promising. The drawback of the algorithm at the moment is speed, in particular of the R-step, which is a well-known issue of multinomial regression models when in the presence of several covariates. Hence, we provide one simulated example with a 4-dimensional categorical covariate, and a much larger real insurance example with two categorical covariates.

6.1. *Synthetic data*

We consider a simulated example where the data genuinely comes from a classical mixture-of-experts model. More precisely, the dataset consists of a total of 2000 observations divided into 4 groups of size 500, each having distributions as follows:

$$\text{Group A} : Y_i \sim \Gamma(\text{shape} = 1, \text{scale} = 3), \quad \text{Group B} : Y_i \sim \Gamma(\text{shape} = 3, \text{scale} = 9),$$
$$\text{Group C} : Y_i \sim \Gamma(\text{shape} = 1, \text{scale} = 9), \quad \text{Group D} : Y_i \sim \Gamma(\text{shape} = 3, \text{scale} = 3).$$

We consider a 5-dimensional PH structure, which was chosen small enough so that if there were no interaction between states, it would not be possible to model the four groups. Indeed, mixtures of five exponential components would not correctly capture the four given distributions.

Subsequently, we employed a homogeneous version of Algorithm 1, where Steps 2 and 6 are suppressed. In other words, we assume that $\lambda(t) = 1, \forall t \geq 0$, and in particular, the tails of both the data and the model are exponentially decaying. The results are as follows[1] :

$$\hat{\boldsymbol{T}} = \begin{pmatrix} -0.349 & 0 & 0 & 0 & 0 \\ 0.303 & -0.303 & 0 & 0 & 0 \\ 0 & 0.162 & -0.553 & 0 & 0.391 \\ 0 & 0 & 0.059 & -0.06 & 0.001 \\ 0 & 0.618 & 0.607 & 0 & -1.225 \end{pmatrix},$$

and the output of the R-step is given in Table 1.

In particular, the coefficients translate into the following initial Markov probabilities for each group:

$$\boldsymbol{\pi}(\text{Group A}) = (0.976,\ 0.000,\ 0.013,\ 0.000,\ 0.011),$$
$$\boldsymbol{\pi}(\text{Group B}) = (0.000,\ 0.007,\ 0.000,\ 0.993,\ 0.000),$$
$$\boldsymbol{\pi}(\text{Group C}) = (0.328,\ 0.094,\ 0.324,\ 0.173,\ 0.080),$$

[1] Here and in the rest of the numerical section, estimates are rounded to three decimal places. This means that some displayed null values may actually be very small but non-zero.

Table 2. Theoretical, observed, and fitted means.

Group	Theoretical	Empirical (= GLM)	PH-MoE
A	3	3.005	3.021
B	27	27.212	26.347
C	9	9.463	10.001
D	9	9.499	9.807

$$\boldsymbol{\pi}\,(\text{Group D}) = (0.000,\ 0.000,\ 0.976,\ 0.000,\ 0.024).$$

When considering the means for each group, given by formula (4), we obtain, upon comparing with a Gamma Generalized Linear Model (GLM), the following Table 2.

Since the PH-MoE does not match the empirical means for each group, as is the case for the GLM, we can observe slight discrepancies between the fitted and observed averages per group. However, if we estimate the dispersion coefficient of the GLM with the average deviance, we may compare not only the mean but the entire distribution of both models. Figure 1 shows the densities for each group for the theoretical and fitted cases, and for the GLM and PH-MoE models. We observe that the risks are better understood if we use the latter model, and consequently, any other measure of performance which is not solely based on the mean will favor the matrix-based method.

6.2. *Insurance data*

We consider the French Motor Third Party Liability (freMTPL) insurance data, contained in the datasets `freMTPLfreq` and `freMTPLsev` in the `CASdatasets` package in `R`. The data consists of risk features corresponding to 413,169 motor insurance policies and the number of claims and their severity. Presently we aim at analyzing only a portion of the total 15,390 claim sizes,[2] mainly for computational power reasons: the multinomial step of the PH-MoE routine can be slow to (or not) converge for large p (say, above 10), n (in the tens of thousands) and d (more than 20 covariates).

To find a sensible and interesting subset of the data, we first observe the top left panel of Figure 2, where we see that the data has a highly pronounced peak in the log scale. Such peak can only be captured by a PH distribution of a huge order ($p > 200$), which makes it unfeasible to fit even without covariates. Thus, we consider only the excesses above the threshold $M = 0.15$, which in insurance terms would correspond to data entering an XL reinsurance contract with retention level M. The excesses are plotted in the top right panel of Figure 2 in the log scale, which are much easier to estimate with a lower dimension. On the bottom panel of Figure 2, we observe the heavy-tailed nature of the excesses and that there is a substantial bias away from strict Pareto behavior (the estimator curves for smaller order statistics).

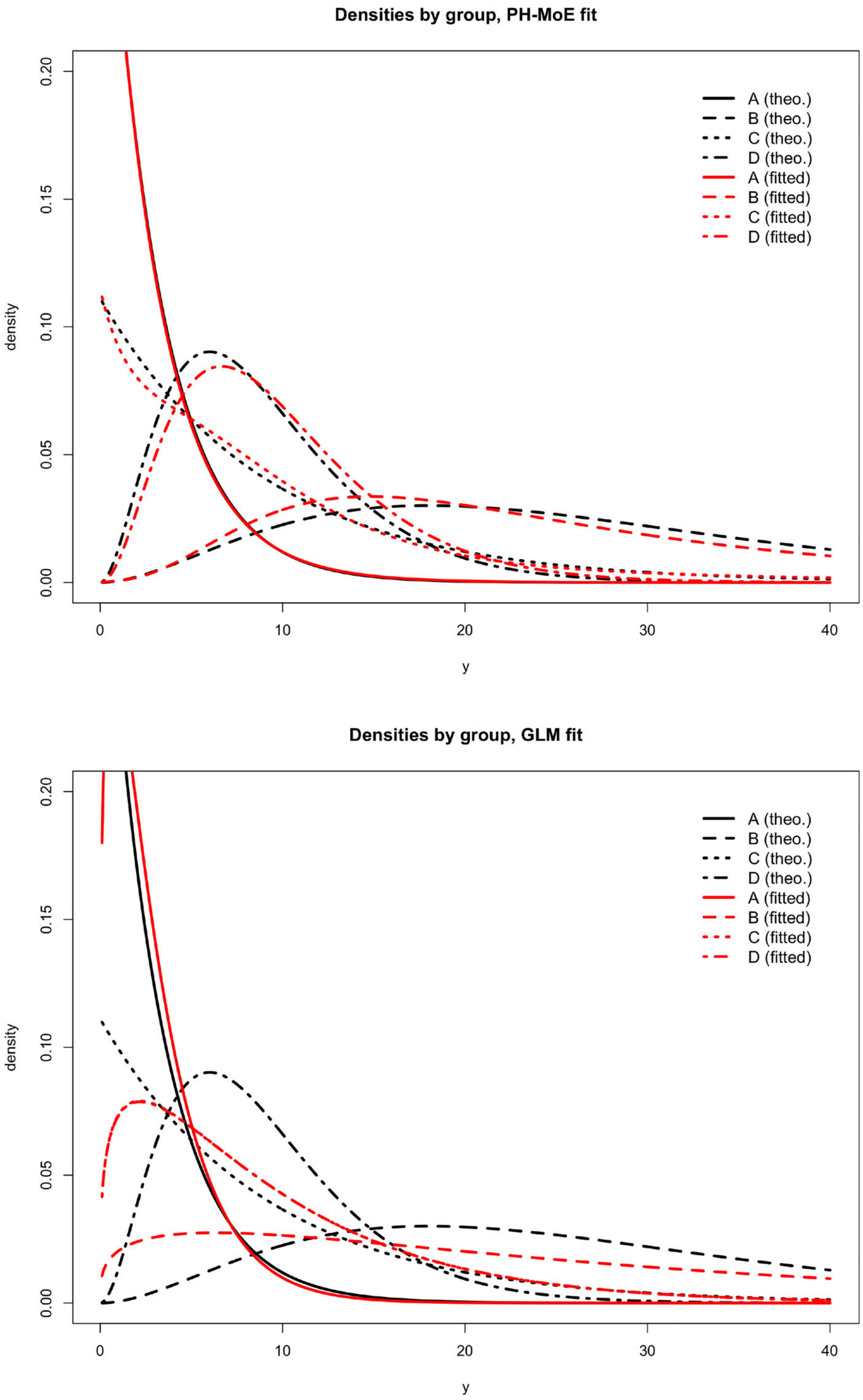
With respect to covariates, we fitted a log-normal regression model and selected the only two covariates which seem to be relevant to predicting the mean[3] :

(1) `Power`: The power of the car, an ordered categorical variable with values: $d, e, f, g, h, i, j, k, l, m, n, o$.

(2) `Region`: The policy region in France, based on the 1970–2015 classification. Possible values associated with the excesses are: Aquitaine, Basse-Normandie, Bretagne, Centre, Haute-Normandie, Ile-de-France, Limousin, Nord-Pas-de-Calais Pays-de-la-Loire, Poitou-Charentes.

[2] Here, we have divided claim severities by the corresponding claim numbers for each policy. For numerical reasons, we also divided the result by 10^4.

[3] In general, insurance covariates provide very small predictive power for severity, in contrast to claim counts, where the performance is usually much better.


Densities by group, PH-MoE fit

Densities by group, GLM fit

Figure 1. Fitted densities for each group of the simulated data, for the PH-MoE (top panel) and GLM (bottom panel) models.

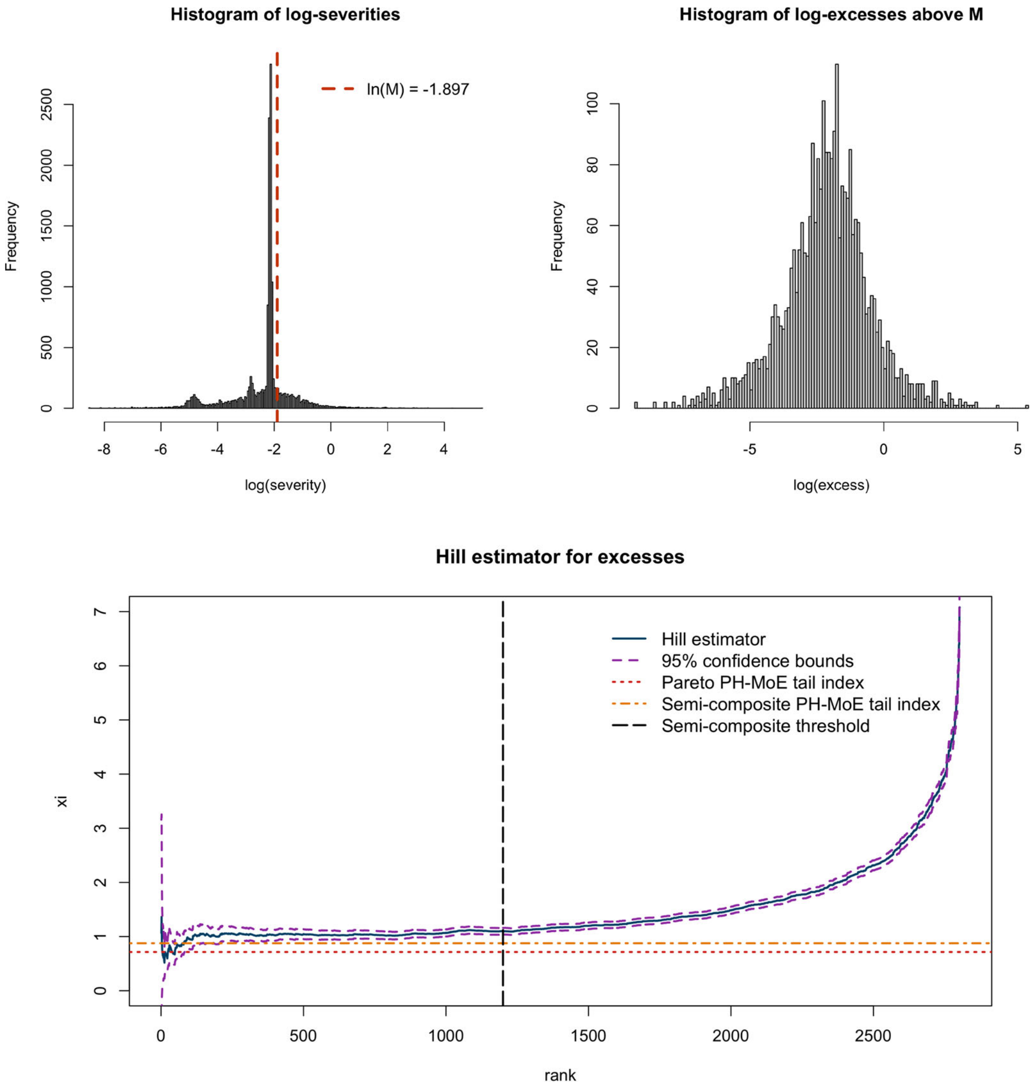


Figure 2. French MTPL full data with selected excess threshold (top left panel), the resulting excesses (top right), and their implied tail index according to the Hill estimator (bottom). For the latter plot, we also overlay the implied tail indices from two different PH-MoE fits.

To illustrate and compare the modeling capabilities of our model, we proceed to estimate different PH-MoE and LRMoE models, the latter being a natural candidate for comparison. More specifically, and to keep the number of parameters for both models similar, we considered three Pareto PH-MoE models of dimensions 3, 4, and 5, and three LRMoE models with 4, 5, and 6 experts. For the PH-MoE models, we employed 1000 EM steps with random initialization of the parameters. On the other hand, the estimation of the LRMoE models was done using the LRMoE `R` package (cf. Tseung et al. 2020) and as per the Vignettes found in https://github.com/UofTActuarial/LRMoE/tree/master/vignettes with 200 CEM steps and expert components automatically chosen with the `cmm_init` function. A summary of the results can be found in Tables 3 and 4.

Although perhaps mathematically more complex, our investigations show that the numerical routines for PH-MoE models are at least on par with those of LRMoE in terms of likelihood performance, and certainly much faster. Note also that the number of EM steps cannot be compared with those of

Table 3. Summary for PH-MoE model for the freMTPL dataset.

	PH-MoE		
Dimension	3	4	5
Log Likelihood	718.38	743.52	759.71
Number of parameters	52	80	110
Computational times	3.14 min	6.24 min	11.88 min

Table 4. Summary for LRMoE model for the freMTPL dataset.

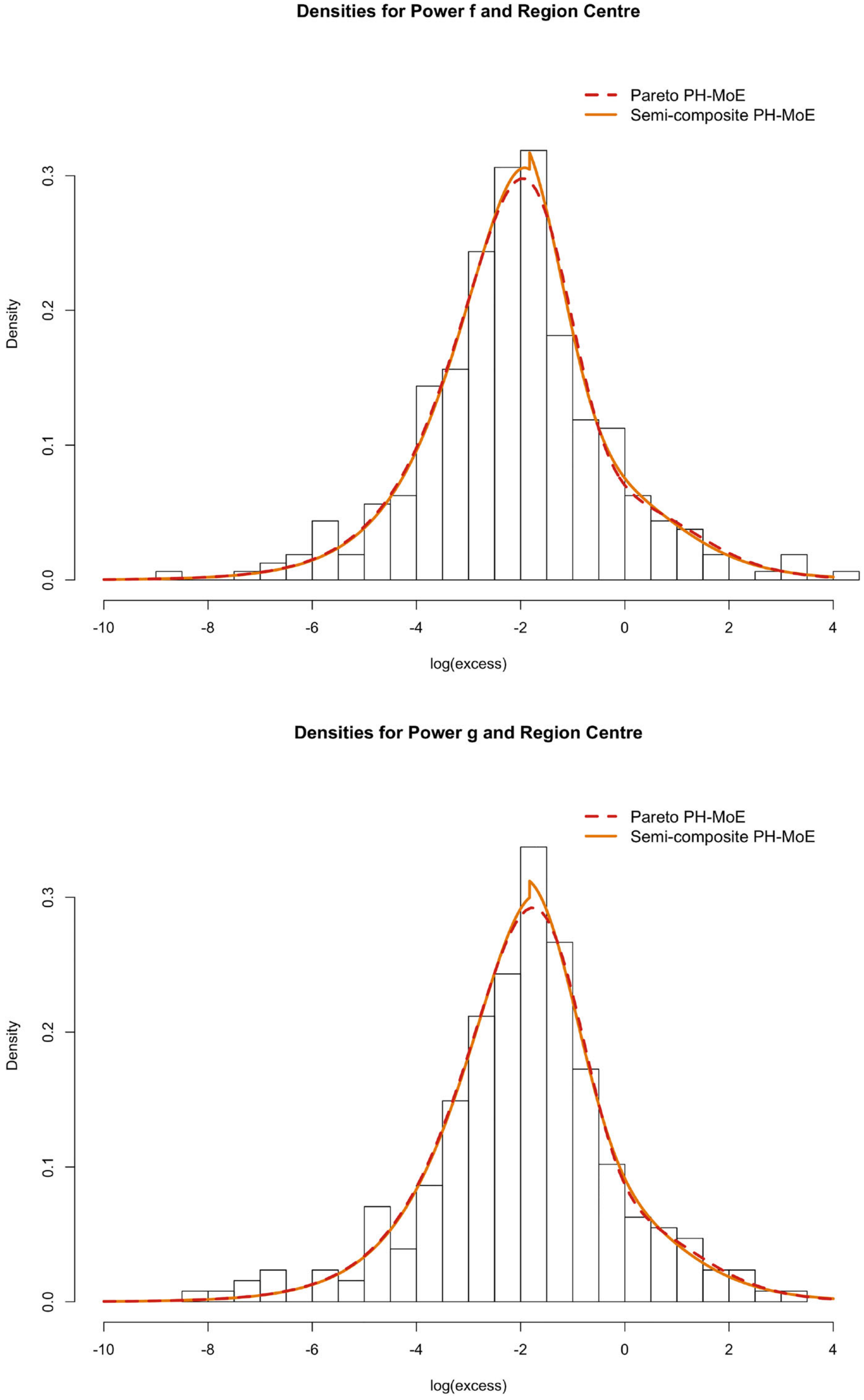
	LRMoE		
Number experts	4	5	6
Log likelihood	728.64	746.64	750.29
Number of parameters	71	94	117
Computational times	49.75 min	1.21 h	1.65 h

the CEM algorithm since the latter is much slower but converges in fewer iterations. This is in line with the thinking of the additional parameters of a PH-MoE model as weak learners rather than actual statistical parameters. A full systematic comparison between PH-MoE, LRMoE, TG-LRMoE, and related MoE models in terms of in-sample and out-of-sample performance is out of the scope of this work. However, we can mention that an advantage of having better computational times, is that we can try different initializations for the EM algorithm. This is highly relevant for the estimation of both models since there is always the possibility of obtaining a local maxima depending on the initial values.

It can be appreciated that the number of parameters involved in both models is relatively large. However, it is crucial to understand that these models can be considered as interpretable machine learning methods rather than concise statistical models. In other words, each additional degree of freedom does not always target a particular distributional feature but instead serves as a *weak learner*, working towards an overall good estimation. Hence, classical information criteria such as AIC and BIC will tend to overpenalize these models and should not be used for model selection. Ideally, an information criterion specifically designed for PH-MoE models (and even for PH variables) would allow for goodness of fit considerations without using the misspecified AIC and BIC criteria, similar to the development of AICC in the context of time series analysis. Regularization through cross-validation is also a natural topic of further study, but out of the scope of the current manuscript.

With the different PH-MoE models above at hand, we now select one to describe our data and present the results. We start by giving some words about the dimension selection of the PH-MoE model, for which we will follow the typical approach of dimension selection for PH distributions. More specifically, given that adding more dimensions always improves the quality of the fit, one typically starts with low dimensions and assesses the benefit/cost of adding extra dimensions. This assessment is usually done using visual aids and/or by looking at changes in the loglikelihood. One aims for a dimension that is a good compromise between the quality of the fit and a reasonable number of parameters. The reason for this approach is the identifiability issues of PH distribution, meaning that the number of free parameters is unknown. However, it is worth mentioning that recent steps towards more statistical-based selection approaches have been recently introduced in the literature. For instance, we can mention the work in Albrecher et al. (2021b).

In our particular case of study, dimension 5 seems to be a good compromise. To support our choice, we also fitted a PH-MoE model of dimension 6, obtaining an increase in the loglikelihood of 8.15, and additional 32 parameters, which was a much smaller likelihood increase than the previous steps. Hence, we decided to stick to a PH-MoE of dimension 5. For completeness, we also fitted a semi-composite PH-MoE with Pareto tail and same dimension 5. In this case, we select the threshold value


Densities for Power f and Region Centre

Densities for Power g and Region Centre

Figure 3. Conditional densities for two select covariates.

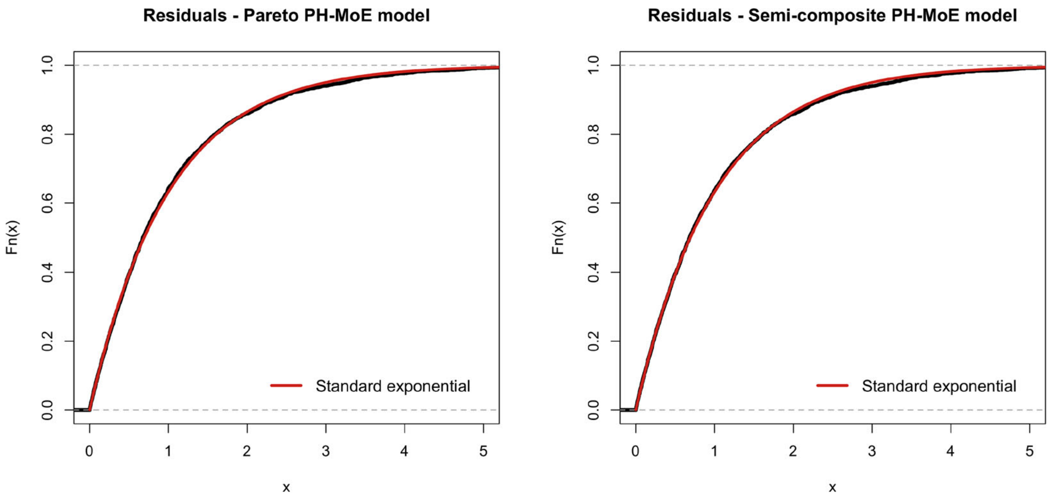


Figure 4. CDF for the residuals of the fitted PH-MoE model against CDF of a standard exponential (left panel), and corresponding plot for the semi-composite PH-MoE model (right panel).

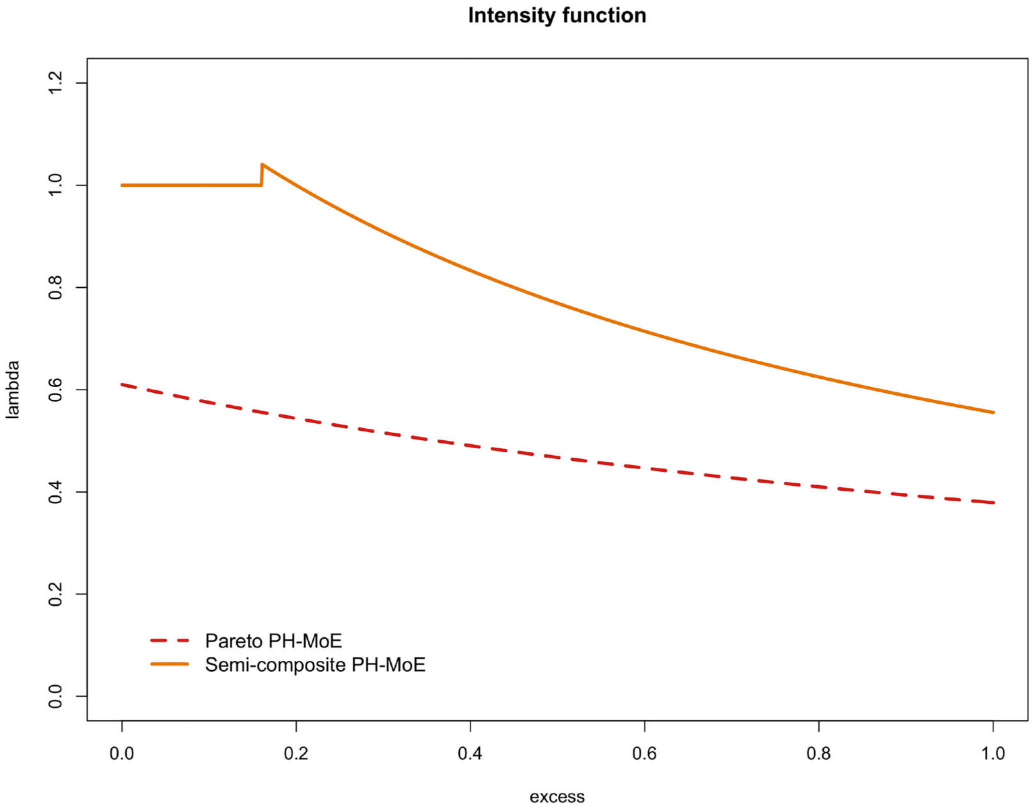


Figure 5. Fitted intensity functions for the PH-MoE models.

y_0 to be at the order statistic number 1200, which is when the Hill plot visually starts to flatten, obtaining a loglikelihood of 759.3. It is worth mentioning that we are selecting the threshold but not the implied tail index. The latter is estimated during the EM algorithm jointly with all other parameters.

The full fitted coefficients $\hat{\boldsymbol{\alpha}}$ and their statistical significance are given in Appendix 1, with the asymptotic properties of the MLE delegated to Appendix 2. In Figure 3, we observe that the densities shift their shapes when varying the covariates, as expected. The semi-composite model has a discontinuity at y_0, which is slightly visible, a feature that is also common in fully composite models. The associated PH-MoE probabilities for these two densities are as follows:

$$\begin{aligned}
\text{Pareto PH} - \text{MoE :}&\boldsymbol{\pi}\,(\text{Power}\,f,\ \text{Region Centre}) = (0.000, 0.000, 0.000, 0.138, 0.862),\\
&\boldsymbol{\pi}\,(\text{Power}\,g,\ \text{Region Centre}) = (0.074, 0.014, 0.695, 0.132, 0.085),\\
&\hat{\boldsymbol{T}} = \begin{pmatrix} -22.119 & 0.000 & 0.005 & 3.580 & 0.041\\ 0.011 & -9.233 & 6.689 & 0 & 2.511\\ 0 & 0.292 & -9.931 & 0 & 0.518\\ 0.022 & 1.285 & 0.064 & -1.402 & 0.026\\ 0.005 & 0.404 & 0.986 & 0 & -13.203 \end{pmatrix},\\
&\hat{\theta} = 1.639.\\
\text{S} - \text{C PH} - \text{MoE :}&\boldsymbol{\pi}\,(\text{Power}\,f,\ \text{Region Centre}) = (0.000, 0.000, 0.000, 0.153, 0.847),\\
&\boldsymbol{\pi}\,(\text{Power}\,g,\ \text{Region Centre}) = (0.104, 0.023, 0.655, 0.138, 0.079).\\
&\hat{\boldsymbol{T}} = \begin{pmatrix} -14.444 & 0.000 & 0.008 & 2.639 & 0.053\\ 0.007 & -5.734 & 4.569 & 0 & 1.146\\ 0 & 0.086 & -5.785 & 0 & 0.479\\ 0.006 & 1.103 & 0.024 & -1.142 & 0.008\\ 0.003 & 0.148 & 1.107 & 0 & -8.351 \end{pmatrix},\\
&\hat{\theta} = 0.961.
\end{aligned}$$

Note that all the five states in both cases communicate, then the resulting tail indices for the PH-MoE models are given by

$$\hat{\xi} = -1/\max\{\Re\,\text{Eigen}(\hat{\boldsymbol{T}})\} = 0.72,\ 0.88, \quad \text{respectively},$$

which, according to the lower panel of Figure 2, are both very reasonable estimates. Further evidence of the quality of the estimation is given in Figure 4, where we observe that the empirical distribution function of the residuals of PH-MoE models (computed as described in Section 4.3) align closely with the distribution function of a standard exponential. This is further supported by applying Kolmogorov-Smirnov tests, for which we obtain a p-value of 0.8717 for the PH-MoE and 0.8401 for the semi-composite PH-MoE. Figure 5 shows how the intensity functions λ behave, which may be considered as an infinitesimal 'environment' time change of the underlying phase-type distribution. Another possible extension of our model is to make these transformations dependent on $\boldsymbol{X}$.

To extend the analysis to the full data and not only excesses, a direction which is promising is to consider special sub-structures of phase-type distributions. In those cases, the EM algorithm becomes simpler and potentially much faster, consequently enabling the analysis of larger models with more phases being fitted to larger data.

7. Conclusion

We have presented a claim severities regression model based on PH distributions, incorporating covariates through the initial probability vector, which can be cast into a mixture-of-experts framework. When combined with an inhomogeneity transform, these regression models span distributions with different tail behaviors and possible multimodality. Furthermore, they are flexible and may converge to fairly general regression model specifications. We have derived an effective estimation procedure based on the EM algorithm and a weighted multinomial regression problem and shown its feasibility on synthetic and real insurance data.

Several questions remain open for further research, such as automatic feature selection procedures or using other machine learning methods to predict the initial probability vector. In addition, the analysis of more than one risk, together with their respective claim frequencies, all together in a global model with the same underlying Markov structure is an interesting research direction.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

MB would like to acknowledge financial support from the Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (Swiss National Science Foundation Project) [grant number 200021_191984] . JY would like to acknowledge financial support from the Swiss National Science Foundation Project [grant number IZHRZ0_180549].

References

Albrecher H. & Bladt M. (2019). Inhomogeneous phase-type distributions and heavy tails. Journal of Applied Probability 56(4), 1044–1064.

Albrecher H., Bladt M., Bladt M. & Yslas J. (2021a). Mortality modeling and regression with matrix distributions. arXiv:2011.03219

Albrecher H., Bladt M. & Muller L. J. (2021b). Penalised likelihood methods for phase-type dimension selection. Preprint.

Albrecher H., Bladt M. & Yslas J. (2022). Fitting inhomogeneous phase-type distributions to data: the univariate and the multivariate case. Scandinavian Journal of Statistics 49(1), 44–77.

Asmussen S. (2003). Applied probability and queues. New York: Springer.

Asmussen S., Nerman O. & Olsson M. (1996). Fitting phase-type distributions via the EM algorithm. Scandinavian Journal of Statistics 23(4), 419–441.

Bladt M. (2021). Phase-type distributions for claim severity regression modeling. ASTIN Bulletin **52**(2), 1–32.

Bladt M. & Nielsen B. F. (2017). Matrix-exponential distributions in applied probability. Springer.

Cox D. R. (1975). Partial likelihood. Biometrika 62(2), 269–276.

Embrechts P., Klüppelberg C. & Mikosch T. (2013). Modelling extremal events: for insurance and finance. New York: Springer.

Fung T. C., Badescu A. L. & Lin X. S. (2019). A class of mixture of experts models for general insurance: theoretical developments. Insurance: Mathematics and Economics 89, 111–127.

Fung T. C., Badescu A. L. & Lin X. S. (2021a). A new class of severity regression models with an application to IBNR prediction. North American Actuarial Journal 25(2), 206–231.

Fung T. C., Tzougas G. & Wuthrich M. (2021b). Mixture composite regression models with multi-type feature selection. arXiv:2103.07200

Grün B. & Miljkovic T. (2019). Extending composite loss models using a general framework of advanced computational tools. Scandinavian Actuarial Journal 2019(8), 642–660.

Hill B. M. (1975). A simple general approach to inference about the tail of a distribution. The Annals of Statistics 3(5), 1163–1174.

Lee S. C. & Lin X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. North American Actuarial Journal 14(1), 107–130.

Lehmann E. L. & Casella G. (2006). Theory of point estimation. Springer Science & Business Media.

Miljkovic T. & Grün B. (2016). Modeling loss data using mixtures of distributions. Insurance: Mathematics and Economics 70, 387–396.

Neuts M. F. (1975). Probability distributions of phase type. In *Liber Amicorum Professor Emeritus H. Florin*. Department of Mathematics, University of Louvian, Belgium. P. 173–206.

Neuts M. F. (1981). Matrix-geometric solutions in stochastic models: an algorithmic approach. Baltimore: The Johns Hopkins University Press.

Reynkens T., Verbelen R., Beirlant J. & Antonio K. (2017). Modelling censored losses using splicing: a global fit strategy with mixed Erlang and extreme value distributions. Insurance: Mathematics and Economics 77, 65–77.

Tseung S. C., Badescu A., Fung T. C. & Lin X. S. (2020). LRMoE: an R package for flexible actuarial loss modelling using mixture of experts regression model. SSRN 3740215.

Tzougas G., Vrontos S. & Frangos N. (2014). Optimal bonus-malus systems using finite mixture models. ASTIN Bulletin 44(2), 417–444.

Wong W. H. (1986). Theory of partial likelihood. The Annals of Statistics 14(1), 88–123.

Yuksel S. E., Wilson J. N. & Gader P. D. (2012). Twenty years of mixture of experts. IEEE Transactions on Neural Networks and Learning Systems 23(8), 1177–1193.

Appendices

Appendix 1. Estimated coefficients for the PH-MoE models

Appendix 2. Inference for phase-type regression models

Inference and goodness of fit can always be done via parametric bootstrap methods. However, re-fitting a PH regression can be too costly. A first approach is the following general-purpose result:

Theorem A.1: *Let* λ, $\boldsymbol{\eta} := (\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{T})$ *be such that the log-density*

$$y \mapsto \log\left[\boldsymbol{\pi}(\boldsymbol{\alpha}) \exp\left(\int_0^y \lambda(s;\boldsymbol{\theta})\,\mathrm{d}s\boldsymbol{T}\right) \boldsymbol{t}\lambda(y;\boldsymbol{\theta})\right], \quad y > 0,$$

satisfies Assumptions (A0)–(A3) of Section 6.3 of Lehmann & Casella (2006) (common supports, identifiable parameters, i.i.d. observations, and true parameters in the interior of the parameter space) and Assumptions (A)–(D) of Section 6.5 of Lehmann & Casella (2006) (existence and finite expectation of third derivatives of log-density, strict positive-definiteness of information matrix, and the representation of the latter in terms of expected double partial derivatives of the log-density).

Then, as the sample size $n \to \infty$, *we have that*

(1) *There exist consistent solutions* $\hat{\boldsymbol{\eta}}_n$ *to the likelihood equations.*

(2) *The following convergence holds:*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}),$$

where $\mathcal{I}$ *is the information matrix.*

(3) *The jth parameter is asymptotically efficient:*

$$\sqrt{n}(\hat{\eta}_{jn} - \eta_j) \xrightarrow{d} \mathcal{N}(0, [\mathcal{I}^{-1}]_{jj}).$$

Proof: The proof translates directly from Theorem 5.1 in Section 6.5 of Lehmann & Casella (2006). ∎

Theorem A.1 is somewhat academic in nature for general PH distributions, because although most of the properties are easy to verify for most parameters and transforms, others are difficult, such as the moment conditions, or near-impossible, such as the identifiability and strict positive-definiteness of the information matrix. Indeed, the latter conditions require that all eigenvalues of $\boldsymbol{T}$ be distinct (although this is not sufficient) and that all parameters be away from the border regions, which is an uncommon scenario when fitting real data.

However, there is still something to be said when it comes to the regression coefficients. Once the EM algorithm has converged, we may consider the parameters $(\boldsymbol{\theta}, \boldsymbol{T})$ as nuisance parameters and thus perform inference on the partial likelihood $\ell(\boldsymbol{\alpha} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{T})$, cf. Cox (1975), see also Wong (1986). As the latter reference suggests, the sub-optimal use of information incurs a loss of efficiency (standard errors should be considered only as lower bounds) which should be weighted against the possible gains in robustness and simplicity of analysis. The usual experiments where the partial likelihood is useful is when the nuisance parameters take values in high-dimensional spaces, making the calculation of the matrix $\mathcal{I}$ difficult and not robust. Since the number of nuisance parameters for a PH-MoE model is at least $p^2 + 1$, it is a clear candidate for benefitting from these tradeoffs.

Another advantage of performing inference on the regression variables alone is that we circumvent making allusion to the possibly non-identifiable parameters of the sub-intensity matrix $\boldsymbol{T}$. In practice, this means that we perform inference during the multinomial regression step (R-step) of Algorithm 1, and use the output to draw conclusions on the statistical significance of the covariates $\boldsymbol{X}$, as well as to perform variable selection.

Table A1. The coefficients associated with state 1 can be deduced from the constraint $\sum_{k=1}^{5} \pi_k(\boldsymbol{X}) = 1$.

State: variable level	Pareto PH-MoE	Semi-composite PH-MoE
2:(Intercept)	0.06(0.51)	0.08(0.52)
2:Powere	−0.58(0.51)	−0.59(0.51)
2:Powerf	−33.49	−25.88
2:Powerg	−1.00(0.52)	−1.03(0.52)*
2:Powerh	0.59(0.46)	0.57(0.46)
2:Poweri	−0.24(0.65)	−0.37(0.65)
2:Powerj	40.39(4.09)***	45.45(7.05)***
2:Powerk	0.27(0.51)	0.23(0.52)
2:Powerl	0.95(0.58)	0.85(0.59)
2:Powerm	−17.44(0.00)***	−17.87
2:Powern	−47.86	−52.91(0.00)***
2:Powero	2.01(1.41)	0.93(0.90)
2:RegionBasse-Normandie	−11.76(152.10)	−16.73(0.00)***
2:RegionBretagne	0.21(0.25)	0.33(0.25)
2:RegionCentre	−0.76(0.29)**	−0.55(0.28)
2:RegionHaute-Normandie	−46.08	−49.11
2:RegionIle-de-France	−1.46(0.39)***	−1.47(0.39)***
2:RegionLimousin	9.97(243.82)	10.79(0.00)***
2:RegionNord-Pas-de-Calais	13.50(120.65)	24.08(10.18)*
2:RegionPays-de-la-Loire	−0.22(0.29)	−0.14(0.29)
2:RegionPoitou-Charentes	−0.33(0.34)	−0.22(0.34)
3:(Intercept)	−94.94(40.27)*	−96.95(1.45)***
3:Powere	38.62(226.78)	62.64(1.02)***
3:Powerf	7.73(4.73)	27.96(2.15)***
3:Powerg	57.42(74.76)	69.91(0.55)***
3:Powerh	−2.44(11.39)	−15.44(0.00)***
3:Poweri	78.55(58.65)	75.11(0.81)***
3:Powerj	48.72(5.22)***	73.89(4.96)***
3:Powerk	−71.15(0.01)***	−98.00(0.00)***
3:Powerl	−102.41(0.98)***	−146.20(0.00)***
3:Powerm	156.56(0.02)***	169.79(0.15)***
3:Powern	−7.52(0.00)***	−7.06(0.00)***
3:Powero	9.92(5.00)*	26.76(2.24)***
3:RegionBasse-Normandie	11.43(0.01)***	−12.22(0.01)***
3:RegionBretagne	−9.48(0.00)***	−23.21(0.14)***
3:RegionCentre	39.76(38.48)	28.88(1.79)***
3:RegionHaute-Normandie	191.25(18.38)***	220.06(0.00)***
3:RegionIle-de-France	88.00(40.35)*	69.88(0.82)***
3:RegionLimousin	48.99(0.00)***	47.03(0.15)***
3:RegionNord-Pas-de-Calais	67.78(40.04)	61.61(6.40)***
3:RegionPays-de-la-Loire	17.61(21.66)	23.22(1.86)***
3:RegionPoitou-Charentes	37.12(38.48)	26.96(1.81)***
4:(Intercept)	0.35(9.79)	9.79(36.37)
4:Powere	−77.73(18.24)***	−101.22(0.00)***
4:Powerf	−33.56(7.75)***	−34.54(27.48)
4:Powerg	−60.72(41.44)	−74.51(13.04)***
4:Powerh	−40.08(7.75)***	−69.93(13.04)***
4:Poweri	−39.41(7.75)***	−69.09(13.04)***
4:Powerj	0.25(4.04)	−24.40(6.00)***
4:Powerk	−106.39(0.00)***	−115.65(0.00)***
4:Powerl	−107.74(20.20)***	−129.44(227.37)
4:Powerm	−7.35(0.02)***	−34.53(0.15)***
4:Powern	−79.21(55.22)	−87.78(2.60)***
4:Powero	−117.79	−114.34(0.00)***
4:RegionBasse-Normandie	99.85(23.37)***	107.02(83.71)
4:RegionBretagne	37.09(6.50)***	57.67(25.35)*
4:RegionCentre	60.95(35.89)	65.00(25.35)*
4:RegionHaute-Normandie	115.60(45.65)*	132.22(141.61)
4:RegionIle-de-France	−25.40(0.00)***	−38.18(0.00)***
4:RegionLimousin	95.60(3.29)***	100.51(16.77)***

(*continued*).

Table A1. Continued.

State: variable level	Pareto PH-MoE	Semi-composite PH-MoE
4:RegionNord-Pas-de-Calais	87.37(26.09)***	96.12(20.10)***
4:RegionPays-de-la-Loire	38.32(6.51)***	58.70(25.35)*
4:RegionPoitou-Charentes	−0.62(0.00)***	−1.92(0.00)***
5:(Intercept)	35.11(7.75)***	36.13(27.47)
5:Powere	−62.31(41.44)	−75.62(13.04)***
5:Powerf	−33.55(7.75)***	−34.53(27.48)
5:Powerg	−62.99(41.44)	−76.76(13.04)***
5:Powerh	−39.53(7.74)***	−69.38(13.04)***
5:Poweri	−40.73(7.75)***	−70.45(13.04)***
5:Powerj	0.34(4.03)	−24.30(6.00)***
5:Powerk	−85.51(149.07)	−102.62(0.28)***
5:Powerl	−142.25(0.00)***	−162.24(0.08)***
5:Powerm	−86.81(0.00)***	−89.54(0.00)***
5:Powern	−120.92(0.00)***	−137.69(0.00)***
5:Powero	−121.06(0.00)***	−120.60(0.00)***
5:RegionBasse-Normandie	68.24(19.46)***	83.73(83.75)
5:RegionBretagne	4.72(1.65)**	33.51(18.47)
5:RegionCentre	28.03(38.04)	40.35(18.47)*
5:RegionHaute-Normandie	95.90(22.90)***	119.57(137.82)
5:RegionIle-de-France	−36.69(7.75)***	−37.44(27.48)
5:RegionLimousin	64.53(3.29)***	77.67(16.79)***
5:RegionNord-Pas-de-Calais	54.28(26.22)*	71.24(16.20)***
5:RegionPays-de-la-Loire	5.64(1.68)***	34.32(18.47)
5:RegionPoitou-Charentes	−0.87(0.56)	−0.73(0.58)

Significance code: ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.