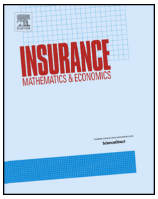
Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

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



Robust claim frequency modeling through phase-type mixture-of-experts regression

Martin Bladt[a,*], Jorge Yslas[b]

[a] *Department of Mathematical Sciences, University of Copenhagen, DK-2100 Copenhagen, Denmark*
[b] *Institute for Financial and Actuarial Mathematics, University of Liverpool, L69 7ZL Liverpool, UK*

ARTICLE INFO

Article history:
Received December 2022
Received in revised form February 2023
Accepted 25 February 2023
Available online 6 March 2023

JEL classification:
C13

Keywords:
Discrete phase-type distributions
Regression modeling
Claim count distributions

ABSTRACT

This paper addresses the problem of modeling loss frequency using regression when the counts have a non-standard distribution. We propose a novel approach based on mixture-of-experts specifications on discrete-phase type distributions. Compared to continuous phase-type counterparts, our approach offers fast estimation via expectation-maximization, making it more feasible for use in real-life scenarios. Our model is both robust and interpretable in terms of risk classes, and can be naturally extended to the multivariate case through two different constructions. This avoids the need for ad-hoc multivariate claim count modeling. Overall, our approach provides a more effective solution for modeling loss frequency in non-standard situations.

1. Introduction

Modeling loss frequency using regression is a common task in insurance and actuarial science. Accurately modeling the frequency and severity of losses is essential for effective risk management and pricing of insurance policies. However, this can be a difficult task when the counts have a non-standard distribution, with zero-inflated-like models (such as in Yip and Yau (2005); Lee (2021); Chen et al. (2019); Zhang et al. (2022)) and machine-learning methods (see Gabrielli (2020); Gao et al. (2022) and Wüthrich and Merz (2022) for a broader overview) being predominant in the recent literature. This is because many commonly used models, such as Poisson or Negative Binomial regression, are not particularly well-suited for non-standard distributions, and ad-hoc methods are often adopted, though with no formal justification. In this paper, we propose a novel approach for addressing this problem by using mixture-of-experts specifications on discrete-phase type distributions. This approach has gained popularity in the last decade, as it is inspired by machine learning and allows for fast and effective modeling of complex data.

Mixture-of-experts models are widely used in machine learning and statistics (Yuksel et al. (2012)), and they are known for their ability to accurately model complex data distributions. These models are particularly useful in situations where the data contains multiple modes or clusters, as they allow each mode to be modeled separately by a different expert component. This flexibility allows mixture-of-experts models to capture the underlying structure of the data more accurately than other types of probabilistic models. The applications of mixture-of-experts models are numerous and diverse and include classification and density estimation. Recently, they have been introduced to actuarial science in Fung et al. (2019).

In this paper, we introduce a series of models, which in the simplest univariate case, falls into the classical mixture-of-experts definition, inheriting their favorable properties. However, a key difference is that our proposal has an additional viewpoint from the point of view of absorption times of Markov processes, which statistically allows for a much more concise (in terms of model components) and effective tailor-made estimation procedure. Furthermore, our model is both robust and interpretable in terms of risk classes, which is important for effective risk management. In addition, the model can be easily extended to the multivariate case through two different constructions, which avoids the need for ad-hoc multivariate claim count modeling. This is important because ad-hoc methods can be

[*] Corresponding author.
E-mail addresses: martinbladt@math.ku.dk (M. Bladt), jorge.yslas-altamirano@liverpool.ac.uk (J. Yslas).

https://doi.org/10.1016/j.insmatheco.2023.02.008

difficult to implement and may not accurately capture the underlying data. It is worth mentioning that Fung et al. (2019) considered the multivariate case as well, though for other much simpler component distributions.

Similarly to continuous phase-type counterparts (which for the multivariate case has been estimated in general in Albrecher et al. (2022) and for a subclass in Bladt (2023)), our approach offers estimation via a (generalized) expectation-maximization (EM), though a key difference is that we deal with matrix powers instead of matrix exponentiation, dramatically reducing computation complexity and thus making it applicable for use in real-life scenarios with large datasets. The EM algorithm is a powerful general-purpose tool for estimating the parameters of probabilistic models that contain latent variables (Dempster et al. (1977)). In the univariate setting, this algorithm works by iteratively updating the estimated values of the latent variables and the model parameters based on the observed data (though some adaptations are necessary for the mixture-of-experts models, which are particularly tricky to fit, see also Chen et al. (1999)). This approach can be extended to the multivariate setting, where it estimates the parameters of our mixture-of-experts specification of discrete phase-type distributions. This allows us to use existing methods for estimating the parameters of multinomial distributions (zero-layer neural networks) to quickly and effectively estimate the parameters of our models. It is also important to note that gradient-based methods can be problematic when applied to these models because of the great deal of variables and non-linearity in the likelihood. Therefore, the EM algorithm provides a more robust and reliable approach for estimating the parameters in this setting. In the absence of regressors, the general class of multivariate discrete phase-type distributions was introduced in Navarro (2018), while in actuarial science, multivariate counts were studied using a subclass of discrete phase-type distributions in He and Ren (2016a) in terms of risk measures (see also Ren and Zitikis (2017) for a different construction) and in He and Ren (2016b) in terms of estimation through an EM algorithm.

We apply the proposed methodology to a real-life data set, the Wisconsin Local Government Property Insurance Fund (LGPIF) data set (previously analyzed in Frees et al. (2016) to compare several standard regression models) to demonstrate its effectiveness in practice. We will compare our approach to existing methods and show that it provides more accurate estimates, with the code to estimate our models being open-source and freely available online.[1] This is mainly because our approach is able to capture the subtleties in the distribution of the data, whereas existing methods are not well-suited for this type of data. Such subtleties are worth modeling when datasets are large.

In the following sections, we will first revisit some of the properties of classical univariate discrete phase-type distributions (Section 2), which are the building blocks of our proposed model. We will then present the multivariate extensions of these distributions and derive estimation procedures for two of the most tractable sub-classes (Section 3). These procedures allow us to estimate the parameters of our model using expectation-maximization. We then introduce estimation of regression models via the mixture-of-experts specification for all models in the paper (Section 4). Finally, we show its practical applicability using real-life data (Section 5) and conclude (Section 6).

2. Discrete phase-type distributions

Discrete distributions are an important tool in statistics and actuarial science, as they provide a means of modeling the count data that is often encountered in these fields. Claim counts, in particular, are an important type of data that can be modeled using discrete distributions. These distributions provide a flexible and powerful means of modeling the complex relationships between the factors that influence claim counts, and can be used to make accurate modeling about present and future claim frequency.

Discrete phase-type (DPH) distributions are a particularly good model for flexibly modeling a wide range of data within the same context. These distributions are a generalization of the traditional discrete distributions, and allow for the modeling of complex behavior of discrete variables. This makes them well-suited for use in actuarial science, where there is high monetary value associated with accurately capturing the behavior of claim frequency.

In this section, we revisit some of the key properties of the classical univariate discrete phase-type distributions, which are the building blocks of every model presented in the remainder of this paper.

Hence, consider a time-homogeneous Markov chain $Z = (Z_n)_{n\in\mathbb{N}_0}$ on a state space $E = \{1, \dots, p, p+1\}$, where states $1, \dots, p$ are transient and $p+1$ is absorbing. In essence, this guarantees that the hitting time of the process to state $p+1$ is almost surely finite. The transition matrix $\boldsymbol{P}$ of Z is then of the form

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{t} \\ \boldsymbol{0} & 1 \end{pmatrix},$$

where $\boldsymbol{T} = (t_{kl})_{k,l=1,\dots,p}$ is a $p \times p$ matrix, called a sub-transition matrix, containing the transition probabilities among the transient states, and $\boldsymbol{t} = (t_1, \dots, t_p)^\top$ is the vector of transition probabilities from the transient states to the absorbing state, also known as the exit probability vector. Note that since the rows of $\boldsymbol{P}$ sum to one, we have the relationship $\boldsymbol{t} = \boldsymbol{e} - \boldsymbol{T}\boldsymbol{e}$, where $\boldsymbol{e}$ denotes the p-dimensional vector of ones. Naturally, the rows' sums of $\boldsymbol{T}$ are possibly below one. Let $\pi_k = \mathbb{P}(Z_0 = k)$, $k = 1, \dots, p, p+1$ and $(\boldsymbol{\pi}, \pi_{p+1})$ be the initial probabilities of the Markov chain, where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$. In what follows, we assume that we cannot start in the absorbing state, so that $\pi_{p+1} = 0$, and then $\boldsymbol{\pi}\boldsymbol{e} = 1$, unless stated otherwise.

Definition 2.1 *(DPH).* The time until absorption

$$Y = \inf\{n \geq 1 : Z_n = p+1\},$$

follows a discrete phase-type distribution with initial distribution $\boldsymbol{\pi}$ and sub-transition matrix $\boldsymbol{T}$, and we write $Y \sim \mathrm{DPH}(\boldsymbol{\pi}, \boldsymbol{T})$.

The density f_Y and cumulative distribution function F_Y of $Y \sim \mathrm{DPH}(\boldsymbol{\pi}, \boldsymbol{T})$ are explicit and given in terms of matrix functions as

$$f_Y(y) = \mathbb{P}(Y = y) = \boldsymbol{\pi}\boldsymbol{T}^{y-1}\boldsymbol{t}, \quad y \in \mathbb{N},$$
$$F_Y(y) = \mathbb{P}(Y \leq y) = 1 - \boldsymbol{\pi}\boldsymbol{T}^{y}\boldsymbol{e}, \quad y \in \mathbb{N}.$$

[1] The repository can be found at https://github.com/martinbladt/matrixdist_1.0, see also Bladt and Yslas (2022a, 2021).

These formulas are reminiscent of the geometric distribution if we replace the matrix with a scalar. Indeed, if $p = 1$ we recover the geometric distribution. Perhaps less obvious is the fact that the Negative Binomial distribution can also be obtained as a special case with identical phases in sequence. The class is closed under mixtures.

Concerning expected values, we have that the first moment of Y is given by

$$\mathbb{E}[Y] = \boldsymbol{\pi}\,(\boldsymbol{I} - \boldsymbol{T})^{-1}\boldsymbol{e}\,.$$

More generally, the κ-th factorial moment of Y, $\kappa \in \mathbb{N}$, is explicit and given by

$$\mathbb{E}[Y(Y-1)\dots(Y-\kappa+1)] = \kappa!\boldsymbol{\pi}\boldsymbol{T}^{\kappa-1}\,(\boldsymbol{I} - \boldsymbol{T})^{-\kappa}\,\boldsymbol{e}\,.$$

Additionally, the probability-generating function G_Y of Y possesses a closed-form expression given by

$$G_Y(z) = \mathbb{E}[z^Y] = z\boldsymbol{\pi}\,(\boldsymbol{I} - z\boldsymbol{T})^{-1}\boldsymbol{t} = \boldsymbol{\pi}\left(z^{-1}\boldsymbol{I} - \boldsymbol{T}\right)^{-1}\boldsymbol{t}\,,$$

and it exists (at least) for $|z| \leq 1$.

These properties make DPH distributions easy to implement, given a *representation*, that is, given $(\boldsymbol{\pi}, \boldsymbol{T})$. Most of the applications of these distributions thus concentrate on finding representations efficiently. As we will see below, using the underlying jump-process is very useful in this regard, and thus expectation-maximization (EM) algorithms are of statistical importance. It is worth noting, however, that for small dimensions, direct gradient-based methods are also effective.

Remark 2.1. Note that in the definition of $Y \sim \text{DPH}(\boldsymbol{\pi}, \boldsymbol{T})$ above, we do not allow for $Y = 0$. However, if we are interested in having a discrete distribution that can take the value of 0, we can simply consider $\pi_{p+1} > 0$ so that $\mathbb{P}(Y = 0) = \pi_{p+1}$. In such a case, we can think of the law of Y as a mixture distribution with density function

$$\pi_{p+1}1\{y = 0\} + (1 - \pi_{p+1})f(y)1\{y > 0\}\,, \quad y \in \mathbb{N}_0\,,$$

where $f(y)$ is the density function of a $\text{DPH}(\boldsymbol{\pi}, \boldsymbol{T})$ distribution as before.

Equally effective, in terms of interpretation and statistical accuracy, is simply modeling $Y + 1$ as a DPH distributed random variable.

The following result from Neuts (1975) shows the robust modeling capabilities of DPH distributions. While a proof is provided here for future reference, the representation is not particularly concise. This result demonstrates the flexibility of DPH distributions, as they can be used to approximate a wide range of other distributions with arbitrary precision.

Theorem 2.2. *Any probability mass function g with finite support on $\mathbb{N}$ is a discrete phase-type.*

Proof. Consider a probability mass function g with support on $\{1, \dots, n\}$ for some $n \in \mathbb{N}$. We denote the corresponding probabilities by $g(i)$. Then, a DPH representation of dimension n for g is the following:

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

$$\boldsymbol{T} = \begin{pmatrix} 0 & (1 - g(1)) & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{1-g(1)-g(2)}{1-g(1)} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \frac{1-g(1)-g(2)-g(3)}{1-g(1)-g(2)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}. \quad \square$$

In particular, we also have that the set of DPH distribution is dense in the set of discrete distributions with support on the entire $\mathbb{N}$.

Example 2.3. Consider the function $x \mapsto \sqrt{x}(1 + \sin(x/2))$ sampled at the points $1, 2, \dots, 35$, and then normalized by the total sum. By specifying a DPH construction as in the above theorem, we obtain a 35-dimensional representation exactly matching the constructed density. Fig. 2.1 shows the construction.

Remark 2.2 *(On the dimension of an approximation).* Generally speaking, when no regression is present, the dimension p of the required DPH distribution should be taken to be smaller than the maximum value of the data.

In the worst-case scenario, DPH distributions work as mixtures, which will target one specific histogram value at a time. However, by using matrix parameters, we allow for jumping back and forth between the states of the underlying Markov chain, whereby increased flexibility is attained for fixed p. In other words, increasing p usually results in a better fit not only for one value in the histogram but entire sections of the distribution. Furthermore, since the number of parameters (we prefer to think of them as weak learners in this context) grows with p^2 (which can be a drawback but also a strength, as is the case when talking about denseness), then it is virtually never the case that a very large dimension is required to obtain a good fit, and often a good model will have a single-digit p.

Also, note that the theoretical denseness construction is not very parsimonious and should not be used for model building. Such construction does not take advantage of the full discrete phase-type potential (namely, the parsimony induced by state interactions).

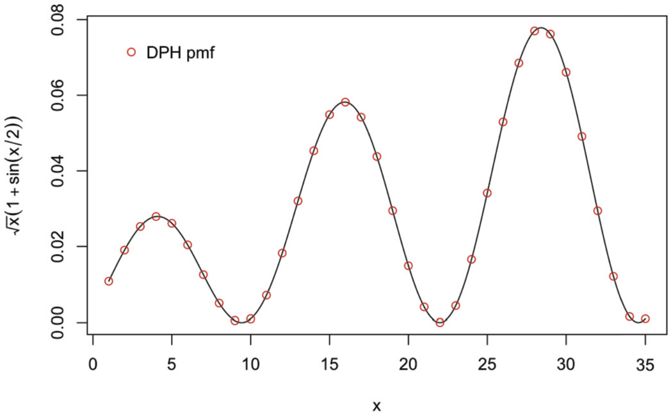


Fig. 2.1. DPH representation of a finite-support distribution.

3. Multivariate discrete phase-type distributions

The goal of this section is twofold. First, it provides an overview of existing notation and results related to multivariate extensions of DPH distributions, which will be useful in the following sections. Second, it presents the first known estimation procedures for two of the most tractable multivariate discrete phase-type distributions. These procedures will serve as the building blocks for the multivariate mixture-of-experts regression models that will be discussed later on.

3.1. The multivariate discrete phase-type class

In this section, we will begin by reviewing the MDPH* class of multivariate DPH distributions introduced in Navarro (2018). This class of distributions is generally considered to be of theoretical importance due to the fact that only certain subclasses are tractable, meaning that they may not have explicit probability mass functions and distribution functions. Consequently, we focus on these tractable subclasses in subsequent sections.

Let $\tau \sim \text{DPH}(\boldsymbol{\pi}, \boldsymbol{T})$ be a discrete phase-type distributed random variable of dimension p with underlying Markov chain $(Z_n)_{n\in\mathbb{N}_0}$. Let $\boldsymbol{r}_j = \left(r_j(1), \ldots, r_j(p)\right)^\top$ be p-dimensional column vectors taking values in $\mathbb{N}_0^p$, $j = 1, \ldots, d$, and let

$$\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_d)$$

be a $p \times d$ matrix, called a *reward matrix*.

Definition 3.1 *(MDPH*).* Set

$$Y^{(j)} = \sum_{m=0}^{\tau-1} r_j\left(Z_m\right),$$

for all $j = 1, \ldots, d$. We then say that the random vector $\boldsymbol{Y} = \left(Y^{(1)}, \ldots, Y^{(d)}\right)$ has a multivariate discrete phase-type distribution of the MDPH* type, and we write $\boldsymbol{Y} \sim \text{MDPH}^*(\boldsymbol{\pi}, \boldsymbol{T}, \boldsymbol{R})$.

In the above definition, we can interpret $r_j(k)$ as the reward obtained while Z is in state k. Thus, $Y^{(j)}$ can be seen as the total reward for component j obtained prior to absorption. In particular, we have that $Y^{(j)}$ is DPH distributed for all $j = 1, \ldots, d$.

Navarro (2018) showed that elements of the MPH* class have explicit expressions for the joint probability-generating function, joint moment-generating function, and joint moments. Moreover, some closure properties of the class were derived, and denseness in the class of distributions with support in $\mathbb{N}^d$ was shown. However, as previously mentioned, and as is the case for the continuous counterparts (the MPH* class), there are no general explicit expressions for the density and distribution functions. In addition, and in contrast to the continuous case,[2] the estimation for this general class of distributions is still an open statistical challenge, which limits their applicability to describe real-life data.

Next, we present two sub-classes of the DMPH* class which have explicit expressions for the joint density and distribution functions, and that preserve the denseness property of the DMPH* class. As an important consequence, these sub-classes allow for estimation via EM algorithms, making them attractive tools for modeling real-life data. We derive the main parts of the algorithms below, delegating the full details to Appendix A. In Section 4, we show how these models can be adapted to work with the regression setting presented there.

3.2. The feed-forward class of multivariate discrete phase-type distributions

In this section, we will present our results for the bivariate case in order to improve the clarity of the presentation. It is possible to extend these results to higher dimensions, although doing so can be challenging from an implementation perspective.

[2] The estimation idea used for the MPH* class (cf., Breuer (2016)) cannot be applied directly, given that the assumption of rows of the reward matrix summing one cannot be imposed in this case.

Thus, consider a time-homogeneous Markov chain $Z = (Z_n)_{n\in\mathbb{N}_0}$ on a state space $E = \{1, \dots, p, p+1\}$, where states $1, \dots, p$ are transient and $p+1$ is absorbing. We now split the set of transit states $\{1, \dots, p\}$ into two sets: $E_1 = \{1, \dots, p_1\}$ and $E_2 = \{p_1+1, \dots, p\}$, $p_1 < p$. Furthermore, we denote by p_2 the number of states in E_2, that is, $p_2 = p - p_1$.

We denote by $\boldsymbol{T}_{11}$ and $\boldsymbol{T}_{22}$ the sub-transition matrices describing the transition among the states in E_1 and E_2, respectively. Additionally, we denote by $\boldsymbol{T}_{12}$ the transition matrix describing movement from E_1 to E_2. Finally, we make the following assumption: the process cannot reach the absorbing state from a state in E_1 and cannot return to a state in E_1 once it has entered E_2.

Thus, the sub-transition matrix of Z is of the form

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ \boldsymbol{0} & \boldsymbol{T}_{22} \end{pmatrix}.$$

Note that under this setup, $\boldsymbol{T}_{11}\boldsymbol{e} + \boldsymbol{T}_{12}\boldsymbol{e} = \boldsymbol{e}$, and the vector of exit probabilities is $\boldsymbol{t} = (\boldsymbol{0}^\top, \boldsymbol{t}_2^\top)^\top$, where $\boldsymbol{t}_2 = \boldsymbol{e} - \boldsymbol{T}_{22}\boldsymbol{e}$.

Next, we assume that the Markov chain can only start in a state of E_1 but not E_2, which imposes the following structure for its initial distribution $\boldsymbol{\pi} = (\boldsymbol{\eta}, \boldsymbol{0})$, where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{p_1})$ with $\boldsymbol{\eta}\boldsymbol{e} = 1$. Now let $Y = \inf\{n \geq 1 : Z_n \in E_2\}$ and $W = \inf\{n \geq 1 : Z_n = p+1\}$.

Definition 3.2 *(fMDPH, two-dimensional).* Let $Y = \inf\{n \geq 1 : Z_n \in E_2\}$ and $W = \inf\{n \geq 1 : Z_n = p+1\}$. Then, we say that $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)}) = (Y, W - Y)$ is bivariate discrete phase-type distributed of the feed-forward type, and we write $\boldsymbol{Y} \sim \text{fMDPH}(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22})$.

Remark 3.1 *(Interpretation of the fMDPH class).* Suppose that two claim frequencies follow the fMDPH distribution. Then we assume that there is a latent underlying risk process that generates the claim numbers for the first component while in a certain environment (the state space E_1). When there is a change of environment (entering state space E_2), the generation of the first component terminates, and the one for the second component starts until the process gets absorbed. The dependence is generated by the fact that the *manner* (or state) in which the first component terminates directly affects how the second one is started.

Standard probabilistic considerations yield the following expressions for the joint density $f_{\boldsymbol{Y}}$ and joint survival function $\overline{F}_{\boldsymbol{Y}}$ of $\boldsymbol{Y}$:

$$f_{\boldsymbol{Y}}(y^{(1)}, y^{(2)}) = \mathbb{P}(Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}) = \boldsymbol{\eta}\boldsymbol{T}_{11}^{y^{(1)}-1}\boldsymbol{T}_{12}\boldsymbol{T}_{22}^{y^{(2)}-1}\boldsymbol{t}_2\,,$$
$$\overline{F}_{\boldsymbol{Y}}(y^{(1)}, y^{(2)}) = \mathbb{P}(Y^{(1)} > y^{(1)}, Y^{(2)} > y^{(2)}) = \boldsymbol{\eta}\boldsymbol{T}_{11}^{y^{(1)}}\boldsymbol{T}_{12}\boldsymbol{T}_{22}^{y^{(2)}}\boldsymbol{e}\,.$$

We can then use the equations above to derive other functionals of $\boldsymbol{Y}$. First, we have that the joint probability-generating function of $\boldsymbol{Y}$ is

$$\begin{aligned}\mathbb{E}\left[z_1^{Y^{(1)}} z_2^{Y^{(2)}}\right] &= z_1 z_2 \boldsymbol{\eta}(\boldsymbol{I} - z_1\boldsymbol{T}_{11})^{-1}\boldsymbol{T}_{12}(\boldsymbol{I} - z_2\boldsymbol{T}_{22})^{-1}\boldsymbol{t}_2 \\ &= \boldsymbol{\eta}(z_1^{-1}\boldsymbol{I} - \boldsymbol{T}_{11})^{-1}\boldsymbol{T}_{12}(z_2^{-1}\boldsymbol{I} - \boldsymbol{T}_{22})^{-1}\boldsymbol{t}_2\,.\end{aligned}$$

Next, the joint factorial moments of $\boldsymbol{Y}$ are given by

$$\mathbb{E}\left[\prod_{j=1}^{2}\prod_{b=1}^{\kappa_j}(Y^{(j)} - b + 1)\right] = \kappa_1!\kappa_2!\boldsymbol{\eta}\boldsymbol{T}_{11}^{\kappa_1-1}\left(\boldsymbol{I} - \boldsymbol{T}_{11}\right)^{-\kappa_1-1}\boldsymbol{T}_{12}\boldsymbol{T}_{22}^{\kappa_2-1}\left(\boldsymbol{I} - \boldsymbol{T}_{22}\right)^{-\kappa_2}\boldsymbol{e}\,,$$

where $\kappa_1, \kappa_2 \in \mathbb{N}$. In particular, we can use this last expression to obtain

$$\mathbb{E}\left[Y^{(1)}Y^{(2)}\right] = \boldsymbol{\eta}\left(\boldsymbol{I} - \boldsymbol{T}_{11}\right)^{-2}\boldsymbol{T}_{12}\left(\boldsymbol{I} - \boldsymbol{T}_{22}\right)^{-1}\boldsymbol{e}\,,$$

which is required to compute the correlation matrix of $\boldsymbol{Y}$.

Finally, we have that the marginals are DPH distributed with $Y^{(1)} \sim \text{DPH}(\boldsymbol{\eta}, \boldsymbol{T}_{11})$ and $Y^{(2)} \sim \text{DPH}(\boldsymbol{\eta}(\boldsymbol{I} - \boldsymbol{T}_{11})^{-1}\boldsymbol{T}_{12}, \boldsymbol{T}_{22})$.

Example 3.3 *(Dependence structures).*

Independence. Consider $Y^{(1)} \sim \text{DPH}(\boldsymbol{\pi}, \boldsymbol{T})$ and $Y^{(2)} \sim \text{DPH}(\boldsymbol{\gamma}, \boldsymbol{S})$ independent. Then, an fMDPH representation for $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)})$ is given by $\boldsymbol{\eta} = (\boldsymbol{\pi}, \boldsymbol{0})$, $\boldsymbol{T}_{11} = \boldsymbol{T}$, $\boldsymbol{T}_{22} = \boldsymbol{S}$, and $\boldsymbol{T}_{12} = \boldsymbol{t}\boldsymbol{\gamma}$.

Perfect positive dependence. Intuitively, positive dependence in an fMDPH distribution is obtained when highly-visited states in E_1 are connected with states in E_2 with the same property via the matrix $\boldsymbol{T}_{12}$. For example, the following fMDPH distribution yields an extreme case.

$$\boldsymbol{\eta} = (0, 1)\,, \quad \boldsymbol{T}_{11} = \begin{pmatrix} 0 & 0 \\ 0.5 & 0 \end{pmatrix}, \quad \boldsymbol{T}_{12} = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad \boldsymbol{T}_{22} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Note that in the example above, the perfect positive dependence is obtained because there is a unique way to move through the Markov chain due to the sparse structure of the matrices with some probabilities equal to 1. This construction easily generalizes to higher dimensions.

Perfect negative correlation. On the other hand, negative dependence is obtained when highly-visited states in E_1 are connected with states in E_2 with low visit probabilities. The following example exhibits perfect negative dependence.

$$\boldsymbol{\eta} = (0, 1)\,, \quad \boldsymbol{T}_{11} = \begin{pmatrix} 0 & 0 \\ 0.5 & 0 \end{pmatrix}, \quad \boldsymbol{T}_{12} = \begin{pmatrix} 0 & 1 \\ 0.5 & 0 \end{pmatrix}, \quad \boldsymbol{T}_{22} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

The construction principle of fMDPH distributions in terms of a single Markov chain allows us to obtain the following result analogous to Theorem 2.2.

Theorem 3.4. *Any joint probability mass function g with finite support on $\mathbb{N}^2$ is a bivariate discrete phase-type of the feed-forward type.*

Proof. Notice that we can always construct an absorbing Markov chain with paths absorbing precisely at the time points given by the support of g and with probabilities matching those of g. The details are omitted. □

We present a simple example illustrating the above statement.

Example 3.5. Consider $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)})$ with joint mass probability function given by $g(1,1) = 0.2$, $g(1,2) = 0.4$, $g(2,1) = 0.1$, and $g(2,2) = 0.3$. Then, an fMDPH representation of this distribution is given by

$$\boldsymbol{\eta} = (1, 0)\,, \quad \boldsymbol{T}_{11} = \begin{pmatrix} 0 & 0.4 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{T}_{12} = \begin{pmatrix} 0.5 & 0.1 \\ 15/16 & 1/16 \end{pmatrix}, \quad \boldsymbol{T}_{22} = \begin{pmatrix} 0 & 0.8 \\ 0 & 0 \end{pmatrix}.$$

Remark 3.2. Note that a further consequence of Theorem 3.4 is that the set of fMDPH distributions is dense in the set of discrete distributions supported on $\mathbb{N}^2$.

We now show that the definition of fMDPH distributions falls into the MDPH* specification of Navarro (2018).

Proposition 3.6 *(fMDPH ⊂ MDPH*). The fDMPH class is contained in the MDPH* class*

Proof. It suffices to see that an MDPH* representation of an element in fMDPH is given by

$$\boldsymbol{\pi} = (\boldsymbol{\eta}, \boldsymbol{0})\,, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ \boldsymbol{0} & \boldsymbol{T}_{22} \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{e} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{e} \end{pmatrix}. \quad \square$$

3.2.1. Estimation in two dimensions

Consider a dataset consisting of N two-dimensional i.i.d. observations

$$\boldsymbol{y}_i = (y_i^{(1)}, y_i^{(2)})\,, \quad i = 1, \dots, N\,,$$

from a $\boldsymbol{Y} \sim \text{fMDPH}(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22})$ distributed random vector. In what follows, we denote this sample as $\overline{\boldsymbol{y}} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\}$.

We first note that we only observe the times to entry to E_2 and to absorption of the underlying Markov chain. Hence, we are in an incomplete-data setup, and the expectation-maximization (EM) algorithm shall be employed. To apply such a method, we first require the complete likelihood, that is, the likelihood assuming that the entire path of the underlying Markov chain is observed.

Let B_k be the number of Markov chains starting in state k, N_{kl} the number of transitions from state k to state l, and N_k the number of Markov chains that exit from state k to the absorbing state. Then, the complete likelihood function for this sample is given by

$$L_c\left(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22}; \overline{\boldsymbol{y}}\right) = \prod_{k=1}^{p_1} \eta_k^{B_k} \prod_{k,l=1}^{p_1} t_{kl}^{N_{kl}} \prod_{k=1}^{p_1} \prod_{l=p_1+1}^{p} t_{kl}^{N_{kl}} \prod_{k,l=p_1+1}^{p} t_{kl}^{N_{kl}} \prod_{k=p_1+1}^{p} t_k^{N_k}\,.$$

This yields the following expression for the complete loglikelihood

$$\begin{aligned} l_c\left(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22}; \overline{\boldsymbol{y}}\right) = {} & \sum_{k=1}^{p_1} B_k \log(\eta_k) + \sum_{k,l=1}^{p_1} N_{kl} \log(t_{kl}) + \sum_{k=1}^{p_1} \sum_{l=p_1+1}^{p} N_{kl} \log(t_{kl}) \\ & + \sum_{k,l=p_1+1}^{p} N_{kl} \log(t_{kl}) + \sum_{k=p_1+1}^{p} N_k \log(t_k)\,. \end{aligned}$$

Then, we need to alternate between an expectation (E) step consisting of calculating the expected value of the complete loglikelihood given the observed sample and a maximization (M) step in which we maximize the complete likelihood with the conditional expectations in place of the actual sufficient statistics.

We first proceed to compute the conditional expectations of the sufficient statistics given the observed data. For such a purpose, we consider a generic data point $\boldsymbol{y} = (y^{(1)}, y^{(2)})$ to perform the calculations. We start with B_k:

$$\begin{aligned} \mathbb{E}[B_k \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{E}\left[1\{Z_0 = k\} \mid \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= \frac{\mathbb{P}\left(Z_0 = k, Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}\right)}{\mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y}\right)} \\ &= \frac{\eta_k \boldsymbol{e}_k{}^{\top} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}\,. \end{aligned}$$

Here, $\boldsymbol{e}_k$ denotes the k-th canonical basis vector in $\mathbb{R}^{p_1}$. In what follows, we use the same notation for canonical basis vectors in real vector spaces of different dimensions.

For N_{kl}, given the structure of the underlying Markov chain, we need to split the computations into three different cases.

Case 1: $k, l \in E_1$. Note that transition between two states in E_1 is not possible if the transition of the chain to E_2 was before time 2. Thus, we have that

$$N_{kl} = 1\{Y^{(1)} \geq 2\} \sum_{m=0}^{Y^{(1)}-2} 1\{Z_m = k, Z_{m+1} = l\}\,.$$

Then,

$$\begin{aligned}
\mathbb{E}[N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}] &= 1\{y^{(1)} \geq 2\} \sum_{m=0}^{y^{(1)}-2} \mathbb{P}\left(Z_m = k, Z_{m+1} = l \mid \boldsymbol{Y} = \boldsymbol{y}\right) \\
&= 1\{y^{(1)} \geq 2\} \sum_{m=0}^{y^{(1)}-2} \frac{\mathbb{P}\left(Z_m = k, Z_{m+1} = l, Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}\right)}{\mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y}\right)} \\
&= 1\{y^{(1)} \geq 2\} \sum_{m=0}^{y^{(1)}-2} \frac{\boldsymbol{\eta} \boldsymbol{T}_{11}^{m} \boldsymbol{e}_k t_{kl} \boldsymbol{e}_l^{\top} \boldsymbol{T}_{11}^{y^{(1)}-(m+1)-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}\,.
\end{aligned}$$

Case 2: $k, l \in E_2$. Similarly to the previous case, we have that

$$\begin{aligned}
\mathbb{E}[N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}] &= 1\{y^{(2)} \geq 2\} \sum_{m=0}^{y^{(2)}-2} \frac{\mathbb{P}\left(Z_{y^{(1)}+m} = k, Z_{y^{(1)}+m+1} = l, Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}\right)}{\mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y}\right)} \\
&= 1\{y^{(2)} \geq 2\} \sum_{m=0}^{y^{(2)}-2} \frac{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{m} \boldsymbol{e}_{k-p_1} t_{kl} \boldsymbol{e}_{l-p_1}^{\top} \boldsymbol{T}_{22}^{y^{(2)}-(m+1)-1} \boldsymbol{t}_2}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}\,.
\end{aligned}$$

Case 3: $k \in E_1$, $l \in E_2$. We note that the corresponding conditional expectation can be interpreted as the probability of going from E_1 to E_2, at time $y^{(1)}$, due to a transition from k to l. Thus,

$$\begin{aligned}
\mathbb{E}[N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{P}\left(Z_{y^{(1)}-1} = k, Z_{y^{(1)}} = l \mid \boldsymbol{Y} = \boldsymbol{y}\right) \\
&= \frac{\mathbb{P}\left(Z_{y^{(1)}-1} = k, Z_{y^{(1)}} = l, Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}\right)}{\mathbb{P}\left(\boldsymbol{Y} = \boldsymbol{y}\right)} \\
&= \frac{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{e}_k t_{kl} \boldsymbol{e}_{l-p_1}^{\top} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}\,.
\end{aligned}$$

For the statistics N_k, we simply have

$$\begin{aligned}
\mathbb{E}[N_k \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{P}\left(Z_{y^{(1)}+y^{(2)}-1} = k \mid \boldsymbol{Y} = \boldsymbol{y}\right) \\
&= \frac{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{e}_{k-p_1} t_k}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y^{(2)}-1} \boldsymbol{t}_2}\,.
\end{aligned}$$

Finally, for a sample of size N, we add the formulas above evaluated at all data points $\boldsymbol{y}_i$, $i = 1, \dots, N$.

The M-step follows straightforwardly by applying a Lagrange multiplier argument and is thus omitted for brevity. The complete EM algorithm is summarized in Algorithm 1.

3.2.2. Extension to higher dimensions

We can extend the construction principle of fMDPH distributions to higher dimensions by simply splitting the set of transit states $\{1, \dots, p\}$ into more subsets, say $E_1, \dots, E_d$ for some $d > 1$. We denote by $\boldsymbol{C}_1, \dots, \boldsymbol{C}_d$ the sub-transition matrices describing the movement of the Markov chain within states in $E_1, \dots, E_d$, respectively. Furthermore, we assume that the Markov chain starts in a state in E_1 and can only move to a state in E_{j+1} if it is currently in a state in E_j. Here, we understand E_{d+1} as $\{p+1\}$. These movements are described by the non-negative matrice $\boldsymbol{D}_1, \dots, \boldsymbol{D}_{d-1}$ satisfying $\boldsymbol{C}_j \boldsymbol{e} + \boldsymbol{D}_j \boldsymbol{e} = \boldsymbol{e}$, $j = 1, \dots, d-1$. Thus, the initial probabilities and sub-transition matrix of the underlying Markov chain are of the form

$$\boldsymbol{\pi} = (\boldsymbol{\eta}, \boldsymbol{0}, \dots, \boldsymbol{0})\,, \quad \boldsymbol{T} = \begin{pmatrix} \boldsymbol{C}_1 & \boldsymbol{D}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_2 & \boldsymbol{D}_2 & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{C}_3 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{C}_d \end{pmatrix}\,.$$

Definition 3.7 *(fMDPH, general dimension).* Let $W_j = \inf\{n \geq 1 : Z_n \in E_{j+1}\}$, $j = 1, \ldots, d$. Then, we say that $\boldsymbol{Y} = (Y^{(1)}, Y^{(2)}, \ldots, Y^{(d)}) = (W_1, W_2 - W_1, \ldots, W_d - W_{d-1})$ follows a fMDPH distribution.

As in the bivariate case, several functionals of $\boldsymbol{Y}$ are explicit in terms of matrix functions. For instance, the joint density is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \boldsymbol{\eta} \boldsymbol{C}_1^{y^{(1)}-1} \boldsymbol{D}_1 \boldsymbol{C}_2^{y^{(2)}-1} \boldsymbol{D}_2 \cdots \boldsymbol{D}_{d-1} \boldsymbol{C}_d^{y^{(d)}-1} \boldsymbol{c}_d \,,$$

with $\boldsymbol{c}_d = \boldsymbol{e} - \boldsymbol{C}_d \boldsymbol{e}$. However, as we can see from the derivations of Algorithm 1, the complexity of an estimation procedure for higher dimensions increases non-linearly, and each dimension needs to be treated individually, which makes it challenging from an implementation point of view.

Hence, we now introduce another subclass of the MDPH*, which turns out to also be a subclass of the fMDPH class, with similar desirable properties, but with an estimation procedure that allows for simple scaling into higher dimensions.

3.3. The mDPH class

Consider d time-homogeneous Markov chains $Z^{(j)} = (Z_n^{(j)})_{n \in \mathbb{N}_0}$, $j = 1, \ldots, d$, on a common state space $E = \{1, \ldots, p, p+1\}$, where states $1, \ldots, p$ are transient and $p+1$ is absorbing. We assume that the dependence structure of these processes is as follows:

$$Z_0^{(j)} = Z_0^{(q)} \,, \quad \forall j, q \in \{1, \ldots, d\} \,,$$

and

$$Z^{(j)} \perp\!\!\!\perp_{Z_0^{(1)}} Z^{(q)} \,, \quad \forall j, q \in \{1, \ldots, d\} \,, \;\; j \neq q \,.$$

In other words, all the Markov chains start at the same state and evolve independently after that. In what follows, we denote by $Z_0 = Z_0^{(1)}$, and $\pi_k = \mathbb{P}(Z_0 = k)$, $k = 1, \ldots, p$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$. Moreover, the sub-transition matrix and vector of exit probabilities associated with $Z^{(j)}$ are denoted by

$$\boldsymbol{T}_j = (t_{kl}^{(j)})_{k,l=1,\ldots,p} \quad \text{and} \quad \boldsymbol{t}_j = \boldsymbol{e} - \boldsymbol{T}_j \boldsymbol{e} = (t_1^{(j)}, \ldots, t_p^{(j)})^\top \,,$$

respectively, for all $j = 1, \ldots, d$.

Definition 3.8 *(mDPH).* Let

$$Y^{(j)} = \inf\{n \geq 1 : Z_n^{(j)} = p+1\} \,, \quad j = 1, \ldots, d \,.$$

Then, we say that the random vector $\boldsymbol{Y} = (Y^{(1)}, \ldots, Y^{(d)})$ is mDPH distributed, and we use the notation

$$\boldsymbol{Y} \sim \text{mDPH}(\boldsymbol{\pi}, \mathcal{T}), \quad \text{with} \quad \mathcal{T} = \{\boldsymbol{T}_1, \ldots, \boldsymbol{T}_d\} \,.$$

Note in particular that $Y^{(j)} \sim \text{DPH}(\boldsymbol{\pi}, \boldsymbol{T}_j)$ for all $j = 1, \ldots, d$.

Remark 3.3 *(Interpretation of the mDPH class).* This class is peculiar in the sense that dependence between claim count variables following an mDPH law is completely determined by the initial state, which is shared by otherwise fully independently evolving sample paths. Though we will see below that this class belongs to the fMDPH class, the way of thinking of the mDPH class is closer to pure mixture models, where the mixture classes partition the distribution into natural risk classes. Within each risk class, the claim counts are independent.

As previously mentioned, an advantage of this class is that it possesses closed-form formulas for several functionals of interest, and they can easily be obtained by conditioning on the starting value of the Markov chains. For instance, the joint distribution function $F_{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ is obtained as follows:

$$\begin{aligned} F_{\boldsymbol{Y}}(\boldsymbol{y}) &= \mathbb{P}\left(Y^{(1)} \leq y^{(1)}, \ldots, Y^{(d)} \leq y^{(d)}\right) \\ &= \sum_{k=1}^{p} \mathbb{P}\left(Y^{(1)} \leq y^{(1)}, \ldots, Y^{(d)} \leq y^{(d)} \mid Z_0 = k\right) \mathbb{P}\left(Z_0 = k\right) \\ &= \sum_{k=1}^{p} \pi_k \prod_{j=1}^{d} \left(1 - \boldsymbol{e}_k^\top \boldsymbol{T}_j^{y^{(j)}} \boldsymbol{e}\right), \quad \boldsymbol{y} \in \mathbb{N}^d. \end{aligned}$$

Similarly, we have that the joint density is given by

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \sum_{k=1}^{p} \pi_k \prod_{j=1}^{d} \boldsymbol{e}_k^\top \boldsymbol{T}_j^{y^{(j)}-1} \boldsymbol{t}_j \,, \quad \boldsymbol{y} \in \mathbb{N}^d \,,$$

the joint probability-generating function is

$$\mathbb{E}\left[z_1^{Y^{(1)}}\cdots z_d^{Y^{(d)}}\right]=\sum_{k=1}^{p}\pi_k\prod_{j=1}^{d}\boldsymbol{e}_k^{\top}(z_j^{-1}\boldsymbol{I}-\boldsymbol{T}_j)^{-1}\boldsymbol{t}_j\,,$$

and the joint factorial moments can be computed as

$$\mathbb{E}\left[\prod_{j=1}^{d}\prod_{b=1}^{\kappa_j}(Y^{(j)}-b+1)\right]=\sum_{k=1}^{p}\pi_k\prod_{j=1}^{d}\kappa_j!\boldsymbol{e}_k^{\top}\boldsymbol{T}_j^{\kappa_j-1}\left(\boldsymbol{I}-\boldsymbol{T}_j\right)^{-\kappa_j}\boldsymbol{e}\,.$$

In particular, we also have that for $q\neq j$

$$\mathbb{E}\left[Y^{(j)}Y^{(q)}\right]=\sum_{k=1}^{p}\pi_k\left(\boldsymbol{e}_k^{\top}\left(\boldsymbol{I}-\boldsymbol{T}_j\right)^{-1}\boldsymbol{e}\right)\left(\boldsymbol{e}_k^{\top}\left(\boldsymbol{I}-\boldsymbol{T}_q\right)^{-1}\boldsymbol{e}\right),$$

which is needed to compute the correlation matrix of $\boldsymbol{Y}$.

Example 3.9 *(Dependence structures).*

Independence. Consider $Y^{(1)}\sim\text{DPH}(\boldsymbol{\nu},\boldsymbol{T})$ and $Y^{(2)}\sim\text{DPH}(\boldsymbol{\gamma},\boldsymbol{S})$ independent and of the same dimension p. Then, an mDPH representation for $\boldsymbol{Y}=(Y^{(1)},Y^{(2)})$ is given by $\boldsymbol{\pi}=(\boldsymbol{\nu}\otimes\boldsymbol{\gamma})\in\mathbb{R}^{p^2}$, and the $p^2\times p^2$ matrices

$$\boldsymbol{T}_1=\begin{pmatrix}\boldsymbol{T} & & & \\ & \boldsymbol{T} & & \\ & & \ddots & \\ & & & \boldsymbol{T}\end{pmatrix},\quad \boldsymbol{T}_2=\begin{pmatrix}\boldsymbol{S} & & & \\ & \boldsymbol{S} & & \\ & & \ddots & \\ & & & \boldsymbol{S}\end{pmatrix}.$$

Perfect positive dependence. One can again construct models with deterministic jumps, leading to highly dependent models, such as

$$\boldsymbol{\pi}=(0.5,0.5)\,,\quad \boldsymbol{T}_1=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix},\quad \boldsymbol{T}_2=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}.$$

Perfect negative dependence. The parameters in this case are given by

$$\boldsymbol{\pi}=(0.5,0.5)\,,\quad \boldsymbol{T}_1=\begin{pmatrix}0 & 0\\ 1 & 0\end{pmatrix},\quad \boldsymbol{T}_2=\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}.$$

The following results show that any other multivariate discrete distribution can be approximated by an mDPH model, which makes them a robust class from a statistical perspective.

Theorem 3.10 *(Densenseness). The class of mDPH distributions is dense in the set of distributions with support on $\mathbb{N}^d$.*

Proof. Consider DPH distributions of the form given in the proof of Theorem 2.2. Observe that these distributions satisfy Condition 3 of Proposition 3.1. in Fung et al. (2019). Finally, note that multivariate finite mixture models with DPH components of this form are particular specifications of mDPH distributions. Then, the result follows by Condition 1 of Proposition 3.1 in Fung et al. (2019). □

Finally, we establish the relationship of this class with respect to the other two multivariate models previously introduced.

Proposition 3.11 *(mDPH ⊂ fMDPH ⊂ MDPH*). The mDPH class is contained in the fMDPH and MDPH* classes.*

Proof. The proof is analogous to the one of Proposition 2.2 in Bladt (2023). □

3.3.1. *Estimation*

Maximum likelihood estimation of the mDPH class can be carried out via an explicit EM algorithm for arbitrary dimensions, which we derive next.

Consider i.i.d. realizations $\overline{\boldsymbol{y}}=\{\boldsymbol{y}_1,\dots,\boldsymbol{y}_N\}$, $\boldsymbol{y}_i=(y_i^{(1)},\dots,y_i^{(d)})$, $i=1,\dots,N$, of a d-dimensional random vector $\boldsymbol{Y}\sim\text{mDPH}(\boldsymbol{\pi},\mathcal{T})$. In order to write down the complete likelihood associated with this sample, we require the following definitions. As before, we let B_k be the total number of Markov chains starting in state k. For each one of the Markov chains $Z^{(j)}$, $j=1,\dots,d$, we let $N_{kl}^{(j)}$ be the number of transitions from state k to state l, and $N_k^{(j)}$ the number of exits from state k to the absorbing state. With these definitions at hand, we have that the complete likelihood is given by

$$L_c\left(\boldsymbol{\pi},\mathcal{T};\boldsymbol{y}\right)=\prod_{k=1}^{p}\pi_k^{B_k}\prod_{j=1}^{d}\prod_{k,l=1}^{p}(t_{kl}^{(j)})^{N_{kl}^{(j)}}\prod_{k=1}^{p}(t_k^{(j)})^{N_k^{(j)}}\,.$$

We proceed to compute the required conditional expectations. We consider a generic data point $\boldsymbol{y} = (y^{(1)}, \ldots, y^{(d)})$ with corresponding joint probability mass function $\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{s=1}^{p} \pi_s \prod_{j=1}^{d} \boldsymbol{e}_s^{\top} \boldsymbol{T}_j^{y^{(j)}-1} \boldsymbol{t}_j$. Then, for B_k we have that

$$\begin{aligned}
\mathbb{E}[B_k \mid \boldsymbol{Y} = \boldsymbol{y}] &= \sum_{j=1}^{d} \mathbb{E}[1\{Z_0^{(j)} = k\} \mid \boldsymbol{Y} = \boldsymbol{y}] \\
&= d\mathbb{P}(Z_0 = k \mid \boldsymbol{Y} = \boldsymbol{y}) \\
&= d\frac{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid Z_0 = k)\mathbb{P}(Z_0 = k)}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})} \\
&= d\frac{\pi_k \prod_{j=1}^{d} \boldsymbol{e}_k^{\top} \boldsymbol{T}_j^{y^{(j)}-1} \boldsymbol{t}_j}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})} .
\end{aligned}$$

Consider now $N_{kl}^{(j)}$. We first note that transition between two states is only possible if $y^{(j)} \geq 2$. Thus,

$$N_{kl}^{(j)} = 1\{Y^{(j)} \geq 2\} \sum_{m=0}^{Y^{(j)}-2} 1\{Z_m^{(j)} = k, Z_{m+1}^{(j)} = l\} .$$

Then,

$$\begin{aligned}
\mathbb{E}[N_{kl}^{(j)} \mid \boldsymbol{Y} = \boldsymbol{y}] &= 1\{y^{(j)} \geq 2\} \sum_{m=0}^{y^{(j)}-2} \mathbb{P}(Z_m^{(j)} = k, Z_{m+1}^{(j)} = l \mid \boldsymbol{Y} = \boldsymbol{y}) \\
&= 1\{y^{(j)} \geq 2\} \sum_{m=0}^{y^{(j)}-2} \frac{\mathbb{P}(Z_m^{(j)} = k, Z_{m+1}^{(j)} = l, \boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})} \\
&= 1\{y^{(j)} \geq 2\} \sum_{m=0}^{y^{(j)}-2} \frac{\sum_{s=1}^{p} \pi_s \prod_{q \neq j} \boldsymbol{e}_s^{\top} \boldsymbol{T}_q^{y^{(q)}-1} \boldsymbol{t}_q \boldsymbol{e}_s^{\top} \boldsymbol{T}_j^{m} \boldsymbol{e}_k t_{kl}^{(j)} \boldsymbol{e}_l^{\top} \boldsymbol{T}_j^{y^{(j)}-(m+1)-1} \boldsymbol{t}_j}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})} \\
&= 1\{y^{(j)} \geq 2\} t_{kl}^{(j)} \frac{\sum_{s=1}^{p} \pi_s \prod_{q \neq j} \boldsymbol{e}_s^{\top} \boldsymbol{T}_q^{y^{(q)}-1} \boldsymbol{t}_q \boldsymbol{e}_l^{\top} \boldsymbol{K}(y^{(j)}; \boldsymbol{e}_s^{\top}, \boldsymbol{T}_j) \boldsymbol{e}_k}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})} ,
\end{aligned}$$

where $\boldsymbol{K}(y; \boldsymbol{\pi}, \boldsymbol{T})$ is defined as

$$\boldsymbol{K}(y; \boldsymbol{\pi}, \boldsymbol{T}) = \sum_{m=0}^{y-2} \boldsymbol{T}^{y-2-m} \boldsymbol{t} \boldsymbol{\pi} \boldsymbol{T}^{m} ,$$

for any vector of initial probabilities $\boldsymbol{\pi}$ and sub-transition matrix $\boldsymbol{T}$.

Finally, for $N_k^{(j)} = 1\{Z_{Y^{(j)}-1}^{(j)} = k\}$, we have that

$$\begin{aligned}
\mathbb{E}[N_k^{(j)} \mid \boldsymbol{Y} = \boldsymbol{y}] &= \mathbb{P}(Z_{y^{(j)}-1}^{(j)} = k \mid \boldsymbol{Y} = \boldsymbol{y}) \\
&= \frac{\mathbb{P}(Z_{y^{(j)}-1}^{(j)} = k, \boldsymbol{Y} = \boldsymbol{y})}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})} \\
&= \frac{\sum_{s=1}^{p} \pi_s \prod_{q \neq j} \boldsymbol{e}_s^{\top} \boldsymbol{T}_q^{y^{(q)}-1} \boldsymbol{t}_q \boldsymbol{e}_s^{\top} \boldsymbol{T}_j^{y^{(j)}-1} \boldsymbol{e}_k t_k^{(j)}}{\mathbb{P}(\boldsymbol{Y} = \boldsymbol{y})} .
\end{aligned}$$

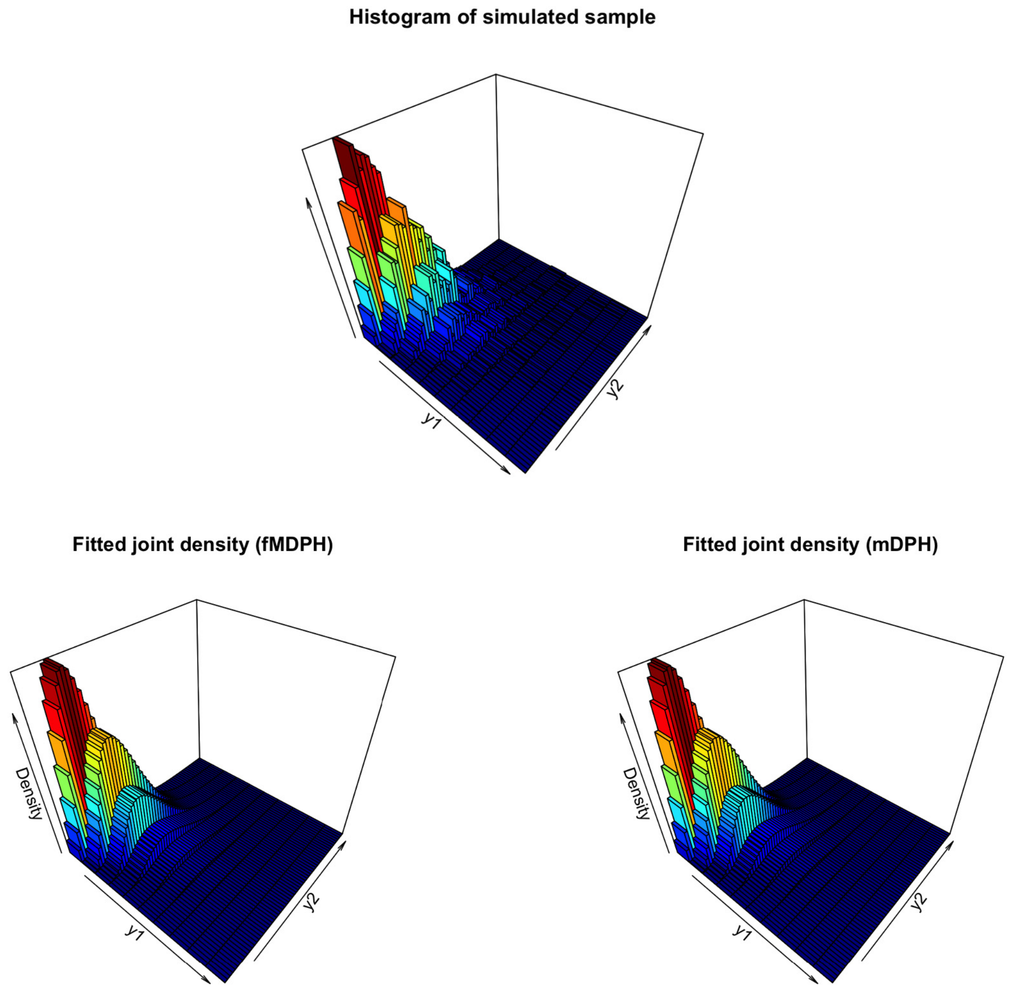
For a sample of size N, we simply sum over the different sample values $\boldsymbol{y}_i$, $i = 1, \ldots, N$, on the formulas above.

The M-step follows easily by applying a Lagrange multiplier argument. We provide the detailed routine in Algorithm 2.

Example 3.12 *(Synthetic data).* We generated an i.i.d. sample of size 5000 of a bivariate discrete random vector with Negative Binomial margins and Gaussian copula. More specifically, the first margin has a dispersion parameter of 3 and a success probability of 0.3, the second margin has a dispersion parameter of 4 with a success probability of 0.7, and the copula has a correlation parameter of 0.5. Then we fitted an fMDPH and an mDPH model to this data set. The aim is to show that the two classes of multivariate DPH distributions introduced before can successfully recover this type of dependence.

For the fMDPH case, we selected a model with $p_1 = p_2 = 6$, and for the mDPH one, we chose a distribution with $p = 6$. We ran both algorithms with 2000 iterations, which was chosen such that changes on the log-likelihoods of both models became negligible, leading to computational times[3] of 8.60 mins for the mDPH model and 19.10 mins fMDPH one. Fig. 3.1 shows that both fitted joint densities successfully recover the shape of the histogram of the simulated sample. This is further supported by Fig. 3.2, where we observe that the

[3] All computations were done in a MacBook Pro Laptop with M1 Pro processor and 32 GB of RAM.


Histogram of simulated sample

Fitted joint density (fMDPH)

Fitted joint density (mDPH)

Fig. 3.1. Histogram of simulated sample versus joint densities of the fitted fMDPH and mDPH models.

marginal behavior is described adequately. Moreover, we have that the sample correlation of the generated sample is 0.4789, which is well approximated by the fMDPH and mDPH models with values of 0.4353 and 0.4379, respectively. Finally, we note that the loglikelihoods of both models (-24,630.56 for the fMDPH model and -24,625.87 for the mDPH model) are above the loglikelihood computed employing the original model (-24,708.38).

Interestingly, the simpler model (recall that mDPH $\subset$ fMDPH) is performing slightly better in loglikelihood, and with fewer overall parameters (114 for the fMDPH model and 78 for the mDPH one). This is because their theoretical equivalences are shown through dimension-augmentation arguments, effectively making the fitted mDPH have a much larger fMDPH equivalent. Moreover, note that the AIC (Akaike's Information Criterion) of the mDPH model is 49,407.74, which outperforms the original model with a corresponding value of 49,426.76. Nevertheless, this is not the case for the fMDPH model with an AIC of 49,489.12. This is an example of simpler models behaving parsimoniously if encoded correctly.

4. Discrete phase-type mixture-of-experts models

The mixture-of-experts approach is a popular method for creating more flexible and powerful regression models. This approach involves combining multiple individual regression models, or "experts" in a way that allows the model to make more accurate predictions. In this section, we propose a new variant of the mixture-of-experts approach that incorporates discrete univariate and multivariate phase-type distributions as defined in the previous sections. This allows the model to more accurately capture the complex, non-linear relationships between the input and output variables, leading to a more flexible regression estimation.

4.1. Univariate model

In the discrete case, we may follow the construction approach from Bladt and Yslas (2022b) to define the PH-MoE class of regression models. More specifically, assume that we have a vector $\boldsymbol{X} \in \mathbb{R}^h$ corresponding to covariate information of the random variable Y, then we endow the underlying Markov chain Z with the initial probabilities depending on the covariate information. Here, we consider the softmax parametrization, which neural networks have popularized.

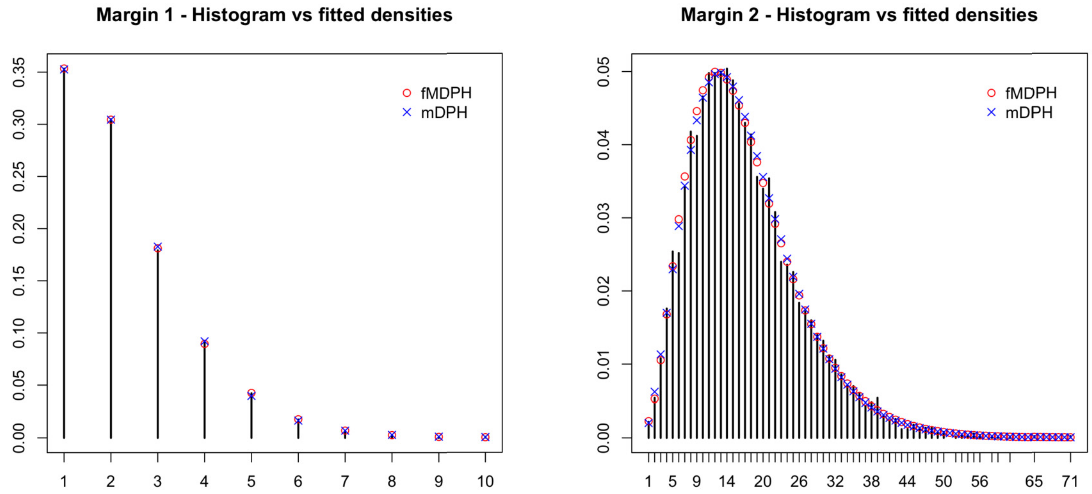


Fig. 3.2. Histograms of margins of the simulated sample versus marginal densities of the fitted fMDPH and mDPH models.

Definition 4.1. We say that the DPH-MoE model with initial probabilities $\boldsymbol{\pi}(\boldsymbol{X};\boldsymbol{\alpha}) = (\pi_k(\boldsymbol{X};\boldsymbol{\alpha}))_{k=1,\dots,p}$ given by

$$\pi_k(\boldsymbol{X};\boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_k)}{\sum_{s=1}^{p}\exp(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\alpha}_s)}\,, \quad k = 1,\dots,p\,, \tag{4.1}$$

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

It is worth remarking that, in many cases, univariate loss modeling can be more effective than multivariate loss modeling because it allows for a more focused and detailed analysis of the variable in question. This can be particularly useful when the goal is to understand the underlying factors contributing to the loss of a particular variable and when the relationships between different variables are not well understood.

Univariate loss modeling may also be more computationally efficient than multivariate loss modeling since it involves dealing with a smaller amount of data. This can be important when the data is very large or when a real-time analysis is required.

4.2. Multivariate model

While univariate loss modeling can be useful for understanding the loss of a single variable, multivariate modeling offers several advantages. For example, multivariate modeling allows for a more comprehensive analysis of the relationships between different variables, which can provide valuable insights into the underlying factors that drive the system. This can be particularly useful in situations where the relationships between variables are not well understood, or where the effects of changes in one variable on the others are unclear.

In addition, multivariate modeling can be more effective at capturing the complex, non-linear relationships that often exist between input variables and output variables by "learning from the other dimensions", see for instance Frees et al. (2010) (and also Frees (2003) for a credibility approach). This can lead to more accurate estimation and better decision-making, which can be critical in various applications.

The definitions of the fMDPH-MoE and mDPH-MoE models are similar to the univariate case, and they consist of considering vectors of initial probabilities depending on the covariate information. In the present case, we again only consider the softmax parametrization.

Remark 4.1 *(Interpretation of the multivariate phase-type MoE models).* The natural interpretation that follows directly from combining the mixture-of-experts and the phase-type concepts is as follows. Risk classes are chosen according to a mixture specification governed by the initial vector, and thereafter, the number of claims follows a multi-state model. Each dimension of the claim count vector collects rewards on specific states of the process until absorption. Thus, the claim evolution is highly dependent on the multinomial parameters, which for high matrix dimensions can become somewhat ambiguous, creating risk classes that are, in a sense, arbitrarily constructed. This problem, although still less pronounced than for Neural Networks, is present in all mixture-of-experts specifications, and a statistical workaround (which sadly does not make the model any more interpretable) would be to introduce regularization.

However, we want to stress that by adding matrix parameters, we aim at making our models much more flexible and competitive for small p, compared to what pure mixture-of-experts approaches provide. See also Tables 3 and 4 in Bladt and Yslas (2022b), for empirical evidence of such dimensionality reduction, in the case of continuous responses.

4.3. Denseness

The denseness of probability distributions is an important property that translates into flexible and powerful statistical modeling. A class of distributions is, loosely speaking, said to be mathematically dense if it can approximate arbitrarily closely other distributions. This property allows a wide range of modeling techniques, including the use of mixture models, convolution, and other advanced techniques.

Mathematically, dense distributions are particularly useful in situations where the data is complex and non-linear, as they allow for the creation of more flexible and accurate models. They can also be used when the data is noisy or incomplete, as there is less focus on the particular structure of any given distribution in the class.

We now proceed to state some definitions and regularity conditions that allow our models to be framed into a correct mathematical concept of denseness, which is not only applicable in the response domain, but also uniformly over regression models, that is, taking the feature space into account.

Definition 4.2. Let $\mathcal{A}$ be the set of possible values of the covariates $\boldsymbol{X}$. A class of regression models $\mathcal{C}(\mathcal{A})$ in $\mathcal{A}$ is the set of conditional distributions given the covariates, that is, each element of $\mathcal{C}(\mathcal{A})$ is a set of probability distributions $\{F(\cdot \mid \boldsymbol{X}=\boldsymbol{x})\,;\, \boldsymbol{x} \in \mathcal{A}\}$.

Definition 4.3. Given a class of regression models $\mathcal{C}_1(\mathcal{A})$, we say that the class of regression models $\mathcal{C}_2(\mathcal{A})$ is dense (uniformly dense) in $\mathcal{C}_1(\mathcal{A})$ if, for each element in $\mathcal{C}_1(\mathcal{A})$, there exists a sequence of regression models in $\mathcal{C}_2(\mathcal{A})$ such that all the associated conditional distributions converge weakly (uniformly weakly) for each $\boldsymbol{x} \in \mathcal{A}$.

Definition 4.4. We say that a feature space $\mathcal{A}$ is regular if it is of the form $\mathcal{A} = \{1\} \times [a,b]^{h-1}$, $a, b \in \mathbb{R}$. In other words, the covariates contain an intercept and are otherwise contained in a hypercube.

Condition 4.5. *A class of regression models $\mathcal{C}(\mathcal{A})$ is said to satisfy the tightness and Lipschitz conditions on $\mathcal{A}$ if for each element $\{F(\cdot \mid \boldsymbol{x})\,;\, \boldsymbol{x} \in \mathcal{A}\}$ in $\mathcal{C}(\mathcal{A})$,*

$$\{F(\cdot \mid \boldsymbol{X}=\boldsymbol{x}); \boldsymbol{x} \in \mathcal{A}\}$$

is a tight family of distributions, and for each $\boldsymbol{y}$, the function

$$\boldsymbol{x} \mapsto F(\boldsymbol{y} \mid \boldsymbol{X}=\boldsymbol{x})$$

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

Proposition 4.6 *(Denseness). Let $\mathcal{C}(\mathcal{A})$ be a class of multivariate frequency regression models satisfying the tightness and Lipschitz conditions on a regular $\mathcal{A}$. Then, the class of mDPH-MoE regression models is uniformly dense in $\mathcal{C}(\mathcal{A})$.*

Proof. The result is a direct consequence of Theorem 3.3 in Fung et al. (2019) by noticing that multivariate LRMoE models with DPH experts of the form given in the proof of Theorem 2.2 are particular instances of mDPH-MoE models. □

Remark 4.2. Given that mDPH ⊂ fMDPH we have that the fMDPH-MoE class also satisfies the denseness property above. Moreover, observe that for $d=1$ the mDPH class retrieves the DPH class. Thus, the set of DPH-MoE regression models satisfies the denseness property for univariate frequency regression models.

4.4. Estimation

Deriving an estimation procedure for the DPH-MoE model via a modified EM algorithm is an important step in developing this powerful and flexible regression technique. The EM algorithm is a widely-used iterative method for estimating the parameters of a statistical model, and has been successfully applied to a variety of different types of models.

By developing a modified EM algorithm specifically for the DPH-MoE model, we may take advantage of the existing statistical methods for the estimation of multinomial models (effectively, a zero-layer neural network), and combine them with the phase-type-specific part of the algorithm. The result is a fast and robust procedure that can be easily implemented in high-level programming languages (though all our examples in this paper were coded in c++).

The derivation of the multivariate classes' algorithms is omitted in this paper for brevity, as they are similar to the estimation procedures previously derived without covariate information. However, the full details of these algorithms can be found in the appendix of this paper. Specifically, refer to Algorithms 4 and 5 for algorithms for the fMDPH-MoE and mDPH-MoE classes, respectively.

Consider an i.i.d. sample of size N from a DPH($\boldsymbol{\pi}(\boldsymbol{X}), \boldsymbol{T}$) model with absorption times $\boldsymbol{y} = (y_1, \ldots, y_N)$ and corresponding paired covariate information $\bar{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$. Now, let $B_k(\boldsymbol{x})$ be the number of Markov chains with initial distribution $\boldsymbol{\pi}(\boldsymbol{x})$ starting in state k, $N_{kl}(\boldsymbol{x})$ the number of transitions from state k to state l conditional on $\boldsymbol{\pi}(\boldsymbol{x})$, and $N_k(\boldsymbol{x})$ the number of chains that exit from state k to the absorbing state given $\boldsymbol{\pi}(\boldsymbol{x})$. Then, the complete likelihood function for this sample is given by

$$\begin{aligned} L_c\left(\boldsymbol{\pi}, \boldsymbol{T}; \boldsymbol{y}, \bar{\boldsymbol{x}}\right) &= \prod_{i=1}^{N} \prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{B_k(\boldsymbol{x}_i)} \prod_{k,l=1}^{p} t_{kl}^{N_{kl}(\boldsymbol{x}_i)} \prod_{k=1}^{p} t_k^{N_k(\boldsymbol{x}_i)} \\ &= \left(\prod_{i=1}^{N} \prod_{k=1}^{p} \pi_i(\boldsymbol{x}_i)^{B_k(\boldsymbol{x}_i)} \right) \prod_{k,l=1}^{p} t_{kl}^{N_{kl}} \prod_{k=1}^{p} t_k^{N_k}, \end{aligned}$$

where

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

However, making use of the complete likelihood above, we can employ an expectation-maximization (EM) algorithm to compute the MLE iteratively. The derivation of the E- and M-steps is similar to the case without covariate information, and the exact formulas are the following:

1) *E-step, conditional expectations:*

$$\mathbb{E}[B_k(\boldsymbol{x}_i) \mid Y = y_i, \boldsymbol{X} = \boldsymbol{x}_i] = \frac{\pi_k(\boldsymbol{x}_i)\boldsymbol{e}_k{}^\top \boldsymbol{T}^{y_i-1}\boldsymbol{t}}{\boldsymbol{\pi}(\boldsymbol{x}_i)\boldsymbol{T}^{y_i-1}\boldsymbol{t}}, \quad i = 1, \ldots, N,$$

$$\mathbb{E}[N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}] = \sum_{i=1}^{N} 1\{y_i \geq 2\} t_{kl} \frac{\boldsymbol{e}_l{}^\top \boldsymbol{K}(y_i; \boldsymbol{\pi}(\boldsymbol{x}_i), \boldsymbol{T})\boldsymbol{e}_k}{\boldsymbol{\pi}(\boldsymbol{x}_i)\boldsymbol{T}^{y_i-1}\boldsymbol{t}},$$

$$\mathbb{E}[N_k \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}] = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\pi}(\boldsymbol{x}_i)\boldsymbol{T}^{y_i-1}\boldsymbol{e}_k}{\boldsymbol{\pi}(\boldsymbol{x}_i)\boldsymbol{T}^{y_i-1}\boldsymbol{t}}.$$

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

$$\hat{t}_{kl} = \frac{\mathbb{E}\left[N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks} \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}\right]}, \quad k, l = 1, \ldots, p,$$

$$\hat{t}_{k} = \frac{\mathbb{E}\left[N_{k} \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks} \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k \mid \boldsymbol{Y} = \boldsymbol{y}, \bar{\boldsymbol{x}}\right]}, \quad k = 1, \ldots, p.$$

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

$$\hat{\boldsymbol{\pi}}(\cdot) = \underset{\boldsymbol{\pi}(\cdot) \in \Delta^{p-1}}{\arg\max} \left(\prod_{i=1}^{N} \prod_{k=1}^{p} \pi_k(\boldsymbol{x}_i)^{\mathbb{E}[B_k(\boldsymbol{x}_i) \mid Y = y_i, \boldsymbol{X} = \boldsymbol{x}_i]} \right),$$

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

The detailed routine can be found in Algorithm 3.

Remark 4.3 *(Data with zeroes).* When dealing with data that contains zeroes, there are at least three possible approaches to using the proposed regression model. The first approach is to simply exclude any data points that contain zeroes, as these points may not provide useful information for the model. This approach can be particularly useful if the proportion of zero-valued data points is relatively small, and if the remaining data is sufficient to accurately model the relationships of interest. In this approach, we may alternatively keep the zeroes and estimate π_{p+1} separately without regression by simply considering the number of zero observations divided by the total number of observations and fit a DPH-MoE model to the remaining positive observation. A similar approach is sometimes used for regression models based on zero-inflated distributions fitted in two steps.

The second approach involves translating the data by one unit, so that all data starts at 1 and then fitting a DPH-MoE model. The interpretability of this approach is not diminished because, by design, multivariate DPH distributions represent the progression of states through time, so the translation simply means starting the clock at a later time.

A third approach is to use a specialized variant of the regression model that is specifically designed to handle data with zeroes. For example, the model could incorporate a zero-inflated component, which allows for the modeling of excess zeros in the data. For instance, assume that $\pi_{p+1}(\boldsymbol{x})$ if of the form $\pi_{p+1}(\boldsymbol{x}) = (1 + \exp(\boldsymbol{x}^\top \boldsymbol{\gamma}))^{-1}$. Then, the procedure splits into fitting a logistic regression to model the probability of obtaining zero or not and fitting a DPH-MoE to the positive data.

Remark 4.4 *(Data with exposures).* The classic approach to dealing with exposure for claim counts is through a log-offset in the regression formula. This results in a multiplicative effect of exposure on the claim frequency whenever a log-link function is used on a Poisson distribution. The multiplicative structure is appealing for practitioners, and it holds quite generally, that is, irrespective of the form of the predictive model: linear, trees, Neural Networks, among others.

In our setting, however, we have moved away from the Poisson specification so that we deal with exposure in a different manner: as a covariate. This approach can improve the accuracy of the predictions, as it allows the model to take into account the effects of the exposure on the outcome of interest. However, this approach does have a small downside, in that it can make the model less interpretable. This can make it more difficult to communicate the results to decision-makers.

Despite this downside, incorporating exposure as a covariate is still a valuable approach, as it does not compromise the statistical performance of the model. The improved accuracy of the predictions can be valuable in a wide range of applications, not to claim counts, and can help to make better-informed decisions. In many cases, the benefits of improved accuracy will outweigh the slight loss of interpretability, making this an effective approach for improving the performance of the regression model.

5. Case study: LGPIF data set

This section considers the Wisconsin Local Government Property Insurance Fund (LGPIF) data set,[4] which was previously described in detail and analyzed in Frees et al. (2016), and also recently considered in Yang and Shi (2019); Jeong et al. (2023). The fund was established to provide property insurance for local government entities and is administered by a government office. Properties covered under this fund include government buildings, vehicles, and equipment. These data provide a good example of a typical multi-line insurance company

[4] The data set can be found at https://sites.google.com/a/wisc.edu/local-government-property-insurance-fund.

Table 5.1
Summary of LGPIF data per entity type.

	City	County	Misc	School	Town
BC	794	328	612	1599	981
IM	784	328	320	1195	775

encountered in practice. The data span the period 2006-2010, as a training set (which is the part used in the in-sample analysis below), and the year 2011 is also available, as a testing set (we use this part for prediction).

Our analysis focuses on the univariate and multivariate settings, allowing us to investigate the data from different perspectives. We showcase the two multivariate regression models that we have introduced in the paper, each of which offers insights into the data and the modeling methodology. Through the analysis, we use the LGPIF data set to illustrate the favorable performance of discrete phase-type regression modeling. Since the data is large and significantly zero-inflated, we decide to use the second method (translation) of Remark 4.3 to deal with zeros.

Claims in the dataset have six different codes, depending on the applicable coverage type. The corresponding codes are BC, IM, C, P, N, and O, and they are all standard codes used in the insurance industry to represent different types of coverage. BC stands for buildings and contents coverage, which provides insurance for buildings and the properties within. IM is an abbreviation for "inland marine" and is used as the coverage code for equipment coverage, which originally belonged to contractors. C represents coverage for the impact of a vehicle with an object, the impact of a vehicle with an attached vehicle, or overturn of a vehicle. P stands for coverage for direct and accidental loss or damage to a motor vehicle, including breakage of glass, loss caused by missiles, falling objects, fire, theft, explosion, earthquake, windstorm, hail, water, flood, malicious mischief or vandalism, riot or civil commotion, or colliding with a bird or animal. N is used as an indication that the coverage is for vehicles of the current model year or 1-2 years prior to the current model year. O is used to indicate that the coverage is for vehicles three or more years prior to the current model year.

We focus our analysis on BC and IM, since they are the ones that lead to the majority of claims in the dataset, roughly 65%. The number of observations within these two coverages is then $10,282$. Regarding explanatory variables, for each coverage, we choose the following: coverage amount (in log-scale, respective means 2.12, -1.34 and variances 4, 3.51), an indicator for no claims in the previous year (respectively, 3796 and 2239 zero values and 1864, 2383 unit values), entity type (City, County, Misc, School, Town), and deductible level (in log-scale, respective means 7.15, 6.54 and variances 1.38, 0.42). See Table 5.1 for the number of claims per entity type. In Frees et al. (2016), BC was also specially highlighted above the rest of the coverages, and the same explanatory variables were used.

5.1. Building and content and contractor's equipment frequency modeling

We initially focus on the univariate task of modeling the frequency of building and content (BC) and compare it with the reference models presented in Frees et al. (2016). Table 5.2 shows the results for the expected versus empirical counts for varying claim numbers. We observe that the DPH-MoE models behave favorably across the entire distribution, which is also supported in Table 5.3 using a Chi-square goodness of fit statistic and Table 5.4[5] in terms of log-likelihood and AIC. In particular, we see that despite the high number of model parameters, the additional information which is extracted from the covariates is still worthwhile, with a matrix dimension of 7 being preferred. Moreover, to assess the predictive performance of the fitted models, we used the testing set and compared the mean predictions with the observed frequency via the root-mean-square error (RMSE) measure, which can be found in Table 5.4. We observe that the DPH-MoE specifications outperform the other models here as well.

An analogous analysis for IM is provided in Tables 5.5, 5.6 and 5.7,[6] where the AIC prefers a dimension of 3. Note that the AIC is a measure of the relative quality of statistical models for a given set of data. It balances the fit of the model (how well the model explains the data) with the complexity of the model (the number of parameters the model uses). It penalizes models that use a large number of parameters, as these models may overfit the data and may not generalize well to new data. The AIC is presently being used to compare different types of models despite the fact that it is known to penalize matrix methods more harshly than other models, since it is a standard criterion for comparing the other models and because the data include covariate information.

5.2. Joint modeling of BC and IM frequencies

In Frees et al. (2016), joint modeling for claim counts is done using a Zero-one-inflated NB for BC and an NB for IM, bonded together with a Gaussian copula. This procedure is standard and widespread in the industry for both discrete (counts) and continuous (severity) response variables. Here, we apply an fMDPH-MoE with dimensions 7 and 3 (which are motivated by the univariate modeling framework) and an mDPH-MoE of dimension 7 (performing the best between similar dimensions). The results of our models are presented in Table 5.8 in terms of expected versus observed counts, and in Table 5.9[7] in terms of likelihood and AIC. Our models show favorable performance compared to the standard method. Interestingly, the smaller dimensional mDPH-MoE performs better than the fMDPH-MoE, indicating that parsimonious solutions may be found with simpler models. Therefore, we recommend trying both methodologies before deciding on one of them.

[5] For the DPH-MoE specifications, we performed 500 iterations of the EM algorithm. The other models were computed using the implementations available at https://sites.google.com/a/wisc.edu/local-government-property-insurance-fund.

[6] 300 EM-steps were employed for both DPH-MoE models.

[7] 350 iterations of the EM algorithms were employed for both multivariate DPH-MoE models. The computational time of the copula model is not displayed, given that it is a two-step fitting procedure.

Table 5.2
Comparison between empirical values and expected values for building and contents frequency.

	Empirical	DPH-MoE (6)	DPH-MoE (7)	DPH-MoE (8)	ZeroinflPoisson	ZeroonePoisson	Poisson	NB	ZeroinflNB	ZerooneNB
0	3976.00	3973.37	3973.94	3974.72	4038.13	3975.40	3709.99	4075.37	4093.70	3996.89
1	997	997.65	996.30	1003.37	754.38	1024.22	1012.27	809.08	791.43	1003.23
2	333	331.19	334.80	329.50	355.93	276.08	417.33	313.36	314.62	280.57
3	136	137.20	135.53	135.31	187.90	146.96	202.29	155.74	157.28	136.75
4	76	67.35	66.14	66.51	106.78	82.05	106.87	88.87	89.61	75.82
5	31	38.20	37.79	37.43	63.84	48.43	60.16	55.48	55.70	46.02
6	19	24.17	24.14	23.32	39.85	30.21	36.54	36.92	36.84	29.85
7	19	16.41	16.50	15.72	26.08	19.85	24.26	25.77	25.55	20.38
8	16	11.62	11.74	11.24	18.02	13.67	17.44	18.66	18.40	14.48
9	5	8.45	8.55	8.38	13.17	9.81	13.22	13.93	13.65	10.63
10	7	6.25	6.34	6.44	10.09	7.27	10.30	10.66	10.39	8.02
11	2	4.69	4.77	5.04	8.01	5.51	8.12	8.34	8.08	6.18
12	4	3.58	3.64	4.01	6.51	4.22	6.43	6.64	6.41	4.86
13	5	2.78	2.82	3.23	5.36	3.25	5.09	5.37	5.16	3.88
14	5	2.20	2.23	2.62	4.44	2.50	4.02	4.40	4.21	3.14
15	2	1.78	1.80	2.15	3.69	1.93	3.18	3.65	3.48	2.57
16	4	1.47	1.48	1.78	3.06	1.48	2.52	3.07	2.91	2.13
17	3	1.24	1.25	1.49	2.53	1.13	2.00	2.60	2.46	1.78
18	1	1.07	1.08	1.27	2.08	0.87	1.60	2.22	2.10	1.50
≥ 19	19	29.15	28.97	26.11	10.17	5.17	16.37	19.88	18.00	11.35

Table 5.3
Goodness of fit statistics for BC.

DPH-MoE (6)	DPH-MoE (7)	DPH-MoE (8)	ZeroinflPoisson	ZeroonePoisson	Poisson	NB	ZeroinflNB	ZerooneNB
24.51	24.18	18.69	154.57	77.06	105.20	88.09	98.40	34.52

Table 5.4
Loglikelihoods, number of parameters, AIC, computational times, and RMSE for BC.

	DPH-MoE (6)	DPH-MoE (7)	DPH-MoE (8)	ZeroinflPoisson	ZeroonePoisson	Poisson	NB	ZeroinflNB	ZerooneNB
Loglik	-4777.52	-4745.51	-4,736.11	-7,013.58	-5,179.37	-7,838.56	-5,102.62	-7,013.58	-4,958.29
NumPar	76	97	120	13	17	9	10	13	18
AIC	9,707.03	9,685.02	9,712.22	14,053.16	10,392.74	15,695.11	10,225.25	14,053.16	9,952.58
Time	8.22 mins	11.38 mins	15.13 min	0.38 secs	19.23 secs	0.02 secs	0.42 secs	0.37 secs	5.93 mins
RMSE	6.2450	6.2872	5.8695	6.7616	6.9701	6.7212	6.8581	6.8635	6.9722

Table 5.5
Comparison between empirical values and expected values for contractor's equipment frequency.

	Empirical	DPH-MoE (3)	DPH-MoE (4)	ZeroinflPoisson	ZeroonePoisson	Poisson	NB	ZeroinflNB	ZerooneNB
0	4386	4384.33	4385.05	4381.66	4383.58	4351.36	4383.53	4384.72	4384.72
1	182	187.98	184.07	189.28	184.50	233.11	191.21	188.28	188.28
2	40	31.75	38.12	35.99	37.17	29.99	31.19	32.56	32.56
3	6	10.57	7.65	10.38	11.43	5.79	9.31	9.72	9.72
4	4	4.28	3.78	3.24	3.66	1.31	3.56	3.67	3.67
5	2	1.79	1.77	1.01	1.16	0.32	1.55	1.56	1.56
6	2	0.75	0.83	0.31	0.36	0.08	0.74	0.72	0.72
≥ 7	0	0.55	0.73	0.13	0.15	0.02	0.89	0.76	0.76

Table 5.6
Goodness of fit statistics for IM.

DPH-MoE (3)	DPH-MoE (4)	ZeroinflPoisson	ZeroonePoisson	Poisson	NB	ZeroinflNB	ZerooneNB
6.96	2.89	13.05	11.21	74.79	7.34	6.50	6.50

Table 5.7
Loglikelihoods, number of parameters, AIC, computational times, and RMSE for IM.

	DPH-MoE (3)	DPH-MoE (4)	ZeroinflPoisson	ZeroonePoisson	Poisson	NB	ZeroinflNB	ZerooneNB
Loglik	-876.26	-862.91	-894.30	-889.61	-919.55	-895.63	-894.30	-889.38
NumPar	25	40	13	17	9	10	13	18
AIC	1,802.12	1,805.82	1,814.60	1,813.22	1,857.10	1,811.26	1,814.60	1,814.76
Time	31.50 secs	1.33 mins	0.12 secs	4.67 secs	0.02 secs	0.27 secs	0.22 secs	9.33 secs
RMSE	0.4843	0.4809	0.4956	0.4933	0.4978	0.4998	0.4984	0.4984

Table 5.8
Comparison between empirical values and expected values for multivariate models.

	Empirical									fMDPH (7,3)									mDPH (7)								
	0	1	2	3	4	5	6	≥ 7	Sum	0	1	2	3	4	5	6	≥ 7	Sum	0	1	2	3	4	5	6	≥ 7	Sum
0	3022	63	7	1	1	0	0	0	3094	3002.65	86.57	10.03	3.17	1.11	0.39	0.14	0.05	3104.11	3012.52	65.74	5.62	0.87	0.22	0.08	0.03	0.01	3085.08
1	810	52	4	0	0	0	0	0	866	822.13	31.90	5.71	1.91	0.67	0.23	0.08	0.03	862.66	820.90	46.64	6.16	1.15	0.30	0.10	0.04	0.02	875.32
2	281	25	7	0	0	0	0	0	313	287.55	16.92	4.13	1.42	0.50	0.17	0.06	0.02	310.78	282.09	28.46	4.36	0.89	0.24	0.08	0.04	0.02	316.18
3	114	12	5	1	0	0	0	0	132	117.02	10.27	2.94	1.02	0.36	0.13	0.04	0.02	131.81	112.87	14.60	2.75	0.63	0.18	0.07	0.03	0.01	131.14
4	68	6	2	0	0	0	0	0	76	54.59	6.75	2.10	0.73	0.26	0.09	0.03	0.01	64.56	53.62	8.24	1.82	0.45	0.14	0.05	0.02	0.01	64.35
5	25	4	2	0	0	0	0	0	31	28.70	4.62	1.51	0.53	0.18	0.06	0.02	0.01	35.63	29.57	5.11	1.24	0.33	0.11	0.05	0.02	0.01	36.44
6	12	4	1	0	0	0	0	0	17	16.50	3.25	1.09	0.38	0.13	0.05	0.02	0.01	21.42	18.07	3.34	0.86	0.25	0.09	0.04	0.02	0.01	22.69
7	10	4	0	0	1	0	1	0	16	10.09	2.33	0.79	0.28	0.10	0.03	0.01	0.00	13.64	11.78	2.25	0.62	0.20	0.08	0.03	0.02	0.01	14.99
8	13	0	1	0	0	0	0	0	14	6.45	1.71	0.59	0.21	0.07	0.03	0.01	0.00	9.07	8.00	1.56	0.46	0.17	0.07	0.03	0.02	0.01	10.32
9	4	0	0	0	1	0	0	0	5	4.27	1.29	0.45	0.16	0.06	0.02	0.01	0.00	6.25	5.61	1.12	0.36	0.14	0.06	0.03	0.01	0.01	7.34
10	2	3	2	0	0	0	0	0	7	2.93	1.00	0.35	0.12	0.04	0.02	0.01	0.00	4.47	4.05	0.83	0.29	0.13	0.06	0.03	0.01	0.01	5.40
11	2	0	0	0	0	0	0	0	2	2.09	0.81	0.29	0.10	0.04	0.01	0.00	0.00	3.33	2.99	0.63	0.25	0.12	0.06	0.03	0.01	0.01	4.09
12	2	1	1	0	0	0	0	0	4	1.56	0.67	0.24	0.08	0.03	0.01	0.00	0.00	2.59	2.26	0.50	0.22	0.11	0.05	0.03	0.01	0.01	3.18
13	4	0	1	0	0	0	0	0	5	1.21	0.58	0.21	0.07	0.03	0.01	0.00	0.00	2.11	1.74	0.41	0.20	0.10	0.05	0.03	0.01	0.01	2.55
14	1	1	1	1	0	1	0	0	5	0.99	0.51	0.18	0.06	0.02	0.01	0.00	0.00	1.78	1.37	0.35	0.18	0.10	0.05	0.02	0.01	0.01	2.09
15	1	0	0	0	1	0	0	0	2	0.84	0.46	0.17	0.06	0.02	0.01	0.00	0.00	1.56	1.10	0.30	0.17	0.09	0.05	0.02	0.01	0.01	1.76
16	2	1	1	0	0	0	0	0	4	0.74	0.43	0.15	0.05	0.02	0.01	0.00	0.00	1.41	0.89	0.27	0.16	0.09	0.05	0.02	0.01	0.01	1.50
17	1	0	1	0	0	1	0	0	3	0.68	0.40	0.14	0.05	0.02	0.01	0.00	0.00	1.30	0.73	0.24	0.16	0.09	0.05	0.02	0.01	0.01	1.31
18	0	0	1	0	0	0	0	0	1	0.62	0.38	0.14	0.05	0.02	0.01	0.00	0.00	1.21	0.61	0.22	0.15	0.09	0.04	0.02	0.01	0.01	1.16
≥ 19	6	6	3	3	0	0	1	0	19	18.28	11.50	4.14	1.46	0.51	0.18	0.06	0.02	36.16	7.65	7.22	6.33	3.74	1.94	0.96	0.47	0.23	28.55
Sum	4380	182	40	6	4	2	2	0		4379.89	182.35	35.35	11.91	4.19	1.48	0.49	0.17		4378.42	188.03	32.36	9.74	3.89	1.74	0.81	0.44	

Table 5.9
Loglikelihoods, number of parameters, and AICs for multivariate models.

	fMDPH (7,3)	mDPH (7)	Gaussian Copula
Loglik	-5,149.94	-5,116.88	-5,493.03
NumPar	145	164	29
AIC	10,589.87	10,561.75	11,044.07
Time	8.01 mins	8.11 mins	-

6. Conclusion

In this paper, we proposed a novel approach for modeling loss frequency using mixture-of-experts specifications on discrete-phase type distributions. Our approach is inspired by machine learning and allows for fast and effective modeling of complex data. We applied the proposed methodology to a real-life data set, the Wisconsin Local Government Property Insurance Fund (LGPIF) data set, and compared it to existing methods, showing that it provides more accurate estimates. We demonstrated that our approach is a more effective solution (both in-sample and out-of-sample) for modeling loss frequency in non-standard situations and is a versatile and practical tool for analyzing a wide range of data.

Declaration of competing interest

There is no competing interest.

Data availability

Data is publicly available by a third party. The source is specified in the text.

Appendix A. Comprehensive list of algorithms

Algorithm 1 EM algorithm for fMDPH.

Input: *Data points* $\overline{\boldsymbol{y}} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\}$ *in* $\mathbb{N}^2$*, with* $\boldsymbol{y}_i = (y_i^{(1)}, y_i^{(2)})$*,* $i = 1, \dots, N$*,and initial parameters* $(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22})$.

1) *E-step:* compute the conditional expectations

$$\mathbb{E}[B_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}] = \sum_{i=1}^{N} \frac{\eta_k \boldsymbol{e}_k{}^{\top} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, \quad k = 1, \dots, p_1,$$

$$\mathbb{E}[N_{kl} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}]$$

$$\begin{cases} \sum_{i=1}^{N} 1\{y_i^{(1)} \geq 2\} t_{kl} \sum_{m=0}^{y_i^{(1)}-2} \frac{\boldsymbol{e}_l^{\top} \boldsymbol{T}_{11}^{y_i^{(1)}-m-2} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2 \boldsymbol{\eta} \boldsymbol{T}_{11}^{m} \boldsymbol{e}_k}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, & k, l = 1, \dots, p_1, \\ \sum_{i=1}^{N} t_{kl} \frac{\boldsymbol{e}_{l-p_1}^{\top} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2 \boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{e}_k}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, & k \leq p_1, l = p_1 + 1, \dots, p, \\ \sum_{i=1}^{N} 1\{y_i^{(2)} \geq 2\} t_{kl} \sum_{m=0}^{y_i^{(2)}-2} \frac{\boldsymbol{e}_{l-p_1}^{\top} \boldsymbol{T}_{22}^{y_i^{(2)}-m-2} \boldsymbol{t}_2 \boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{m} \boldsymbol{e}_{k-p_1}}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, & k, l = p_1 + 1, \dots, p, \end{cases}$$

$$\mathbb{E}[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}] = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{e}_{k-p_1}}{\boldsymbol{\eta} \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, \quad k = p_1 + 1, \dots, p.$$

2) *M-step:* Let

$$\hat{\eta}_k = \frac{\mathbb{E}\left[B_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}{N}, \quad k = 1, \dots, p_1,$$

$$\hat{t}_{kl} = \frac{\mathbb{E}\left[N_{kl} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}, \quad k = 1, \dots, p_1, \ l = 1, \dots, p,$$

$$\hat{t}_{kl} = \frac{\mathbb{E}\left[N_{kl} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}{\sum_{s=p_1+1}^{p} \mathbb{E}\left[N_{ks} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right] + \mathbb{E}\left[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}, \quad k, l = p_1 + 1, \dots, p,$$

$$\hat{t}_k = \frac{\mathbb{E}\left[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}{\sum_{s=p_1+1}^{p} \mathbb{E}\left[N_{ks} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right] + \mathbb{E}\left[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}, \quad k = p_1 + 1, \dots, p,$$

$$\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \dots, \hat{\eta}_{p_1}), \ \hat{\boldsymbol{T}}_{11} = (\hat{t}_{kl})_{k,l=1,\dots,p_1},$$

$$\hat{\boldsymbol{T}}_{12} = (\hat{t}_{k(l+p_1)})_{k=1,\dots,p_1, l=1,\dots,p_2}, \ \hat{\boldsymbol{T}}_{22} = (\hat{t}_{(k+p_1)(l+p_1)})_{k,l=1,\dots,p_2},$$

3) Update the current parameters to $(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22}) = (\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{T}}_{11}, \hat{\boldsymbol{T}}_{12}, \hat{\boldsymbol{T}}_{22})$. Return to step 1 unless a stopping rule is satisfied.

Output: *Fitted representation* $(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22})$.

Algorithm 2 EM algorithm for mDPH.

```
Input: Data points $\overline{\boldsymbol{y}} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\}$ in $\mathbb{N}^d$, with $\boldsymbol{y}_i = (y_i^{(1)}, \dots, y_i^{(d)})$, $i = 1, \dots, N$, and initial parameters $(\boldsymbol{\pi}, \mathcal{T})$.
  1) E-step: Compute the conditional expectations
       $\mathbb{E}[B_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}] = d \sum_{i=1}^{N} \frac{\pi_k \prod_{j=1}^{d} \boldsymbol{e}_k^\top \boldsymbol{T}_j^{y_i^{(j)}-1} \boldsymbol{t}_j}{\sum_{s=1}^{p} \pi_s \prod_{j=1}^{d} \boldsymbol{e}_s^\top \boldsymbol{T}_q^{y_i^{(j)}-1} \boldsymbol{t}_j}$ ,
       $\mathbb{E}[N_{kl}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}] = \sum_{i=1}^{N} 1\{y_i^{(j)} \geq 2\} t_{kl}^{(j)} \frac{\sum_{s=1}^{p} \pi_s \prod_{q \neq j} \boldsymbol{e}_s^\top \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q \boldsymbol{e}_l^\top \boldsymbol{K}(y_i^{(j)}; \boldsymbol{e}_s^\top, \boldsymbol{T}_j) \boldsymbol{e}_k}{\sum_{s=1}^{p} \pi_s \prod_{q=1}^{d} \boldsymbol{e}_s^\top \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q}$ ,
       $\mathbb{E}[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}] = \sum_{i=1}^{N} t_k^{(j)} \frac{\sum_{s=1}^{p} \pi_s \prod_{q \neq j} \boldsymbol{e}_s^\top \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q \boldsymbol{e}_s^\top \boldsymbol{T}_j^{y_i^{(j)}-1} \boldsymbol{e}_k}{\sum_{s=1}^{p} \pi_s \prod_{q=1}^{d} \boldsymbol{e}_s^\top \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q}$ .
  2) M-step: Let
       $\hat{\pi}_k = \frac{\mathbb{E}\left[B_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}{d \cdot N}, \quad k = 1, \dots, p$ ,
       $\hat{t}_{kl}^{(j)} = \frac{\mathbb{E}\left[N_{kl}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right] + \mathbb{E}\left[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}, \quad k, l = 1, \dots, p, \; j = 1, \dots, d$ ,
       $\hat{t}_k^{(j)} = \frac{\mathbb{E}\left[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right] + \mathbb{E}\left[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}\right]}, \quad k = 1, \dots, p, \; j = 1, \dots, d$ .
  3) Update the current parameters to $(\boldsymbol{\pi}, \mathcal{T}) = (\hat{\boldsymbol{\pi}}, \hat{\mathcal{T}})$. Return to step 1 unless a stopping rule is satisfied.
Output: Fitted representation $(\boldsymbol{\pi}, \mathcal{T})$.
```

Algorithm 3 EM algorithm for DPH-MoE (Softmax parametrization).

```
Input: Positive integer data points $\boldsymbol{y} = (y_1, \dots, y_N)$, covariates $\overline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, and initial parameters $(\boldsymbol{\alpha}, \boldsymbol{T})$.
  1) Mixture specification: Set
       $\pi_k(\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha}_k)}{\sum_{s=1}^{p} \exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha}_s)}, \quad i = 1, \dots, N, \; k = 1, \dots, p$ .
  2) E-step: Compute the statistics
       $\mathbb{E}[B_k(\boldsymbol{x}_i) \mid Y = y_i, \boldsymbol{X} = \boldsymbol{x}_i] = \frac{\pi_k(\boldsymbol{x}_i) \boldsymbol{e}_k{}^\top \boldsymbol{T}^{y_i-1} \boldsymbol{t}}{\boldsymbol{\pi}(\boldsymbol{x}_i) \boldsymbol{T}^{y_i-1} \boldsymbol{t}}, \quad i = 1, \dots, N$ ,
       $\mathbb{E}[N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}] = \sum_{i=1}^{N} 1\{y_i \geq 2\} t_{kl} \frac{\boldsymbol{e}_l{}^\top \boldsymbol{K}(y_i; \boldsymbol{\pi}(\boldsymbol{x}_i), \boldsymbol{T}) \boldsymbol{e}_k}{\boldsymbol{\pi}(\boldsymbol{x}_i) \boldsymbol{T}^{y_i-1} \boldsymbol{t}}$ ,
       $\mathbb{E}[N_k \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}] = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\pi}(\boldsymbol{x}_i) \boldsymbol{T}^{y_i-1} \boldsymbol{e}_k}{\boldsymbol{\pi}(\boldsymbol{x}_i) \boldsymbol{T}^{y_i-1} \boldsymbol{t}}$ .
  3) M-step: Let
       $\hat{t}_{kl} = \frac{\mathbb{E}\left[N_{kl} \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks} \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}\right]}, \quad k, l = 1, \dots, p$ ,
       $\hat{t}_k = \frac{\mathbb{E}\left[N_k \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks} \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k \mid \boldsymbol{Y} = \boldsymbol{y}, \overline{\boldsymbol{x}}\right]}, \quad k = 1, \dots, p$ .
  4) R-step: Maximize the weighted multinomial logistic regression
       $\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^{(p \times h)}}{\arg\max} \sum_{i=1}^{N} \sum_{k=1}^{p} \mathbb{E}[B_k(\boldsymbol{x}_i) \mid Y = y_i, \boldsymbol{X} = \boldsymbol{x}_i] \log(\pi_k(\boldsymbol{x}_i; \boldsymbol{\alpha}))$ ,
     and set
       $\hat{\pi}_k(\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) = \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\alpha}}_k)}{\sum_{s=1}^{p} \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\alpha}}_s)}, \quad i = 1, \dots, N, \; k = 1, \dots, p$ .
  5) Update the current parameters to $(\boldsymbol{\alpha}, \boldsymbol{T}) = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{T}})$. Return to step 1 unless a stopping rule is satisfied.
Output: Fitted representation $(\boldsymbol{\alpha}, \boldsymbol{T})$.
```

Algorithm 4 EM algorithm for fMDPH-MoE (Softmax parametrization).

```
Input: Data points $\overline{\boldsymbol{y}} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\}$ in $\mathbb{N}^2$, with $\boldsymbol{y}_i = (y_i^{(1)}, y_i^{(2)})$, $i = 1, \dots, N$, covariates $\overline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, and initial parameters $(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22})$.
1) Mixture specification: Set
    $\eta_k(\boldsymbol{x}_i) = \eta_k(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha}_k)}{\sum_{s=1}^{p_1} \exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha}_s)}, \quad i = 1, \dots, N, \ k = 1, \dots, p_1 .$
2) E-step: Compute the statistics
    $\mathbb{E}[B_k \mid \boldsymbol{Y} = \boldsymbol{y}_i, \boldsymbol{X} = \boldsymbol{x}_i] = \frac{\eta_k(\boldsymbol{x}_i) \boldsymbol{e}_k{}^\top \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}{\boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, \quad i = 1, \dots, N, \ k = 1, \dots, p_1 .$
    $\mathbb{E}[N_{kl} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}]$
    $\begin{cases} \sum_{i=1}^{N} 1\{y_i^{(1)} \geq 2\} t_{kl} \sum_{m=0}^{y_i^{(1)}-2} \frac{\boldsymbol{e}_l^\top \boldsymbol{T}_{11}^{y_i^{(1)}-m-2} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2 \boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{m} \boldsymbol{e}_k}{\boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, & k, l = 1, \dots, p_1, \\ \sum_{i=1}^{N} t_{kl} \frac{\boldsymbol{e}_{l-p_1}^\top \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2 \boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{e}_k}{\boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, & k = 1, \dots, p_1, \ l = p_1 + 1, \dots, p \\ \sum_{i=1}^{N} 1\{y_i^{(2)} \geq 2\} t_{kl} \sum_{m=0}^{y_i^{(2)}-2} \frac{\boldsymbol{e}_{l-p_1}^\top \boldsymbol{T}_{22}^{y_i^{(2)}-m-2} \boldsymbol{t}_2 \boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{m} \boldsymbol{e}_{k-p_1}}{\boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, & k, l = p_1 + 1, \dots, p, \end{cases}$
    $\mathbb{E}[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}] = \sum_{i=1}^{N} t_k \frac{\boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{e}_{k-p_1}}{\boldsymbol{\eta}(\boldsymbol{x}_i) \boldsymbol{T}_{11}^{y_i^{(1)}-1} \boldsymbol{T}_{12} \boldsymbol{T}_{22}^{y_i^{(2)}-1} \boldsymbol{t}_2}, \quad k = p_1 + 1, \dots, p,$
3) M-step: Let
    $\hat{t}_{kl} = \frac{\mathbb{E}\left[N_{kl} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}{\sum_{s=1}^{p} \mathbb{E}\left[N_{ks} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}, \quad k = 1, \dots, p_1, \ l = 1, \dots, p,$
    $\hat{t}_{kl} = \frac{\mathbb{E}\left[N_{kl} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}{\sum_{s=p_1+1}^{p} \mathbb{E}\left[N_{ks} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}, \quad k, l = p_1 + 1, \dots, p,$
    $\hat{t}_k = \frac{\mathbb{E}\left[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}{\sum_{s=p_1+1}^{p} \mathbb{E}\left[N_{ks} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}, \quad k = p_1 + 1, \dots, p,$
4) R-step: Maximize the weighted multinomial logistic regression
    $\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \in \overline{\mathbb{R}}^{(p_1 \times h)}}{\arg\max} \sum_{i=1}^{N} \sum_{k=1}^{p_1} \mathbb{E}[B_k(\boldsymbol{x}_i) \mid \boldsymbol{Y} = \boldsymbol{y}_i, \boldsymbol{X} = \boldsymbol{x}_i] \log(\eta_k(\boldsymbol{x}_i; \boldsymbol{\alpha})),$
   and set
    $\hat{\eta}_k(\boldsymbol{x}_i) = \eta_k(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) = \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\alpha}}_k)}{\sum_{s=1}^{p} \exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\alpha}}_s)}, \quad i = 1, \dots, N, \ k = 1, \dots, p_1 .$
5) Update the current parameters to $(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22}) = (\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{T}}_{11}, \hat{\boldsymbol{T}}_{12}, \hat{\boldsymbol{T}}_{22})$. Return to step 1 unless a stopping rule is satisfied.
Output: Fitted representation $(\boldsymbol{\eta}, \boldsymbol{T}_{11}, \boldsymbol{T}_{12}, \boldsymbol{T}_{22})$.
```

Algorithm 5 EM algorithm for mDPH-MoE (Softmax parametrization).

```
***Input***: Data points $\overline{\boldsymbol{y}} = \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_N\}$ in $\mathbb{N}^d$, with $\boldsymbol{y}_i = (y_i^{(1)}, \dots, y_i^{(d)})$, $i = 1, \dots, N$, covariates $\overline{\boldsymbol{x}} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, and initial parameters $(\boldsymbol{\pi}, \mathcal{T})$.
  1) Mixture specification: Set
       $\pi_k(\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_k)}{\sum_{s=1}^p \exp(\boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\alpha}_s)}, \quad i = 1, \dots, N, \; k = 1, \dots, p.$
  2) E-step: Compute the statistics
       $\mathbb{E}[B_k \mid \boldsymbol{Y} = \boldsymbol{y}_i, \boldsymbol{X} = \boldsymbol{x}_i] = d \frac{\pi_k(\boldsymbol{x}_i) \prod_{j=1}^d \boldsymbol{e}_k^{\top} \boldsymbol{T}_j^{y_i^{(j)}-1} \boldsymbol{t}_j}{\sum_{s=1}^p \pi_s \prod_{j=1}^d \boldsymbol{e}_s^{\top} \boldsymbol{T}_j^{y_i^{(j)}-1} \boldsymbol{t}_j}, \quad i = 1, \dots, N, \; k = 1, \dots, p.$
       $\mathbb{E}[N_{kl}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}] = \sum_{i=1}^N 1\{y_i^{(j)} \geq 2\} t_{kl}^{(j)} \frac{\sum_{s=1}^p \pi_s(\boldsymbol{x}_i) \prod_{q \neq j} \boldsymbol{e}_s^{\top} \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q \boldsymbol{e}_l^{\top} \boldsymbol{K}(y_i^{(j)}; \boldsymbol{e}_s^{\top}, \boldsymbol{T}_j) \boldsymbol{e}_k}{\sum_{s=1}^p \pi_s(\boldsymbol{x}_i) \prod_{q=1}^d \boldsymbol{e}_s^{\top} \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q},$
       $\mathbb{E}[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}] = \sum_{i=1}^N t_k^{(j)} \frac{\sum_{s=1}^p \pi_s(\boldsymbol{x}_i) \prod_{q \neq j} \boldsymbol{e}_s^{\top} \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q \boldsymbol{e}_s^{\top} \boldsymbol{T}_j^{y_i^{(j)}-1} \boldsymbol{e}_k}{\sum_{s=1}^p \pi_s(\boldsymbol{x}_i) \prod_{q=1}^d \boldsymbol{e}_s^{\top} \boldsymbol{T}_q^{y_i^{(q)}-1} \boldsymbol{t}_q}.$
  3) M-step: Let
       $\hat{t}_{kl}^{(j)} = \frac{\mathbb{E}\left[N_{kl}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}{\sum_{s=1}^p \mathbb{E}\left[N_{ks}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}, \quad k, l = 1, \dots, p, \; j = 1, \dots, d,$
       $\hat{t}_k^{(j)} = \frac{\mathbb{E}\left[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}{\sum_{s=1}^p \mathbb{E}\left[N_{ks}^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right] + \mathbb{E}\left[N_k^{(j)} \mid \overline{\boldsymbol{Y}} = \overline{\boldsymbol{y}}, \overline{\boldsymbol{x}}\right]}, \quad k = 1, \dots, p, \; j = 1, \dots, d,$
  4) R-step: Maximize the weighted multinomial logistic regression
       $\hat{\boldsymbol{\alpha}} = \operatorname*{arg\,max}_{\boldsymbol{\alpha} \in \mathbb{R}^{(p \times h)}} \sum_{i=1}^N \sum_{k=1}^p \mathbb{E}[B_k(\boldsymbol{x}_i) \mid \boldsymbol{Y} = \boldsymbol{y}_i, \boldsymbol{X} = \boldsymbol{x}_i] \log(\pi_k(\boldsymbol{x}_i; \boldsymbol{\alpha})),$
     and set
       $\pi_k(\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) = \frac{\exp(\boldsymbol{x}_i^{\mathsf{T}} \hat{\boldsymbol{\alpha}}_k)}{\sum_{s=1}^p \exp(\boldsymbol{x}_i^{\mathsf{T}} \hat{\boldsymbol{\alpha}}_s)}, \quad i = 1, \dots, N, \; k = 1, \dots, p.$
  5) Update the current parameters to $(\boldsymbol{\pi}, \mathcal{T}) = (\hat{\boldsymbol{\pi}}, \hat{\mathcal{T}})$. Return to step 1 unless a stopping rule is satisfied.
***Output***: Fitted representation $(\boldsymbol{\pi}, \mathcal{T})$.
```

References

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.

Bladt, M., 2023. A tractable class of multivariate phase-type distributions for loss modeling. North American Actuarial Journal. Forthcoming.

Bladt, M., Yslas, J., 2021. matrixdist: an R package for inhomogeneous phase-type distributions. arXiv preprint. arXiv:2101.07987.

Bladt, M., Yslas, J., 2022. matrixdist: Statistics for Matrix Distributions. R package version 1.1.5.

Bladt, M., Yslas, J., 2022b. Phase-type mixture-of-experts regression for loss severities. Scandinavian Actuarial Journal, 1–27.

Breuer, L., 2016. A semi-explicit density function for Kulkarni's bivariate phase-type distribution. Stochastic Models 32 (4), 632–642.

Chen, K., Huang, R., Chan, N.H., Yau, C.Y., 2019. Subgroup analysis of zero-inflated Poisson regression model with applications to insurance data. Insurance. Mathematics & Economics 86, 8–18.

Chen, K., Xu, L., Chi, H., 1999. Improved learning algorithms for mixture of experts in multiclass classification. Neural Networks 12 (9), 1229–1252.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, Methodological 39 (1), 1–22.

Frees, E.W., 2003. Multivariate credibility for aggregate loss models. North American Actuarial Journal 7 (1), 13–37.

Frees, E.W., Lee, G., Yang, L., 2016. Multivariate frequency-severity regression models in insurance. Risks 4 (1), 4.

Frees, E.W.J., Meyers, G., Cummings, A.D., 2010. Dependent multi-peril ratemaking models. ASTIN Bulletin: The Journal of the IAA 40 (2), 699–726.

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

Gabrielli, A., 2020. A neural network boosted double overdispersed Poisson claims reserving model. ASTIN Bulletin: The Journal of the IAA 50 (1), 25–60.

Gao, G., Wang, H., Wüthrich, M.V., 2022. Boosting Poisson regression models with telematics car driving data. Machine Learning 111 (1), 243–272.

He, Q.-M., Ren, J., 2016a. Analysis of a multivariate claim process. Methodology and Computing in Applied Probability 18 (1), 257–273.

He, Q.-M., Ren, J., 2016b. Parameter estimation of discrete multivariate phase-type distributions. Methodology and Computing in Applied Probability 18 (3), 629–651.

Jeong, H., Tzougas, G., Fung, T.C., 2023. Multivariate claim count regression model with varying dispersion and dependence parameters. Journal of the Royal Statistical Society. Series A. Statistics in Society 186 (1), 61–83.

Lee, S.C., 2021. Addressing imbalanced insurance data through zero-inflated Poisson regression with boosting. ASTIN Bulletin: The Journal of the IAA 51 (1), 27–55.

Navarro, A.C., 2018. Order statistics and multivariate discrete phase-type distributions. PhD thesis. DTU Lyngby.

Neuts, M.F., 1975. Probability distributions of phase type. In: Liber Amicorum Prof. Emeritus H. Florin.

Ren, J., Zitikis, R., 2017. CMPH: a multivariate phase-type aggregate loss distribution. Dependence Modeling 5 (1), 304–315.

Wüthrich, M.V., Merz, M., 2022. Statistical Foundations of Actuarial Learning and Its Applications. Springer Nature.

Yang, L., Shi, P., 2019. Multiperil rate making for property insurance using longitudinal data. Journal of the Royal Statistical Society. Series A. Statistics in Society 182 (2), 647–668.

Yip, K.C., Yau, K.K., 2005. On modeling claim frequency data in general insurance with extra zeros. Insurance. Mathematics & Economics 36 (2), 153–163.

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.

Zhang, P., Pitt, D., Wu, X., 2022. A new multivariate zero-inflated hurdle model with applications in automobile insurance. ASTIN Bulletin: The Journal of the IAA 52 (2), 393–416.