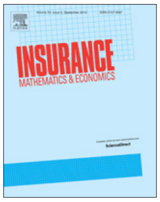
Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

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



A class of mixture of experts models for general insurance: Theoretical developments

Tsz Chai Fung, Andrei L. Badescu [*], X. Sheldon Lin

Department of Statistical Sciences, University of Toronto, 100 St George Street, Toronto, ON M5S 3G3, Canada

ARTICLE INFO

Article history:
Received November 2018
Received in revised form September 2019
Accepted 26 September 2019
Available online 4 October 2019

Keywords:
Claim frequency and severity modeling
Denseness theory
Mixture of experts models
Multivariate regression analysis
Neural network

ABSTRACT

In the Property and Casualty (P&C) ratemaking process, it is critical to understand the effect of policyholders' risk profile to the number and amount of claims, the dependence among various business lines and the claim distributions. To include all the above features, it is essential to develop a regression model which is flexible and theoretically justified. Motivated by the issues above, we propose a class of logit-weighted reduced mixture of experts (LRMoE) models for multivariate claim frequencies or severities distributions. LRMoE is interpretable, as it has two components: Gating functions, which classify policyholders into various latent sub-classes; and Expert functions, which govern the distributional properties of the claims. Also, upon the development of denseness theory in regression setting, we can heuristically interpret the LRMoE as a "fully flexible" model to capture any distributional, dependence and regression structures subject to a denseness condition. Further, the mathematical tractability of the LRMoE is guaranteed since it satisfies various marginalization and moment properties. Finally, we discuss some special choices of expert functions that make the corresponding LRMoE "fully flexible". In the subsequent paper (Fung et al., 2019b), we will focus on the estimation and application aspects of the LRMoE.

1. Introduction

Modeling insurance claim frequencies and severities is crucial in ratemaking for property and casualty (P&C) insurance (Frees et al., 2016), but it is not coming without its challenges. Policyholders are highly heterogeneous since they possess different risk characteristics. Hence, insurers should understand the risk drivers and perform risk classification based on the policyholders' information (covariates) obtained to design adequate tariff structures. Insurance products are sometimes sold in a bundled form, so that a policyholder may have claims from multiple business lines due to the same event. Furthermore, since it is impossible to collect all policyholders' information, unobserved heterogeneity still exists among policyholders having the same observed risk characteristics, leading to complicated distributional phenomena, such as multimodality and over-dispersion of claim frequencies and severities.

The afore-mentioned issues have been widely studied in actuarial literature through parametric modeling approach. The covariate influence is commonly incorporated through a Generalized Linear Model (GLM) regression framework, where the claim frequency or severity follows an exponential family of distributions with the mean function expressed in terms of linear combinations of the covariates (De Jong and Heller, 2008). To improve the flexibility, GLM can be extended to the Generalized Additive Model (GAM) that allows for non-linear or even interactive covariates effects. To model the dependence structures among business lines, a common shock regression model (Bermúdez and Karlis, 2011) is introduced to capture the effect that some accidents trigger multiple types of claims at the same time. To capture a full range of covariance structure, the use of copulas is becoming increasingly popular, see for example Shi and Valdez (2014) and Frees et al. (2016) for the insurance applications of copulas to the dependence modeling. To model the unobserved heterogeneity, Yip and Yau (2005), for example, propose a class of zero-inflated regression models to capture the excess of zeros for claim frequency data. Assuming particular structures on the claim frequencies or severities, these parametric models are usually mathematically tractable, interpretable and easily computable.

In insurance applications, the data characteristics can sometimes be very complicated, therefore it is hard to know what assumptions should be imposed prior to the modeling process. In this case, it may be challenging to specify a parametric model that fits all the underlying data features. To this end, some recent actuarial papers (see e.g. Quan and Valdez, 2018, Wüthrich, 2018 and Diao and Weng, in press) make use of non-parametric machine learning techniques in the insurance context. Various

[*] Corresponding author.
E-mail address: badescu@utstat.toronto.edu (A.L. Badescu).

https://doi.org/10.1016/j.insmatheco.2019.09.007

machine learning tools, including neural networks and random forests, are summarized comprehensively by Wuthrich and Buser (2019) with potential applications on modeling and predicting insurance claim counts. Being highly flexible in capturing the covariates influence, these models are effective in mitigating model misspecification risk where the fitted model may not adequately resemble the data.

Alternative to the above two very distinctive approaches for modeling multivariate insurance loss frequencies or severities, recently there is also a growth of interest in mixture-based semi-parametric modeling. See for instance Lee and Lin (2012), Badescu et al. (2015) and Miljkovic and Grün (2016). On one hand, mixture-based models act like parametric models that contain well-specified functional forms. On the other hand, the number of mixture components can be adjusted as a hyperparameter to control the distributional complexity, making them very flexible in catering for multimodality and over-dispersion.

This paper contributes to multivariate regression modeling under the mixture-based modeling framework. Motivated by the flexibility of the class of expert models (that is theoretically justified by e.g. Jiang and Tanner, 1999 and Nguyen et al., 2016), we start with the class of Generalized Mixture of Experts (GMoE) models, which is first introduced by Jacobs et al. (1991), as a candidate model for multivariate claim frequencies/severities regression. GMoE consists of two major components: a gating function, which governs how an individual is classified into different latent subgroups; and an expert function, which describes either the frequency or the severity distributions for an individual belonging to a particular subgroup. Both functions are influenced by the covariates. Being closely related to neural networks, GMoE has been applied to a wide variety of research areas, including social science (Gormley et al., 2008) and natural science (Übeyli, 2005). It also covers a wide range of flexible class of models, including finite mixtures of GLMs, which is applied to general insurance frequency modeling (see e.g. Bermúdez, 2009 and Badescu et al., 2015). However, with an excessive number of parameters and an over-complicated structure, GMoE is undesirable in computational feasibility and is likely to cause over-fitting, hindering its usefulness in the context of general insurance.

To address these issues, in this paper we propose a class of logit-weighted reduced MoE (LRMoE), a GMoE that removes the regression relationships for the expert functions and assumes an exponential linear form of regression for the gating functions. With such a model simplification, the number of parameters of the LRMoE is significantly reduced compared to that of the GMoE. Under this model, each policyholder belongs to one of the unobserved homogeneous subgroups and the probability that a policyholder belongs to a certain subgroup depends on his/her risk characteristics. The regression coefficients for the gating functions represent the impact of subgroup assignments on the risk characteristics.

Apart from interpretability, a key motivation of introducing the LRMoE for insurance predictive modeling is that its versatility can be theoretically justified by denseness theory. Denseness property guarantees the existence of a model within the class of LRMoE that resembles well the input data, potentially avoiding the need of ad-hoc model selection procedures where multiple classes of models are fitted by trial and error in order to obtain a model that adequately represents the data. The actuarial literature discussing the denseness theory is scarce. In severity modeling (without covariates), the multivariate mixture of Erlang model proposed by Lee and Lin (2012) is dense in the space of all positive continuous multivariate distributions, meaning that any severity distributions can be approximated arbitrarily accurately by the Erlang mixture model. Similarly, the denseness property is also satisfied by a class of Phase-type (PH) or log-PH distributions Asmussen et al. (1996) and Ahn et al. (2012). However, finding a frequency model that has the denseness property is a more difficult task because many commonly used actuarial frequency models (such as Poisson and Negative Binomial distributions) are not designed to capture under-dispersion. Incorporating covariates into the model makes the denseness problem even more challenging and it will be completely formulated in this paper.

We first justify the use of the proposed simplified model instead of the more complicated GMoE. In this paper, we show that, under very mild conditions, the class of LRMoE is dense in the space of the GMoE, so the model flexibility is not impeded when GMoE is simplified to LRMoE. A remarkable feature of the LRMoE is that it can capture any regression patterns (including non-linear patterns and covariates interactions) involved in the GMoE even if the LRMoE contains only linear regressions. Since the LRMoE is a flexible model with the simplest possible model structure, it is a parsimonious model. Since the LRMoE mitigates the overfitting risk without sacrificing its flexibility, we expect that the LRMoE is robust, meaning that it provides stable fitting results and good predictive power for any type of data.

We further illustrate the advantage of using the class of LRMoE to a practical insurance application, where the data to be calibrated possesses complicated features. Under some suitably chosen expert functions, the class of LRMoE is furthermore dense in the space of any frequency/severity regression distributions, meaning that any regression true models (subject to mild restrictions) can be approximated arbitrarily closely by the LRMoE. In such cases, we refer to our proposed LRMoE class as to a "full flexible" class of models that can cater for any distributional, dependence and regression patterns. By choosing such expert functions, the proposed model becomes fully data-driven. In practice, regardless of the complexities of the model generating the input data, the characteristics of the calibrated model will be highly synchronous to that of the input data.

Our proposed class of LRMoE is desirable in terms of mathematical tractability, which is important in insurance applications in terms of premium and risk measure calculations. Firstly, it is closed under response marginalization, i.e., the marginal frequency/severity for each claim type still follows a univariate LRMoE. Secondly, it is closed under covariate marginalization, meaning that even if some important covariates are missing, the resulting model can still be expressed in the form of mixture of experts. The marginalization properties facilitate computing various useful quantities related to the proposed model more efficiently. Thirdly, various moments and measures of association under the LRMoE can be written in a simplified form that involves only the quantities corresponding to the individual expert functions, making them easily computable.

In a subsequent paper (Fung et al., 2019b), we apply the LRMoE with Erlang Count expert functions to solve the estimation and application problems for multivariate insurance claim frequency regression. A fitting algorithm is developed for efficient model calibration, while the effectiveness of the algorithm and the flexibility of the proposed model are verified through several simulation studies. We conclude that the proposed model is able to adequately fit the complicated structures implied by a real automobile insurance dataset.

The paper is structured as follows. In Section 2, we define and interpret the class of GMoE and LRMoE respectively. The use of LRMoE over GMoE as a multivariate insurance claim regression model is further justified. Section 3 defines "denseness" in the context of multivariate regression problems and proves several denseness properties possessed by the class of LRMoE. Other desirable properties such as marginalization and moment properties are discussed in Section 4. Section 5 provides a few choices of expert functions in the attempt of modeling frequency or severity distributions. By checking the denseness condition for each expert function, we can evaluate whether or not an LRMoE with such an expert function can achieve "full flexibility" in modeling. Finally, in Section 6, we summarize our findings, discuss some practical aspects of the proposed model and provide a brief description of future work.

2. Model description

In this section, we define and interpret the proposed class of mixture of experts (MoE) models for multivariate frequency or severity regression. Denote $\boldsymbol{Y} = (Y_1, \ldots, Y_K)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_K)^T$ respectively the multivariate response frequency or severity random column vector and corresponding realization. We make the following assumption for frequency random vectors with support $\{0, 1, \ldots\}^K$.

Assumption 2.1. For a frequency random vector, its probability mass function (pmf) is strictly positive on $\{0, 1, \ldots\}^K$.

For severity random vectors with support $(0, \infty)^K$, we impose the following assumption.

Assumption 2.2. For a severity random vector, there exists a continuous cumulative distribution function (cdf).

We also define $\boldsymbol{x} = (x_0, x_1, \ldots, x_P)^T$ as the covariates column vector. Restricting $x_0 = 1$, only "with-intercept" regression models are considered. In the context of insurance, K may represent the number of attributes for an insurance contract. To begin with, we introduce a generalized form of MoE (GMoE), which is also described by Grun and Leisch (2008) as a finite mixture model with concomitant variables.

Definition 2.1. Under the GMoE, the cumulative probability distribution (cdf) of $\boldsymbol{Y}$ is given by

$$\begin{aligned} H^*(\boldsymbol{y}; \boldsymbol{x}) &:= H^*(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, g) \\ &= \sum_{j=1}^{g} \pi_j^*(\boldsymbol{x}; \boldsymbol{\alpha}^*) \prod_{k=1}^{K} F_k(y_k; \boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*)), \end{aligned} \tag{1}$$

where g is the number of latent classes, $\pi_j^*(\boldsymbol{x}; \boldsymbol{\alpha}^*)$ (called the gating network/function) is the mixing weight for the jth class, $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_1^*, \ldots, \boldsymbol{\alpha}_g^*)$ are the parameters for the regressions of the mixing weights; $\boldsymbol{\beta}^* = \{\boldsymbol{\beta}_{jk}^*; j \in \{1, \ldots, g\}, k \in \{1, \ldots, K\}\}$ are the parameters for the regressions of $\boldsymbol{\theta}_{jk}^*$, which are themselves the parameters of the univariate cdf $F_k(y_k; \boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*))$ (called the expert network/function).

From Eq. (1), the pmf/pdf of $\boldsymbol{Y}$ is

$$\begin{aligned} h^*(\boldsymbol{y}; \boldsymbol{x}) &:= h^*(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, g) \\ &= \sum_{j=1}^{g} \pi_j^*(\boldsymbol{x}; \boldsymbol{\alpha}^*) \prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*)), \end{aligned} \tag{2}$$

where we also call the univariate pmf/pdf $f_k(y_k; \boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*))$ as the expert function.

Remark 2.1. The number of parameters of the GMoE varies with the assumptions on the functional forms of $\pi_j^*(\boldsymbol{x}; \boldsymbol{\alpha}^*)$ and $\boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*)$. Suppose that the gating function is in an exponential linear form (will be discussed in Definition 2.2), f_k is an m-parameter expert function and the regressions in the expert functions are under the GLM framework (linear regression on only one of the m parameters in the expert functions). Then, we have $\boldsymbol{\beta}_{jk}^* \in \mathbb{R}^{P+1}$, $\boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*) \in \mathbb{R}^m$ and $\boldsymbol{\alpha}_j^* \in \mathbb{R}^{P+1}$. The number of regression parameters for gating functions, regression parameters for expert functions and non-regression parameters for expert functions are respectively $g \times (P + 1)$, $g \times K \times (P + 1)$ and $g \times K \times (m - 1)$. As shown in Section 5, note that commonly used expert functions usually contain a small number of parameters (mostly $m \leq 2$). Therefore, the total number of parameters for the GMoE ($N_{GMoE} = g \times (P+1) + g \times K \times (P+m)$) is usually dominated by the term $g \times K \times P$, especially when there are many covariates.

Note that the above model does not impose any assumptions on the functional forms of $\pi_j^*(\boldsymbol{x}; \boldsymbol{\alpha}^*)$ and $\boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*)$. In other words, we do not restrict any regression patterns (e.g. linear regression) on either the mixing weights or the expert functions. The only assumption of GMoE is the conditional independence among the marginal responses $Y_1, \ldots, Y_K$. One may attempt to extend it by introducing a dependence structure (e.g. copula) on the class-dependent distribution. However, the denseness property in Section 3 will show that the model flexibility is already sufficient under such an assumption. Also, such an extension will impede the model's mathematical tractability, making the conditional independence assumption well justified.

The GMoE contains a wide range of highly flexible classes of models for insurance modeling. For severity modeling, the multivariate Erlang mixtures model proposed by Lee and Lin (2012), which is flexible to model any positive continuous multivariate distributions, is a very special case of the GMoE with $P = 0$ (no covariate influence) and Erlang-distributed expert function. For frequency modeling, the multivariate Pascal mixture regression model introduced by Badescu et al. (2015), which is versatile to deal with a wide range of over-dispersed distributional features and dependence structures, is also a special choice of the GMoE with linear regressions on the expert functions, but without regressions on the mixing weights.

Remark 2.2. One may attempt to extend the GMoE to hierarchical mixture of experts (HME) introduced by Jordan and Jacobs (1992), which consists of two levels of expert networks. McLachlan and Peel (2000) state that HME can enhance the flexibility of MoE through an increase of the level of experts, but Proposition 4.2 of this paper suggests that HME can already be represented in the form of MoE under exponential linear gating functions, so the flexibility of the class of GMoE is ensured.

Despite its flexibility, the excessive complexity of the GMoE greatly reduces its usefulness for general insurance applications. From Remark 2.1, the number of parameters for the GMoE contains the term $g \times K \times P$, a product of 3 quantities. If more complicated features outside the GLM framework (e.g. non-linear regressions) are incorporated on the expert functions, GMoE involves even more parameters. Such a large number of parameters will complicate the model interpretation and cause troubles in model fitting and model selection. Motivated by this issue, we propose a class of reduced-form MoE (RMoE) models, which is a special choice of the GMoE.

Definition 2.2. Under the RMoE, the cdf of $\boldsymbol{Y}$ is given by

$$H(\boldsymbol{y}; \boldsymbol{x}) := H(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Theta}, g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \prod_{k=1}^{K} F_k(y_k; \boldsymbol{\theta}_{jk}), \tag{3}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ and $\boldsymbol{\alpha}_j = (\alpha_{j0} \ldots, \alpha_{jP})^T \in \mathbb{R}^{P+1}$ are the weight regression parameters; $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{jk}; j = 1, \ldots, g,\ k = 1, \ldots, K\}$ and $\boldsymbol{\theta}_{jk} \in \mathbb{R}^m$ are the parameters associated to the expert functions. If the mixing weight $\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})$ follows an exponential linear gating function

$$\pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{exp\{\boldsymbol{\alpha}_j^T \boldsymbol{x}\}}{\sum_{j'=1}^{g} exp\{\boldsymbol{\alpha}_{j'}^T \boldsymbol{x}\}}, \qquad j = 1, \ldots, g, \tag{4}$$

then the resulting model is called the logit-weighted reduced Mixture of Experts Models (LRMoE).

From Eq. (3), the pmf/pdf of $\boldsymbol{Y}$ is

$$h(\boldsymbol{y};\boldsymbol{x}) := h(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\Theta},g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x};\boldsymbol{\alpha}) \prod_{k=1}^{K} f_k(y_k;\boldsymbol{\theta}_{jk}). \tag{5}$$

The above model can be interpreted in an insurance context as the following. The whole population of policyholders is classified into g unobserved subgroups. The claim behavior varies among different subgroups but it is homogeneous for policyholders within a subgroup. The probability that a policyholder belongs to subgroup j ($\pi_j(\boldsymbol{x};\boldsymbol{\alpha})$) depends on the covariates, so that policyholders with more undesirable characteristics (e.g. younger driver and poorer car model in automobile insurance) are more likely to belong to a subgroup with high risk profile.

The choice of exponential linear gating function is motivated by its connection to multivariate logistic regression. For each policyholder, the log-probability ratio between subgroup j_1 and subgroup j_2 is given by

$$\log\left(\frac{\pi_{j_1}(\boldsymbol{x};\boldsymbol{\alpha})}{\pi_{j_2}(\boldsymbol{x};\boldsymbol{\alpha})}\right) = (\boldsymbol{\alpha}_{j_1}^T - \boldsymbol{\alpha}_{j_2}^T)\boldsymbol{x}.$$

Therefore, larger regression coefficient α_{jp} (the $(p+1)st$ element of $\boldsymbol{\alpha}_j$, $p = 1,2,\ldots,P$) represents higher chance for an individual to be classified as subgroup j when x_p is large.

Compared to the GMoE, the LRMoE imposes two specific restrictions: The mixing weights take a specific linear regression form and the expert functions do not consider regressions. The number of parameters for the LRMoE is $N_{LRMoE} = g \times (P+1) + g \times K \times m$, which is reduced by the dominating term $g \times K \times P$ compared to that for the GMoE. Moreover, since the regressions on a distribution of a non-exponential family are usually computationally intensive, removing the regression relationships in the expert functions can offer us a greater variety of choices of expert functions for model fitting with reasonable computational costs. Therefore, the remaining concern about the LRMoE is: how does the reduced form of MoE affect the model flexibility? In the following section, we will demonstrate that the LRMoE in the form of Eq. (5) can approximate the GMoE in Eq. (2) arbitrarily closely under very mild conditions, even if the LRMoE consists of linear regressions only while the GMoE consists of non-linear regressions. The impact of imposing extra assumptions to the model versatility is indeed minimal.

3. Denseness property

Model flexibility is an important criteria for a good model. It is desirable that the model can capture a wide range of characteristics of multivariate regression distributions, so that data generated from the fitted model will be highly synchronous to the input data in model fitting perspective, even if the fitted model is not the true model. In mathematical perspective, such a desirable property is called "denseness".

The denseness problem is very complicated when regression is incorporated, because it requires the ability for the model to cater for any distribution and regression patterns, including but not limiting to any kinds of interactions among covariates and the non-linear relationships between the response variable and the covariates. The main goal of this section is to show several denseness properties of the LRMoE.

3.1. Definition of denseness

We first define "denseness" mathematically. We start with a class of distributions (without considering regression).

Definition 3.1. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two classes of distributions. $\mathcal{C}_1$ is dense in $\mathcal{C}_2$ if and only if for every $F \in \mathcal{C}_2$, there exists a sequence of $\{G_n\}_{n=1,2,\ldots}$ with $G_n \in \mathcal{C}_1$ for $n = 1,2,\ldots$ such that $G_n \xrightarrow{\mathcal{D}} F$ as $n \to \infty$, where $\xrightarrow{\mathcal{D}}$ means weakly convergence or convergence in distribution.

Non-technically speaking, the above definition implies that $\mathcal{C}_1$ is at least as flexible as $\mathcal{C}_2$ in modeling perspective, because any probability distribution in $\mathcal{C}_2$ can be approximated by a distribution in $\mathcal{C}_1$ at any precision.

Remark 3.1. One may argue that weak convergence of $\{G_n\}_{n=1,2,\ldots}$ to F does not mean there exists a G_n that can approximate F at any precision, because weak convergence is only a kind of pointwise convergence. One may suggest uniform convergence, which means that for all $\epsilon > 0$, there exists a G_n such that $\sup_{\boldsymbol{y}} |G_n(\boldsymbol{y}) - F(\boldsymbol{y})| < \epsilon$. In this case, the whole distribution (body and tail) of G_n can approximate that of F within the precision level ϵ (in an absolute scale), which can be freely adjusted. If G_n and F are either frequency or severity probability distributions under Assumption 2.1 or 2.2, it can be easily proved that weakly convergence already implies uniform convergence. The proof details are discussed in Appendix A.

In regression settings, denseness problem has not been investigated, which leads us to introduce a new definition of denseness. Such a definition should be tightly related to the full flexibility of a class of models to cater for a broad range of both distributional and regressional patterns. We first define the term "regression distribution".

Definition 3.2. A class of regression distributions $\mathcal{C}(\mathcal{A})$ (where $\mathcal{A}$ is the support of the covariates $\boldsymbol{x}$) is a set, where each element $F(\mathcal{A}) := \{F(\cdot;\boldsymbol{x}); \boldsymbol{x} \in \mathcal{A}\}$ in $\mathcal{C}(\mathcal{A})$ is itself a set of probability distributions.

Then, a new definition of denseness in regression settings is introduced as follows.

Definition 3.3. Let $\mathcal{A}$ be the support of the covariates $\boldsymbol{x}$. Also, denote $\mathcal{C}_1(\mathcal{A})$ and $\mathcal{C}_2(\mathcal{A})$ as two classes of regression distributions. $\mathcal{C}_1(\mathcal{A})$ is dense in $\mathcal{C}_2(\mathcal{A})$ if and only if for every $F(\mathcal{A}) \in \mathcal{C}_2(\mathcal{A})$, there exists a sequence of regression distributions $\{G_n(\mathcal{A})\}_{n=1,2,\ldots}$ with $G_n(\mathcal{A}) \in \mathcal{C}_1(\mathcal{A})$ for $n = 1,2,\ldots$ such that for every $\boldsymbol{x} \in \mathcal{A}$, $G_n(\boldsymbol{y};\boldsymbol{x}) \xrightarrow{\mathcal{D}} F(\boldsymbol{y};\boldsymbol{x})$ as $n \to \infty$. If the convergence $G_n(\boldsymbol{y};\boldsymbol{x}) \to F(\boldsymbol{y};\boldsymbol{x})$ is uniform on $\boldsymbol{x} \in \mathcal{A}_{\boldsymbol{y}}$ for any $\boldsymbol{y}$, where $\mathcal{A}_{\boldsymbol{y}}$ is the set of $\boldsymbol{x}$ such that $\boldsymbol{y}$ is a continuity point of $F(\boldsymbol{y};\boldsymbol{x})$, then $\mathcal{C}_1(\mathcal{A})$ is uniformly dense in $\mathcal{C}_2(\mathcal{A})$.

Remark 3.2. Note that if $G_n(\boldsymbol{y};\boldsymbol{x})$ and $F(\boldsymbol{y};\boldsymbol{x})$ are frequency distributions, then under Assumption 2.1 the set of continuity points is $\mathcal{S}_c := (\mathbb{R}\backslash\{0,1,\ldots\})^K$, regardless of $\boldsymbol{x}$. Hence, $\mathcal{A}_{\boldsymbol{y}}$ under Definition 3.3 is $\mathcal{A}$ if $\boldsymbol{y} \in \mathcal{S}_c$ and $\mathcal{A}_{\boldsymbol{y}}$ is null otherwise. For severity distributions, all points in the Euclidean space are continuity points under Assumption 2.2, so $\mathcal{A}_{\boldsymbol{y}} = \mathcal{A}$ for any $\boldsymbol{x}$. Overall, $\mathcal{A}_{\boldsymbol{y}}$ is either $\mathcal{A}$ or null, so the continuity point issue suggested in Definition 3.3 can be ignored.

Definition 3.3 is a direct extension of Definition 3.1. Under such a definition, $\mathcal{C}_1(\mathcal{A})$ is flexible in capturing any distributional and regressional characteristics of $\mathcal{C}_2(\mathcal{A})$, because convergence of distribution is required for any choices of covariates $\boldsymbol{x} \in \mathcal{A}$.

3.2. Denseness in the class of GMoE

In this subsection, we provide theoretical justifications on the flexibility of the LRMoE in the form of Eq. (5). We will prove that under several mild conditions that are negligible in practical insurance applications, the LRMoE, which assumes linear regression patterns on the mixing weights, is already versatile enough to cater for any distributional and regressional (including non-linear regressions) patterns of the GMoE in the form of Eq. (2). Being a flexible yet simple model, the proposed LRMoE is deemed to be parsimonious. Model parsimony is a crucial feature as it ensures that the model has a good explanatory predictive power. The mathematical formulations for such a denseness property are as follows.

Let $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}_0(\mathcal{A})$ be two classes of regression distributions. Each element in $\mathcal{G}(\mathcal{A})$ (or in $\mathcal{G}_0(\mathcal{A})$) is a GMoE regression distribution in the form of Eq. (1) (or a LRMoE regression distribution in the form of Eq. (3)). Precisely, $\mathcal{G}(\mathcal{A})=\{H^*(\boldsymbol{y};\mathcal{A},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,g); g\in\mathbb{N},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*\}$ with $H^*(\boldsymbol{y};\mathcal{A},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,g):=\{H^*(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*,g);\boldsymbol{x}\in\mathcal{A}\}$, and $\mathcal{G}_0(\mathcal{A})=\{H(\boldsymbol{y};\mathcal{A},\boldsymbol{\alpha},\boldsymbol{\Theta},g); g\in\mathbb{N},\boldsymbol{\alpha},\boldsymbol{\Theta}\}$ with $H(\boldsymbol{y};\mathcal{A},\boldsymbol{\alpha},\boldsymbol{\Theta},g):=\{H(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\Theta});\boldsymbol{x}\in\mathcal{A}\}$, where H^* and H are the distribution functions corresponding to Eqs. (1) and (3) respectively, sharing the same, fixed and known expert functions $\{f_k\}_{k=1,\ldots,K}$ and dimension K. Here, $\mathcal{A}\subseteq\mathbb{R}^{P+1}$ and $\boldsymbol{x}$ is a $(P+1)$-dimensional vector. Since the LRMoE is a subclass of the GMoE, we have $\mathcal{G}_0(\mathcal{A})\subseteq\mathcal{G}(\mathcal{A})$ and hence it is trivial that $\mathcal{G}(\mathcal{A})$ is dense in $\mathcal{G}_0(\mathcal{A})$. We aim to prove the converse result: $\mathcal{G}_0(\mathcal{A})$ is uniformly dense in $\mathcal{G}(\mathcal{A})$ under certain mild conditions. To begin with, we want to demonstrate the possibility to construct an LRMoE which provides a one-to-one correspondence between the covariates combination and the subgroup assignment. Hence, the following technical lemma is first introduced. The rigorous proof is presented in Appendix B.

Lemma 3.1. *Suppose that $\boldsymbol{x}\in\{1\}\times\{m_1,\ldots,m_L\}^P$ with $m_1<m_2<\cdots<m_L$, and define $h^{(\boldsymbol{l})}(\boldsymbol{x})=\lambda_0^{(\boldsymbol{l})}+\lambda_1^{(\boldsymbol{l})}x_1+\cdots+\lambda_P^{(\boldsymbol{l})}x_P$ for $\boldsymbol{l}=(l_1,\ldots,l_P)^T\in\{1,\ldots,L\}^P$. Also define $\phi:\{m_1,\ldots,m_L\}\mapsto\{1,\ldots,L\}$ as a function with $\phi(m_l)=l$ for every $l\in\{1,\ldots,L\}$. Denote $\boldsymbol{\phi}:\{1\}\times\{m_1,\ldots,m_L\}^P\mapsto\{1,\ldots,L\}^P$ as a function with $\boldsymbol{\phi}(\boldsymbol{x})=(\phi(x_1),\ldots,\phi(x_P))^T$ for every $(x_1,\ldots,x_P)^T\in\{m_1,\ldots,m_L\}^P$. Then, we can construct $\{(\lambda_0^{(\boldsymbol{l})},\ldots,\lambda_P^{(\boldsymbol{l})})^T;\boldsymbol{l}\in\{1,\ldots,L\}^P\}$ as the parameters of $\{h^{(\boldsymbol{l})}(\boldsymbol{x});\boldsymbol{l}\in\{1,\ldots,L\}^P\}$ such that $\boldsymbol{\phi}(\boldsymbol{x})=\mathrm{argmax}_{\boldsymbol{l}\in\{1,\ldots,L\}^P}\{h^{(\boldsymbol{l})}(\boldsymbol{x})\}$ for every $\boldsymbol{x}$.*

The above lemma can be interpreted as follows. Consider an L^P-component LRMoE, and $\boldsymbol{l}$ in the above lemma represents the component label. Note that L^P is also the number of possible combinations for the covariates. Corresponding to each component $\boldsymbol{l}$, we first construct a function $h^{(\boldsymbol{l})}(\boldsymbol{x})$ that is linear on $\boldsymbol{x}$. Lemma 3.1 suggests that for every covariates combination $\boldsymbol{x}$, there exists a one-to-one corresponding component $\boldsymbol{l}$ (i.e. $\boldsymbol{l}=\boldsymbol{\phi}(\boldsymbol{x})$) such that $h^{(\boldsymbol{l})}(\boldsymbol{x})$ is greater than the remaining (L^P-1) linear functions $h^{(\boldsymbol{l}')}(\boldsymbol{x})$ $(\boldsymbol{l}'\neq\boldsymbol{l})$. It leads to another lemma.

Lemma 3.2. *Construct the component weight $\pi_{\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha})=\exp\{nh^{(\boldsymbol{l})}(\boldsymbol{x})\}/\sum_{\boldsymbol{l}'\in(1,\ldots,L)^P}\exp\{nh^{(\boldsymbol{l}')}(\boldsymbol{x})\}$ for the L^P-component LRMoE, which is also an exponential gating function since $nh^{(\boldsymbol{l})}(\boldsymbol{x})$ is still linear on $\boldsymbol{x}$. Then, it follows that*

$$\pi_{\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha})=\left(\sum_{\boldsymbol{l}'\in(1,\ldots,L)^P}\exp\{n(h^{(\boldsymbol{l}')}(\boldsymbol{x})-h^{(\boldsymbol{l})}(\boldsymbol{x}))\}\right)^{-1}\xrightarrow{n\to\infty}1\{\boldsymbol{l}=\boldsymbol{\phi}(\boldsymbol{x})\}. \tag{6}$$

Proof. Consider for $\boldsymbol{l}=\boldsymbol{\phi}(\boldsymbol{x})$, then $n(h^{(\boldsymbol{l}')}(\boldsymbol{x})-h^{(\boldsymbol{l})}(\boldsymbol{x}))\xrightarrow{n\to\infty}-\infty$ if $\boldsymbol{l}'\neq\boldsymbol{l}$ and $\to0$ if $\boldsymbol{l}'=\boldsymbol{l}$, so $\pi_{\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha})\to1$; for $\boldsymbol{l}\neq\boldsymbol{\phi}(\boldsymbol{x})$, then there exists an $\boldsymbol{l}\neq\boldsymbol{\phi}(\boldsymbol{x})$ such that $n(h^{(\boldsymbol{l}')}(\boldsymbol{x})-h^{(\boldsymbol{l})}(\boldsymbol{x}))\xrightarrow{n\to\infty}\infty$, and hence $\pi_{\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha})\to0$. ■

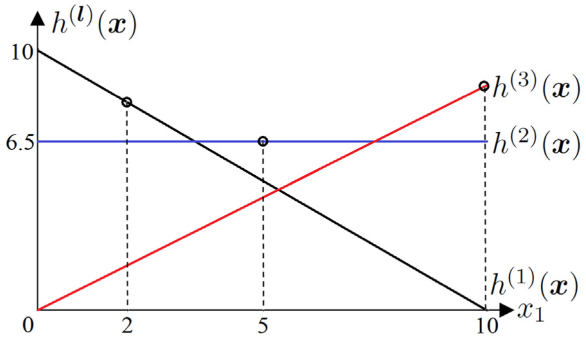


Fig. 1. $h^{(l)}(\boldsymbol{x})$ versus the covariate x_1 for the numerical example.

Lemma 3.2 shows that the proposed LRMoE is flexible enough to assign individuals with different covariates combinations into arbitrarily different subgroups, providing a critical foundation for the subsequent theorems to justify the proposed model's flexibility.

Before we state our main theorems, we would like to demonstrate a simple numeric example to facilitate understandability of Lemmas 3.1 and 3.2. Consider $\mathcal{A}=\{1\}\times\{2,5,10\}$ such that there is only one covariate $(P=1)$ taking $L=3$ possible values $m_1=2$, $m_2=5$ and $m_3=10$. Note that we have $\phi(2)=1$, $\phi(5)=2$ and $\phi(10)=3$. We construct the parameters of the linear functions $\{h^{(l)}(\boldsymbol{x});l\in\{1,2,3\}\}$ according to the scheme proposed by the proof of Lemma 3.1 (Appendix B) as follows.

$$\begin{aligned}&\lambda_0^{(1)}=10>\lambda_0^{(2)}=6.5>\lambda_0^{(3)}=0,\\&\lambda_1^{(1)}=-1<\lambda_1^{(2)}=0<\lambda_1^{(3)}=13/15.\end{aligned} \tag{7}$$

Then, these parameters result in the following linear functions $h^{(l)}(\boldsymbol{x})$.

$$h^{(1)}(\boldsymbol{x})=10-x_1,\quad h^{(2)}(\boldsymbol{x})=6.5,\quad h^{(3)}(\boldsymbol{x})=(13/15)x_1. \tag{8}$$

The functions $h^{(l)}(\boldsymbol{x})$ are also plotted in Fig. 1. When $x_1=2$, $x_1=5$ and $x_1=10$, it is observed that $h^{(1)}(\boldsymbol{x})$, $h^{(2)}(\boldsymbol{x})$ and $h^{(3)}(\boldsymbol{x})$ respectively take the greatest value, so Lemma 3.1 is verified. We then introduce a 3-component LRMoE with component weights constructed in the way proposed by Lemma 3.2. The subgroup assignment probabilities across various n are displayed in Table 1. We see that for sufficiently large n, different covariate values are classified into different subgroups with almost certainty, verifying Lemma 3.2.

The main denseness results for the LRMoE are shown in the following two theorems. Note that the results hold for the expert networks F_k corresponding to either a frequency or severity random variable. Theorem 3.1 assumes that the covariates have a finite support, but no restrictions are imposed on the expert functions. On the other hand, Theorem 3.2 allows a continuous compact space for the support of the covariates, but certain mild conditions are imposed on the expert functions.

Theorem 3.1. *If $\mathcal{A}=\{1\}\times\{m_1,\ldots,m_L\}^P$, then $\mathcal{G}_0(\mathcal{A})$ is uniformly dense in $\mathcal{G}(\mathcal{A})$.*

Proof. The rigorous proof is displayed in Appendix C, but here we roughly sketch the proof idea. Given a fixed covariates combination $\boldsymbol{x}$, the GMoE regression distribution results to a g-component finite mixture model. With a total of L^P possible combinations for the covariates, each GMoE regression distribution contains a total of L^P finite mixture distributions. Motivated by the flexibility for the LRMoE to assign component weights (Lemma 3.2), we now construct a $g\times L^P$-component LRMoE that assigns/classifies

Table 1
The component weights $\pi_l(\boldsymbol{x}; \boldsymbol{\alpha})$ constructed under Lemma 3.2.

	$n = 0.1$			$n = 1$			$n = 10$		
$\pi_l(\boldsymbol{x}; \boldsymbol{\alpha})$	$x_1 = 2$	$x_1 = 5$	$x_1 = 10$	$x_1 = 2$	$x_1 = 5$	$x_1 = 10$	$x_1 = 2$	$x_1 = 5$	$x_1 = 10$
$l = 1$	**0.418**	0.323	0.189	**0.816**	0.167	0.000	**1.000**	0.000	0.000
$l = 2$	0.359	**0.375**	0.362	0.182	**0.748**	0.103	0.000	**1.000**	0.000
$l = 3$	0.223	0.302	**0.449**	0.002	0.086	**0.897**	0.000	0.000	**1.000**

each (different) covariates combination to the (different) g components that constitute the corresponding finite mixture model resulted from the GMoE. Therefore, each GMoE can be well approximated by the LRMoE at any precision. ■

Since the theorem above restricts the covariates $\boldsymbol{x}$ to a finite support, it is useful when all the covariates are categorical. In general insurance, a wide range of explanatory variables that affect policyholders' claim behavior, such as gender, location/region and vehicle brand/type, are categorical. Even if the covariates are continuous in nature (e.g. age), it is still common and practical to categorize the explanatory variables before fitting a model. See the data structure used in Bermúdez (2009) as an example. In the case where continuous covariates are considered, the following theorem is needed.

Theorem 3.2. *Suppose that $\mathcal{A} = \{1\} \times [m_{\min}, m_{\max}]^P$, F_k and π_j^* are Lipschitz continuous on $\boldsymbol{x} \in \mathcal{A}$ $\forall k = 1, \ldots, K$, $\forall j = 1, \ldots, g < \infty$, $\forall \boldsymbol{y}$ and for any fixed parameters settings under the GMoE. Then, $\mathcal{G}_0(\mathcal{A})$ is uniformly dense in $\mathcal{G}(\mathcal{A})$.*

Proof. The rigorous proof is displayed in Appendix D and the proof idea is as follows. We first partition $[m_{\min}, m_{\max}]$ into L intervals to obtain L partition points (namely $m_1, \ldots, m_L$). Then for the covariates space $\mathcal{A}_0 := \{1\} \times \{m_1, \ldots, m_L\}^P \subseteq \mathcal{A}$, we can construct an LRMoE arbitrarily closely approximating the GMoE on $\mathcal{A}_0$, using the idea from Theorem 3.1. When L is sufficiently large (i.e. the partitioning intervals are small), for every $\boldsymbol{x} \in \mathcal{A}$, we can find an element $\tilde{\boldsymbol{x}} \in \mathcal{A}_0$ such that $\boldsymbol{x}$ is very close to $\tilde{\boldsymbol{x}}$. Through the Lipschitz assumptions, the resulting distributions given $\boldsymbol{x}$ are very close to those given $\tilde{\boldsymbol{x}}$. Therefore, the accurate approximations of the LRMoE to the GMoE can be extended to the continuous covariates space $\mathcal{A}$. ■

The extra conditions required in Theorem 3.2 are indeed very mild. The Lipschitz-continuity restrictions, which avoid $H^*(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*, g)$ changing too fast under infinitesimal change in $\boldsymbol{x}$, are purely mathematical and hence they are not any issues in practice. In practical regression models, all functions and parameters can be safely assumed to be continuously differentiable with respect to $\boldsymbol{x}$. This already guarantees the Lipschitz-continuity conditions when $\boldsymbol{x}$ is bounded. The finite bound for $\mathcal{A}$ is also not a concern in practice, because only finite number of data can be obtained in reality and the covariates for each data point are always finite. Even if the covariates are unbounded in nature (e.g. age), one can easily perform a one-to-one transformation on the unbounded covariate so that it falls within a bounded interval.

Remark 3.3. One may concern about the assumption that all covariates take the same range (the domains of each covariate are $\{m_1, \ldots, m_L\}$ in Theorem 3.1 and $[m_{\min}, m_{\max}]$ in Theorem 3.2 respectively). In practice, different covariates obviously have different ranges. For example, the domain of variable "age" can be $\{18, 19, \ldots, 100\}$ but that of variable "car fuel" can be $\{0, 1\}$ (diesel or gasoline). In this case, the covariates domain $\mathcal{A}_0 := \{1\} \times \{18, 19, \ldots, 100\} \times \{0, 1\}$ does not satisfy the assumption of Theorem 3.1, but we can extend the domain and choose $\mathcal{A} = \{1\} \times \{0, 1, 18, 19, \ldots, 100\}^2$ such that the result of Theorem 3.1 holds. Further, as a direct consequence of Definition 3.3, denseness on a larger domain $\mathcal{A}$ implies denseness on a smaller domain $\mathcal{A}_0$, because we need to check the weak convergence over a smaller region of $\boldsymbol{x}$ to show the denseness property on $\mathcal{A}_0$. The above argument can be easily generalized to any number of covariates P, so that having multiple covariates with varying ranges would not affect the denseness properties.

3.3. Denseness in the space of non-negative regression distributions

Although the versatility of the LRMoE is already well justified based on the theorems proposed in the previous subsection, it can be still far from a "full flexibility" – the ability to capture any distributional, dependence and regressional structures. This subsection derives stronger results compared to Theorems 3.1 and 3.2 – the denseness property of the LRMoE in the space of any frequency or severity regression distributions. With such a denseness property, we can be confident that for any datasets we fit, the fitted model will share similar characteristics as the input data. We aim to investigate the necessary and sufficient conditions for the expert functions f_k such that the denseness property holds.

Motivated by the broad applicability of finite mixture models to claim frequency/severity modeling in general insurance and the denseness properties of finite mixture models studied in actuarial literature (see e.g. Lee and Lin, 2012), we start with investigating the denseness conditions of finite mixture models. Then, the connections between the denseness properties of finite mixture models and that of the LRMoE are demonstrated.

Mathematically, define $\mathcal{H}_k$ $(k = 1, \ldots, K)$ and $\mathcal{H}$ as the classes of univariate and multivariate finite mixture models respectively, such that $\mathcal{H}_k = \{h_k(y_k; \boldsymbol{\pi}, \boldsymbol{\theta}_k, g) = \sum_{j=1}^{g} \pi_j f_k(y_k; \boldsymbol{\theta}_{jk}); \boldsymbol{\pi}, \boldsymbol{\theta}_k, g\}$ and $\mathcal{H} = \{h(\boldsymbol{y}; \boldsymbol{\pi}, \boldsymbol{\theta}, g) = \sum_{j=1}^{g} \pi_j \prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{jk}); \boldsymbol{\pi}, \boldsymbol{\theta}, g\}$. In finite mixture models, f_k (or F_k) is called the component function instead of the expert function named by LRMoE. The denseness condition is as follows.

Proposition 3.1. *The following statements are equivalent:*

1. *$\mathcal{H}$ is dense in the space of multivariate frequency (or severity) distributions.*
2. *$\mathcal{H}_k$ is dense in the space of univariate frequency (or severity) distributions for every $k = 1, \ldots, K$.*
3. *For every $k = 1, \ldots, K$ and $q \in \mathbb{Q}_k$, there exists a sequence of parameters $\{\boldsymbol{v}_q^{(n)}\}_{n=1,2,\ldots}$ such that $F_k(\cdot; \boldsymbol{v}_q^{(n)}) \xrightarrow{\mathcal{D}} q$ as $n \to \infty$, where $\mathbb{Q}_k = \{0, 1, 2, \ldots\}$ so that F_k corresponds to a frequency random variable (or $\mathbb{Q}_k = (0, \infty)$ so that F_k corresponds to a severity random variable).*

Proof. The rigorous proof is displayed in Appendix E but the proof idea is simple. For frequency distributions, if the component function is able to converge to any degenerate distributions (Statement 3 holds), then the mixture of all the degenerate functions with arbitrarily set component weights can obviously recover any frequency distributions. For severity distributions, we can partition the space $\mathbb{Q}_k$ into many small intervals and apply similar yet slightly more complicated approximation techniques as in the frequency case. ■

Proposition 3.1 shows that the class of finite mixture models is fully flexible if and only if the component functions F_k can converge to any degenerate distributions. For the LRMoE, the following theorem shows that the denseness conditions on the expert functions F_k is the same as that suggested in Proposition 3.1, subject to some mild conditions.

Theorem 3.3. *Let $\mathcal{G}_1(\mathcal{A})$ be a class of multivariate frequency (or severity) regression distributions. For each element $G^*(\mathcal{A}) \in \mathcal{G}_1(\mathcal{A})$ where $G^*(\mathcal{A}) := \{G^*(\cdot;\boldsymbol{x}); \boldsymbol{x} \in \mathcal{A}\}$, $\{G^*(\cdot;\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{A}}$ is tight and $G^*(\boldsymbol{y};\boldsymbol{x})$ is Lipschitz continuous on $\boldsymbol{x} \in \mathcal{A}$ for every $\boldsymbol{y}$. Also, assume that $\mathcal{A} = \{1\} \times [m_{\min}, m_{\max}]^P$. Let $\mathcal{G}_0(\mathcal{A})$ be the LRMoE with expert functions F_k, which also correspond to frequency (or severity) random variables. Then, $\mathcal{G}_0(\mathcal{A})$ is uniformly dense in $\mathcal{G}_1(\mathcal{A})$ if and only if Statement 3 of Proposition 3.1 is satisfied for the expert functions.*

Proof. The rigorous proof is presented in Appendix F and the proof idea is as follows. Similar to the proof of Theorem 3.2, we partition $[m_{\min}, m_{\max}]$ into L intervals with partition points $m_1, \ldots, m_L$. We first denote the covariates subspace $\mathcal{A}_0 := \{1\} \times \{m_1, \ldots, m_L\}^P$ and consider a regression distribution $G^*(\mathcal{A}_0)$ that consists of a total of L^P distributions. If Statement 3 of Proposition 3.1 holds, then there exists a sufficiently large g_ϵ such that (a number of L^P) g_ϵ-component finite mixture models can accurately approximate the corresponding L^P distributions (up to an error bound ϵ). Then, construct a $g_\epsilon \times L^P$-component LRMoE and use the same strategy as the proof of Theorem 3.1. Such an LRMoE will be very "close" to the regression distribution G^* on $\mathcal{A}_0$. Based on the Lipschitz continuity assumption, now we can implement the same strategy as the proof of Theorem 3.2 to extend the covariates space from $\mathcal{A}_0$ to $\mathcal{A}$. The tightness assumption is required in order that the weak convergence is uniform on $\mathcal{A}$, but the mathematical procedures are complicated so we put the details into the appendix. ■

Remark 3.4. The tightness condition required by Theorem 3.3 is indeed very mild. Let $H_k^*(\cdot;\boldsymbol{x})$ be the kth marginal distribution of $G^*(\cdot;\boldsymbol{x})$ with the corresponding random variable Y_k^*. Unless the tail of Y_k^* is artificially heavy, in practice it is safe to expect and assume that $E\left[\log(Y_k^*+1);\boldsymbol{x}\right]$ is finite and continuous on $\boldsymbol{x}$. Since $\boldsymbol{x}$ is bounded, $\{E\left[\log(Y_k^*+1);\boldsymbol{x}\right]\}_{\boldsymbol{x}\in\mathcal{A}}$ is also bounded. Then, Theorem 3.2.8 of Durrett (2010) shows that $\{H_k^*(\cdot;\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{A}}$ is tight for every $k = 1, \ldots, K$. From this, basic probabilistic arguments can show that $\{G^*(\cdot;\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{A}}$ is also tight.

To sum up, Proposition 3.1 and Theorem 3.3 demonstrate the connection between the denseness property of finite mixture models and that of the LRMoE. Such a connection is Statement 3 of Proposition 3.1, which is the possibility of the component/expert functions to be arbitrarily close to any degenerate distributions. Also, it is related to the ability of the expert functions to capture model under-dispersion, which will be discussed in Section 5 through a series of examples.

3.4. Limitations of denseness theory

In Sections 3.2 and 3.3, we have presented two major denseness properties for the LRMoE: the denseness in the space of the GMoE, which justifies the parsimony of the LRMoE, and the denseness in the space of any frequency/severity regression distributions, which illustrates the potential "full" versatility of the LRMoE to capture very complicated data characteristics and alleviates model misspecification risk prevalent in parametric modeling framework. These are the key theoretical motivations for us to bring the MoE framework for the insurance predictive modeling and to propose a reduced structure from the GMoE.

Despite the above desirable features, the denseness property does not guarantee the convergence rate, so there is no control on the number of components g. For instance, the proof of Theorem 3.1 constructs an LRMoE involving a large number of components ($g \times L^P$), which is obviously parameter inefficient and practically unfeasible. Several existing actuarial papers have already revealed the limitations of mixture-type models to fit extremely heavy-tailed distributions despite the denseness properties. For example, Verbelen et al. (2015) and Fung et al. (2019a) find that when the mixture of Erlang distribution (relatively light tailed) is fitted to data with extreme tail behavior, a prohibitively large number of components will be obtained that obviously overfits the tail and fails to extrapolate the tail-heaviness. As a result, the parameter efficiency and model fitting performance of the LRMoE can be very different among various choices of expert functions even if they all satisfy the denseness condition.

To address such a practical concern, it is essential to develop an efficient fitting algorithm and control the number of parameters, which will be presented in Fung et al. (2019b). We find empirically through various simulation studies and a real data analysis that (depending on the complexity of the data structures) only 5 to 36 components are sufficient to cater for very complicated features, including heterogeneous distribution (e.g. over- and under-dispersion) and dependence (e.g. positive and negative correlation) structures across business lines, as well as non-linear regression patterns with covariates interactions, verifying the applicability of denseness properties. Also, encountering data with extreme values, it may be critical to consider heavy-tailed distributions as the expert functions for the LRMoE (Section 5), or to transform the dataset before applying the LRMoE with light-tailed expert functions.

Another limitation of the denseness properties is that the denseness condition (Statement 3 of Proposition 3.1) is indeed not a mild restriction. Some commonly used expert functions, such as Poisson distribution, do not satisfy such a condition. To preserve the maximum benefits (i.e. "full flexibility" to capture complex structures) enjoyed by the proposed LRMoE, we may need to consider alternative less popular choices of expert functions, which will be discussed in detail in Section 5.

4. Other model properties

4.1. Marginalization properties

Closure under marginalization is a crucial desirable property because it directly relates to the mathematical tractability and interpretability of the model. For multivariate regression distributional models, we need to consider two types of marginalizations: response marginalization and covariates marginalization.

Proposition 4.1 (*Response Marginalization*)**.** *LRMoE in the form of Eqs.* (3) and (5) *is closed under response marginalization, i.e., each response marginal Y_k still follows a univariate LRMoE with expert function f_k. Furthermore, any p-variate ($p < K$) response marginal is still a p-variate LRMoE.*

Proof. For the sake of conciseness, only univariate case is presented. Let $H_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\Theta},g)$ be the distribution of Y_k. By Eq. (3), we have

$$H_k(\boldsymbol{y};\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\Theta},g) = \lim_{y_{k'}\to\infty;k'\neq k} \sum_{j=1}^{g} \pi_j(\boldsymbol{x};\boldsymbol{\alpha}) \prod_{k'=1}^{K} F_{k'}(y_{k'};\boldsymbol{\theta}_{jk'})$$

$$= \sum_{j=1}^{g} \pi_j(\boldsymbol{x};\boldsymbol{\alpha}) F_k(y_k;\boldsymbol{\theta}_{jk}). \quad ■$$

Consider a population of policyholders, each has their own covariates characteristics. If a policyholder is sampled from the whole population, the covariates $\boldsymbol{x}$ will become random. It is assumed that $\boldsymbol{x}$ follows a distribution function $W(\boldsymbol{x})$ with density $w(\boldsymbol{x})$. Then, the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{x}$ is given by

$$h(\boldsymbol{y}, \boldsymbol{x}; \boldsymbol{\alpha}, \boldsymbol{\Theta}, g) = h(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\Theta}, g) w(\boldsymbol{x}).$$

In practice, it is common that some significant covariates are missing in regression analysis because of the impossibility to collect all features of policyholders. Motivated by this, it is important to derive the distribution of $\boldsymbol{Y}$ conditioned on the observed covariates $\boldsymbol{x}^c$ ($h(\boldsymbol{y}; \boldsymbol{x}^c) := h(\boldsymbol{y}; \boldsymbol{x}^c, \boldsymbol{\alpha}, \boldsymbol{\Theta}, g)$), where $\boldsymbol{x}^c$ is a subset of the complete covariates $\boldsymbol{x}$. In particular, if $\boldsymbol{x}^c = x_0 = 1$, the distribution obtained is unconditional (i.e. no covariate information). Also denote $\boldsymbol{x}^u := \boldsymbol{x} \backslash \boldsymbol{x}^c$ as the missing covariates. This is called covariates marginalization.

Before that, we demonstrate the following general distributional result that links two-layer HME (hierarchical MoE) to LRMoE. We will show that covariate marginalization is a special case of the following result. As a by-product, such result also shows the motivation of using the LRMoE instead of the HME, a more complicated class of models. This is described in Remark 2.2. HME is itself an MoE written in the form

$$h(\boldsymbol{y}; \boldsymbol{x}) = \sum_{l=1}^{L} \pi_l(\boldsymbol{x}; \boldsymbol{\beta}) b_l(\boldsymbol{y}), \tag{9}$$

where π_l is still an exponential gating function and the expert function $b_l(\boldsymbol{y})$ is itself also an LRMoE with

$$b_l(\boldsymbol{y}) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}^{(l)}) \prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{jk}^{(l)}). \tag{10}$$

Proposition 4.2. *HME in the form of Eq. (9) can be expressed as LRMoE in the form of Eq. (5).*

Proof. From Eqs. (9) and (10), we have equation given in Box I. where $Q = \{j_1, \ldots, j_L = 1, \ldots, g; l_1, \ldots, l_L = 1, \ldots, L\}$ and $Q' = \{j'_1, \ldots, j'_L = 1, \ldots, g; l'_1, \ldots, l'_L = 1, \ldots, L\}$. Note that the last equality results from the label switching between l_1 and l_q. Therefore, $h(\boldsymbol{y}; \boldsymbol{x})$ is still an LRMoE with exponential gating functions. ■

Proposition 4.2 leads to the following marginalization result.

Proposition 4.3 (*Covariates Marginalization*)**.** *If $\boldsymbol{x}$ follows a discrete distribution $W(\boldsymbol{x})$ with a finite support and the covariates $\boldsymbol{x}$ are independent of each other, then the LRMoE in the form of Eqs. (3) and (5) is closed under covariates marginalization, i.e., $h(\boldsymbol{y}; \boldsymbol{x}^c)$ is still in an LRMoE with exponential linear gating functions for any dimensions and combinations of $\boldsymbol{x}^c$. Without any distributional restrictions on $\boldsymbol{x}$, $h(\boldsymbol{y}; \boldsymbol{x}^c)$ is still a g-component RMoE, but the gating functions are not necessarily exponential.*

Proof. Let D^u be the support of $\boldsymbol{x}^u$. Regarding the first statement, let $w(\boldsymbol{x}^u)$ and $w(\boldsymbol{x}^c)$ be the pmf of $\boldsymbol{x}^u$ and $\boldsymbol{x}^c$ respectively. We have

$$h(\boldsymbol{y}; \boldsymbol{x}^c) = \sum_{\boldsymbol{x}^u \in D^u} \frac{h(\boldsymbol{y}, \boldsymbol{x})}{w(\boldsymbol{x}^c)} = \sum_{j=1}^{g} \sum_{\boldsymbol{x}^u \in D^u} w(\boldsymbol{x}^u) \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \prod_{k=1}^{K} f_k(y_k, \boldsymbol{\theta}_{jk}). \tag{11}$$

Now, compare Eq. (11) to Eqs. (9) and (10). Choose L as the number of elements of D^u. Also, select $\beta_{\boldsymbol{x}^u,p} = [\log w(\boldsymbol{x}^u)] 1\{p = 0\}$ where $\beta_{\boldsymbol{x}^u,p}$ is the pth element of $\boldsymbol{\beta}_{\boldsymbol{x}^u}$. Further, we can write $\pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) = \exp\{\boldsymbol{\alpha}_j^{(\boldsymbol{x}^u)^T} \boldsymbol{x}^c\} / \sum_{j'=1}^{g} \exp\{\boldsymbol{\alpha}_{j'}^{(\boldsymbol{x}^u)^T} \boldsymbol{x}^c\} = \pi_j(\boldsymbol{x}^c; \boldsymbol{\alpha}^{(\boldsymbol{x}^u)})$ and choose $\alpha_{jp}^{(\boldsymbol{x}^u)} = \boldsymbol{\alpha}_j^{u\,T} \boldsymbol{x}^u 1\{p = 0\} + \alpha_{jp}^c$, where $\boldsymbol{\alpha}_j^u$ and $\boldsymbol{\alpha}_j^c$ are regression coefficients corresponding to $\boldsymbol{x}^u$ and $\boldsymbol{x}^c$ respectively, and α_{jp}^c is the pth element of $\boldsymbol{\alpha}_j^c$. Finally, choosing $\boldsymbol{\theta}_{jk}^{(\boldsymbol{x}^u)} = \boldsymbol{\theta}_{jk}$, Eq. (11) can be written as

$$h(\boldsymbol{y}; \boldsymbol{x}^c) = \sum_{\boldsymbol{x}^u \in D^u} \pi_{\boldsymbol{x}^u}(\boldsymbol{x}^c; \boldsymbol{\beta}) \sum_{j=1}^{g} \pi_j(\boldsymbol{x}^c; \boldsymbol{\alpha}^{(\boldsymbol{x}^u)}) \prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{jk}^{(\boldsymbol{x}^u)}),$$

which is in the form of Eqs. (9) and (10). Regarding the second statement, we have

$$\begin{aligned} h(\boldsymbol{y}; \boldsymbol{x}^c) &= \int_{\boldsymbol{x}^u \in D^u} h(\boldsymbol{y}; \boldsymbol{x}) dW(\boldsymbol{x}^u; \boldsymbol{x}^c) \\ &= \sum_{j=1}^{g} \int_{\boldsymbol{x}^u \in D^u} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) dW(\boldsymbol{x}^u; \boldsymbol{x}^c) \prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{jk}), \end{aligned}$$

where $W(\boldsymbol{x}^u; \boldsymbol{x}^c)$ is the distribution of $\boldsymbol{x}^u$ conditioned on $\boldsymbol{x}^c$. Hence, it is still a g-component LRMoE with non-exponential gating functions $\tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha}) := \int_{\boldsymbol{x}^u \in D^u} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) dW(\boldsymbol{x}^u; \boldsymbol{x}^c)$. ■

4.2. Moments and common measures of association

In this subsection, we show that the moments and commonly used measures of association, such as Kendall's tau and Spearman's rho, can be expressed in a simpler form which can be computed more easily under the class of LRMoE. Assume that $\boldsymbol{Y} = (Y_1, \ldots, Y_K)^T$ follows the LRMoE in the form of Eq. (5). Also, denote $Y_1^{(j)}, \ldots, Y_K^{(j)}$ as independent random variables with pdf/pmf $f_1(\cdot; \boldsymbol{\theta}_{j1}), \ldots, f_K(\cdot; \boldsymbol{\theta}_{jK})$ (i.e. the jth-component expert functions of $Y_1, \ldots, Y_K$) respectively. Further, aligning with the notations adopted in the proof of Proposition 4.3, we denote $\tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha}) := \int_{\boldsymbol{x}^u \in D^u} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) dW(\boldsymbol{x}^u; \boldsymbol{x}^c)$ as the non-exponential gating function, where $W(\boldsymbol{x}^u; \boldsymbol{x}^c)$ is the distribution of $\boldsymbol{x}^u$ conditioned on $\boldsymbol{x}^c$.

Proposition 4.4. *For $k = 1, \ldots, K$, let $h_k(\cdot)$ be a function where $E[h_k(Y_k^{(j)})] < \infty$ for every $j = 1, \ldots, g$. Then we have*

$$E\left[\prod_{k=1}^{K} h_k(Y_k) \middle| \boldsymbol{x}\right] = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \prod_{k=1}^{K} E\left[h_k(Y_k^{(j)})\right]. \tag{12}$$

Further, $E\left[\prod_{k=1}^{K} h_k(Y_k) | \boldsymbol{x}^c\right]$ can be obtained through replacing $\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})$ by $\tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha})$ in Eq. (12). Let $\mu_k^{(j)} = E\left[Y_k^{(j)}\right]$ and $\sigma_k^{(j)2} = Var\left[Y_k^{(j)}\right]$. For $k, k_1, k_2 = 1, \ldots, K$, some specific moments are given by

$$\begin{aligned} E\left[Y_k | \boldsymbol{x}\right] &= \sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \mu_k^{(j)}; \quad Var\left[Y_k | \boldsymbol{x}\right] \\ &= \sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) (\sigma_k^{(j)2} + \mu_k^{(j)2}) - \left(\sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \mu_k^{(j)}\right)^2; \end{aligned} \tag{13}$$

$$\begin{aligned} Cov\left[Y_{k_1}, Y_{k_2} | \boldsymbol{x}\right] &= \sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \mu_{k_1}^{(j)} \mu_{k_2}^{(j)} \\ &\quad - \left(\sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \mu_{k_1}^{(j)}\right) \left(\sum_{j=1}^{g} \pi_j(\boldsymbol{x}; \boldsymbol{\alpha}) \mu_{k_2}^{(j)}\right). \end{aligned} \tag{14}$$

$$
\begin{aligned}
h(\boldsymbol{y}; \boldsymbol{x}) &= \sum_{l=1}^{L}\sum_{j=1}^{g} \frac{\exp\{(\boldsymbol{\beta}_l + \boldsymbol{\alpha}_j^{(l)})^T\boldsymbol{x}\}}{\sum_{l'=1}^{L}\sum_{j'=1}^{g}\exp\{(\boldsymbol{\beta}_{l'} + \boldsymbol{\alpha}_{j'}^{(l)})^T\boldsymbol{x}\}} \prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{jk}^{(l)})\\
&= \sum_{j=1}^{g} \frac{\sum_{l=1}^{L}\exp\{(\boldsymbol{\beta}_l + \boldsymbol{\alpha}_j^{(l)})^T\boldsymbol{x}\}\prod_{l^*\neq l}\sum_{l'=1}^{L}\sum_{j'=1}^{g}\exp\{(\boldsymbol{\beta}_{l'} + \boldsymbol{\alpha}_{j'}^{(l^*)})^T\boldsymbol{x}\}\prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{jk}^{(l)})}{\prod_{l^*=1}^{L}\sum_{l'=1}^{L}\sum_{j'=1}^{g}\exp\{(\boldsymbol{\beta}_{l'} + \boldsymbol{\alpha}_{j'}^{(l^*)})^T\boldsymbol{x}\}}\\
&= \frac{\left(\dfrac{\sum_{q=1}^{L}\sum_{l_q=1}^{L}\sum_{j_q=1}^{g}\exp\{(\boldsymbol{\beta}_{l_q} + \boldsymbol{\alpha}_{j_q}^{(q)})^T\boldsymbol{x}\}1\{l_q = q\}}{\prod_{l^*\neq q}\sum_{l'=1}^{L}\sum_{j'=1}^{g}\exp\{(\boldsymbol{\beta}_{l'} + \boldsymbol{\alpha}_{j'}^{(l^*)})^T\boldsymbol{x}\}\prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{j_q k}^{(q)})}\right)}{\prod_{l^*=1}^{L}\sum_{l'=1}^{L}\sum_{j'=1}^{g}\exp\{(\boldsymbol{\beta}_{l'} + \boldsymbol{\alpha}_{j'}^{(l^*)})^T\boldsymbol{x}\}}\\
&= \frac{\sum_{Q}\exp\{\sum_{l^*=1}^{L}(\boldsymbol{\beta}_{l_{l^*}} + \boldsymbol{\alpha}_{j_{l^*}}^{(l^*)})^T\boldsymbol{x}\}\sum_{q=1}^{L}\prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{j_q k}^{(q)})1\{l_q = q\}}{\sum_{Q'}\exp\{\sum_{l^*=1}^{L}(\boldsymbol{\beta}_{l'_{l^*}} + \boldsymbol{\alpha}_{j'_{l^*}}^{(l^*)})^T\boldsymbol{x}\}}\\
&= \sum_{Q} \frac{\exp\{\sum_{l^*=1}^{L}(\boldsymbol{\beta}_{l_{l^*}} + \boldsymbol{\alpha}_{j_{l^*}}^{(l^*)})^T\boldsymbol{x}\}}{\sum_{Q'}\exp\{\sum_{l^*=1}^{L}(\boldsymbol{\beta}_{l'_{l^*}} + \boldsymbol{\alpha}_{j'_{l^*}}^{(l^*)})^T\boldsymbol{x}\}} \prod_{k=1}^{K} f_k(y_k; \boldsymbol{\theta}_{j_{l_1}k}^{(l_1)}),
\end{aligned}
$$

Box I.

$E\left[Y_k|\boldsymbol{x}^c\right]$, $Var\left[Y_k|\boldsymbol{x}^c\right]$ and $Cov\left[Y_{k_1}, Y_{k_2}|\boldsymbol{x}^c\right]$ are obtained through replacing $\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})$ by $\tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha})$.

Proof. Let $\boldsymbol{Z} = (Z_1, \ldots, Z_g)$ be a latent random vector, where $Z_j = 1$ when $\boldsymbol{Y}$ belongs to the jth component and $Z_j = 0$ otherwise for $j = 1, \ldots, g$. We have

$$
\begin{aligned}
E\left[\prod_{k=1}^{K} h_k(Y_k)\Big|\boldsymbol{x}\right] &= E\left[E\left[\prod_{k=1}^{K} h_k(Y_k)\Big|\boldsymbol{x}, \boldsymbol{Z}\right]\right]\\
&= E\left[\sum_{j=1}^{g}\prod_{k=1}^{K} E\left[h_k(Y_k^{(j)})\right]1\{Z_j = 1\}\right]\\
&= \sum_{j=1}^{g}\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})\prod_{k=1}^{K} E\left[h_k(Y_k^{(j)})\right].
\end{aligned}
$$

Also, the expression of $E\left[\prod_{k=1}^{K} h_k(Y_k)|\boldsymbol{x}^c\right]$ results from the covariates marginalization property (Proposition 4.3). The specific moments are easy to obtain after choosing several suitable $h_k(\cdot)$. ■

Proposition 4.4 shows that many commonly used moments for the LRMoE can be expressed analytically in terms of the moments corresponding to the individual expert functions, making LRMoE mathematically and computationally tractable. One may also be interested in the connections between the moment properties and the denseness properties under the LRMoE. We first introduce the following results on the characteristics of dispersion measures: dispersion ratio and coefficient of variation (CV), which are popular measures of the extent of variability for frequency and severity distributions respectively.

Proposition 4.5. *Let $\mathcal{D}_k(\boldsymbol{x}) = Var(Y_k|\boldsymbol{x})/E(Y_k|\boldsymbol{x})$ and $CV_k(\boldsymbol{x}) = SD(Y_k|\boldsymbol{x})/E(Y_k|\boldsymbol{x})$ be the dispersion ratio and the CV of Y_k conditioned on $\boldsymbol{x}$ respectively, where $SD(Y_k|\boldsymbol{x}) = \sqrt{Var(Y_k|\boldsymbol{x})}$. Also, define $\mathcal{D}_k^{(j)} = {\sigma_k^{(j)}}^2/\mu_k^{(j)}$ and $CV_k^{(j)} = \sigma_k^{(j)}/\mu_k^{(j)}$ as the dispersion ratio and CV of $Y_k^{(j)}$ respectively. Further, denote $\mathcal{D}_k^{\min} = \min_{j=1,\ldots,g}\mathcal{D}_k^{(j)}$ and $CV_k^{\min} = \min_{j=1,\ldots,g} CV_k^{(j)}$ as the minimum dispersion ratio and CV across all components. We have the following results.*

$$
\begin{aligned}
&Var(Y_k|\boldsymbol{x}) \geq \sum_{j=1}^{g}\pi_j(\boldsymbol{x}; \boldsymbol{\alpha}){\sigma_k^{(j)}}^2;\\
&Var(Y_k|\boldsymbol{x}^c) \geq E\left(Var(Y_k|\boldsymbol{x})|\boldsymbol{x}^c\right) \geq \sum_{j=1}^{g}\tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha}){\sigma_k^{(j)}}^2;
\end{aligned}
\tag{15}
$$

$$
\begin{aligned}
&SD(Y_k|\boldsymbol{x}) \geq \sum_{j=1}^{g}\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})\sigma_k^{(j)};\\
&SD(Y_k|\boldsymbol{x}^c) \geq E\left(SD(Y_k|\boldsymbol{x})|\boldsymbol{x}^c\right) \geq \sum_{j=1}^{g}\tilde{\pi}_j(\boldsymbol{x}^c; \boldsymbol{\alpha})\sigma_k^{(j)};
\end{aligned}
\tag{16}
$$

$$
\mathcal{D}_k(\boldsymbol{x}) \geq \mathcal{D}_k^{\min}; \qquad \mathcal{D}_k(\boldsymbol{x}^c) \geq \inf_{\boldsymbol{x}^u\in D^u}\mathcal{D}_k(\boldsymbol{x}) \geq \mathcal{D}_k^{\min}; \tag{17}
$$

$$
CV_k(\boldsymbol{x}) \geq CV_k^{\min}; \qquad CV_k(\boldsymbol{x}^c) \geq \inf_{\boldsymbol{x}^u\in D^u} CV_k(\boldsymbol{x}) \geq CV_k^{\min}. \tag{18}
$$

Proof. Eq. (15) can be obtained through simple algebraic manipulations on Eq. (13) and the usage of the conditional variance formula. Using Eqs. (15), (17) can be obtained as follows.

$$
\mathcal{D}_k(\boldsymbol{x}) \geq \frac{\sum_{j=1}^{g}\pi_j(\boldsymbol{x}; \boldsymbol{\alpha}){\sigma_k^{(j)}}^2}{\sum_{j=1}^{g}\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})\mu_k^{(j)}} = \frac{\sum_{j=1}^{g}\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})\mu_k^{(j)}\mathcal{D}_k^{(j)}}{\sum_{j=1}^{g}\pi_j(\boldsymbol{x}; \boldsymbol{\alpha})\mu_k^{(j)}} \geq \mathcal{D}_k^{\min}, \tag{19}
$$

$$
\begin{aligned}
\mathcal{D}_k(\boldsymbol{x}^c) &\geq \frac{\int_{\boldsymbol{x}^u\in D^u}\mathrm{Var}(Y_k|\boldsymbol{x})dW(\boldsymbol{x}^u; \boldsymbol{x}^c)}{\int_{\boldsymbol{x}^u\in D^u}E(Y_k|\boldsymbol{x})dW(\boldsymbol{x}^u; \boldsymbol{x}^c)}\\
&= \frac{\int_{\boldsymbol{x}^u\in D^u}E(Y_k|\boldsymbol{x})\mathcal{D}_k(\boldsymbol{x})dW(\boldsymbol{x}^u; \boldsymbol{x}^c)}{\int_{\boldsymbol{x}^u\in D^u}E(Y_k|\boldsymbol{x})dW(\boldsymbol{x}^u; \boldsymbol{x}^c)} \geq \inf_{\boldsymbol{x}^u\in D^u}\mathcal{D}_k(\boldsymbol{x}).
\end{aligned}
\tag{20}
$$

Eq. (16) results from Eq. (15) and Jensen's inequality. Then, Eq. (18) follows using similar techniques as Eqs. (19) and (20). ■

Eq. (17) shows that the dispersion measure of LRMoE is bounded below by that of the underlying expert functions. Let $\mathcal{D}_k(\boldsymbol{\theta_k})$ and $\mathrm{CV}_k(\boldsymbol{\theta_k})$ be the dispersion ratio and CV of a random variable with pdf/pmf $f_k(y_k; \boldsymbol{\theta}_k)$ respectively. If a dispersion measure of the expert function is bounded below by a positive constant among all parameters (e.g. $\inf_{\theta_k}\mathcal{D}_k(\boldsymbol{\theta_k}) = c > 0$), then the corresponding LRMoE is unable to capture distributions with even lower dispersion measures. Dispersion measures are important factors to be considered in both frequency and severity modeling. Many widely used models, such as Poisson and Exponential distributions, do not capture under-dispersion. Under-dispersed modeling highly relates to the denseness condition of the class of LRMoE on the expert functions (see Theorem 3.3). Recall denseness requires that the expert functions can be arbitrarily close to any degenerate distributions, which always have zero dispersion measures (except for the case where $y_k = 0$). Therefore, the key of model flexibility of LRMoE is the versatility for the expert functions to capture a broad range of under-dispersions. More details will be discussed in the next section with a series of motivating examples.

The correlation coefficient, a commonly used measure of dependence, can be obtained analytically from Eqs. (13) and (14). Since correlation coefficient assumes linear relationships and is sensitive to outliers, one may suggest other measures of association to tackle these problems. Among those, Kendall's tau and Spearman's rho are two of the most widely used measures. The results regarding the expressions of these measures are as follows.

Proposition 4.6. *Let $\tau_{k_1,k_2}(\boldsymbol{x})$ and $\rho_{k_1,k_2}(\boldsymbol{x})$ respectively be the Kendall's tau and Spearman's rho between Y_{k_1} and Y_{k_2} conditioned on the covariates $\boldsymbol{x}$. They are given by*

$$\begin{aligned}
&\tau_{k_1,k_2}(\boldsymbol{x})\\
&:= P\left[(Y_{k_1}-Y'_{k_1})(Y_{k_2}-Y'_{k_2})>0\right] - P\left[(Y_{k_1}-Y'_{k_1})(Y_{k_2}-Y'_{k_2})<0\right]\\
&= \sum_{j=1}^{g}\sum_{j'=1}^{g}\pi_j(\boldsymbol{x};\boldsymbol{\alpha})\pi_{j'}(\boldsymbol{x};\boldsymbol{\alpha})\left[P(Y_{k_1}^{(j)}>Y_{k_1}^{(j')})-P(Y_{k_1}^{(j)}<Y_{k_1}^{(j')})\right]\\
&\qquad\qquad\qquad\times\left[P(Y_{k_2}^{(j)}>Y_{k_2}^{(j')})-P(Y_{k_2}^{(j)}<Y_{k_2}^{(j')})\right],
\end{aligned}$$

$$\begin{aligned}
\rho_{k_1,k_2}(\boldsymbol{x}) &:= 3\left(P\left[(Y_{k_1}-Y'_{k_1})(Y_{k_2}-Y''_{k_2})>0\right]\right.\\
&\quad\left.-P\left[(Y_{k_1}-Y'_{k_1})(Y_{k_2}-Y''_{k_2})<0\right]\right)\\
&= 3\sum_{j=1}^{g}\sum_{j'=1}^{g}\sum_{j''=1}^{g}\pi_j(\boldsymbol{x};\boldsymbol{\alpha})\pi_{j'}(\boldsymbol{x};\boldsymbol{\alpha})\pi_{j''}(\boldsymbol{x};\boldsymbol{\alpha})\\
&\quad\times\left[P(Y_{k_1}^{(j)}>Y_{k_1}^{(j')})-P(Y_{k_1}^{(j)}<Y_{k_1}^{(j')})\right]\\
&\quad\times\left[P(Y_{k_2}^{(j)}>Y_{k_2}^{(j'')})-P(Y_{k_2}^{(j)}<Y_{k_2}^{(j'')})\right],
\end{aligned}$$

where Y'_k and Y''_k are iid copies of Y_k. Similarly, $\tau_{k_1,k_2}(\boldsymbol{x}^c)$ and $\rho_{k_1,k_2}(\boldsymbol{x}^c)$ can be obtained through replacing $\pi_j(\boldsymbol{x};\boldsymbol{\alpha})$ by $\tilde{\pi}_j(\boldsymbol{x}^c;\boldsymbol{\alpha})$. The probabilities $P(Y_k^{(j)}>Y_k^{(j')})$ can be computed as: $P(Y_k^{(j)}>Y_k^{(j')})=\sum_{y=1}^{\infty}F_k(y-1;\boldsymbol{\theta}_{j'k})f_k(y;\boldsymbol{\theta}_{jk})$ for frequency expert functions or $P(Y_k^{(j)}>Y_k^{(j')})=\int_0^{\infty}F_k(y;\boldsymbol{\theta}_{j'k})dF_k(y;\boldsymbol{\theta}_{jk})$ for severity expert functions.

Proof. For conciseness, we only derive $\tau_{k_1,k_2}(\boldsymbol{x})$. Let $\boldsymbol{Z}=(Z_1,\ldots,Z_g)$ and $\boldsymbol{Z}'=(Z'_1,\ldots,Z'_g)$ be independent latent random vectors, where $Z_j=1$ (or $Z'_j=1$) when $\boldsymbol{Y}$ (or $\boldsymbol{Y}'=(Y'_1,\ldots,Y'_K)$ belongs to the jth component and $Z_j=0$ (or $Z'_j=0$) otherwise for $j=1,\ldots,g$. We have

$$\begin{aligned}
&\tau_{k_1,k_2}(\boldsymbol{x}) = E\left[P\left[(Y_{k_1}-Y'_{k_1})(Y_{k_2}-Y'_{k_2})>0|\boldsymbol{Z},\boldsymbol{Z}'\right]\right.\\
&\quad\left.-P\left[(Y_{k_1}-Y'_{k_1})(Y_{k_2}-Y'_{k_2})<0|\boldsymbol{Z},\boldsymbol{Z}'\right]\right]\\
&= \sum_{j=1}^{g}\sum_{j'=1}^{g}\pi_j(\boldsymbol{x};\boldsymbol{\alpha})\pi_{j'}(\boldsymbol{x};\boldsymbol{\alpha})\Big[P\left((Y_{k_1}^{(j)}-Y_{k_1}^{(j')})(Y_{k_2}^{(j)}-Y_{k_2}^{(j')})>0\right)\\
&\quad-P\left((Y_{k_1}^{(j)}-Y_{k_1}^{(j')})(Y_{k_2}^{(j)}-Y_{k_2}^{(j')})<0\right)\Big].
\end{aligned}$$

The result follows by the independence between $(Y_{k_1}^{(j)}-Y_{k_1}^{(j')})$ and $(Y_{k_2}^{(j)}-Y_{k_2}^{(j')})$. ■

5. Specific choices of expert functions

So far, we have considered a class of LRMoE without specifying the functional form of the expert function. In applications, it is necessary to choose an expert function prior to the model fitting process, so we discuss some of the possible choices of expert functions throughout this section.

While the proposed LRMoE possesses several important desirable properties, we should be aware of its shortcoming discussed in Section 3.4 that not all expert functions make the LRMoE "fully flexible". Therefore, it is crucial to study various choices of expert functions. In particular, it is desirable that the denseness condition (Statement 3 of Proposition 3.1) is fulfilled, ensuring the versatility of the corresponding LRMoE. We first provide the following proposition, which will facilitate the checking of the denseness condition.

Proposition 5.1. *If for every $k=1,\ldots,K$ and $q\in\mathbb{Q}_k$, there exists a sequence of parameters $\{\boldsymbol{\nu}_q^{(n)}\}_{n=1,2,\ldots}$ such that $E(Y_k^{(n)})\to q$ and $Var(Y_k^{(n)})\to 0$ as $n\to\infty$, where $Y_k^{(n)}$ is a frequency (or severity) random variable with cdf $F_k(y_k;\boldsymbol{\nu}_q^{(n)})$ and $\mathbb{Q}_k$ is defined in Proposition 3.1, then Statement 3 of Proposition 3.1 holds.*

Proof. For every $k=1,\ldots,K$ and $q\in\mathbb{Q}_k$, we have

$$E\left[(Y_k^{(n)}-q)^2\right]=\mathrm{Var}\left[Y_k^{(n)}\right]+\left(E\left[Y_k^{(n)}-q\right]\right)^2\to 0.$$

Therefore, $Y_k^{(n)}$ converges to q in the L^2-norm. From basic probability theory, this implies that $Y_k^{(n)}$ converges to q in distribution. ■

Corollary 5.1. *Let $\mathcal{D}(Y_k^{(n)})=Var(Y_k^{(n)})/E(Y_k^{(n)})$ and $CV(Y_k^{(n)})=SD(Y_k^{(n)})/E(Y_k^{(n)})$ be the dispersion ratio and CV of $Y_k^{(n)}$ respectively. If for every $k=1,\ldots,K$ and $q\in\mathbb{Q}_k$, there exists a sequence of parameters $\{\boldsymbol{\nu}_q^{(n)}\}_{n=1,2,\ldots}$ such that $E(Y_k^{(n)})\to q$ and $\mathcal{D}(Y_k^{(n)})\to 0$ (or $CV(Y_k^{(n)})\to 0$) as $n\to\infty$, then Statement 3 of Proposition 3.1 holds.*

Corollary 5.1 suggests that expert functions that can capture any highly under-dispersed distributions will satisfy the denseness condition. Although the reverse of Corollary 5.1 is not always true (i.e. it is still theoretically possible to construct a series of over-dispersed distributions converging to degeneracy), in practice most commonly used expert functions not satisfying the condition in Corollary 5.1 fail to fulfill the denseness condition.

Several choices of expert functions (for both frequency/severity distributions) are discussed in the following subsections. For notational convenience, we use Y_k as a random variable following an expert function (instead of the LRMoE) in the following examples. Also, we assume that the same class of expert functions f_k are used across marginals, i.e., if f_1 is a Gamma expert function, then $f_2,\ldots,f_k$ are all Gamma expert functions. One can easily extend the results by allowing for different classes of expert functions across marginals.

5.1. Frequency distributions

For ratemaking purpose, it is legitimate that only the aggregate claim amounts (severities) for individual policyholders are modeled, because these represent directly the amount to be paid by the insurers. However, it is still desirable for insurance companies to also keep track on the claim frequencies because this can help the insurers to fulfill regulatory requirements and to get more insights on the claim characteristics that are useful in decision-making process. Details about the arguments for modeling claim frequencies can be found in Frees et al. (2016).

In frequency modeling, since over-dispersed data are more common in practice, the most used frequency models are associated with over-dispersed distributions and are not designed to cater for under-dispersions. Therefore, such frequency expert functions do not fulfill the denseness condition, raising a challenge for us to identify expert functions that make the LRMoE "fully flexible".

5.1.1. The $(a, b, 0)$ and $(a, b, 1)$ classes of distributions

The $(a, b, 0)$ class of distributions is a class of frequency models commonly used in actuarial practice, because it contains several popular frequency models (such as Poisson and Negative Binomial distributions) and has the following simple recursive relationships on the pmf $p_y := P(Y_k = y)$:

$$\frac{p_y}{p_{y-1}} = a + \frac{b}{y}, \qquad y = 1, 2, 3, \ldots,$$

where a and b are constants and $\sum_{y=0}^{\infty} p_y = 1$. It can be shown that only Poisson, Negative Binomial and Binomial distributions satisfy the above $(a, b, 0)$ relationship. Since Binomial distribution always has a finite support, which does not fit into our Assumption 2.1 for frequency distributions, it is not within the scope of this paper.

Negative binomial (NB) distribution is an $(a, b, 0)$ class model with $a = \beta/(1+\beta)$ and $b = (r-1)\beta/(1+\beta)$, where $r > 0$ and $\beta > 0$ are the parameters of the NB distribution with $E(Y_k) = r\beta$ and $\mathrm{Var}(Y_k) = r\beta(1+\beta)$. Since the dispersion ratio $\mathcal{D}(Y_k) = 1 + \beta > 1$, NB is suitable to model over-dispersed data. In the context of general insurance, such model can explain the unobserved heterogeneity of the claim behavior among policyholders because it can be expressed as a mixed Poisson distribution. Its pmf is given by

$$f_k(y; r, \beta) := p_y = \binom{y+r-1}{y}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^y; \qquad y = 0, 1, \ldots, \quad r > 0 \text{ and } \beta > 0. \tag{21}$$

Since the dispersion ratio of NB distribution is bounded below by 1, it is not a model that captures data under-dispersion. To demonstrate that the denseness condition is not satisfied, we first show that

$$\frac{p_{y+1}/p_y}{p_y/p_{y-1}} = \frac{y(y+r)}{(y+1)(y+r-1)} \geq \frac{1}{2} \tag{22}$$

for $y \geq 1$. We now assume that $p_y \geq 1/2$, then we have

$$\frac{p_{y+1}/p_y}{p_y/p_{y-1}} \leq \frac{(1-p_y-p_{y-1})p_{y-1}}{p_y^2} \leq 4\left(\frac{1}{2} - p_{y-1}\right)p_{y-1} \leq \frac{1}{4}, \qquad y = 1, 2, \ldots,$$

which contradicts Eq. (22). Therefore, p_y is bounded above by 1/2 for any parameter settings (for $y = 1, 2, \ldots$) and the denseness condition fails.

Note that other $(a, b, 0)$ class distributions like Poisson, Geometric and Pascal distributions are special or limiting cases of Negative Binomial distribution. From the result above, we can easily show that denseness condition also fails for such classes of expert functions.

In practice, it is common that most policyholders never file any claims, so the probability of zero claims is very large. This motivates extending the $(a, b, 0)$ class of distributions to the $(a, b, 1)$ class so that excess zeros can be captured. Under the $(a, b, 1)$ class, Eq. (21) only holds for $y = 2, 3, \ldots$ and p_0 can be freely adjusted. Note that the $(a, b, 1)$ class contains much richer types of distributions, including any zero-modified forms of the $(a, b, 0)$ class distributions, extended truncated NB (ETNB) and logarithmic distributions. Using similar arguments as for the $(a, b, 0)$ class distributions, we can easily show that they still do not satisfy the denseness condition.

With simple closed-form expressions, the usage of the aforementioned class of models as the expert functions allows easy implementation and transparent interpretation. On the other hand, if our main goal is to achieve 'full flexibility" in frequency modeling using the LRMoE framework, we may need to pursue alternative less popular classes of expert functions, which are presented in the following subsections.

5.1.2. Conway–Maxwell–Poisson (CMP) distribution: $Y_k \sim$ CMP (λ, ν)

To allow for more flexible frequency modeling, Conway and Maxwell (1962) propose Conway–Maxwell–Poisson (CMP) distribution, an extension to Poisson distribution, to cater for both over-dispersion and under-dispersion. Its pmf is

$$f_k(y; \lambda, \nu) = \frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda, \nu)}, \quad Z(\lambda, \nu) = \sum_{j=0}^{\infty}\frac{\lambda^j}{(j!)^\nu}; \qquad y = 0, 1, \ldots, \quad \lambda > 0 \text{ and } \nu > 0.$$

Note that ν is called the dispersion parameter. For $\nu > 1$, $\nu = 1$ and $\nu < 1$, the model exhibits under-, equil- and over-dispersion respectively. To check the denseness condition, we choose a sequence of parameters $\lambda^{(n)} = (q+0.5)^n$ and $\nu^{(n)} = n$ for each $q = 0, 1, \ldots$. Then, we have

$$\begin{aligned}
f_k(q; \lambda^{(n)}, \nu^{(n)}) &= \left(\sum_{j=0}^{\infty}\frac{(q+0.5)^{nj}/(j!)^n}{(q+0.5)^{nq}/(q!)^n}\right)^{-1} \\
&= \left(\sum_{j=0}^{\infty}\left(\frac{(q+0.5)^{(j-q)}q!}{j!}\right)^n\right)^{-1}, \\
\lim_{n\to\infty} f_k(q; \lambda^{(n)}, \nu^{(n)}) &= \left(\sum_{j=0}^{\infty}\lim_{n\to\infty}\left(\frac{(q+0.5)^{(j-q)}q!}{j!}\right)^n\right)^{-1} \\
&= \left(\sum_{j=0}^{\infty} 1\{j = q\}\right)^{-1} = 1.
\end{aligned} \tag{23}$$

Note that the second equality of Eq. (23) arises from the fact that $(q+0.5)^{(j-q)}q!/j! < 1$ when $j \neq q$ and $= 1$ when $j = q$. Hence, denseness condition is satisfied for CMP expert functions. Despite this, CMP has some undesirable characteristics that hinder its applications to claim frequency modeling. There exists no physical interpretations for CMP. Also, The normalizing constant $Z(\lambda, \nu)$ of the pmf consists of a summation with infinite terms, causing high computational costs for model calibrations, especially when we consider the LRMoE with CMP expert functions. These motivate us to consider the following classes of frequency expert functions, which are constructed based on a transformation of severity distributions.

5.1.3. Renewal count model

Renewal Count Model, which can potentially capture both under- and over-dispersions, is introduced by Winkelmann (1995). The frequency distribution is modeled through the waiting times $\{\tau_s; s \in \mathbb{N}\}$ between the $(s-1)th$ and the sth event. Also, let $\nu_s = \sum_{s'=1}^{s}\tau_{s'}$ be the time of occurrence of the sth event. Then, the number of events occurring up to time T is given by

$N_T = \sup_{s\in\mathbb{N}}\{s;\, \nu_s \leq T\}$. We assume that $\{\tau_s\}_{s=1,2,\ldots}$ iid following a severity distribution with pdf $g_\tau(t;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter. Without much loss of generality, we assume $T = 1$. Then, the corresponding Renewal Count expert function is written as

$$\begin{aligned} f_k(y;\boldsymbol{\theta}) &= P(\nu_y \leq 1,\, \nu_{y+1} > 1;\boldsymbol{\theta}) \\ &= P(\nu_y \leq 1;\boldsymbol{\theta}) - P(\nu_{y+1} \leq 1;\boldsymbol{\theta}) \\ &= g_{\tau^{*y}}(1;\boldsymbol{\theta}) - g_{\tau^{*(y+1)}}(1;\boldsymbol{\theta}); \qquad y = 0,1,2,\ldots, \end{aligned} \tag{24}$$

where $g_{\tau^{*y}}$ is the pdf of the y-fold convolution of τ_s. Renewal Count Model has a physical interpretation in general insurance, when we consider claims as an arrival process but we only observe the total number of claims within a time period for each policyholder. To check the denseness condition, the following theorem is introduced.

Theorem 5.1. *For Renewal Count Model, if g_τ satisfies the denseness condition (i.e. for all $q > 0$, there exists a sequence of parameters $\{\boldsymbol{\theta}^{(n)}\}_{n=1,2,\ldots}$ such that $\tau^{(n)} \xrightarrow{\mathcal{D}} q$, where $\tau^{(n)}$ has pdf $g_\tau(t;\boldsymbol{\theta}^{(n)})$), then the corresponding Renewal Count expert functions also satisfy the denseness condition.*

Proof. For any $q \in \{0,1,\ldots,\}$, choose $\boldsymbol{\theta}^{(n)}$ such that $\tau^{(n)} \xrightarrow{\mathcal{D}} 1/(q+0.5)$. Then, $\nu_q^{(n)} := \sum_{s'=1}^{q}\tau_{s'}^{(n)} \xrightarrow{\mathcal{D}} q/(q+0.5) < 1$ and $\nu_{q+1}^{(n)} := \sum_{s'=1}^{q+1}\tau_{s'}^{(n)} \xrightarrow{\mathcal{D}} (q+1)/(q+0.5) > 1$, where $\tau_{s'}^{(n)}$ is an iid copy of $\tau^{(n)}$. By the definition of convergence in distribution, we have $P(\nu_q \leq 1;\boldsymbol{\theta}^{(n)}) \to 1$ and $P(\nu_{q+1} \leq 1;\boldsymbol{\theta}^{(n)}) \to 0$ as $n \to \infty$. From Eq. (24), we have $f_k(q;\boldsymbol{\theta}^{(n)}) \xrightarrow{n\to\infty} 1$, so the result follows. ■

From Theorem 5.1 and as shown by the following subsection (Section 5.2) that many severity expert functions indeed satisfy the denseness condition, we can see that many Renewal Count Models, such as Gamma Count Model and Weibull Count Model, satisfy the denseness condition. Among these models, we find that Erlang Count Model is desirable in terms of mathematical tractability. It has a closed-form expression for the pmf:

$$f_k(y;\boldsymbol{\theta}) = e^{-\beta}\sum_{b=0}^{m-1}\frac{\beta^{my+b}}{(my+b)!}, \qquad y = 0,1,2,\ldots, \tag{25}$$

where $\boldsymbol{\theta} = (m,\beta)$, $m \in \{1,2,\ldots\}$ is the shape parameter and $\beta > 0$ is the rate parameter. The applications of LRMoE with Erlang Count expert function to multivariate insurance claim frequency regression will be discussed in the subsequent paper (Fung et al., 2019b).

5.1.4. Discretized severity model

Another approach to transform a severity distribution to a frequency distribution is to discretize a severity random variable. Let $\tilde{Y}_k$ be the severity random variable with density $g_k(y;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter. Denote Y_k as the discretized version of $\tilde{Y}_k$, i.e. $Y_k = \lfloor \tilde{Y}_k \rfloor$, where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x. Then, the Discretized Severity expert function is given by

$$f_k(y;\boldsymbol{\theta}) = G_k(y+1;\boldsymbol{\theta}) - G_k(y;\boldsymbol{\theta}); \qquad y = 0,1,2,\ldots, \tag{26}$$

where G_k is the cdf of $\tilde{Y}_k$. To check the denseness condition for such frequency expert functions, we have the following theorem:

Theorem 5.2. *For Discretized Severity Model, if g_k satisfies the denseness condition (see the definition of denseness condition in Theorem 5.1), then the corresponding Discretized Severity expert functions also satisfy the denseness condition.*

Proof. For any $q \in \{0,1,\ldots,\}$, choose $\boldsymbol{\theta}^{(n)}$ such that $\tilde{Y}_k^{(n)} \xrightarrow{\mathcal{D}} (q+0.5)$, where $\tilde{Y}_k^{(n)}$ has pdf $g_k(y,\boldsymbol{\theta}^{(n)})$. By the definition of convergence in distribution, we have $G_k(q+1;\boldsymbol{\theta}^{(n)}) \to 1$ and $G_k(q;\boldsymbol{\theta}^{(n)}) \to 0$ as $n \to \infty$. From Eq. (26), $f_k(q;\boldsymbol{\theta}^{(n)}) \xrightarrow{n\to\infty} 1$, so the result follows. ■

From Theorem 5.2 and the following subsection, we can see that Discrete Gamma Model, Discrete Weibull Model etc. satisfy the denseness condition. Many of them have closed-form pmf, making them desirable in terms of mathematical tractability. One shortcoming of Discretized Severity Model is that they do not have a good interpretation in the context of general insurance. Instead, such models are more meaningful in survival analysis, especially when the lifetime of an individual is measured and recorded only in a discrete (e.g. monthly) basis (Chakraborty and Chakravarty, 2012).

5.2. Severity distributions

In the following examples, we will show that unlike frequency expert functions, a wide range of commonly used severity expert functions satisfy the denseness condition. In other words, the denseness condition is less restrictive in the context of severity regression.

5.2.1. Gamma distribution: $Y_k \sim G(m,\theta)$

Gamma distribution has long been popular in modeling light-to-medium-tailed insurance losses. Its density function with shape parameter m and scale parameter θ is given by:

$$f_k(y;m,\theta) = \frac{1}{\Gamma(m)\theta^m}y^{m-1}e^{-y/\theta}; \qquad y > 0, m > 0 \text{ and } \theta > 0,$$

where $\Gamma(z) = \int_0^\infty u^{z-1}e^{-u}du$ is a Gamma function. The moments are given by $E(Y_k) = m\theta$ and $\text{Var}(Y_k) = m\theta^2$. Therefore, Gamma distribution covers a full range of dispersion ratios $\mathcal{D}(Y_k) = \theta$. Let $Y_k^{(n)} \sim G(m^{(n)},\theta^{(n)})$. For each $q > 0$, set the parameter sequence $m^{(n)} = nq$ and $\theta^{(n)} = 1/n$. Then, we have $E(Y_k^{(n)}) = q$ for all n and also $\text{Var}(Y_k^{(n)}) \to 0$. From Proposition 5.1, denseness condition of LRMoE holds for Gamma expert functions.

Remark 5.1. Because of the analytical tractability of the class of Erlang-based distributions, which corresponds to Gamma distribution with the shape parameter m being an integer, Lee and Lin (2012) proposed multivariate Erlang mixture models as a suitable candidate to model correlated insurance losses. In such model, it is further assumed that the shape parameter θ is constant across all marginals and mixture components. By Theorem 2.1 of Lee and Lin (2012), Proposition 3.1 and Theorem 3.3, denseness property for the corresponding class of LRMoE holds.

5.2.2. Weibull Distribution: $Y_k \sim W(m,\theta)$

Weibull distribution allows for a more flexible tail modeling than Gamma distribution. Its density function with shape parameter m and scale parameter θ is given by

$$f_k(y;m,\theta) = \frac{m}{\theta}\left(\frac{y}{\theta}\right)^{m-1}e^{-(y/\theta)^m}; \qquad y > 0, m > 0 \text{ and } \theta > 0.$$

If $m > 1$, Weibull distribution will have a lighter tail than Gamma distribution, and vice versa. The moments are given by $E(Y_k) = \theta\Gamma(1+1/m)$ and $\text{Var}(Y_k) = \theta^2\left[\Gamma(1+2/m) - (\Gamma(1+1/m))^2\right]$. By choosing a sequence of parameters $m^{(n)} = n$ and $\theta^{(n)} = q$ for each $q > 0$, it is easy to see that Weibull expert functions satisfy the denseness condition.

5.2.3. *Log-normal distribution: $Y_k \sim LN(\mu, \sigma^2)$*

Log-normal distribution is suitable in modeling heavier-tailed losses. Its density with mean parameter μ and variance parameter σ^2 is

$$f_k(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}y}e^{(\log y-\mu)^2/2\sigma^2}; \quad y > 0,\ \mu \in \mathbb{R} \text{ and } \sigma^2 > 0.$$

The moments are $E(Y_k) = \exp\{\mu + \sigma^2/2\}$ and $\mathrm{Var}(Y_k) = (\exp\{\sigma^2\} - 1)\exp\{2\mu + \sigma^2\}$. Choosing a sequence of parameters $\mu^{(n)} = \log q$ and $\sigma^{2(n)} = 1/n$ for each $q > 0$, denseness condition follows.

5.2.4. *Inverse burr distribution: $Y_k \sim IBurr(\tau, \theta, \gamma)$*

In modeling catastrophic losses, we need a distribution that can cater for very heavy tails. Inverse Burr distribution is a potential suitable candidate because it has a polynomial tail, which aligns well with extreme value theory. Its pdf/cdf with shape parameters τ and γ and scale parameter θ are

$$f_k(y;\tau,\theta,\gamma) = \frac{\tau\gamma(y/\theta)^{\tau\gamma}}{y(1+(y/\theta)^{\gamma})^{\tau+1}};$$

$$F_k(y;\tau,\theta,\gamma) = \left[1+\left(\frac{y}{\theta}\right)^{-\gamma}\right]^{-\tau}; \quad y,\tau,\theta,\gamma > 0.$$

Note that $E(Y_k) = \infty$ when $0 < \gamma < 1$, meaning a possibly for Inverse Burr LRMoE to capture infinite mean models. For its denseness property, we check F_k directly with a sequence of parameters $\tau^{(n)} = 1$, $\theta^{(n)} = q$ and $\gamma^{(n)} = n$ for each q. Then, $F_k(y;\tau^{(n)},\theta^{(n)},\gamma^{(n)}) \xrightarrow{n\to\infty} 1\{y \geq q\} - 0.5 \times 1\{y = q\}$. Since $y = q$ is not a continuity point, denseness condition holds.

5.2.5. *Exponential distribution: $Y_k \sim Exp(\lambda)$*

Exponential distribution is a simple, mathematically tractable and interpretable model for insurance loss modeling. However, this one-parameter expert function is not flexible enough in terms of dispersion modeling. With the density function $f_k(y;\lambda) = \lambda\exp\{-\lambda y\}$, the CV is always constant $\mathrm{CV}(Y_k) = 1$. Hence, the condition in Corollary 5.1 fails. This motivates us to prove that denseness condition is violated under exponential expert function. Standard calculus shows that $\sup_{\lambda>0} f_k(y;\lambda) = e^{-1}/y$ for all $y > 0$, meaning that the density is bounded above by a finite and continuous curve on $y \in (0,\infty)$. Therefore, for a sufficiently small interval $Q \subseteq (0,\infty)$, $P(Y_k \in Q) < 0.5$ for any λ, showing that denseness condition is not satisfied.

5.2.6. *Type II Pareto distribution: $Y_k \sim Pareto(\alpha, \theta)$*

Similar to Inverse Burr distribution, Pareto distribution also consists of polynomial tail. Therefore, it is an alternative candidate model for extreme tail risks. Its density with shape parameter α and scale parameter θ is

$$f_k(y;\alpha,\theta) = \frac{\alpha\theta^{\alpha}}{(y+\theta)^{\alpha+1}}; \qquad y > 0, \alpha > 0 \text{ and } \theta > 0.$$

Also, we have $E(Y_k) = \theta/(\alpha - 1)$ for $\alpha > 1$ and $\mathrm{Var}(Y_k) = \theta^2\alpha/(\alpha-1)^2(\alpha-2)$ for $\alpha > 2$. Assuming finite moments, the CV is $\mathrm{CV}(Y_k) = \alpha/(\alpha-2)$, which is bounded below by 1. Failing to fulfill the condition in Corollary 5.1, Pareto expert function is not flexible enough to cater for under-dispersions. For denseness condition, since f_k is a decreasing function of y regardless of the parameters, we have $P(Y_k \in (0,1)) \geq P(Y_k \in (1,2))$ and hence $P(Y_k \in (1,2)) \leq 0.5$ for all $\alpha, \theta > 0$. Therefore, Pareto experts fail to fulfill denseness condition.

6. Concluding remarks

In this paper, we propose a class of logit-weighted reduced mixture of experts (LRMoE) models for multivariate insurance claim frequency or severity regression. It acquires a natural interpretation that the classification of latent homogeneous subgroups is affected by the policyholder's risk profile. We also formulate the denseness property with regard to regression and demonstrate that LRMoE may be interpreted as a "fully flexible" model under some appropriate choices of the expert functions. With various marginalization properties and with simplified form of expressions for moments and common measures of association, LRMoE is mathematically and computationally tractable.

In Fung et al. (2019b), we consider how the proposed model can fit a real insurance dataset consisting of multivariate claim frequencies for each policyholder with covariates. As discussed in Section 5.1.3, the Erlang Count distribution is chosen as the expert function for LRMoE because it satisfies the denseness condition and it is mathematically tractable. An Expectation–Conditional Maximization (ECM) algorithm is developed to fit the proposed model to the dataset so that model parameters can be estimated. Through several simulation studies, the efficiency of the proposed algorithm and the flexibility of the proposed model are verified. The proposed model also fits very well a real automobile insurance dataset, which has rather complicated data characteristics.

With respect to insurance severity modeling, it is common in practice to see excessive zeros. In this case, the LRMoE can be extended to incorporate zero-inflated components for the severity expert functions, so that the resulting expert function is a combination of discrete and continuous distributions. Having proved the denseness properties for both frequency and severity models under the proposed LRMoE, similar denseness properties still hold for such an extension.

Acknowledgment

This work was partly supported by Natural Sciences and Engineering Research Council of Canada [RGPIN 284246, RGPIN-2017-06684].

Appendix A. Relationship between pointwise and uniform convergence

This section proves the assertion stated in Remark 3.1 that weakly convergence implies uniform convergence for frequency or severity distributions under Assumption 2.1 or 2.2. We first recall from Lemma 3.2 of Rao (1962) that weakly convergence implies uniform convergence for any distributions with continuous cdf. Under Assumption 2.2, such a result holds naturally for severity distributions. For frequency distributions, we first define (multivariate frequency) random vectors $\boldsymbol{Y}_n \sim G_n$ and $\boldsymbol{Y} \sim F$ with $G_n \xrightarrow{\mathcal{D}} F$, where G_n and F are defined in accordance to Remark 3.1. We now introduce perturbed random vectors $\tilde{\boldsymbol{Y}}_n := \boldsymbol{Y}_n - \boldsymbol{U}_n \sim \tilde{G}_n$ and $\tilde{\boldsymbol{Y}} := \boldsymbol{Y} - \boldsymbol{U} \sim \tilde{F}$ where $\boldsymbol{U}_n$ and $\boldsymbol{U}$ iid follow Uniform distribution on $[0,1]^K$. Note that $\tilde{\boldsymbol{Y}}_n$ and $\tilde{\boldsymbol{Y}}$ are continuous random vectors. Basic probability theory yields $\tilde{G}_n \xrightarrow{\mathcal{D}} \tilde{F}$. We now have

$$\begin{aligned}\sup_{\boldsymbol{y}\in\mathbb{R}^K}|G_n(\boldsymbol{y}) - F(\boldsymbol{y})| &= \sup_{\boldsymbol{y}\in\mathbb{N}^K}|G_n(\boldsymbol{y}) - F(\boldsymbol{y})| \\ &= \sup_{\boldsymbol{y}\in\mathbb{N}^K}|\tilde{G}_n(\boldsymbol{y}) - \tilde{F}(\boldsymbol{y})| \\ &\leq \sup_{\boldsymbol{y}\in\mathbb{R}^K}|\tilde{G}_n(\boldsymbol{y}) - \tilde{F}(\boldsymbol{y})| \to 0 \text{ as } n \to \infty,\end{aligned}$$

where $\mathbb{N} = \{0,1,\ldots\}$ corresponds to a set of natural numbers and the convergence resulted from the uniform convergence property by Lemma 3.2 of Rao (1962). Therefore, weakly convergence implies uniform convergence for frequency distributions as well.

Appendix B. Proof of Lemma 3.1

Denote $h^{(l,p)}(x_p) = \lambda_0^{(l,p)} + \lambda_p^{(l,p)} x_p$ for $l = 1, 2, \ldots, L$ and $p = 1, 2, \ldots, P$. For each p, choose $\lambda_0^{(l,p)}$ and $\lambda_p^{(l,p)}$ by the following scheme:

1. Arbitrarily choose $\lambda_0^{(1,p)}$ and $\lambda_p^{(1,p)}$.
2. Choose $\lambda_p^{(l,p)} > \lambda_p^{(l-1,p)}$ for $l = 2, 3, \ldots, L$.
3. Choose $\lambda_0^{(l,p)}$ satisfying $\lambda_0^{(l-1,p)} + \lambda_p^{(l-1,p)}(\frac{m_{l-1}+m_l}{2}) = \lambda_0^{(l,p)} + \lambda_p^{(l,p)}(\frac{m_{l-1}+m_l}{2})$ for $l = 2, 3, \ldots, L$.

Under the scheme above, we have

$$\begin{aligned}
&h^{(l,p)}(x_p) - h^{(l-1,p)}(x_p)\\
&= \left(\lambda_0^{(l,p)} + \lambda_p^{(l,p)} x_p\right) - \left(\lambda_0^{(l-1,p)} + \lambda_p^{(l-1,p)} x_p\right)\\
&= \lambda_0^{(l,p)} - \lambda_0^{(l-1,p)} + \left(\lambda_p^{(l,p)} - \lambda_p^{(l-1,p)}\right)\left(\frac{m_{l-1}+m_l}{2}\right)\\
&\quad + \left(\lambda_p^{(l,p)} - \lambda_p^{(l-1,p)}\right)\left(x_p - \frac{m_{l-1}+m_l}{2}\right)\\
&= \left(\lambda_p^{(l,p)} - \lambda_p^{(l-1,p)}\right)\left(x_p - \frac{m_{l-1}+m_l}{2}\right),
\end{aligned}$$

which is greater than zero if $x_p \geq m_l$ or equivalently $l \leq \phi(x_p)$, and is small than zero if $l \geq \phi(x_p) + 1$. As a function of l, $h^{(l,p)}(x_p)$ increases as l until l reaches $\phi(x_p)$ and decreases afterwards. Hence, we have $argmax_{l\in\{1,\ldots,L\}}\{h^{(l,p)}(x_p)\} = \phi(x_p)$ for $p = 1, 2, \ldots, P$. Now, we construct $h^{(\boldsymbol{l})}(\boldsymbol{x}) = \sum_{p'=1}^{P} h^{(l'_p,p')}(x'_p) = (\sum_{p'=1}^{P}\lambda_0^{(l_{p'},p')}) + \lambda_1^{(l_1,1)}x_1 + \cdots + \lambda_P^{(l_P,P)}x_P$. Comparing with the form $h^{(\boldsymbol{l})}(\boldsymbol{x}) = \lambda_0^{(\boldsymbol{l})} + \lambda_1^{(\boldsymbol{l})}x_1 + \cdots + \lambda_P^{(\boldsymbol{l})}x_P$ stated in the lemma, we can now choose $\lambda_0^{(\boldsymbol{l})} = \sum_{p'=1}^{P}\lambda_0^{(l_{p'},p')}$ and $\lambda_p^{(\boldsymbol{l})} = \lambda_p^{(l_p,p)}$ for $p = 1, 2, \ldots, P$. Finally, we complete the proof by considering

$$\begin{aligned}
&argmax_{\boldsymbol{l}\in\{1,\ldots,L\}^P}\{h^{(\boldsymbol{l})}(\boldsymbol{x})\}\\
&= argmax_{\boldsymbol{l}\in\{1,\ldots,L\}^P}\{\sum_{p=1}^{P} h^{(l_p,p)}(x_p)\}\\
&= (argmax_{l\in\{1,\ldots,L\}}\{h^{(l,1)}(x_1)\}, \ldots, argmax_{l\in\{1,\ldots,L\}}\{h^{(l,P)}(x_P)\})^T\\
&= (\phi(x_1), \ldots, \phi(x_P))^T,
\end{aligned}$$

where the second equality follows because the each term of the summation $\sum_{p=1}^{P} h^{(l_p,p)}(x_p)$ depends only on individual component l_p instead of $\boldsymbol{l}$. Maximizing this summation with respect to $\boldsymbol{l}$ is equivalent to maximizing each individual term $h^{(l_p,p)}(x_p)$ with respect to l_p.

Appendix C. Proof of Theorem 3.1

Given a fixed parameters setting in the GMoE, set $\boldsymbol{\theta}_{(j,\boldsymbol{l}),k} = \boldsymbol{\theta}_{jk}^*(m_{\boldsymbol{l}}; \boldsymbol{\beta}_{jk}^*)$, where $m_{\boldsymbol{l}}$ is defined by $m_{\boldsymbol{l}} = (m_{l_1}, \ldots, m_{l_P})^T$ and $\boldsymbol{l} \in \{1, \ldots, L\}^P$. Note that while in Eq. (1) $\boldsymbol{\theta}_{jk}^*$ is defined on a $(P+1)$-dimensional vector $\boldsymbol{x}$, we here have defined it on a P-dimensional vector $m_{\boldsymbol{l}}$ by making a slight abuse of notation (writing $\boldsymbol{\theta}_{jk}^*((1, m_{\boldsymbol{l}}^T)^T; \boldsymbol{\beta}_{jk}^*)$ as $\boldsymbol{\theta}_{jk}^*(m_{\boldsymbol{l}}; \boldsymbol{\beta}_{jk}^*)$) for notational convenience. Consider the following sequence of $g \times L^P$-component LRMoE for $n = 1, 2, \ldots$:

$$\begin{aligned}
H^{(n)}(\boldsymbol{y}; \boldsymbol{x}) &:= H(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\alpha}^{(n)}, \boldsymbol{\Theta}, g)\\
&= \sum_{j=1}^{g}\sum_{\boldsymbol{l}\in\{1,\ldots,L\}^P}\pi_{j,\boldsymbol{l}}(\boldsymbol{x}; \boldsymbol{\alpha}^{(n)})\prod_{k=1}^{K} F_k(y_k; \boldsymbol{\theta}_{(j,\boldsymbol{l}),k}), \qquad (27)
\end{aligned}$$

where $\pi_{j,\boldsymbol{l}}(\boldsymbol{x}; \boldsymbol{\alpha}^{(n)}) = \exp\{\boldsymbol{\alpha}_{(j,\boldsymbol{l})}^{(n)T}\boldsymbol{x}\}/\sum_{j'=1}^{g}\sum_{\boldsymbol{l}'\in 1,\ldots,L^P}\exp\{\boldsymbol{\alpha}_{(j',\boldsymbol{l}')}^{(n)T}\boldsymbol{x}\}$. Now, choose $\boldsymbol{\alpha}_{(j,\boldsymbol{l})}^{(n)} = n\boldsymbol{\lambda}^{(\boldsymbol{l})} + \log\pi_j^*(m_{\boldsymbol{l}}; \boldsymbol{\alpha}^*)(1\ 0\ldots 0)^T$, where $\boldsymbol{\lambda}^{(\boldsymbol{l})} := (\lambda_0^{(\boldsymbol{l})}, \ldots, \lambda_P^{(\boldsymbol{l})})^T$ is chosen based on the scheme proposed in the proof of Lemma 3.1 and π_j^* is the GMoE gating function defined in Eq. (1) (note that for notational convenience we write $\pi_j^*((1, m_{\boldsymbol{l}}^T)^T; \boldsymbol{\alpha}^*)$ as $\pi_j^*(m_{\boldsymbol{l}}; \boldsymbol{\alpha}^*)$ as well). Then we have

$$\begin{aligned}
&\pi_{j,\boldsymbol{l}}(\boldsymbol{x}; \boldsymbol{\alpha}^{(n)})\\
&= \exp\{\boldsymbol{\alpha}_{(j,\boldsymbol{l})}^{(n)T}\boldsymbol{x}\}/\sum_{j'=1}^{g}\sum_{\boldsymbol{l}'\in\{1,\ldots,L\}^P}\exp\{\boldsymbol{\alpha}_{(j',\boldsymbol{l}')}^{(n)T}\boldsymbol{x}\}\\
&= \pi_j^*(m_{\boldsymbol{l}}; \boldsymbol{\alpha}^*)\frac{\exp\{n\boldsymbol{\lambda}^{(\boldsymbol{l})T}\boldsymbol{x}\}}{\sum_{\boldsymbol{l}'\in\{1,\ldots,L\}^P}\exp\{n\boldsymbol{\lambda}^{(\boldsymbol{l}')T}\boldsymbol{x}\}}\\
&\to \pi_j^*(m_{\boldsymbol{l}}; \boldsymbol{\alpha}^*)1\{\boldsymbol{l} = \boldsymbol{\phi}(\boldsymbol{x})\} = \pi_j^*(m_{\boldsymbol{l}}; \boldsymbol{\alpha}^*)1\{\boldsymbol{x} = m_{\boldsymbol{l}}\} \ \ as\ n \to \infty,
\end{aligned}$$

where the convergence is directly followed by Lemma 3.2 since $\boldsymbol{\lambda}^{(\boldsymbol{l})T}\boldsymbol{x}$ above is equivalent to $h^{(\boldsymbol{l})}(\boldsymbol{x})$ in Lemma 3.2. Finally, we have

$$\begin{aligned}
&\lim_{n\to\infty} H^{(n)}(\boldsymbol{y}; \boldsymbol{x})\\
&= \sum_{j=1}^{g}\sum_{\boldsymbol{l}\in\{1,\ldots,L\}^P}\lim_{n\to\infty}\pi_{j,\boldsymbol{l}}(\boldsymbol{x}; \boldsymbol{\alpha}^{(n)})\prod_{k=1}^{K} F_k(y_k; \boldsymbol{\theta}_{(j,\boldsymbol{l}),k})\\
&= \sum_{j=1}^{g}\sum_{\boldsymbol{l}\in\{1,\ldots,L\}^P}\pi_j^*(m_{\boldsymbol{l}}; \boldsymbol{\alpha}^*)1\{\boldsymbol{x} = m_{\boldsymbol{l}}\}\prod_{k=1}^{K} F_k(y_k; \boldsymbol{\theta}_{(j,\boldsymbol{l}),k})\\
&= \sum_{j=1}^{g}\pi_j^*(\boldsymbol{x}; \boldsymbol{\alpha}^*)\prod_{k=1}^{K} F_k(y_k; \boldsymbol{\theta}_{jk}^*(\boldsymbol{x}; \boldsymbol{\beta}_{jk}^*)).
\end{aligned}$$

Appendix D. Proof of Theorem 3.2

Partition $[m_{\min}, m_{\max}]$ into L identical intervals with $\Delta x = (m_{\max} - m_{\min})/L$. Define $m_{l_p} = m_{\min} + (l_p - 1/2)\Delta x$ ($l_p = 1, \ldots, L$) as the midpoint of each small interval and a sequence of LRMoE $H^{(n)}(\boldsymbol{y}; \boldsymbol{x})$ expressed in the same way as Eq. (27), with $\boldsymbol{\theta}_{(j,\boldsymbol{l}),k} = \boldsymbol{\theta}_{jk}^*(m_{\boldsymbol{l}}; \boldsymbol{\beta}_{jk}^*)$, and $\pi_{j,\boldsymbol{l}}(\boldsymbol{x}; \boldsymbol{\alpha}^{(n)}) := \pi_j^*(m_{\boldsymbol{l}}; \boldsymbol{\alpha}^*)\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}$, where $\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)} = \exp\{n\boldsymbol{\lambda}^{(\boldsymbol{l})T}\boldsymbol{x}\}/\sum_{\boldsymbol{l}'\in\{1,\ldots,L\}^P}\exp\{n\boldsymbol{\lambda}^{(\boldsymbol{l}')T}\boldsymbol{x}\}$ and $m_{\boldsymbol{l}}$ is defined at the same way as the proof of Theorem 3.1 (Appendix C).

We choose $\boldsymbol{\lambda}^{(\boldsymbol{l})}$ using the scheme presented in the proof of Lemma 3.1 with $\lambda_p^{(l,p)} = l$ and $\lambda_0^{(l,p)} = -l(m_{\min} - \Delta x/2) - l^2\Delta x/2$. Define $D(\boldsymbol{x}) = \{\boldsymbol{l}; \boldsymbol{l} \in \{1, \ldots, L\}^P, |m_{l_p} - x_p| \leq \Delta x\ \forall p \in \{1, \ldots, P\}\}$ and $D'(\boldsymbol{x}) = \{\boldsymbol{l}; \boldsymbol{l} \in \{1, \ldots, L\}^P, \exists p \in \{1, \ldots, P\}\ s.t.\ |m_{l_p} - x_p| > \Delta x\}$. We first aim to find $\boldsymbol{\lambda}^{(\boldsymbol{l}')T}\boldsymbol{x} - \boldsymbol{\lambda}^{(\boldsymbol{l})T}\boldsymbol{x}$ if: we choose $\boldsymbol{l}'$ such that $x_p - m_{l'_p} := \delta_p$ and $|\delta_p| \leq \Delta x/2$ for every $p = 1, \ldots, P$; and for $\boldsymbol{l}$, $\boldsymbol{l} \in D'(\boldsymbol{x})$. Note that $\boldsymbol{l}'$ always exist. We have

$$\begin{aligned}
&h^{(l'_p,p)}(x_p) = -\frac{l'^2_p}{2}\Delta x + l'_p(l'_p\Delta x + \delta_p);\\
&h^{(l_p,p)}(x_p) = -\frac{l^2_p}{2}\Delta x + l_p(l'_p\Delta x + \delta_p);\\
&h_p^* := h^{(l'_p,p)}(x_p) - h^{(l_p,p)}(x_p) = \frac{1}{2}(l'_p - l_p)^2\Delta x + (l'_p - l_p)\delta_p. \qquad (28)
\end{aligned}$$

We evaluate h_p^* in two cases. In the first case, $|m_{l_p} - x_p| \leq \Delta x$, then $|l'_p - l_p| = 0$ or 1. In both cases, $h_p^* \geq 0$. In the second case, $|m_{l_p} - x_p| > \Delta x$, then $|l'_p - l_p| \geq 1$. If $|l'_p - l_p| \geq 2$, then $h_p^* \geq \Delta x$. Otherwise if $|l'_p - l_p| = 1$, $\text{sign}(l'_p - l_p) = \text{sign}(\delta_p)$ must hold and so $h_p^* \geq \Delta x/2$. Since $\boldsymbol{l} \in D'(\boldsymbol{x})$, there must exist at least one p such that $|m_{l_p} - x_p| > \Delta x$. Therefore, $\boldsymbol{\lambda}^{(\boldsymbol{l}')T}\boldsymbol{x} - \boldsymbol{\lambda}^{(\boldsymbol{l})T}\boldsymbol{x} = h^{(\boldsymbol{l}')}(\boldsymbol{x}) - h^{(\boldsymbol{l})}(\boldsymbol{x}) = \sum_{p=1}^{P} h_p^* \geq \Delta x/2$. Note that such a bound does not depend on $\boldsymbol{x}$.

Using the above result, for $\boldsymbol{l} \in D'(\boldsymbol{x})$, we have $\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)} \leq \exp\{-n(\boldsymbol{\lambda}^{(\boldsymbol{l}')T}\boldsymbol{x} - \boldsymbol{\lambda}^{(\boldsymbol{l})T}\boldsymbol{x})\} \leq \exp\{-n\Delta x/2\}$. Reformulating mathematically, $\forall\epsilon > 0$, whenever n is sufficiently large, $\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)} < \epsilon/L^P$ for every $\boldsymbol{l} \in D'(\boldsymbol{x})$ and $\boldsymbol{x} \in \{1\} \times [m_{\min}, m_{\max}]^P$. Finally, we have

$$|H^{(n)}(\boldsymbol{y}; \boldsymbol{x}) - H^*(\boldsymbol{y}; \boldsymbol{x})|$$

$$
\begin{aligned}
&\leq |\sum_{j=1}^{g}\sum_{\boldsymbol{l}\in D(\boldsymbol{x})}\pi_j^*(m_{\boldsymbol{l}};\boldsymbol{\alpha}^*)\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}\prod_{k=1}^{K}F_k(y_k;\boldsymbol{\theta}_{jk}^*(m_{\boldsymbol{l}};\boldsymbol{\beta}_{jk}^*))-H^*(\boldsymbol{y};\boldsymbol{x})|\\
&\quad+\sum_{j=1}^{g}\sum_{\boldsymbol{l}\in D'(\boldsymbol{x})}\pi_j^*(m_{\boldsymbol{l}};\boldsymbol{\alpha}^*)\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}\prod_{k=1}^{K}F_k(y_k;\boldsymbol{\theta}_{jk}^*(m_{\boldsymbol{l}};\boldsymbol{\beta}_{jk}^*))\\
&\leq |\sum_{j=1}^{g}\sum_{\boldsymbol{l}\in D(\boldsymbol{x})}\pi_j^*(m_{\boldsymbol{l}};\boldsymbol{\alpha}^*)\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}\prod_{k=1}^{K}F_k(y_k;\boldsymbol{\theta}_{jk}^*(m_{\boldsymbol{l}};\boldsymbol{\beta}_{jk}^*))\\
&\quad-\sum_{j=1}^{g}\sum_{\boldsymbol{l}\in D(\boldsymbol{x})}\pi_j^*(\boldsymbol{x};\boldsymbol{\alpha}^*)\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}\prod_{k=1}^{K}F_k(y_k;\boldsymbol{\theta}_{jk}^*(\boldsymbol{x};\boldsymbol{\beta}_{jk}^*))|\\
&\quad+|H^*(\boldsymbol{y};\boldsymbol{x})\sum_{\boldsymbol{l}\in D(\boldsymbol{x})}\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}-H^*(\boldsymbol{y};\boldsymbol{x})|+\epsilon\\
&\leq \kappa_1 Pg\Delta x+\kappa_2(\boldsymbol{y})KP\Delta x+2\epsilon.
\end{aligned}
$$

The last inequality holds because of the Lipschitz-continuity properties of F_k and π_j^*. Note that κ_2 depends on $\boldsymbol{y}$, but is finite for every $\boldsymbol{y}$. Also, such upper bound of error does not depend on $\boldsymbol{x}$. Hence, choosing $\Delta x\downarrow 0$, $\epsilon\downarrow 0$ and (for each chosen Δx and ϵ) sufficiently large n, the probability distribution of the LRMoE converges to that of the GMoE uniformly on $\boldsymbol{x}$.

Appendix E. Proof of Proposition 3.1

We only prove the univariate case ($2\leftrightarrow 3$) for conciseness and the arguments can be extended to the multivariate case. To simplify the notations, we drop the subscript k throughout the whole proof. For example, $\mathbb{Q}_k$ is written as $\mathbb{Q}$, f_k is written as f and $\boldsymbol{\theta}_k$ is written as $\boldsymbol{\theta}$. We first consider the discrete (frequency) distributions.

$3\rightarrow 2$: $\forall\epsilon>0$, $\exists\mathbb{Q}^0\subseteq\mathbb{Q}$ with finite support such that $\sum_{q\in\mathbb{Q}^0}\xi(q)\geq 1-\epsilon$, where $\xi(\cdot)$ represents the probability masses of any univariate discrete distribution with support $\mathbb{Q}$. We approximate the density ξ by the finite mixture $h(y;\boldsymbol{\pi},\boldsymbol{\theta}^{(n)},g):=\sum_{q\in\mathbb{Q}^0\cup q^*}\pi_q f(y;\boldsymbol{\theta}_q^{(n)})$ with $(\|\mathbb{Q}^0\|+1)$ components. Choosing $\pi_q=\xi(q)$ for $q\in\mathbb{Q}^0$, $\pi_{q^*}=1-\sum_{q\in\mathbb{Q}^0}\xi(q)$ where $q^*\notin\mathbb{Q}^0$ (note that we can arbitrarily choose q^* as well as it falls outside the set $\mathbb{Q}^0$) and suitable parameters $\boldsymbol{\theta}_q^{(n)}$ such that $f(y;\boldsymbol{\theta}_q^{(n)})\xrightarrow{n\rightarrow\infty}1\{y=q\}$, we have

$$
\begin{aligned}
&\lim_{n\rightarrow\infty}|\xi(y)-h(y;\boldsymbol{\pi},\boldsymbol{\theta}^{(n)},g)|\\
&=\lim_{n\rightarrow\infty}|\xi(y)-\sum_{q\in\mathbb{Q}^0\cup q^*}\pi_q f(y;\boldsymbol{\theta}_q^{(n)})|\\
&=|\xi(y)-\sum_{q\in\mathbb{Q}^0}\xi(q)1\{y=q\}-(1-\sum_{q\in\mathbb{Q}^0}\xi(q))1\{y=q^*\}|\\
&\leq\xi(y)(1-1\{y\in\mathbb{Q}^0\})+(1-\sum_{q\in\mathbb{Q}^0}\xi(q))\leq 2\epsilon.
\end{aligned}
$$

The result follows since ϵ is arbitrary.

$2\rightarrow 3$: Assume that Statement 3 is false. Then, $\exists\delta>0$ and $q\in\mathbb{Q}$ such that for every parameters setting $\boldsymbol{\theta}$, $f(q;\boldsymbol{\theta})<1-\delta$. Since $h(q;\boldsymbol{\pi},\boldsymbol{\theta},g)<1-\delta$, h cannot approximate $\xi(y)=1\{y=q\}$ arbitrarily accurately.

For the continuous (severity) case, we have the following

$3\rightarrow 2$: Denote $\varXi(\cdot)$ as a continuous univariate cdf. We want to approximate $\varXi(\cdot)$ by the finite mixture distribution $H(y;\boldsymbol{\pi},\boldsymbol{\theta}^{(n)},g):=\sum_{j=1}^{g}\pi_j F(y;\boldsymbol{\theta}_j^{(n)})$ with $g=(\epsilon^{-1}-1)$ components, where $\epsilon>0$ is chosen as an arbitrarily small quantity. We choose $\pi_j=\epsilon$ for $j=1,\ldots,g-1$ and $\pi_g=2\epsilon$. We further select suitable parameters $\boldsymbol{\theta}_j^{(n)}$ for every $j=1,\ldots,g$ such that $F(\cdot;\boldsymbol{\theta}_j^{(n)})\xrightarrow{\mathcal{D}}\varXi^{-1}(j\epsilon)$, where $\varXi^{-1}(a)=\inf\{y:\varXi(y)\geq a\}$. Then, we have

$$
\begin{aligned}
&\lim_{n\rightarrow\infty}|\varXi(y)-H(y;\boldsymbol{\pi},\boldsymbol{\theta}^{(n)},g)|\\
&=\lim_{n\rightarrow\infty}|\varXi(y)-\sum_{j=1}^{g}\pi_j F(y;\boldsymbol{\theta}_j^{(n)})|\\
&=|\varXi(y)-\sum_{j=1}^{g}\epsilon 1\{y\geq\varXi^{-1}(j\epsilon)\}+\sum_{j=1}^{g}c_j\epsilon 1\{y=\varXi^{-1}(j\epsilon)\}\\
&\quad-\epsilon 1\{y\geq\varXi^{-1}(1-\epsilon)\}|\\
&\leq|\varXi(y)-\sum_{j:\varXi(y)\geq j\epsilon}\epsilon|+\epsilon+\epsilon=|\varXi(y)-\epsilon\left\lfloor\frac{\varXi(y)}{\epsilon}\right\rfloor|+2\epsilon\leq 3\epsilon,
\end{aligned}
$$

where $\lfloor\cdot\rfloor$ represents the floor function, the term $c_j\epsilon 1\{y=\varXi^{-1}(j\epsilon)\}$ in the first equality is an adjustment term as $F(y;\boldsymbol{\theta}_j^{(n)})\rightarrow 1\{y\geq\varXi^{-1}(j\epsilon)\}$ may not hold true when $y=\varXi^{-1}(j\epsilon)$ (discontinuity point), and c_j is either zero or one. Choosing arbitrarily large n and small ϵ, we are able to approximate $\varXi(y)$ by the finite mixture distribution $H(y;\boldsymbol{\pi},\boldsymbol{\theta}^{(n)},g)$, and hence the result follows.

$2\rightarrow 3$: Assume that Statement 3 is false. Then, $\exists\delta>0$, $\omega>0$ and $y\in\mathbb{Q}$ such that for every parameters setting $\boldsymbol{\theta}$, $F(y+\omega;\boldsymbol{\theta})-F(y-\omega;\boldsymbol{\theta})<1-\delta$. Then, choose $\varXi(\cdot)$ which satisfies $\varXi(y+\omega)-\varXi(y-\omega)=1$. If Statement 2 is true, then $\exists\boldsymbol{\pi}$, $\boldsymbol{\theta}^{(n)}$ and $g^{(n)}$ such that $\sum_{j=1}^{g^{(n)}}\pi_j F(y+\omega;\boldsymbol{\theta}^{(n)})\rightarrow\varXi(y+\omega)$ and $\sum_{j=1}^{g^{(n)}}\pi_j F(y-\omega;\boldsymbol{\theta}^{(n)})\rightarrow\varXi(y-\omega)$ as $n\rightarrow\infty$. Taking a difference, we have $\sum_{j=1}^{g^{(n)}}\pi_j(F(y+\omega;\boldsymbol{\theta}^{(n)})-F(y-\omega;\boldsymbol{\theta}^{(n)}))\rightarrow\varXi(y+\omega)-\varXi(y-\omega)=1$. However, the $\sum_{j=1}^{g^{(n)}}\pi_j(F(y+\omega;\boldsymbol{\theta}^{(n)})-F(y-\omega;\boldsymbol{\theta}^{(n)}))<1-\delta$ if Statement 3 is false, leading to a contradiction, so the result follows.

Appendix F. Proof of Theorem 3.3

For the “only if” condition, if Statement 3 of Proposition 3.1 fails, then Proposition 3.1 states that the class of multivariate finite mixture model with component functions F_k is not dense in the space of multivariate frequency (or severity) distributions. Since the class of finite mixture models and the space of multivariate distributions are special cases of $\mathcal{G}_0(\mathcal{A})$ and $\mathcal{G}_1(\mathcal{A})$ respectively with $\mathcal{A}=\{1\}$ (or $P=1$), the statement that $\mathcal{G}_0(\mathcal{A})$ is uniformly dense in $\mathcal{G}_1(\mathcal{A})$ is disproved just by using this special example.

For the “if” condition, we first consider frequency case. Partition $[m_{\min},m_{\max}]$ and define $m_{\boldsymbol{l}}$ at the same way as the proof of Theorem 3.1 (Appendix C). Denote $g^*(\boldsymbol{y};\boldsymbol{x})$ as the density of G^*. $G^*(\mathcal{A})$ is tight implies that for all $\epsilon'>0$, there exists a sequence of integers $s_1,\ldots,s_U$ such that $\sum_{\boldsymbol{u}=\{1,\ldots,U\}^K}g^*(s_{\boldsymbol{u}};\boldsymbol{x})\geq 1-\epsilon'$ $\forall\boldsymbol{x}$, where $s_{\boldsymbol{u}}=(s_{u_1},\ldots,s_{u_K})^T$. $\boldsymbol{u}^*$ is defined such that $s_{\boldsymbol{u}^*}\notin\{s_1,\ldots,s_U\}^K$. Also, define a $(U^K+1)\times L^P$-component LRMoE as follows.

$$
H^{(t,n)}(\boldsymbol{y};\boldsymbol{x})=\sum_{\boldsymbol{u}\in\{1,\ldots,U\}^K\cup\boldsymbol{u}^*}\sum_{\boldsymbol{l}\in\{1,\ldots,L\}^P}\pi_{\boldsymbol{u},\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha}^{(n)})\prod_{k=1}^{K}F_k(y_k;\boldsymbol{\theta}_{(\boldsymbol{u},\boldsymbol{l}),k}^{(t)}).
$$

The problem is to approximate $G^*(\boldsymbol{y};\boldsymbol{x})$ by $H^{(t,n)}(\boldsymbol{y};\boldsymbol{x})$. We can choose suitable regression coefficients $\boldsymbol{\alpha}^{(n)}$ such that the expert functions are in the form of $\pi_{\boldsymbol{u},\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha}^{(n)})=[g^*(s_{\boldsymbol{u}};m_{\boldsymbol{l}})1\{\boldsymbol{u}\in\{1,\ldots,U\}^K\}+(1-\sum_{\boldsymbol{u}'=\{1,\ldots,U\}^K}g^*(s_{\boldsymbol{u}'};m_{\boldsymbol{l}}))1\{\boldsymbol{u}=\boldsymbol{u}^*\}]\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}$, where $\gamma_{\boldsymbol{l},\boldsymbol{x}}^{(n)}$ is defined at the same way as the proof in Theorem 3.2 and $\boldsymbol{u}^*\notin\{1,\ldots,U\}^K$. Also note that we have written $g^*(s_{\boldsymbol{u}};(1,m_{\boldsymbol{l}}^T)^T)$ as $g^*(s_{\boldsymbol{u}};m_{\boldsymbol{l}})$ for notational convenience. Since Statement 3 of Proposition 3.1 is satisfied, $\boldsymbol{\theta}_{(\boldsymbol{u},\boldsymbol{l}),k}^{(t)}$ can be chosen such that $F_k(y_k; \boldsymbol{\theta}^{(t)}_{(\boldsymbol{u},\boldsymbol{l}),k}) \xrightarrow{t\to\infty} 1\{y_k \geq s_{u_k}\}$. Defining $D(\boldsymbol{x})$ and $D'(\boldsymbol{x})$ at the same way and using similar approaches as the proof in Theorem 3.2, we have for sufficiently large n:

$$
\begin{aligned}
&\lim_{t\to\infty} |H^{(t,n)}(\boldsymbol{y};\boldsymbol{x}) - G^*(\boldsymbol{y};\boldsymbol{x})| \\
&\leq | \sum_{\boldsymbol{u}\in\{1,\dots,U\}^K} \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} g^*(s_{\boldsymbol{u}}; m_{\boldsymbol{l}}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \prod_{k=1}^K 1\{y_k \geq s_{u_k}\} - G^*(\boldsymbol{y};\boldsymbol{x})| \\
&\quad + \sum_{\boldsymbol{u}\in\{1,\dots,U\}^K} \sum_{\boldsymbol{l}\in D'(\boldsymbol{x})} g^*(s_{\boldsymbol{u}}; m_{\boldsymbol{l}}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \prod_{k=1}^K 1\{y_k \geq s_{u_k}\} \\
&\quad + \sum_{\boldsymbol{l}\in\{1,\dots,L\}^P} (1 - \sum_{\boldsymbol{u}'=\{1,\dots,U\}^K} g^*(s_{\boldsymbol{u}'}; m_{\boldsymbol{l}})) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \prod_{k=1}^K 1\{y_k \geq s_{u^*_k}\} \\
&\leq | \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} \sum_{y'_1=s_1}^{\min\{y_1,s_U\}} \cdots \sum_{y'_K=s_1}^{\min\{y_K,s_U\}} g^*(\boldsymbol{y}'; m_{\boldsymbol{l}}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} - G^*(\boldsymbol{y};\boldsymbol{x})| + \epsilon + \epsilon' \\
&\leq | \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} G^*(\boldsymbol{y}; m_{\boldsymbol{l}}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} - G^*(\boldsymbol{y};\boldsymbol{x})| + \epsilon + 2\epsilon' \\
&\leq | \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} G^*(\boldsymbol{y}; \boldsymbol{x}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} - G^*(\boldsymbol{y};\boldsymbol{x})| + \epsilon + \kappa_2(\boldsymbol{y}) P \Delta x + 2\epsilon' \\
&\leq 2\epsilon + \kappa_2(\boldsymbol{y}) P \Delta x + 2\epsilon',
\end{aligned}
$$

where the extra ϵ' of the third inequality resulted from the tightness property of $G^*(\boldsymbol{y};\boldsymbol{x})$ and the extra $\kappa_2(\boldsymbol{y})P\Delta x$ of the fourth inequality is based on the Lipschitz-continuity of $G^*(\boldsymbol{y};\boldsymbol{x})$ w.r.t. $\boldsymbol{x}$. It suffices to show that $H^{(t,n)}(\boldsymbol{y};\boldsymbol{x})$ uniformly converges to $\boldsymbol{x}$ as $t \to \infty$. Write $H^{(t,n)}(\boldsymbol{y};\boldsymbol{x}) = \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} a(t,\boldsymbol{l})\gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}}$, where $a(\cdot,\cdot)$ is a function depending on t but not on $\boldsymbol{x}$. Since $0 \leq \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \leq 1$ and $\sum_{\boldsymbol{l}\in D(\boldsymbol{x})} \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \leq 1$, we have the upper error bound $|H^{(t,n)}(\boldsymbol{y};\boldsymbol{x}) - H^{(\infty,n)}(\boldsymbol{y};\boldsymbol{x})| \leq \sup_{\boldsymbol{l}\in D(\boldsymbol{x})} |a(t,\boldsymbol{l}) - a(\infty,\boldsymbol{l})|$. As $D(\boldsymbol{x})$ takes a finite support and the upper error bound does not depend on $\boldsymbol{x}$, the result follows.

We then consider the severity case. Partition $[m_{\min}, m_{\max}]$ and define any notations at the same way as frequency case. Tightness of $G^*(\mathcal{A})$ implies $\forall\epsilon' > 0$, $\exists s_0, s_U$ such that $G^*([(s_U,\dots,s_U)^T,(s_0,\dots,s_0)^T];\boldsymbol{x}) \geq 1-\epsilon'$ for all $\boldsymbol{x}$. Here, for two vectors $\boldsymbol{a} = (a_1,\dots,a_K)^T$, $\boldsymbol{b} = (b_1,\dots,b_K)^T$ and a random vector $\boldsymbol{Y}^*$ following cdf G^*, we define $G^*([\boldsymbol{a},\boldsymbol{b}]) := P(\boldsymbol{Y}^* \in [\boldsymbol{a},\boldsymbol{b}])$, where $[\boldsymbol{a},\boldsymbol{b}] = [a_1,b_1] \times \cdots \times [a_K,b_K]$. Also, partition $[s_0,s_U]$ into U identical intervals such that $\Delta s = (s_U - s_0)/U$ and define $s_u = s_0 + u\Delta s$ for $u = 1,\dots,U$. Further, $\boldsymbol{u}^*$ is defined such that $s_{\boldsymbol{u}^*} \notin \{s_1,\dots,s_U\}^K$. Define the LRMoE and choose the corresponding component weights:

$$
H^{(t,n)}(\boldsymbol{y};\boldsymbol{x}) = \sum_{\boldsymbol{u}\in\{1,\dots,U\}^K \cup \boldsymbol{u}^*} \sum_{\boldsymbol{l}\in\{1,\dots,L\}^P} \pi_{\boldsymbol{u},\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha}^{(n)}) \prod_{k=1}^K F_k(y_k; \boldsymbol{\theta}^{(t)}_{(\boldsymbol{u},\boldsymbol{l}),k}),
$$

$$
\begin{aligned}
&\pi_{\boldsymbol{u},\boldsymbol{l}}(\boldsymbol{x};\boldsymbol{\alpha}^{(n)}) \\
&= \Big[G^*([s_{\boldsymbol{u}}, s_{\boldsymbol{u}} - \Delta s\mathbf{1}]; m_{\boldsymbol{l}}) 1\{\boldsymbol{u} \in \{1,\dots,U\}^K\} \\
&\qquad + \big(1 - G^*([(s_U,\dots,s_U)^T,(s_0,\dots,s_0)^T]; m_{\boldsymbol{l}})\big) 1\{\boldsymbol{u} = \boldsymbol{u}^*\} \Big] \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}},
\end{aligned}
$$

where $\mathbf{1}$ is a K-dimensional column vector of ones. Also, choose $\boldsymbol{\theta}^{(t)}_{(\boldsymbol{u},\boldsymbol{l}),k}$ such that $F_k(y_k; \boldsymbol{\theta}^{(t)}_{(\boldsymbol{u},\boldsymbol{l}),k}) \overset{\mathcal{D}}{\to} 1\{y_k \geq s_{u_k}\}$ as $t \to \infty$. For sufficiently large n, we have

$$
\begin{aligned}
&\lim_{\Delta s\to 0} \lim_{t\to\infty} |H^{(t,n)}(\boldsymbol{y};\boldsymbol{x}) - G^*(\boldsymbol{y};\boldsymbol{x})| \\
&\leq \lim_{\Delta s\to 0} | \sum_{\boldsymbol{u}\in\{1,\dots,U\}^K} \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} G^*([s_{\boldsymbol{u}}, s_{\boldsymbol{u}} - \Delta s\mathbf{1}]; m_{\boldsymbol{l}}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \prod_{k=1}^K 1\{y_k \geq s_{u_k}\} \\
&\quad - G^*(\boldsymbol{y};\boldsymbol{x})| \\
&\quad + \lim_{\Delta s\to 0} \sum_{\boldsymbol{u}\in\{1,\dots,U\}^K} \sum_{\boldsymbol{l}\in D'(\boldsymbol{x})} G^*([s_{\boldsymbol{u}}, s_{\boldsymbol{u}} - \Delta s\mathbf{1}]; m_{\boldsymbol{l}}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \prod_{k=1}^K 1\{y_k \geq s_{u_k}\} \\
&\quad + \sum_{\boldsymbol{l}\in\{1,\dots,L\}^P} \big(1 - G^*([(s_U,\dots,s_U)^T,(s_0,\dots,s_0)^T]; m_{\boldsymbol{l}})\big) \\
&\quad \times \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \prod_{k=1}^K 1\{y_k \geq s_{u^*_k}\} \\
&\quad + \lim_{\Delta s\to 0} \sum_{k=1}^K \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} G^*_k([y_k, y_k - \Delta s]; m_{\boldsymbol{l}}) \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} \\
&\leq | \sum_{\boldsymbol{l}\in D(\boldsymbol{x})} G^*([(s_0,\dots,s_0)^T, (\min\{y_1,s_U\},\dots,\min\{y_K,s_U\})^T]; m_{\boldsymbol{l}}) \\
&\quad \times \gamma^{(n)}_{\boldsymbol{l},\boldsymbol{x}} - G^*(\boldsymbol{y};\boldsymbol{x})| + \epsilon + \epsilon' \\
&\leq 2\epsilon + \kappa_2(\boldsymbol{y}) P \Delta x + 2\epsilon',
\end{aligned}
$$

where G^*_k is the kth marginal cdf of G^*. The last term of the first inequality comes from the fact that $F_k(y_k; \boldsymbol{\theta}^{(t)}_{(\boldsymbol{u},\boldsymbol{l}),k}) \to 1\{y_k \geq s_{u_k}\}$ may not be true when $y_k = s_{u_k}$. Such term, however, will converge to zero as $\Delta s \to 0$. The first term for the second inequality results from Dominated Convergence Theorem. Moreover, it is easy to see that $H^{(t,n)}(\boldsymbol{y};\boldsymbol{x})$ uniformly converges to $\boldsymbol{x}$ as $\Delta s \to 0$ and $t \to \infty$ using the same technique as the frequency case, so the result follows.

References

Ahn, S., Kim, J.H., Ramaswami, V., 2012. A new class of models for heavy tailed distributions in finance and insurance risk. Insurance Math. Econom. 51 (1), 43–52.

Asmussen, S., Nerman, O., Olsson, M., 1996. Fitting phase-type distributions via the EM algorithm. Scand. J. Stat. 23 (4), 419–441.

Badescu, A.L., Lin, X.S., Tang, D., Valdez, E.A., 2015. Multivariate Pascal mixture regression models for correlated claim frequencies. Available in SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2618265.

Bermúdez, L., 2009. A priori ratemaking using bivariate Poisson regression models. Insurance Math. Econom. 44 (1), 135–141.

Bermúdez, L., Karlis, D., 2011. Bayesian multivariate Poisson models for insurance ratemaking. Insurance Math. Econom. 48 (2), 226–236.

Chakraborty, S., Chakravarty, D., 2012. Discrete gamma distributions: properties and parameter estimations. Comm. Statist. Theory Methods 41 (18), 3301–3324.

Conway, R.W., Maxwell, W.L., 1962. A queuing model with state dependent service rates. J. Ind. Eng. 12 (2), 132–136.

De Jong, P., Heller, G.Z., 2008. Generalized Linear Models for Insurance Data. Cambridge University Press,

Diao, L., Weng, C., 2019. Regression tree credibility model. N. Am. Actuar. J. 23 (2), 169–196.

Durrett, R., 2010. Probability: Theory and Examples. Cambridge University Press.

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

Fung, T.C., Badescu, A.L., Lin, X.S., 2019a. Multivariate Cox hidden Markov models with an application to operational risk. Scand. Actuar. J. 2019 (8), 686–710.

Fung, T.C., Badescu, A.L., Lin, X.S., 2019b. A class of mixture of experts models for general insurance: application to correlated claim frequencies. Astin Bull. 49 (3), 647–688.

Gormley, I.C., Murphy, T.B., et al., 2008. A mixture of experts model for rank data with applications in election studies. Ann. Appl. Stat. 2 (4), 1452–1477.

Grun, B., Leisch, F., 2008. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. J. Stat. Softw. 28 (4), 1–35.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., 1991. Adaptive mixtures of local experts. Neural Comput. 3 (1), 79–87.

Jiang, W., Tanner, M.A., 1999. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. Ann. Statist. 987–1011.

Jordan, M.I., Jacobs, R.A., 1992. Hierarchies of adaptive experts. In: Advances in Neural Information Processing Systems. pp. 985–992.

Lee, S.C.K., Lin, X.S., 2012. Modeling dependent risks with multivariate Erlang mixtures. Astin Bull. 42 (1), 153–180.

McLachlan, G., Peel, D., 2000. Finite Mixture Models. Wiley Series in Probability and Statistics.

Miljkovic, T., Grün, B., 2016. Modeling loss data using mixtures of distributions. Insurance Math. Econom. 70, 387–396.

Nguyen, H.D., Lloyd-Jones, L.R., McLachlan, G.J., 2016. A universal approximation theorem for mixture-of-experts models. Neural Comput. 28 (12), 2585–2593.

Quan, Z., Valdez, E.A., 2018. Predictive analytics of insurance claims using multivariate decision trees. Depend. Model. 6 (1), 377–407.

Rao, R., 1962. Relations between weak and uniform convergence of measures with applications. Ann. Math. Stat. 33 (2), 659–680.

Shi, P., Valdez, E.A., 2014. Multivariate negative binomial models for insurance claim counts. Insurance Math. Econom. 55, 18–29.

Übeyli, E.D., 2005. A mixture of experts network structure for breast cancer diagnosis. J. Med. Syst. 29 (5), 569–579.

Verbelen, R., Gong, L., Antonio, K., Badescu, A.L., Lin, X.S., 2015. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. Astin Bull. 45 (3), 729–758.

Winkelmann, R., 1995. Duration dependence and dispersion in count-data models. J. Bus. Econom. Statist. 13 (4), 467–474.

Wüthrich, M.V., 2018. Neural networks applied to chain–ladder reserving. Eur. Actuar. J. 8 (2), 407–436.

Wuthrich, M.V., Buser, C., 2019. Data Analytics for Non-Life Insurance Pricing. Swiss Finance Institute Research Paper (16–68).

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