North American Actuarial Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaaj20

A New Class of Severity Regression Models with an Application to IBNR Prediction

Tsz Chai Fung, Andrei L. Badescu & X. Sheldon Lin

To cite this article: Tsz Chai Fung, Andrei L. Badescu & X. Sheldon Lin (2021) A New Class of Severity Regression Models with an Application to IBNR Prediction, North American Actuarial Journal, 25:2, 206-231, DOI: 10.1080/10920277.2020.1729813

To link to this article: https://doi.org/10.1080/10920277.2020.1729813

North American Actuarial Journal, 25(2), 206–231, 2021

ISSN: 1092-0277 print / 2325-0453 online
DOI: 10.1080/10920277.2020.1729813

Routledge
Taylor & Francis Group

Check for updates

A New Class of Severity Regression Models with an Application to IBNR Prediction

Tsz Chai Fung, Andrei L. Badescu, and X. Sheldon Lin
Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Insurance loss severity data often exhibit heavy-tailed behavior, complex distributional characteristics such as multimodality, and peculiar links between policyholders' risk profiles and claim amounts. To capture these features, we propose a transformed Gamma logit-weighted mixture of experts (TG-LRMoE) model for severity regression. The model possesses several desirable properties. The TG-LRMoE satisfies the denseness property that warrants its full versatility in capturing any distribution and regression structures. It may effectively extrapolate a wide range of tail behavior. The model is also identifiable, which further ensures its suitability for statistical inference. To make the TG-LRMoE computationally tractable, an expectation conditional maximization (ECM) algorithm with parameter penalization is developed for efficient and robust parameter estimation. The proposed model is applied to fit the severity and reporting delay components of a European automobile insurance dataset. In addition to obtaining excellent goodness of fit, the proposed model is shown to be useful and crucial for adequate prediction of incurred but not reported (IBNR) reserves through out-of-sample testing.

1. INTRODUCTION

Loss severity regression modeling is a fundamental yet challenging problem in various actuarial areas, including insurance ratemaking, reserving, and risk management. Insurance data often exhibit heavy-tailed behavior, where the extreme losses in the tail are often the most impactful to the insurers. In addition, unobserved heterogeneity among losses because of the impossibility of collecting all loss information may cause complex distributional phenomena such as multimodality. Further, it is sometimes difficult to determine an appropriate regression form, especially when nonlinear patterns between covariates and loss severities are implied by some insurance dataset. Building a suitable model that includes all of the above features facilitates actuaries' decision-making processes. On the other hand, despite extensive studies on severity regression modeling in the actuarial literature, there is still no consensus on the "best" model.

Tail behavior of insurance loss data is commonly modeled by traditional heavy-tailed distributions such as Burr, log-Gamma, and generalized Pareto distributions (GPD), which are all well aligned with extreme value theory (EVT). In regression setting, generalized linear models (GLMs) with a generalized beta distribution of the second kind (Frees and Valdez 2008), which contain a broad class of heavy-tailed distributions such as lognormal and Burr, are widely adopted.

To cater to the mismatch between the body behavior and the tail behavior, a relatively more flexible modeling approach is the use of a composite distribution or more generally a spliced distribution (Klugman, Panjer, and Willmot 2012), which subdivides the losses into two or more intervals and models losses of different intervals with different distributions. Actuarial literature on various choices of composite models includes, for example, Pigeon and Denuit (2011), Scollnik and Sun (2012), Bakar et al. (2015), Calderín-Ojeda and Kwok (2016), Reynkens et al. (2017), and Grün and Miljkovic (2019). The splicing technique was also extended in regression setting by, for example, Gan and Valdez (2018).

With regards to distributional multimodality, finite mixture models have recently emerged in the actuarial literature. Lee and Lin (2010) proposed a mixture of Erlang distribution for loss severity modeling that is dense in the space of positive continuous distributions, meaning that it possesses a full flexibility to fit any complex distributions. Verbelen et al. (2015) extended the use of Erlang mixture by fitting it to censored and truncated data. However, restricting an Erlang component function in the mixture may lead to an excessive number of components required to suitably fit the target distribution,

Address correspondence to Andrei L. Badescu, Department of Statistical Sciences, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3, Canada. E-mail: badescu@utstat.toronto.edu

especially if the target distribution is heavy-tailed and the Erlang distribution is relatively light-tailed. For example, Verbelen et al. (2015) demonstrated that a prohibitively large number of components is required to fit a GPD using the Erlang mixture, so the tail heaviness of the GPD cannot be adequately extrapolated. As such, Miljkovic and Grün (2016) and Blostein and Miljkovic (2019) presented finite mixtures with various combinations of distributions, including heavy-tailed distributions such as lognormal and Burr, to alleviate the above issue. Through Danish fire losses and Secura Re datasets, they found better fitting performance and detected fewer mixture components when heavy-tailed component functions (instead of Gamma or Erlang) were used. Alternatively, considering the concept of data transformation, the log phase-type distribution proposed by Ahn, Kim, and Ramaswami (2012), which also possesses denseness properties, is suitable for modeling heavy-tailed data.

In Fung, Badescu, and Lin (2019b), the logit-weighted reduced mixture of experts model (LRMoE) was proposed as an alternative statistical tool for frequency or severity regression. The LRMoE can be regarded as a regression version of a finite mixture model, where the regression links are incorporated only through the component weights (called gating functions) but not the component distributions (called expert functions). This ensures model parsimony but still preserves model versatility. The paper also highlights the importance of choosing a suitable expert function to effectively fit both the body and the tail of data. By choosing an Erlang count expert function, Fung, Badescu, and Lin (2019a) demonstrated the success of the resulting LRMoE in capturing complex features of a real automobile insurance frequency dataset. As such, the fitted model is potentially useful for insurance ratemaking.

This article contributes to the LRMoE framework in the context of severity regression. We propose a transformed Gamma LRMoE (TG-LRMoE), where the observed data are first manipulated through a Box-Cox transformation (Box and Cox 1964). The transformed data are then modeled by the LRMoE with a Gamma expert function. The single parameter introduced by the Box-Cox transformation controls the tail heaviness of the TG-LRMoE.

The proposed TG-LRMoE possesses several important desirable properties. First, it possesses the denseness property in regression setting, guaranteeing its full flexibility in catering to any distribution and regression structures. This makes the proposed model fully data driven and ensures that the fitted model will be highly synchronous to the input data. Secondly, involving only one extra transformation parameter, the TG-LRMoE covers a broad range of tail behavior, including light-tailed distributions such as Gamma and Weibull distributions, as well as heavy-tailed distributions such as the Burr distribution. Thirdly, the identifiability property of the proposed model ensures its suitability for statistical inference. Finally, it is possible to develop a stable, robust, and efficient algorithm for model calibration (i.e., an expectation conditional maximization [ECM] algorithm similar to that presented by Fung, Badescu, and Lin [2019a]), making the proposed model computationally tractable.

After justifying theoretically the plausibility of the TG-LRMoE, we demonstrate its ability to fit complex heavy-tailed severity regression distributions and its usefulness in solving reserving problems by analyzing an European automobile insurance dataset. For property and casualty insurance companies, a major reserving problem is to appropriately estimate the incurred but not reported (IBNR) reserves. Though the traditional triangular approach (e.g., chain ladder method) aggregates claim data into a run-off triangle and makes inferences based on the summary data (see, e.g., Wüthrich and Merz [2008] for a comprehensive summary of techniques), the evolution of computational and information technologies allows insurers to evaluate reserves more accurately at an individual claim level, commonly called a microlevel modeling framework. Recent contributions to microlevel reserving include, for example, Antonio and Plat (2014), Badescu, Lin, and Tang (2016), Verrall and Wüthrich (2016), Wüthrich (2018) and Badescu et al. (2019), which require modeling three components: frequency, severity, and reporting delay. The insurance dataset we obtained includes individual claim information as well as detailed policyholder features for each contract. This extra information (covariates) may enable us to predict IBNR even more accurately under the microlevel reserving framework. With the use of the proposed TG-LRMoE regression model for the severity and reporting delay components, not only do we obtain excellent goodness of fit for both components, but we also provide reasonable predictions of the IBNR.

This article is structured as follows. The next section formulates the proposed TG-LRMoE as a flexible severity regression model. Section 3 presents three crucial desirable properties of the proposed model: denseness, tail flexibility, and model identifiability. The computational aspect of the proposed model is discussed in Section 4, where the ECM algorithm is proposed to obtain the maximum a posteriori (MAP) estimates of the parameters. We leverage the proposed model to a real insurance data in Section 5, demonstrating its usefulness to adequately predict IBNR reserves under a microlevel reserving framework. Section 6 summarizes our findings and discusses some potential future research directions.

2. THE TG-LRMOE REGRESSION MODEL

In this section, we propose the TG-LRMoE as a flexible severity regression model. Suppose that there are a total of n mutually independent insurance claims. Denote $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ and $\boldsymbol{y} = (y_1, ..., y_n)^T$ respectively as the claim severity column

vector (response variable) and the corresponding realization. For each claim $i \in \{1, ..., n\}$, we also define $\boldsymbol{x}_i = (x_{i0}, ..., x_{iP})^T$ (with $x_{i0} = 1$) as the information relating to claim i (covariates). It is assumed that the claim severities $Y_1, ..., Y_n$ are mutually independent.

In the insurance context, it insurance claim severity distributions are often very heavy-tailed. Rather than using existing approaches such as traditional heavy-tailed distributions (such as GPD) and composite models, this article proposes an alternative approach that transforms the claim severities $\boldsymbol{Y}$ and models the transformed severities $\tilde{\boldsymbol{Y}} := \left(\tilde{Y}_1, ..., \tilde{Y}_n\right)^T$ by lighter-tailed distributions. We first define a Box-Cox transformation on Y_i:

$$\tilde{Y}_i = \frac{(1 + Y_i)^\gamma - 1}{\gamma}, \qquad \gamma > 0, \tag{2.1}$$

where γ is a parameter controlling the tail heaviness of the distribution of $\tilde{Y}_i$. $\tilde{Y}_i$ has a lighter tail than Y_i when $\gamma < 1$ and vice versa. In addition, $\tilde{Y}_i \to \log(1 + Y_i)$ as $\gamma \to 0$. We will discuss how the above transformation enables flexible tail modeling in Subsection 3.1.

We then model $\tilde{Y}_i$ given $\boldsymbol{x}_i$ through the LRMoE. Its probability density function (pdf) is given by

$$h_{\tilde{Y}_i|\boldsymbol{x}_i}(\tilde{y}_i; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{\Psi}, g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f\left(\tilde{y}_i; \boldsymbol{\psi}_j\right), \qquad \tilde{y}_i > 0, \tag{2.2}$$

where g is the number of latent classes, $\boldsymbol{\Psi} = \left(\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_g\right)$ are the parameters of the expert functions, $\pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) = \exp\{\boldsymbol{\alpha}_j^T \boldsymbol{x}_i\} / \sum_{j'=1}^{g} \exp\{\boldsymbol{\alpha}_{j'}^T \boldsymbol{x}_i\}$ is the mixing weight for the jth class (gating function), and the regression parameters for the mixing weights are $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_g)$ and $\boldsymbol{\alpha}_j = (\alpha_{j0}, ..., \alpha_{jP})^T \in \mathbb{R}^{P+1}$. We choose a Gamma expert function so that $\boldsymbol{\Psi} = (\boldsymbol{m}, \boldsymbol{\theta}), \boldsymbol{\psi}_j = \left(m_j, \theta_j\right)$, and

$$f\left(\tilde{y}_i; \boldsymbol{\psi}_j\right) := f\left(\tilde{y}_i; m_j, \theta_j\right) = \frac{\tilde{y}_i^{m_j - 1} e^{-\tilde{y}_i/\theta_j}}{\Gamma(m_j) \theta_j^{m_j}}, \qquad \tilde{y}_i, m_j, \theta_j > 0, \tag{2.3}$$

where $\boldsymbol{m} = (m_1, ..., m_g)$ and $\boldsymbol{\theta} = \left(\theta_1, ..., \theta_g\right)$ are, respectively, the shape and scale parameters of gamma distribution. Under simple probabilistic arguments, the pdf of Y_i given $\boldsymbol{x}_i$ can also be derived:

$$h_{Y_i|\boldsymbol{x}_i}(y_i; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}, \gamma, g) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f\left(\tilde{y}_i(\gamma); m_j, \theta_j\right)(1 + y_i)^{\gamma - 1}, \qquad y_i, m_j, \theta_j, \gamma > 0, \tag{2.4}$$

where $\tilde{y}_i(\gamma) = \left((1 + y_i)^\gamma - 1\right)/\gamma$ is a function of γ. Note that Y_i follows the LRMoE with transformed Gamma distributions (TGDs) as the expert functions, where the pdf of the TGD is given by

$$\tilde{f}(y; m, \theta, \gamma) = f\left(\tilde{y}(\gamma); m, \theta\right)(1 + y)^{\gamma - 1}. \tag{2.5}$$

A nice property of the proposed model is its interpretability. The model classifies each claim into one of the g unobservable subgroups. Claim severity distributions vary among subgroups but are homogeneous within a subgroup. Depending on the characteristics (covariates $\boldsymbol{x}_i$) of the claim, each claim has different probabilities of being classified into different subgroups. The subgroup assignments are governed by the regression coefficients $\boldsymbol{\alpha}$ of the gating function. A large positive regression coefficient α_{jp} represents a higher chance for a claim to be classified as subgroup j when x_{ip} is large.

3. DESIRABLE PROPERTIES

3.1. Denseness Property

To justify theoretically the full flexibility of the proposed TG-LRMoE to capture any distribution and regression patterns, we need to show that it satisfies the denseness property proposed by Fung, Badescu, and Lin (2019b). The denseness property ensures that the data generated from the fitted model will be highly synchronous to the input data regardless of the dataset's

characteristics. We first revisit the relevant definitions proposed by Fung, Badescu, and Lin (2019b), dropping the subscript i for $\boldsymbol{x}_i$ (only in this subsection) for a cleaner presentation:

Definition 1. *(Regression distribution) 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.*

Definition 2. *(Denseness property in the context of univariate regression distributions) Let $\mathcal{A}$ be the support of the covariates $\boldsymbol{x}$. In addition, 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 all $F(\mathcal{A}) \in \mathcal{C}_2(\mathcal{A})$ there exists a sequence of regression distributions $\{G_n(\mathcal{A})\}_{n=1,2,...}$ with $G_n(\mathcal{A}) \in \mathcal{C}_1(\mathcal{A})$ for $n = 1, 2, ...$ such that for all $\boldsymbol{x} \in \mathcal{A}, G_n(y;\boldsymbol{x}) \xrightarrow{\mathcal{D}} F(y;\boldsymbol{x})$ as $n \to \infty$. If the convergence $G_n(y;\boldsymbol{x}) \to F(y;\boldsymbol{x})$ is uniform on $\boldsymbol{x} \in \mathcal{A}_y$ for any y, where $\mathcal{A}_y$ is the set of $\boldsymbol{x}$ such that y is a continuity point of $F(y;\boldsymbol{x})$, then $\mathcal{C}_1(\mathcal{A})$ is uniformly dense in $\mathcal{C}_2(\mathcal{A})$.*

We now display the denseness properties of the proposed TG-LRMoE under some very mild assumptions suggested by Fung, Badescu, and Lin (2019b) where the proofs are exhibited in Appendix A.1.

Property 1. *Let $\mathcal{G}_1(\mathcal{A})$ be a class of univariate 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 $G^*(y;\boldsymbol{x})$ is Lipschitz continuous on $\boldsymbol{x} \in \mathcal{A}$ for all y. Assume that $\mathcal{A} = \{1\} \times [m_{\min}, m_{\max}]^P$, where $m_{\min}$ and $m_{\max}$ are finite. Then, the class of TG-LRMoE defined in Equation (2.4) with covariates $\boldsymbol{x} \in \mathcal{A}$ is uniformly dense in $\mathcal{G}_1(\mathcal{A})$.*

The following property provides a slightly stronger result than Property 1, suggesting that the denseness property still holds even if we fix the parameter γ, which governs the tail of the distribution. In other words, the flexibility of the proposed TG-LRMoE is not mainly contributed by the parameter γ.

Property 2. *Under the same settings and assumptions as Property 1, the class of TG-LRMoE with a fixed $\gamma > 0$ and covariates $\boldsymbol{x} \in \mathcal{A}$ is uniformly dense in $\mathcal{G}_1(\mathcal{A})$.*

3.2. Tail Heaviness

The denseness properties of the proposed TG-LRMoE ensure its full flexibility to capture any structures, including the tail behavior of any distributions. However, there is a serious practical concern: there is no control on the number of latent classes g. From a general insurance perspective, the claim severity distributions usually exhibit heavy-tailed behavior. Fitting such distributions using the LRMoE with light-tailed expert functions may require a prohibitively large g, causing overfitting problems. See section 3.4 of Fung, Badescu, and Lin (2019b), which discusses the limitation of denseness theory, for more details.

Therefore, it is essential to theoretically justify the effectiveness of the proposed TG-LRMoE in capturing a wide range of tail behaviors. Then, a small number of latent classes will be able to effectively fit a heavy-tailed dataset. We will first show that the parameter γ predominantly affects the tail of the TG-LRMoE. Adjusting γ, we will show that the proposed model covers both light-tailed distributions (e.g., Weibull distributions) and very heavy-tailed distributions that can be connected to extreme value theory (e.g., Burr distributions). A well-known definition to compare the tail heaviness between two distributions is revisited in Definition A.1 of Appendix A.2.

Denote $Y_i^{(j)}$ ($j \in \{1, ..., g\}$) as a random variable with pdf $f^{(j)}$. Assume that Y_i follows a g-component LRMoE with pdf in the form of Equation (2.2) and with expert functions $f^{(1)}, ..., f^{(g)}$. The tail property for the LRMoE is as follows, with the proof displayed in Appendix A.2.

Property 3. *If $Y_i^{(j)}$ ($j \in \{1, ..., g\}$ with $g < \infty$) has one of the heaviest tails among $Y_i^{(1)}, ..., Y_i^{(g)}$ (i.e., $Y_i^{(j)}$ has a heavier tail or one similar to $Y_i^{(j')}$ for every $j' \in \{1, ..., g\}$), then both $Y_i|\boldsymbol{x}$ and $Y_i|\boldsymbol{x}^c$ have tails similar to $Y^{(j)}$, where the observed covariates $\boldsymbol{x}^c$ are a subset of the complete covariates $\boldsymbol{x}$.*

The property above shows that unless $g = \infty$ (which is impractical), the LRMoE fails to extrapolate any tails heavier or lighter than that of expert functions. As a result, under finite g, the choice of expert functions completely determines how effective the corresponding LRMoE can cater to different tail behaviors.

To evaluate the tail of the proposed TG-LRMoE, it suffices to analyze the tail behavior of the TGD with the pdf displayed in Equation (2.5). We first introduce the following property to gain insights on how the parameters (m, θ, γ) affects the tail of the TGD.

Property 4. *Let* $\mathcal{F} = \{\tilde{f}(\cdot; m, \theta, \gamma); m, \theta, \gamma\}$ *be the class of TGD. Then, there exists a total ordering of* $\mathcal{F}$ *such that* $\tilde{f}(\cdot; m, \theta, \gamma) \prec \tilde{f}(\cdot; m^*, \theta^*, \gamma^*)$ *implies that* $\tilde{f}(\cdot; m^*, \theta^*, \gamma^*)$ *has a heavier tail than* $\tilde{f}(\cdot; m, \theta, \gamma)$.

The proof of the above property is presented in Appendix A.2, which shows that γ is the most important parameter affecting the tail of the TGD, followed by θ and finally m. A smaller γ means a heavier tail and vice versa. In contrast, recall from Property 2 that γ is an unimportant parameter for the denseness property of the TG-LRMoE. As a result, the role of the parameter γ, which mainly governs the tail of the distribution, is very distinctive to that of the parameters m and θ, which mainly govern the body.

After understanding the role of the tail parameter γ, it is crucial to demonstrate that the TG-LRMoE covers a very broad range of tail behavior by varying γ. Table A.1 in Appendix A.2 compares the tail of the TGD compared to that of various commonly used severity distributions. It shows that under the limit $\gamma \to 0$, the TGD can capture very heavy-tailed distributions such as Burr distributions (including Pareto distributions as a special case) that have a polynomial tail. The TGD can also cater to lighter tailed severity distributions, such as Weibull distributions (with shape parameters $k > 1$) when γ is greater than one. Further, we want to show how the TGD can be connected to EVT. When $\gamma \to 0$, the TGD converges to the log-Gamma distribution (LGD) with pdf

$$\tilde{f}(y; m, \theta, 0) := \lim_{\gamma \to 0} \tilde{f}(y; m, \theta, \gamma) = \frac{1}{\Gamma(m)\theta^m} \frac{[\log(1+y)]^{m-1}}{(1+y)^{1+1/\theta}}. \tag{3.1}$$

Note that the LGD has a finite sth moment ($s > 0$) only when $\theta < 1/s$. To connect the LGD to the EVT, we recall what is meant by a regularly varying distribution: A distribution with survival function $S(y)$ is regularly varying with index $\rho > 0$ if $\lim_{y \to \infty} S(y\lambda)/S(y) = \lambda^{-\rho}$ for any $\lambda > 0$.

Under the EVT, any regularly varying distributions are in the maximum domain of attraction of the Fre´chet distribution, which belongs to a class of the generalized extreme value distributions. Such a class of distributions has many desirable properties useful for insurance applications; see, for example, Embrechts, Klüppelberg, and Mikosch (1997) and Ahn, Kim, and Ramaswami (2012) for further discussions. We have the following property for the LGD.

Property 5. *The LGD with the pdf defined by Equation (3.1) is a regularly varying distribution with index* $\rho = 1/\theta$.

As a result, the proposed TG-LRMoE is also a regularly varying distribution with index $\rho = 1/\theta$ as $\gamma \to 0$, implying that it can effectively capture heavy tails of the distributions that are well aligned with the EVT.

3.3. Model Identifiability

In addition to model flexibility, it is desirable that the proposed model is identifiable to ensure that model fitting is unique and to prevent different interpretations for a fitted model. However, as pointed out by, for example, Jiang and Tanner (1999) and Fung, Badescu, and Lin (2019a), identifiability generally fails for the class of LRMoE in the form of Equation (2.2), because the model is invariant under a permutation (i.e., $(\boldsymbol{\alpha}_j, \boldsymbol{\psi}_j) \mapsto (\boldsymbol{\alpha}_{c(j)}, \boldsymbol{\psi}_{c(j)})$ where $\{c(1), ..., c(g)\}$ is a permutation of $\{1, ..., g\}$) or a translation (i.e., $\boldsymbol{\alpha}_j \mapsto \boldsymbol{\alpha}_j + \boldsymbol{\delta}$ where $\boldsymbol{\delta}$ is a column vector with length $P+1$). However, we can still show that the proposed TG-LRMoE is identifiable up to translation and permutation, meaning that any model unidentifiability only the result of translational and permutational invariance properties. Before that, we recall the definition of identifiability for the general class of LRMoE from Fung, Badescu, and Lin (2019a).

Definition 3. *Let* $\mathcal{G}$ *be the class of LRMoE with the pdf in the form of Equation (2.2). Each element* $G_{\boldsymbol{\Phi}, g} \in \mathcal{G}$ *is a regression distribution with covariates* $\boldsymbol{x}_i \in \Omega$, *parameter setting* $\boldsymbol{\Phi} = (\boldsymbol{\alpha}, \boldsymbol{\Psi})$, *and the number of latent classes* g, *where* $\Omega \subseteq \mathbb{R}^{P+1}$ *is the support of* $\boldsymbol{x}_i$. *A subclass* $\bar{\mathcal{G}} \subseteq \mathcal{G}$ *is identifiable up to translation and permutation whenever* $G_{\boldsymbol{\Phi}^*, g^*}, G_{\boldsymbol{\Phi}, g} \in \bar{\mathcal{G}}$, $(\boldsymbol{\alpha}^*_{j_1}, \boldsymbol{\psi}^*_{j_1}) \neq (\boldsymbol{\alpha}^*_{j_2}, \boldsymbol{\psi}^*_{j_2})$ *for all* $j_1 \neq j_2 \in \{1, ..., g^*\}$ *and* $(\boldsymbol{\alpha}_{j_1}, \boldsymbol{\psi}_{j_1}) \neq (\boldsymbol{\alpha}_{j_2}, \boldsymbol{\psi}_{j_2})$ *for all* $j_1 \neq j_2 \in \{1, ..., g\}$, *if*

$$\sum_{j=1}^{g^*} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}^*) f(y_i; \boldsymbol{\psi}^*_j) = \sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f(y_i; \boldsymbol{\psi}_j) \tag{3.2}$$

for all $\boldsymbol{x}_i \in \Omega$ *and* $y_i > 0$ *(in other words, the pdf of* $G_{\boldsymbol{\Phi}, g}$ *matches with that of* $G_{\boldsymbol{\Phi}^*, g^*}$*), it implies that* $g^* = g$ *and* $(\boldsymbol{\alpha}^*_j, \boldsymbol{\psi}^*_j) = (\boldsymbol{\alpha}_{c(j)} + \boldsymbol{\delta}, \boldsymbol{\psi}_{c(j)})$ *for* $j = 1, ..., g$, *where* $\{c(1), ..., c(g)\}$ *is a permutation of* $\{1, ..., g\}$ *and* $\boldsymbol{\delta}$ *is a vector that is constant across all* $j = 1, ..., g$.

The identifiability property with TGD as a special choice of expert funcction is as follows, with the proof shown in Appendix A.3:

Property 6. *The TG-LRMoE (with pdf in the form of Eq. [2.4]) is identifiable up to translation and permutation, subject to the restriction that* $(m_1, \theta_1), ..., (m_g, \theta_g)$ *are distinct and* $\boldsymbol{\Omega}$ *spans* $\mathbb{R}^{P+1}$.

4. PARAMETER ESTIMATION

This section presents an ECM algorithm to efficiently calibrate the model parameters. Estimating parameters of the LRMoE severity distributions imposes an extra challenge that the likelihood function may be unbounded (e.g., for the TG-LRMoE, the likelihood may diverge to infinity when $m_j \to \infty$ and $\theta_j \to 0$ for some j). As a result, maximum likelihood estimations may lead to a spurious model mentioned by McLachlan and Peel (2000), where some of the fitted components (expert functions) have very small variances. To this end, the ECM algorithm presented below for the proposed TG-LRMoE (Eq. [2.4]) is different from the algorithm proposed by Fung, Badescu, and Lin (2019a), in the sense that it further penalizes parameters taking extreme values through finding the AP estimates of the parameters.

Suppose that there are n independent observations $\{(Y_i, \boldsymbol{x}_i); i = 1, ..., n\}$. Hereafter, denote $\boldsymbol{y} = (y_1, ..., y_n)$ and $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ as all observed response variables and covariates respectively. We are to estimate the parameters $\boldsymbol{\Phi} = (\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}, \gamma)$, while g is fixed at each ECM run. We also restrict $\boldsymbol{\alpha}_g = \boldsymbol{0}$ for translational invariance concern. Then, the observed data log-likelihood is given by

$$l(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{n} \log \left[\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i; \boldsymbol{\alpha}) f\left(\tilde{y}_i(\gamma); m_j, \theta_j\right) (1 + y_i)^{\gamma - 1} \right]. \tag{4.1}$$

To penalize parameters taking extreme values, we adopt a Bayesian approach to set up a prior distribution for each parameter. Note that under the Bayesian approach, it makes little sense to be concerned about choosing "accurate" prior distributions. Instead, they should be kept simple to minimize their impacts on the computational burden of the ECM algorithm. To this end, we have chosen the following prior distributions for the parameters.

1. For the regression coefficients $\boldsymbol{\alpha}$, we set $\alpha_{jp} \sim N\left(0, \sigma_{jp}^2\right)$ for $j = 1, ..., g-1$ and $p = 0, ..., P$. This avoids the possibility for the fitted model that certain claim features have a probability of (almost) 1 or 0 being classified to a particular subclass, which is indeed an over-fitting. Also, choosing a Normal prior, the concavity of Equation (4.8) (an important part of the CM step that will be discussed in Subsection 4.2) is conserved, so its optimization still always converges to a global maximum.
2. For the shape and size parameters, we set $m_j \sim Gamma\left(\nu_j^{(1)}, \lambda_j^{(1)}\right)$ and $\theta_j \sim Gamma\left(\nu_j^{(2)}, \lambda_j^{(2)}\right)$ for $j = 1, ..., g$ to prevent spurious fitted models. We will show in Subsection 4.2 that the optimal scale parameter θ_j can be expressed as an analytical form (Eq. [4.13]) if a Gamma prior is chosen.
3. We choose not to penalize γ because this parameter does not result in spurious model or overfitting issue.

It is assumed that the prior distributions of all parameters are mutually independent. In addition, we note that σ_{jp}, $\nu_j^{(1)}$, $\lambda_j^{(1)}$, $\nu_j^{(2)}$, $\lambda_j^{(2)}$ are all fixed numbers chosen prior to each ECM run. The observed data posterior log-likelihood is then given by

$$\begin{aligned} l^{\text{pos}}(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}) &= \log \left[\frac{\left(\prod_{i=1}^{n} h_{Y_i|\boldsymbol{x}_i}(y_i; \boldsymbol{x}_i, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}, \gamma, g)\right) p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta})}{p(\boldsymbol{y}; \boldsymbol{x})} \right], \\ &= l(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}) + \log p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) + \text{const.}, \end{aligned} \tag{4.2}$$

where

$$p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) = \sum_{j=1}^{g-1} \sum_{p=0}^{P} \log p_1(\alpha_{jp}) + \sum_{j=1}^{g} \log p_2(m_j) + \sum_{j=1}^{g} \log p_3(\theta_j), \tag{4.3}$$

$p(\cdot)$ represents the joint prior distribution of the parameters and $p_1(\cdot), p_2(\cdot),$ and $p_3(\cdot)$ are the marginal priors. Now, we introduce a latent random vector $\boldsymbol{Z_i} = (Z_{i1}, ..., Z_{ig})^T \sim Multi_g\big(1, \{\pi_1(\boldsymbol{x}_i;\boldsymbol{\alpha}), ..., \pi_g(\boldsymbol{x}_i;\boldsymbol{\alpha})\}\big)$ such that $Z_{ij} = 1$ if the observation y_i comes from the jth component and $Z_{ij} = 0$ otherwise for $i = 1, ..., n$. The complete data posterior log-likelihood is given by

$$\begin{aligned} l^{\text{pos}}(\boldsymbol{\Phi};\boldsymbol{y},\boldsymbol{x},\boldsymbol{Z}) = &\sum_{i=1}^{n}\sum_{j=1}^{g} Z_{ij}\big(\log \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) + \log f\big(\tilde{y}_i(\gamma); m_j, \theta_j\big)\big) \\ &+ (\gamma - 1)\sum_{i=1}^{n}\log(1 + y_i) + \log p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) + \text{const.}, \end{aligned} \tag{4.4}$$

4.1. E Step

In the lth iteration of the E step, we take an expectation of the complete data posterior log-likelihood given the observed data

$$\begin{aligned} Q\Big(\boldsymbol{\Phi};\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Phi}^{(l-1)}\Big) &= E\Big[l^{\text{pos}}(\boldsymbol{\Phi};\boldsymbol{y},\boldsymbol{x},\boldsymbol{Z})\big|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Phi}^{(l-1)}\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(l)}\left(\log \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) + (m_j - 1)\log \tilde{y}_i(\gamma) - \frac{\tilde{y}_i(\gamma)}{\theta_j} - m_j \log \theta_j - \log(m_j - 1)!\right) \\ &\quad + (\gamma - 1)\sum_{i=1}^{n}\log(1 + y_i) - \sum_{j=1}^{g-1}\sum_{p=0}^{P}\frac{\alpha_{jp}^2}{2\sigma_{jp}^2} \\ &\quad + \sum_{j=1}^{g}\left(\Big(\nu_j^{(1)} - 1\Big)\log m_j - \frac{m_j}{\lambda_j^{(1)}}\right) + \sum_{j=1}^{g}\left(\Big(\nu_j^{(2)} - 1\Big)\log \theta_j - \frac{\theta_j}{\lambda_j^{(2)}}\right) + \text{const.}, \end{aligned} \tag{4.5}$$

where $z_{ij}^{(l)}$ is given by the following for $i = 1, ..., n$ and $j = 1, ..., g$:

$$z_{ij}^{(l)} = E\Big[Z_{ij}\big|\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Phi}^{(l-1)}\Big] = \frac{\pi_j\big(\boldsymbol{x}_i;\boldsymbol{\alpha}^{(l-1)}\big) f\Big(y_i;\boldsymbol{\theta}_{jk}^{(l-1)}\Big)}{\sum_{j'=1}^{g}\pi_{j'}\big(\boldsymbol{x}_i;\boldsymbol{\alpha}^{(l-1)}\big) f\Big(y_i;\boldsymbol{\theta}_{j'k}^{(l-1)}\Big)}. \tag{4.6}$$

4.2. CM Step

In the CM step, we update the parameters $\boldsymbol{\Phi}^{(l-1)}$ such that $Q\Big(\boldsymbol{\Phi}^{(l)};\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Phi}^{(l-1)}\Big) \geq Q\Big(\boldsymbol{\Phi}^{(l-1)};\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Phi}^{(l-1)}\Big)$. Note that $Q\Big(\boldsymbol{\Phi};\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Phi}^{(l-1)}\Big)$ can be decomposed as the following:

$$Q\Big(\boldsymbol{\Phi};\boldsymbol{y},\boldsymbol{x},\boldsymbol{\Phi}^{(l-1)}\Big) = Q^{(l)}(\boldsymbol{\alpha}) + \sum_{j=1}^{g} S_j^{(l)}\big(m_j, \theta_j, \gamma\big) + T^{(l)}(\gamma), \tag{4.7}$$

where

$$Q^{(l)}(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(l)}\log \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) - \sum_{j=1}^{g-1}\sum_{p=0}^{P}\frac{\alpha_{jp}^2}{2\sigma_{jp}^2}, \tag{4.8}$$

$$\begin{aligned} S_j^{(l)}\big(m_j, \theta_j, \gamma\big) = &\sum_{i=1}^{n} z_{ij}^{(l)}\left((m_j - 1)\log \tilde{y}_i(\gamma) - \frac{\tilde{y}_i(\gamma)}{\theta_j} - m_j \log \theta_j - \log(m_j - 1)!\right) \\ &+ \Big(\nu_j^{(1)} - 1\Big)\log m_j - \frac{m_j}{\lambda_j^{(1)}} + \Big(\nu_j^{(2)} - 1\Big)\log \theta_j - \frac{\theta_j}{\lambda_j^{(2)}}, \end{aligned} \tag{4.9}$$

and

$$T^{(l)}(\gamma) = (\gamma - 1)\sum_{i=1}^{n} \log(1 + y_i) + \text{const.}, \tag{4.10}$$

We first update the parameters $\boldsymbol{\alpha}^{(l-1)}$ such that $Q^{(l)}(\boldsymbol{\alpha}^{(l)}) \geq Q^{(l)}(\boldsymbol{\alpha}^{(l-1)})$. To do so, we sequentially (for $j = 1, ..., g-1$) maximize $Q^{(l)}\left(\boldsymbol{\alpha}_1^{(l)}, ..., \boldsymbol{\alpha}_{j-1}^{(l)}, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_{j+1}^{(l-1)}, ..., \boldsymbol{\alpha}_g^{(l-1)}\right)$ (note: $\boldsymbol{\alpha}_g^{(l)} = \boldsymbol{0}$) with respective to $\boldsymbol{\alpha}_j$ to obtain $\boldsymbol{\alpha}_j^{(l)}$, which can be achieved through the iteratively reweighted least squares (IRLS) approach (Jordan and Jacobs 1994). The IRLS requires performing the following iterations until convergence; see section 4.2 of Fung, Badescu, and Lin (2019a) for details:

$$\boldsymbol{\alpha}_j \leftarrow \boldsymbol{\alpha}_j - \left(\frac{\partial^2 Q^{(l)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j^T}\right)^{-1} \frac{\partial Q^{(l)}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_j}. \tag{4.11}$$

Note that $Q^{(l)}(\boldsymbol{\alpha})$ is a concave function, so the above IRLS algorithm always converges to a global maximum. Then, we update the parameters $\left(\boldsymbol{m}^{(l-1)}, \boldsymbol{\theta}^{(l-1)}\right)$ through maximizing $S_j^{(l)}\left(m_j, \theta_j, \gamma^{(l-1)}\right)$ with respect to (m_j, θ_j) for each $j = 1, ..., g$. We have

$$m_j^{(l)} = \underset{m_j > 0}{\text{argmax}} \; S_j^{(l)}\left(m_j, \tilde{\theta}_j^{(l)}(m_j), \gamma^{(l-1)}\right); \qquad \theta_j^{(l)} = \tilde{\theta}_j^{(l)}\left(m_j^{(l)}\right), \tag{4.12}$$

where

$$\tilde{\theta}_j^{(l)}(m_j) = \frac{\lambda_j^{(2)}}{2}\left(\left(\nu_j^{(2)} - 1\right) - m_j \sum_{i=1}^{n} z_{ij}^{(l)} + \sqrt{\left(m_j \sum_{i=1}^{n} z_{ij}^{(l)} - \left(\nu_j^{(2)} - 1\right)\right)^2 + \frac{4}{\lambda_j^{(2)}} \sum_{i=1}^{n} z_{ij}^{(l)} \tilde{y}_i\left(\gamma^{(l)}\right)}\right) \tag{4.13}$$

is obtained by taking a derivative of $S_j^{(l)}\left(m_j, \theta_j, \gamma^{(l-1)}\right)$ with respect to θ_j and setting it as zero. Note that univariate numerical optimization is required to compute $m_j^{(l)}$ in Equation (4.12). Finally, we obtain $\gamma^{(l)}$ through maximizing $\sum_{j=1}^{g} S_j^{(l)}\left(m_j^{(l)}, \theta_j^{(l)}, \gamma\right) + T^{(l)}(\gamma)$ numerically with respect to γ.

The full CM step described above ensures that $Q\left(\boldsymbol{\Phi}^{(l)}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right) \geq Q\left(\boldsymbol{\Phi}^{(l-1)}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right)$. Further, Dempster, Laird, and Rubin (1977) and Meng and Rubin (1993) suggested that the ECM algorithm for the MAP estimation preserves all of the desirable convergence properties as the standard EM algorithm for the maximum likelihood estimation. For example, the observed data posterior log-likelihood is monotone nondecreasing for each iteration and finally converges to a local maximum. The E step and CM step are iterated until the change in the observed data posterior log-likelihood is smaller than a tolerance threshold of 10^{-3} or the maximum number of iterations of 200 is reached.

4.3. Initialization and Parameter Adjustments

Good initialization is essential for fast convergence of the proposed ECM algorithm. We first initialize γ using a simple ad hoc method. Note that if Y_i follows the TG-LRMoE, then $\log S\left(\tilde{y}_i(\gamma); \boldsymbol{m}, \boldsymbol{\theta}\right) = o\left(\tilde{y}_i(\gamma)\right) - c^* \tilde{y}_i(\gamma)$, where $\lim_{y \to \infty} o(y)/y = 0$, c^* is a constant, and the log-survival function of $\tilde{Y}_i = \left((1 + Y_i)^{\gamma} - 1\right)/\gamma$ is given by $\log S$. Trying a wide range of γ, we plot (for each γ) $\log \hat{S}\left(\tilde{y}_i(\gamma)\right)$ (empirical survival function) against $\tilde{y}_i(\gamma)$. We choose the initial tail parameter $\gamma^{(0)}$ such that the plot looks asymptotically linear.

Then, $\boldsymbol{\alpha}$, $\boldsymbol{m}$, and $\boldsymbol{\theta}$ are initialized using a method similar to the clusterized method of moments proposed by Gui, Huang, and Lin (2018). First, perform K-means clustering on $\tilde{y}_i(\gamma)$ with g clusters, which yields the clustering mean $\{\mu_j^{\text{cluster}}\}_{j=1,...,g}$, variance $\left\{\left(\sigma_j^{\text{cluster}}\right)^2\right\}_{j=1,...,g}$ and weights $\{\pi_j^{\text{cluster}}\}_{j=1,...,g}$ (the proportion of observations classified in cluster j). Second, set

$m_j^{(0)} = \left(\mu_j^{\text{cluster}}/\sigma_j^{\text{cluster}}\right)^2$ and $\theta_j^{(0)} = \left(\sigma_j^{\text{cluster}}\right)^2/\mu_j^{\text{cluster}}$ to match the first two moments for each cluster. Third, set $\alpha_{j0}^{(0)} = \log\left(\pi_j^{\text{cluster}}/\pi_g^{\text{cluster}}\right)$ and $\alpha_{jp}^{(0)} = 0$ for $p > 0$.

Finally, to determine the optimal g, we try a wide range of g and find one that optimizes the Akaike information criterion (AIC) or the Bayesian information criterion (BIC).

Remark 1. During the model selection procedure, it is important to understand the strengths and weaknesses of AIC and BIC. With a heavier penalty on model complexity, BIC usually yields a model with fewer components that is more interpretable. Kuha (2004) also showed that BIC is consistent in selecting the true model, but this is of serious practical concern because it is impossible to identify the true model in any real datasets. Instead, BIC often over-penalizes the model complexity, which may lead to inferior predictive power. On the other hand, AIC is designed to select models that optimize the predictive power. These properties will be empirically demonstrated in the real data analysis (Section 5).

5. APPLICATION TO IBNR PREDICTION

We demonstrate the usefulness of the proposed TG-LRMoE regression model for IBNR prediction by analyzing a real insurance dataset. Subsection 5.1 outlines the main features of the real insurance dataset we have collected. Because it contains very detailed policyholder information for each contract and claim, we are able to incorporate covariates under the individual reserving approach, which may enable us to understand how the claims are influenced by the individual information and to evaluate the IBNR reserves more accurately. Subsection 5.2 identifies three components to be modeled for IBNR prediction, namely, severity, reporting delay, and frequency. The model fitting aspects are discussed in Subsection 5.3. Though the TG-LRMoE was chosen for severity regression modeling, we have applied a standard Poisson process GLM framework for frequency modeling because severity modeling is the main focus of this article. Some possible extensions on the Poisson process for frequency modeling, such as time series models, are discussed briefly in Section 6 as one of our potential research directions. In addition, we will show how the reporting delay can be modeled through a slightly modified version of the TG-LRMoE. Finally, we present the out-of-sample IBNR prediction in Subsection 5.4, which also analyzes the importance of including covariates in determining accurately the IBNR predictive distributions and understanding policyholders' characteristics for unreported claims.

5.1. Data Overview

The dataset was supplied by a European major automobile insurer. It contains 594,908 third-party liability insurance contracts during the observation period from January 1, 2007 to December 31, 2017, where the number of in-force contracts (total exposure) over time is displayed in the left panel of Figure 1. For each contract, the contract number, starting date, ending date and various policyholder features (see Table 1 for detailed descriptions of the covariates) are recorded. Among all contracts, 28,256 claims are incurred and reported on or before December 31, 2017, where the exposure-adjusted weekly number of claims is plotted in the right panel of Figure 1. The claim frequencies slightly decline over time. In addition, because the exposure is increasing over time, the exposure-adjusted frequencies gradually become less fluctuating. For each claim, the contract number, loss date, reporting date (if available), settlement date, and total amount paid are recorded. For unsettled claims, the insurer also provides the case reserve estimates (the expected future payments) based on detailed claim-specific information. The total incurred loss of a claim is then the sum of the amount paid and the case reserve estimate.

To evaluate the predictive power through the out-of-sample test (Subsection 5.4), we set a validation date τ of December 31, 2012 and divide the dataset into two parts: an in-sample training set containing all 199,730 contracts in-force and 9,608 claims reported between January 1, 2007, and December 31, 2012 and an out-of-sample (OS) test set containing claims reported between January 1, 2013, and December 31, 2017.

5.2. Modeling Framework

This subsection discusses the modeling framework for the aforementioned insurance dataset. Suppose that the development of the lth claim of the kth contract is described as a triplet $\left(T_l^{(k)}, U_l^{(k)}, Z_l^{(k)}\right)$, where $T_l^{(k)}$ is the accident time, $U_l^{(k)}$ is the reporting delay (in days), and $Z_l^{(k)}$ is the development process after the claim is reported. We also define $N_k^a(t) = \sum_{l=1}^{\infty} 1\{T_l^{(k)} \le t\}, N_k^r(t) = \sum_{l=1}^{\infty} 1\{T_l^{(k)} \le t, T_l^{(k)} + U_l^{(k)} \le \tau\}$, and $N_k^u(t) = \sum_{l=1}^{\infty} 1\{T_l^{(k)} \le t, T_l^{(k)} + U_l^{(k)} > \tau\}$ as the total claim, reported claim, and IBNR claim count processes for the kth contract, where $1\{\cdot\}$ is an indicator function and $N_k^a(t) = N_k^r(t) + N_k^u(t)$.

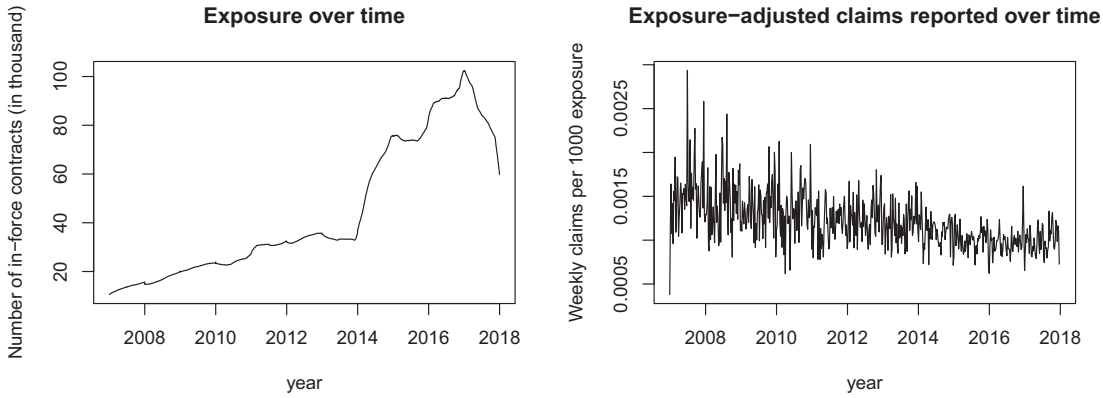

FIGURE 1. Number of In-Force Contracts and Weekly Number of Claims Reported (Exposure Adjusted) versus Time.

TABLE 1
Summary of the Covariates for the kth Contract

Variable	Description	Type	Levels
x_{k1}	Policyholder age	Discrete	
x_{k2}	Car age	Discrete	
x_{k3}	Car fuel	Categorical	Diesel: $x_{k3} = 1$
			Gasoline: $x_{k3} = 0$
x_{k4}–x_{k7}	Geographical location	Categorical	Region I: $x_{k4} = 1$
			Region II: $x_{k5} = 1$
			Region III: $x_{k6} = 1$
			Region IV: $x_{k7} = 1$
			Capital: $x_{k4} = x_{k5} = x_{k6} = x_{k7} = 0$
x_{k8}–x_{k9}	Car brand class	Categorical	Class A: $x_{k8} = 1$
			Class B: $x_{k9} = 1$
			Class C: $x_{k8} = x_{k9} = 0$
x_{k10}	Contract type	Categorical	Renewal contract: $x_{k10} = 1$
			New contract: $x_{k10} = 0$

We follow the framework of Norberg (1993) and Antonio and Plat (2014), which models the claim process through a position dependent marked Poisson process. For demonstrative purposes, the following assumptions are proposed for the dataset:

1. The developments of each contract $\{T_l^{(k)}, U_l^{(k)}, Z_l^{(k)}\}_{l=1,2,...}$ are independent of each other for $k = 1, 2, ..., m$, where m is the total number of contracts.
2. Conditioned on the covariates of the kth contract $\boldsymbol{x}_k$, the claim arrival of contract k follows a Poisson process with intensity measure $\lambda(t|\boldsymbol{x}_k) = \omega_k \exp\{\boldsymbol{x}_k^T \boldsymbol{\beta}\}$ and the associated mark distribution $P_{U,Z|\boldsymbol{x}_k} := P_{U|\boldsymbol{x}_k} P_{Z|U,\boldsymbol{x}_k}$, where ω_k is the exposure of contract k and $\boldsymbol{\beta}$ corresponds to regression coefficients.

Remark 2. Although the claim frequencies shown in the right panel of Figure 1 slightly decline over time, we will show in Subsection 5.3.3 and Appendix C.2 that such a time trend can be adequately explained by the change in insurance portfolio over time (e.g., lower proportion of young drivers). In other words, the impact of time on the claim frequencies is insignificant after controlling for the policyholder covariates, so it is reasonable to exclude the time effect in the Poisson process.

Using the results of Norberg (1993) and Antonio and Plat (2014), the reported claim process and the IBNR claim process for contract i are independent Poisson processes with measures

$$\lambda(dt|\boldsymbol{x}_k) P_{U|\boldsymbol{x}_k}(\tau - t) 1\{t \in [0, \tau]\} \cdot \frac{P_{U|\boldsymbol{x}_k}(du) 1\{u \leq \tau - t\}}{P_{U|\boldsymbol{x}_k}(\tau - t)} \cdot P_{Z|u,\boldsymbol{x}_k}(dz) \tag{5.1}$$

on $\mathcal{C}^r := \{(t,u,z) : t \le \tau, t+u \le \tau\}$ and

$$\lambda(dt|\boldsymbol{x}_k)\big(1 - P_{U|\boldsymbol{x}_k}(\tau - t)\big)1\{t \in [0,\tau]\} \cdot \frac{P_{U|\boldsymbol{x}_k}(du)1\{u > \tau - t\}}{1 - P_{U|\boldsymbol{x}_k}(\tau - t)} \cdot P_{Z|u,\boldsymbol{x}_k}(dz) \tag{5.2}$$

on $\mathcal{C}^u := \{(t,u,z) : t \le \tau, t+u > \tau\}$.

Denote $\{t_l^{(k)r}, u_l^{(k)r}, z_l^{(k)r}\}_{l=1,\ldots,n_k^r(\tau);k=1,\ldots,m}$ as the observed (reported) claims, where $n_k^r(\tau)$ is a realization of $N_k^r(\tau)$. Upon a trivial extension to Antonio and Plat (2014), the likelihood of the observed claims can be expressed as

$$\mathcal{L} \propto g_1(Z) \times g_2(U) \times g_3(N^r), \tag{5.3}$$

where the component $g_1(Z)$ corresponding to the development process is given by

$$g_1(Z) = \prod_{k=1}^{m} \prod_{l=1}^{n_k^r(\tau)} P_{Z|u_l^{(k)r},\boldsymbol{x}_k}\left(dz_l^{(k)r}\right), \tag{5.4}$$

the component $g_2(U)$ corresponding to the reporting delay is given by

$$g_2(U) = \prod_{k=1}^{m} \prod_{l=1}^{n_k^r(\tau)} \frac{P_{U|\boldsymbol{x}_k}\left(du_l^{(k)r}\right)}{P_{U|\boldsymbol{x}_k}\left(\tau - t_l^{(k)r}\right)}, \tag{5.5}$$

and the component $g_3(N^r)$ corresponding to the claim frequency is given by

$$g_3(N^r) = \prod_{k=1}^{m} \left[\prod_{l=1}^{n_k^r(\tau)} \lambda\left(t_l^{(k)r}|\boldsymbol{x}_k\right) P_{U|\boldsymbol{x}_k}\left(\tau - t_l^{(k)r}\right)\right] \exp\left(-\int_0^\tau \lambda(t|\boldsymbol{x}_k) P_{U|\boldsymbol{x}_k}(\tau - t)dt\right). \tag{5.6}$$

5.3. Model Fitting

This subsection proposes the specific models for the development process Z and the reporting delay U and also estimates all of the parameters such that $\mathcal{L}$ in Equation (5.3) is near to its maximum. However, direct maximization of $\mathcal{L}$ is difficult because of its complicated form; that is, $g_3(N^r)$ is affected by U. Instead, we adopt a two-step approach similar to Badescu et al. (2019): The development process loglikelihood function $g_1(Z)$ and the reporting delay likelihood $g_2(U)$ are first (sub-)optimized separately. Then, given that the parameters involved in $g_2(U)$ are obtained and fixed, we maximize the frequency likelihood $g_3(N^r)$. The two-step approach allows us to calibrate the three components separately.

5.3.1. Severity

For the sake of predicting IBNR, we need to model three components: frequency, reporting delay, and severity. To model the severity component, we treat the total amount of payments of each claim $\{Y_i\}_{i=1,\ldots,n}$ as the development processes (originally denoted as $\{Z_l^{(k)}\}_{l=1,\ldots,n_k^r(\tau);k=1,\ldots,m}$), where $n = \sum_{k=1}^m n_k^r(\tau)$ is the total number of claims observed during the in-sample period. Combining with the distributions of the remaining two components, the IBNR predictive distribution will be obtained. Note that because of the settlement delay, not all claims reported have been fully paid and settled, so some of the claim severities Y_i are not fully certain or observed until the valuation date τ. However, after a claim is reported, the insurer usually has very detailed case-by-case information about the accident, providing a case reserve estimate for each reported but not settled claim. For simplicity, we assume that the case reserve estimates are accurate, so we do not differentiate between the actual total payments and the incurred losses (i.e., amounts paid before the valuation date plus the case reserve estimates).

Remark 3. We admit that the above assumption may be quite strong in practice, so one may be concerned about the impact if the assumption is not satisfied. Because only about 5% of the 9,608 reported claims in our training set are not fully settled until the valuation date τ, we do not expect that inaccurate case reserve estimations would bring significant distortion to the severity

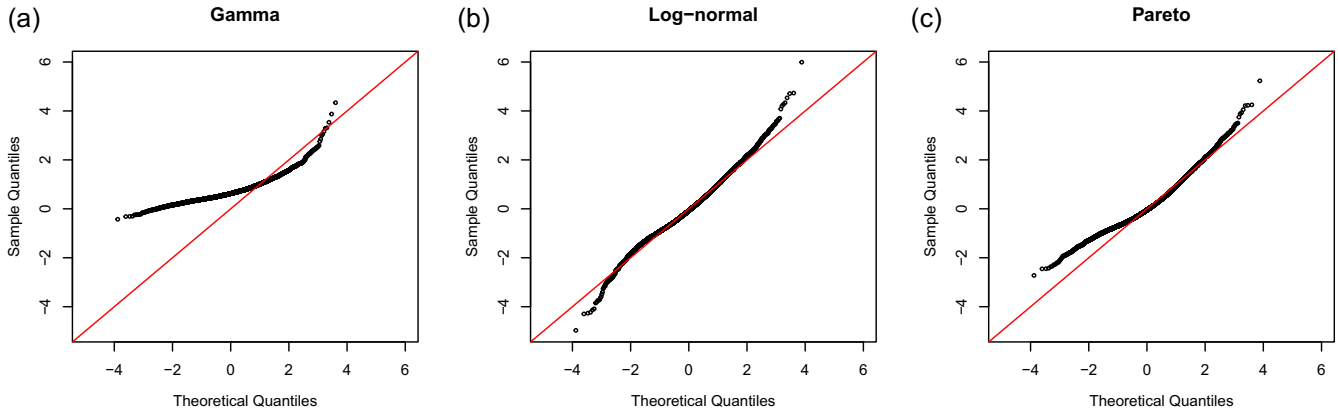

FIGURE 2. Q-Q Normal Plots for the Normalized Residuals under (a) Gamma, (b) Log-normal, and (c) Pareto GLMs.

TABLE 2
Model Selection Statistics and p Values of Goodness-of-Fit Statistics under Various GLM Models and the Proposed LRMoE

	Model selection statistics			p Values of goodness-of-fit statistics		
	Log-likelihood	AIC	BIC	Kolmogorov-Smirnov test	χ^2 test	Anderson-Darling test
Gamma GLM	–90,974.7	181,975.3	182,068.5	$<10^{-7}$	$<10^{-7}$	$<10^{-7}$
Log-normal GLM	–89,303.9	178,633.8	178,727.1	$<10^{-7}$	$<10^{-7}$	$<10^{-7}$
Pareto GLM	–89,936.3	179,898.5	179,991.7	$<10^{-7}$	$<10^{-7}$	$<10^{-7}$
TG-LRMoE (nine components)	**–88,766.4**	**177,762.8**	178,587.4	.9809	.7110	1.0000
TG-LRMoE (four components)	–88,927.1	177,944.1	**178,266.8**	.6943	.2183	.7116

Note: The bold numbers represent the largest value of the Log-likelihood and the smallest AIC and BIC values.

loss distribution. Alternatively, one may discard the unsettled claims for severity modeling, but preliminary analysis shows that settlement delay is correlated to claim severity. Doing so introduces slight bias and leads to underestimations of loss severities.

We first perform a preliminary analysis by fitting Gamma (light-tailed), lognormal (heavy-tailed), and Pareto (extreme-tailed). GLMs, which are widely adopted severity regression models in practice, to the severity dataset. We select all of the variables (x_{i1}–x_{i10}) in Table 1 and the transformed reporting delay $x_{i11} := \log(1+u_i)$ as the covariates. With slight abuse of notations, we here denote $\boldsymbol{x}_i$ as the covariates of the contract that arise at the ith claim. The model density functions are as follows:

- Gamma GLM: $h^G(y_i;\boldsymbol{x}_i,\boldsymbol{\beta},m) = \frac{y_i^{m-1}e^{-y_i/\theta}}{\Gamma(m)\theta^m}$, where $\log\theta = \boldsymbol{x}_i^T\boldsymbol{\beta}$;
- Lognormal GLM: $h^{LN}(y_i;\boldsymbol{x}_i,\boldsymbol{\beta},\sigma) = \frac{1}{y_i\sigma\sqrt{2\pi}}e^{-\frac{(\log y_i-\mu)^2}{2\sigma^2}}$, where $\mu = \boldsymbol{x}_i^T\boldsymbol{\beta}$;
- Pareto GLM: $h^P(y_i;\boldsymbol{x}_i,\boldsymbol{\beta},\alpha) = \frac{\alpha\lambda^\alpha}{(y_i+\lambda)^{\alpha+1}}$, where $\log\lambda = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

To access the goodness of fit, we do the residual analysis and perform three different goodness of fit tests. The goal is to test the null hypothesis (H_0) that the severity data are generated from the fitted model against the alternative hypothesis (H_1) that H_0 is false. The fitted cumulative distribution functions (residuals) $\hat{H}_i := H^q\left(Y_i;\boldsymbol{x}_i,\hat{\boldsymbol{\Phi}}\right)$ are computed for $i = 1,\ldots,n$, where $q \in \{G, LN, P\}$, and $\hat{\boldsymbol{\Phi}}$ are the corresponding fitted model parameters. If H_0 is true, then $\hat{H}_i \sim U[0,1]$. As a result, $\{\hat{H}_i\}_{i=1,\ldots,n}$ is compared to $U[0,1]$ through Q-Q normal plots and several goodness-of-fit statistics (Kolmogorov-Smirnov test, chi-square test [using 200 equiprobable intervals], and Anderson-Darling test). The results are displayed in Figure 2 and

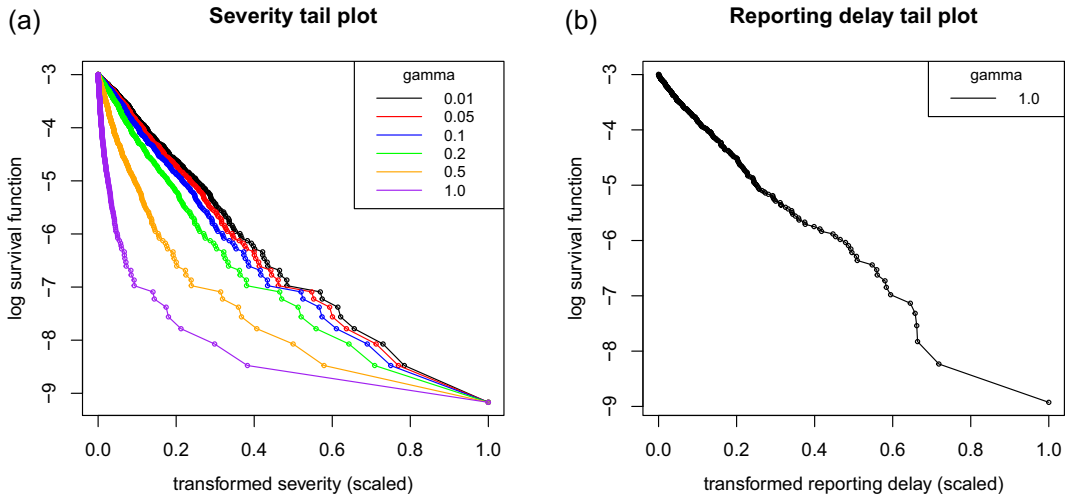

FIGURE 3. Log $\hat{S}(\tilde{y}_i(\gamma, \boldsymbol{m}, \boldsymbol{\theta})$ Plotted against $\tilde{y}_i(\gamma)$ for (a) Severity and (b) Reporting Delay. *Note:* Only 5% of the data points corresponding to the right tail are considered.

Table 2, where the p values can be directly obtained using the functions `ks.test, chisq.test` (under `stats` package), and `ad.test` (under `ADGofTest` package) in `R`. Among the three distributions, the lognormal GLM provides a relatively better fitting, but all three goodness-of-fit statistics still report extremely small p values (10^{-7}), so there is substantial room for improvement of the fitting performance.

Y_i is now modeled through the proposed TG-LRMoE, using the same set of variables (x_{i1}–x_{i11}) as the covariates $\boldsymbol{x}_i$. Maximizing $g_1(Z)$ in Equation (5.4) is equivalent to maximizing the observed log-likelihood in Equation (4.1), but this may lead to a spurious model as discussed in Section 4. We therefore optimize the penalized log-likelihood in Equation (4.2) using the proposed ECM algorithm. We select and justify the penalization hyper-parameters as follows:

- We choose $\sigma_{j0} = 3$ and $\sigma_{jp} = 2/\left(\max_{i=1,\ldots,n}\{x_{ip}\} - \min_{i=1,\ldots,n}\{x_{ip}\}\right)$ for $j = 1, \ldots, g-1$ and $p = 1, \ldots, P$. Under such a choice, each covariate may influence the relative probability of a claim being classified to a subgroup up to approximately a factor of e^2, which is quite large. The weights among different subgroups may span roughly a factor of e^3, which is also large.
- We choose $\nu_j^{(1)} = \nu_j^{(2)} = 1$ and $\lambda_j^{(1)} = \lambda_j^{(2)} = 500$ for $j = 1, \ldots, g$. Therefore, the prior is an exponential distribution with the pdf decaying very slowly.

The prior distributions chosen above can be regarded as weak priors, which allow for minimal distortions to the fitted model. Therefore, the fitted model remains predominantly driven by the data instead of the prior distributions.

To apply the proposed ECM algorithm, we first initialize the parameters in accordance with Subsection 4.3. Figure 3(a) implements the simple ad hoc method to initialize γ. We find that the TG-LRMoE can adequately capture the tail of the severity data when $\gamma \leq 0.2$ (the plots look asymptotically linear), reflecting a very heavy tail. Because we cannot see much difference between $\gamma = 0.01$ and $\gamma = 0.2$, we perform multiple initializations ($\gamma \in \{0.01, 0.05, 0.1, 0.2\}$) and choose one that finally yields the highest posterior observed log-likelihood. The remaining parameters are then initialized using the clusterized method of moments.

According to the AIC and BIC, the optimal fitted model contains nine and four components, respectively. The fitted parameters and the related quantities are displayed in Table C.1 in Appendix C.1. For the AIC model, components 1 and 4 have relatively large subgroup (transformed) means, possibly representing types of serious accidents generally resulting to large claims. The right tail of the severity distribution is mostly governed by components 4 and 5, as reflected by relatively large $\hat{\theta}_j$ (recall Properties 3 and 4). These two components may correspond to two types of accidents that sometimes bring unexpected huge claims. Further, the regression coefficients in Table C.1 provide some insights on the influence of covariates. For example, positive $\hat{\alpha}_{j3}$ for $j = 1$ and $j = 4$ means that claims involving diesel vehicles ($x_{i3} = 1$) are more likely assigned to subgroups 1 and 4, which generally involve greater claim amounts. Similarly, positive large $\hat{\alpha}_{j11}$ for $j = 4$ and $j = 5$ means that the tail heaviness of the severity distribution is positively related to the reporting delay. More details on the covariate influence will be discussed later in this subsection. For the BIC model, similar model interpretations can be attained.

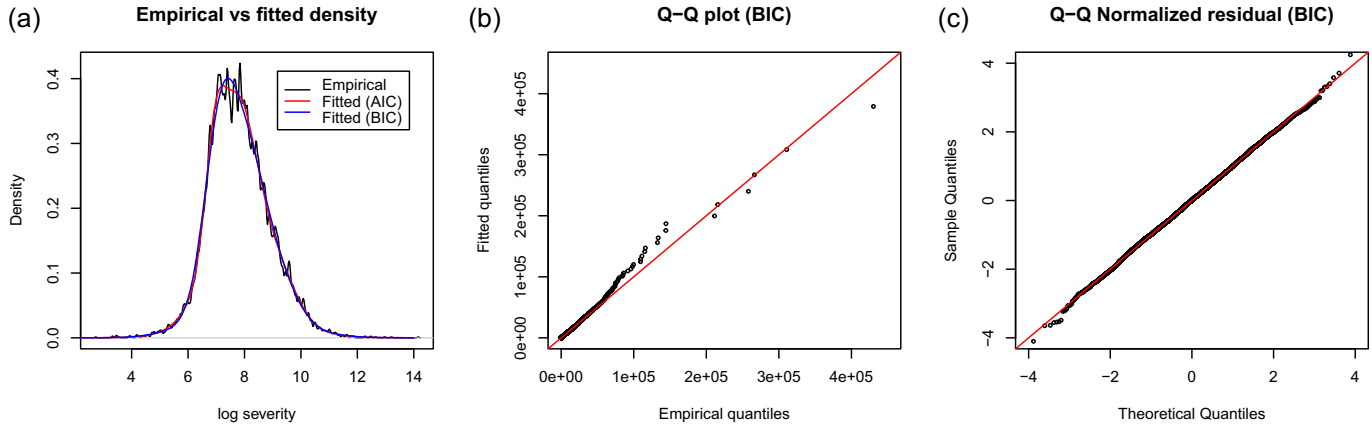

FIGURE 4. (a) Empirical and Fitted Log-Transformed Density Plot. *Note:* Empirical density is generated by kernel method with small bandwidth. (b) Q-Q Plot for Severities (BIC Model). (c) Q-Q Normal Plot for the Normalized Residuals (BIC Model).

To evaluate the fitting performance, the log-likelihood, AIC, and BIC of the proposed fitted models are compared to the fitted GLM models. The results displayed in Table 2 support the use of the TG-LRMoE over the GLM. Residual analysis and goodness-of-fit tests are also performed on the fitted TG-LRMoE and the p values are presented in the same table. Because all of the p values obtained are greater than the threshold .05, all three goodness-of-fit tests do not reject our fitted models. We further compare the empirical density function to the fitted density function and present the Q-Q plots in Figure 4 (note: only the BIC model is demonstrated for conciseness, because very similar plots are obtained for the AIC model). Overall, the proposed TG-LRMoE well captures the distributional structure of the data. One concern is the ability of the proposed mixture-type model to extrapolate the tail heaviness implied by the data. In some cases, even if the fitted model closely follows the observed data points, the tail may be overfitted by a large amount of mixture components, where each of those is specially fitting one (or at most a few) data point on the right tail. Such a problem does not exist in our fitted model. The number of components is reasonably controllable, and even for the more complex AIC model the smallest subgroup weight is 0.019, which is equivalent to about $0.019 \times 9608 = 180$ data points, meaning that none of the components are specifically fitting very few data points.

Remark 4. From Table 2, a p value of 1 is obtained by the Anderson-Darling test under the nine-component TG-LRMoE, which would happen when the fitted model very closely approximates the distribution of the empirical severity data. Therefore, it is important to check the nine-component model and ensure that it does not suffer from overfitting. Figure 4(a) shows that the density of the fitted nine-component model is very smooth compared to the empirical density, addressing the concern of overfitting.

We also study the influence of covariates to the claim severities using the visualization tools (under the non-parametric approach) similar to Fung, Badescu, and Lin (2019a).

The left panel of Figure 5 displays how the reporting delay impacts the 50%, 75%, and 95% quantiles of claim severities. The 95% confidence intervals, which consider parameter uncertainties, are obtained by parametric bootstrap: For $b = 1, ..., B$ (we choose $B = 500$), we simulate the responses $\boldsymbol{y}^{(b)} := \left(y_1^{(b)}, ..., y_n^{(b)}\right)$ and refit $\left(\boldsymbol{y}^{(b)}, \boldsymbol{x}\right)$ using the proposed TG-LRMoE using the fitted model parameters $\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{\theta}}, \hat{\gamma}\right)$ as the initialized values. From the plots, both AIC and BIC fitted models identify a positive relationship between claim size and reporting delay. The 95% quantile fitted curves grow faster than the 50% quantile curve when reporting delay increases, meaning that reporting delay affects the tail more than the body of the distribution. The AIC model captures a sharp spike in the average claim severities when the reporting delay is very long (i.e., $x_{i11} > 5$ or reporting delay > 150 days), but this complex feature is missed by the BIC model, which involves fewer model parameters.

The violin plot in the right panel of Figure 5 shows how geographical location affects the claim severities. It is revealed that policyholders in region III or the capital region file larger claims on average than those in other regions.

The visualization of the impacts of other covariates are not presented for conciseness, we find that "young driver," "car age of about 5 years," "diesel car," "car classified in class C," and "new insurance contract" significantly contribute to larger claim sizes.

The importance of penalizing the parameters (maximizing the MAP instead of the likelihood in the ECM algorithm) are further investigated. We briefly summarize the findings as follows:

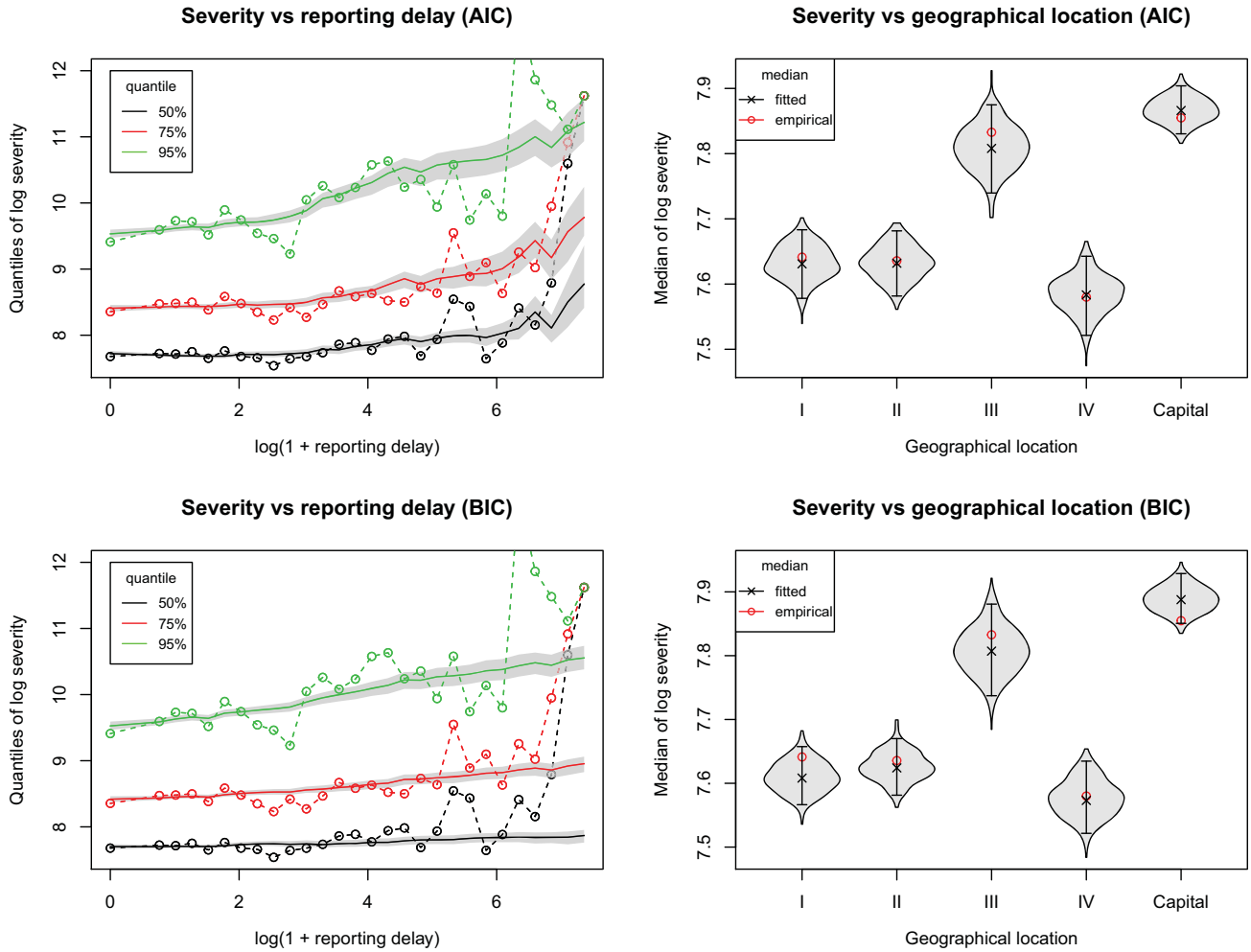

FIGURE 5. (Left) Quantiles of Log Severity versus Reporting Delay Evaluated Using the Nonparametric Approach. *Note:* Dotted curve: empirical pattern; solid curve: fitted pattern; shaded region: 95% confidence interval. (Right) Median of Log Severity versus Geographical Location via a Violin Plot with 95% Confidence Interval Shown by a Bar.

1. Without penalizing the regression parameters $\boldsymbol{\alpha}$, some fitted regression coefficients become extremely large in magnitude ($|\hat{\alpha}_{jp}| > 20$). The numerically unstable results imply that the fitted model overfits one or a few policyholder features.
2. Without penalizing the scale parameters $\boldsymbol{m}$ and scale parameters $\boldsymbol{\theta}$, the fitted model shows very little difference when we fix $g = 9$. In contrast, when we choose a larger number of components (let's say $g = 16$), the fitted model with penalization still behaves properly, but fitting without penalization generates a spurious model (i.e., we find a component with $\hat{m}_j > 10,000$ and $\hat{\theta}_j < 0.001$).

5.3.2. Reporting Delay

To appropriately model the reporting delay distribution, we need to consider $g_2(U)$ in Equation (5.5), a likelihood function involving the right-truncated reporting delays of all $n = 9,608$ observed claims. However, direct maximization of $g_2(U)$ is difficult because the denominator $P_{U|\boldsymbol{x}_i}(\tau - t_i^r)$ (t_i^r is the loss date of the ith claim), which is the probability of truncation, varies among the observed claims. To alleviate such a computational issue, we discard some data points where the losses occur after a specified date τ_0. Choosing December 31, 2011, as τ_0, the truncation probabilities $P_{U|\boldsymbol{x}_i}(\tau - t_i^r)$ are very close to one for the remaining $n^r = 7,524$ data points because we observe that more than 99% of the claims are reported within a year. Therefore, it suffices to consider the ordinary (untruncated) likelihood function for the 7,524 reporting delay observations.

In addition to the data truncation issue, the observed reporting delays are interval censored (i.e., the number of days [integers] instead of continuous values are observed). It is inappropriate to model the reporting delay directly through a continuous

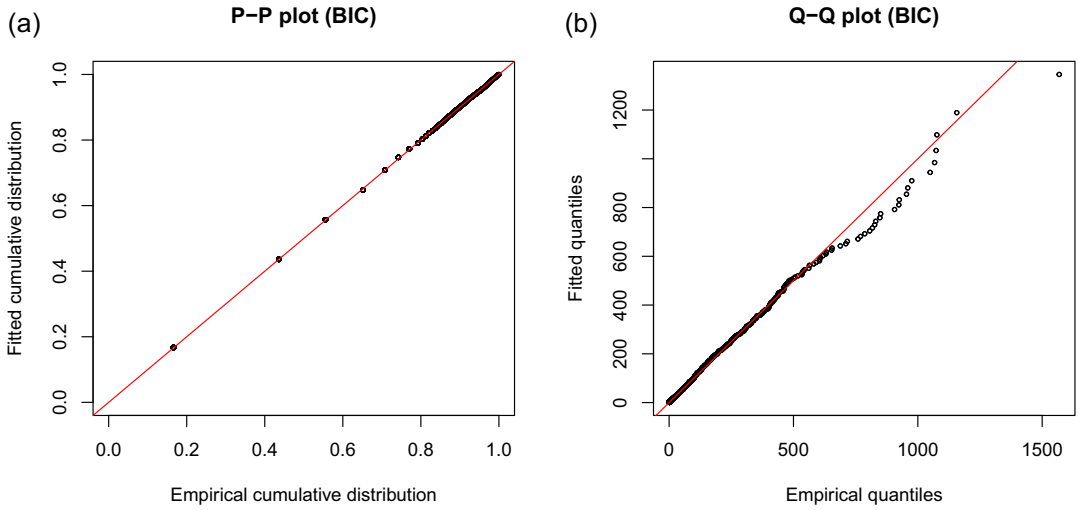

FIGURE 6. (a) P-P Plot and (b) Q-Q Plot for Reporting Delay under the BIC Model.

distribution. We therefore assume that the observed reporting delay U_i (given the covariates x_{i1}–x_{i10} in Table 1) follows an interval-censored version of TG-LRMoE (i.e., the uncensored delay $\tilde{U}_i$ follows the TG-LRMoE and $U_i = \lfloor \tilde{U}_i \rfloor$ where $\lfloor \cdot \rfloor$ is a floor function) for $i = 1, ..., n^r$. For technical ease we also fix $\gamma = 1$, which is an appropriate assumption as demonstrated by an asymptotically linear plot in Figure 3(a). The parameter calibration procedures, which are derived in Appendix B, are very similar to those presented in Section 4.

The AIC and BIC fitted models contain eight and four components, respectively, where the fitted parameters are displayed in Table C.3 in Appendix C.1. Under the AIC fitted model as an example, a large proportion (around 70%) of claims belong to components 2 to 6, representing several types of accidents or claims that are usually quickly reported. In contrast, components 1 and 7 correspond to claims with substantial reporting delays. For the evaluations of the goodness of fit, both fitted models produce similar P-P and Q-Q plots. The plots under the BIC model are displayed in Figure 6. Both plots reveal that the censored version of TG-LRMoE provides excellent fit to both the body and the tail of the reporting delay data.

5.3.3. Frequency

After fitting the severity and reporting delay distributions, the remaining piece is to calibrate the frequency model; that is, to maximize $g_3(N^r)$ in Equation (5.6). This is equivalent to performing a Poisson GLM with response variables being the number of claims reported up to the valuation date τ, covariates being all of the variables displayed in Table 1, and offsets being $\log\left[\omega_k \int_0^\tau P_{U|\boldsymbol{x}_k}(\tau - t)dt\right]$ for each contract k. Note that $\int_0^\tau P_{U|\boldsymbol{x}_k}(\tau - t)dt$ can be numerically computed based on the fitted model of the previous subsection. The fitting procedure can be easily implemented by standard statistical software, such as the `glm` function in `R`. Because severity modeling is our main focus, the frequency fitted parameters are displayed in Table C.5 (left panel) in Appendix C.2. In addition observing in the right panel of Figure 1 that the exposure-adjusted claim frequencies slightly decline over time, we examine the time effect by including policyholder contract date as a covariate and refit the Poisson GLM. Table C.5 (right panel) reveals that the impact of such a time covariate is statistically insignificant after controlling for other covariates, so the use of homogeneous Poisson GLM is reasonable.

5.4. Model Prediction and Out-of-Sample Test

After obtaining the fitted models, it is important to predict the aggregate IBNR at the valuation date of December 31, 2012 and examine the predictive power through OS testing. The predictive distribution of the aggregate IBNR is obtained by simulations, requiring the following steps for each contract $k = 1, ..., m$:

1. Generate the number of claims for the contract $n_k^a(\tau)$ from the fitted frequency GLM described in Subsection 5.3.3.
2. For $l = 1, ..., n_k^a(\tau)$, simulate the accident date of the lth claim $t_{kl}^{\text{acc}} \sim \text{Uniform}[t_k^{\text{start}}, \min\{t_k^{\text{end}}, \tau\}]$, where t_k^{start} is the contract start date (start of the exposure period) and t_k^{end} is the contract end date.
3. For $l = 1, ..., n_k^a(\tau)$, simulate the reporting dalay t_{kl}^{rdelay} from the fitted interval-censored version of TG-LRMoE shown in Subsection 5.3.2.

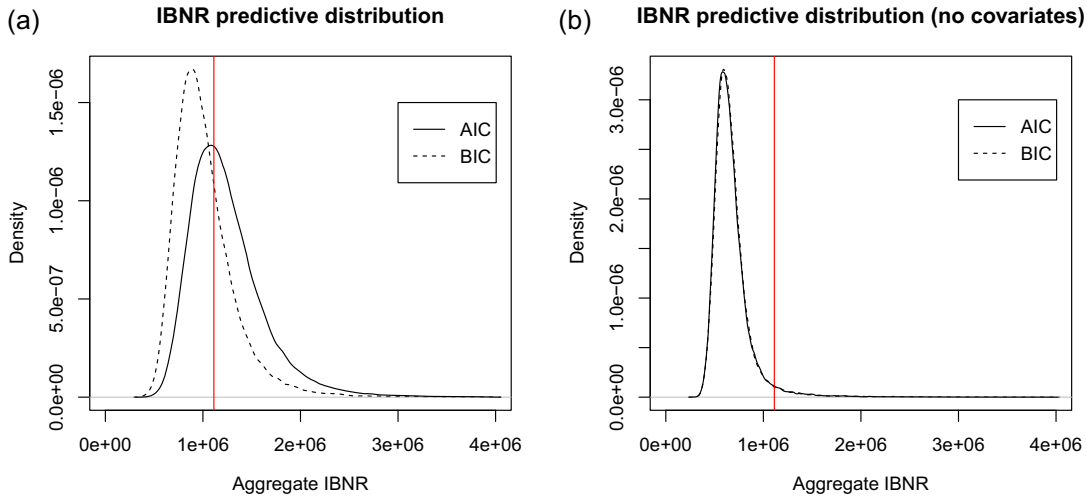

FIGURE 7. Predictive Distribution of the IBNR (a) with Covariates and (b) without Covariates. *Note:* (right panel). The vertical line is the realized aggregate IBNR based on the OS dataset.

TABLE 3
Summary Statistics of the IBNR Predictions

		CTE		VaR			
	Mean	70%	95%	95%	99.5%	Realized	p Value
With covariates (AIC)	1.250	1.718	2.418	1.966	3.023	≥1.112	.836
With covariates (BIC)	1.026	1.406	2.011	1.609	2.485	≥1.112	.597
Without covariates (AIC)	0.676	0.899	1.412	0.992	1.903	≥1.112	.063
Without covariates (BIC)	0.687	0.927	1.568	1.000	2.224	≥1.112	.066

Note: The mean, CTE, VaR, and realized value are expressed in millions.

4. For $l = 1, ..., n_k^a(\tau)$, simulate the unreported amount for each claim y_{kl}^{IBNR}, which is generated from the fitted TG-LRMoE in Subsection 5.3.1 if $t_{kl}^{\text{acc}} + t_{kl}^{\text{rdelay}} > \tau$ and is equal to zero if $t_{kl}^{\text{acc}} + t_{kl}^{\text{rdelay}} \leq \tau$.

The aggregate predicted amount of IBNR claims is given by $y_{\text{agg}}^{\text{IBNR}} = \sum_{k=1}^{m} \sum_{l=1}^{n_k^a(\tau)} y_{kl}^{\text{IBNR}}$. Repeating the above procedure 50,000 times, we display the predictive distribution of the aggregate IBNR in Figure 7(a). The vertical line is the total IBNR claims from the test set. Note that only information up to 2017 is available, so the true IBNR may be larger than the value displayed. Yet, a reporting delay of 5+ years is extremely rare, so the impact of missing information is minimal. From the figure, though our proposed modeling framework (under both AIC and BIC fitted models) provides realistic predictions on the IBNR, the BIC model produces a significantly smaller IBNR estimate than the AIC model. As discussed in Subsection 5.3.1 and displayed in the left panel of Figure 5, a sharp spike in the average claim severities for very long reporting delay is captured by the AIC model but not by the BIC model. Because the reporting delay of a simulated IBNR claim is usually long (this will be explained later in this subsection; see also Table 3), the claim severity for a simulated IBNR claim may be underestimated using the BIC. The IBNR prediction result is consistent with Remark 1, which suggests that that AIC model may result in better forecasts to the future.

Because a full predictive distribution can be obtained by simulation, we can also present the summary statistics in Table 4, where several risk quantities (e.g., conditional tail expectation [CTE] and Value at Risk [VaR]) may represent the regulatory/solvency capital requirement for an insurance company. The two-sided p values are computed as $2\min\{P\left(Y_{\text{agg}}^{\text{IBNR}} > y_{\text{agg}}^{\text{realized}}\right), P\left(Y_{\text{agg}}^{\text{IBNR}} < y_{\text{agg}}^{\text{realized}}\right)\}$, where $Y_{\text{agg}}^{\text{IBNR}}$ is the simulated IBNR (random variable) and $y_{\text{agg}}^{\text{realized}} = 1.112$ millions is the realized IBNR. The probabilities can be easily estimated as we have simulated $Y_{\text{agg}}^{\text{IBNR}}$ 50,000 times.

TABLE 4
Mean Value of Each Covariate for All Simulated IBNR Claims (Simulated from the AIC Fitted Model) and for All Claims (Based on the Empirical Training Set)

	Mean value/proportion		
	IBNR claims	All claims	Difference (%)
Policyholder age	46.113	45.994	0.3
Car age	4.429	3.726	18.9
Car fuel			
>Diesel	0.471	0.410	14.8
>Gasoline	0.529	0.590	–10.3
Geographical location			
>Region I	0.131	0.221	–40.6
>Region II	0.160	0.227	–29.8
>Region III	0.112	0.107	4.7
>Region IV	0.089	0.138	–35.6
>Capital	0.508	0.307	65.7
Car brand class			
>Class A	0.247	0.268	–7.7
>Class B	0.394	0.387	1.8
>Class C	0.359	0.345	3.9
Contract type			
>Renewal	0.644	0.572	12.7
>New	0.356	0.428	–17.0
Reporting delay (days)	308.583	17.839	1629.8

By evaluating the IBNR predictive distribution without covariates, we also highlight the importance of including covariates and using regression models for adequate IBNR predictions. We use the same models described in Subsections 5.3.1 to 5.3.3 and the same afore-mentioned simulation procedures, yet excluding any covariates for model fittings and simulations. The resulting IBNR predictive distribution is exhibited in Figure 7(b). The realized IBNR is on the right tail of the predictive distribution, providing evidence that the IBNR may be underestimated if covariates are excluded.

To investigate the above issue, we trace the covariate distributions of the simulated IBNR claims and compare them to the empirical covariates distributions of all claims from the training set. To do so, in the aforementioned simulation steps, we also record the covariates corresponding to each simulated IBNR claim. The comparison of covariates between the simulated IBNR claims and the empirical claims is summarized in Table 3. It is observed that the average reporting delay of IBNR claims (309 days) is much longer than that of empirical observed claims (18 days). This is because claims with longer reporting delays are less likely reported before the valuation date and hence are more likely to become IBNR claims. Further, the left panel of Figure 5 shows that claims with longer reporting delays are expected to result in larger claim amounts. Apart from the reporting delay issue, the geographical distribution of the IBNR claims differs a lot from that of the empirical claims. Though less than one-third of the observed claims come from the capital region, more than half of the simulated unreported claims occur in such a region. The right panel of Figure 5 also reveals that claims from the capital region are on average more severe than those from other regions. Overall, because the unreported claims are more likely to consist of undesirable features that lead to more severe claims, exclusion of the covariates in model fitting causes negative bias to the IBNR predictions.

6. CONCLUDING REMARKS

In this article, we introduce the TG-LRMoE as a flexible severity regression model. The denseness property warrants its full flexibility to capture complex distribution and regression structures, and the transformation parameter allows the model to effectively extrapolate a broad range of tail behaviors. The identifiability property also makes the proposed model suitable for statistical inference. We then present an ECM algorithm for efficient model calibration. Such an algorithm also imposes penalization on the parameters to prevent spurious fitted models. The proposed model is then adopted to model the severity and

reporting delay components of a European automobile insurance dataset, and both AIC and BIC fitted models are carefully analyzed. Several goodness-of-fit tests suggest that the proposed model (regardless of using AIC or BIC) provides a much better fit compared to various GLM models. On the other hand, the AIC model better captures the underlying complex relationships between reporting delay and claim severity than the BIC model. We also demonstrate the usefulness and the importance of the proposed model to appropriately predict the IBNR by conducting OS testing.

Though severity regression modeling is the main focus of this article, we admit that there are several ways to improve the microlevel reserving modeling framework presented in this article. Some of the potential research directions are discussed as follows:

Firstly, to avoid distortions to the emphasis of this article, we follow the framework of Antonio and Plat (2014) and use a simple Poisson GLM for frequency modeling. Obviously there are many possible enhancements to such a simple frequency model. For example, Badescu et al. (2019) considered a Cox hidden Markov model to capture the serial dependence of the claim counts. Extending the Cox hidden Markov model to incorporate covariates may result in a frequency model highly flexible in capturing complex distribution, regression, and serial dependence structures.

Secondly, for the model fitting in Subsection 5.3, the two-step approach adopted in this article requires discarding some reporting delay data points and hence it suffers from a bias–variance trade-off. Therefore, it is worthwhile to consider a one-step EM approach proposed by Verbelen et al. (2018): modeling frequency and reporting delay simultaneously. Though such a study works on frequency modeling including only time-related covariates, we plan to extend it to incorporate policyholder information as well as the loss severities.

Thirdly, for a particular policyholder, a claim severity record in the past may provide some predictive power to the severity of next claim. This feature can be captured through a random effects model introduced by Laird and Ware (1982). We plan to extend the proposed class of LRMoE to incorporate the random effects among policyholders.

The R code that specifically implements the results of the TG-LRMoE is available upon request. Understanding the importance of allowing practitioners to easily implement the proposed LRMoE, we are currently building a comprehensive and user-friendly R package with detailed documentation on software implementation (Tseung et al. 2020). This will allow researchers and actuaries to choose and test a wide variety of frequency and severity distribution functions. In addition to catering to other characteristics such as multivariate distributions, zero-inflated models, and incomplete data, the `R` package will potentially extend the actuarial application of LRMoE.

ACKNOWLEDGMENTS

The authors thank the editor and two anonymous referees for their valuable comments and suggestions.

FUNDING

The authors acknowledge the financial support provided by the Committee on Knowledge Extension Research (CKER) of the Society of Actuaries. This work is also partly supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. Tsz Chai Fung further acknowledges support by the SOA Hickman Scholars Program.

REFERENCES

Ahn, S., J. H. Kim, and V. Ramaswami. 2012. A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics and Economics* 51 (1):43–52. doi:10.1016/j.insmatheco.2012.02.002

Antonio, K., and R. Plat. 2014. Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal* 2014 (7):649–69. doi:10.1080/03461238.2012.755938

Badescu, A. L., T. Chen, X. S. Lin, and D. Tang. 2019. A marked Cox model for the number of IBNR claims: Estimation and application. *ASTIN Bulletin* 49 (3):709–39. doi:10.1017/asb.2019.15

Badescu, A. L., X. S. Lin, and D. Tang. 2016. A marked Cox model for the number of IBNR claims: Theory. *Insurance: Mathematics and Economics* 69: 29–37. doi:10.1016/j.insmatheco.2016.03.016

Bakar, S. A., N. Hamzah, M. Maghsoudi, and S. Nadarajah. 2015. Modeling loss data using composite models. *Insurance: Mathematics and Economics* 61: 146–54. doi:10.1016/j.insmatheco.2014.08.008

Blostein, M., and T. Miljkovic. 2019. On modeling left-truncated loss data using mixtures of distributions. *Insurance: Mathematics and Economics* 85: 35–46. doi:10.1016/j.insmatheco.2018.12.001

Box, G. E. P., and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 26 (2):211–43. doi:10.1111/j.2517-6161.1964.tb00553.x

Calderín-Ojeda, E., and C. F. Kwok. 2016. Modeling claims data with composite Stoppa models. *Scandinavian Actuarial Journal* 2016 (9):817–36. doi:10.1080/03461238.2015.1034763

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1):1–22. doi:10.1111/j.2517-6161.1977.tb01600.x

Embrechts, P., C. Klüppelberg, and T. Mikosch. 1997. *Modelling extremal events: For insurance and finance*. New York: Springer.

Frees, E. W., and E. A. Valdez. 2008. Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103 (484):1457–69. doi:10.1198/016214508000000823

Fung, T. C., A. L. Badescu, and X. S. Lin. 2019a. A class of mixture of experts models for general insurance: Application to correlated claim frequencies. *ASTIN Bulletin* 49 (3):647–88. doi:10.1017/asb.2019.25

Fung, T. C., A. L. Badescu, and X. S. Lin. 2019b. A class of mixture of experts models for general insurance: Theoretical developments. *Insurance: Mathematics and Economics* 89, 111–27. doi:10.1016/j.insmatheco.2019.09.007

Gan, G., and E. A. Valdez. 2018. Fat-tailed regression modeling with spliced distributions. *North American Actuarial Journal* 22 (4):554–73. doi:10.1080/10920277.2018.1462718

Grün, B., and T. Miljkovic. 2019. Extending composite loss models using a general framework of advanced computational tools. *Scandinavian Actuarial Journal* 2019 (8):642–60. doi:10.1080/03461238.2019.1596151

Gui, W., R. Huang, and X. S. Lin. 2018. Fitting the Erlang mixture model to data via a GEM-CMM algorithm. *Journal of Computational and Applied Mathematics* 343:189–205. doi:10.1016/j.cam.2018.04.032

Jiang, W., and M. A. Tanner. 1999. On the identifiability of mixtures-of-experts. *Neural Networks* 12 (9):1253–58. doi:10.1016/S0893-6080(99)00066-0

Jordan, M. I., and R. A. Jacobs. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6 (2):181–214. doi:10.1162/neco.1994.6.2.181

Klugman, S. A., H. H. Panjer, and G. E. Willmot. 2012. *Loss models: From data to decisions*, Vol. 715. Hoboken, NJ: John Wiley & Sons, Inc..

Kuha, J. 2004. AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods & Research* 33 (2):188–229. doi:10.1177/0049124103262065

Laird, N. M., and J. H. Ware. 1982. Random-effects models for longitudinal data. *Biometrics* 38 (4):963–74. doi:10.2307/2529876

Lee, S. C. K., and X. S. Lin. 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal* 14 (1):107–30. doi:10.1080/10920277.2010.10597580

McLachlan, G., and D. Peel. 2000. *Finite mixture models* Hoboken, NJ: John Wiley & Sons, Inc..

Meng, X.-L., and D. B. Rubin. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* 80 (2):267–78. doi:10.1093/biomet/80.2.267

Miljkovic, T., and B. Grün. 2016. Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics* 70:387–96. doi:10.1016/j.insmatheco.2016.06.019

Norberg, R. 1993. Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin* 23 (1):95–115. doi:10.2143/AST.23.1.2005103

Pigeon, M., and M. Denuit. 2011. Composite lognormal–Pareto model with random threshold. *Scandinavian Actuarial Journal* 2011 (3):177–92. doi:10.1080/03461231003690754

Reynkens, T., R. Verbelen, J. Beirlant, and K. Antonio. 2017. Modelling censored losses using splicing: A global fit strategy with mixed Erlang and extreme value distributions. *Insurance: Mathematics and Economics* 77:65–77. doi:10.1016/j.insmatheco.2017.08.005

Scollnik, D. P., and C. Sun. 2012. Modeling with Weibull-Pareto models. *North American Actuarial Journal* 16 (2):260–72. doi:10.1080/10920277.2012.10590640

Teicher, H. 1963. Identifiability of finite mixtures. *The Annals of Mathematical Statistics* 34 (4):1265–9. doi:10.1214/aoms/1177703862

Tseung, C. L. N. S., A. L. Badescu, T. C. Fung, and X. S. Lin. 2020. LRMoE: An interactive R package for flexible actuarial loss modelling using mixture of experts regression model. Working paper, University of Toronto.

Verbelen, R., K. Antonio, G. Claeskens, and J. Crevecoeur. 2018. An EM algorithm to model the occurrence of events subject to a reporting delay. Working paper, University of Amsterdam.

Verbelen, R., L. Gong, K. Antonio, A. L. Badescu, and X. S. Lin. 2015. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin* 45 (3):729–58. doi:10.1017/asb.2015.15

Verrall, R., and M. Wüthrich. 2016. Understanding reporting delay in general insurance. *Risks* 4 (3):25. doi:10.3390/risks4030025

Wüthrich, M. V. 2018. Machine learning in individual claims reserving. *Scandinavian Actuarial Journal* 2018 (6):465–80. doi:10.1080/03461238.2018.1428681

Wüthrich, M. V., and M. Merz. 2008. *Stochastic claims reserving methods in insurance*, Vol. 435. Chichester, England: John Wiley & Sons.

Discussions on this article can be submitted until January 1, 2022. The authors reserve the right to reply to any discussion. Please see the Instructions for Authors found online at http://www.tandfonline.com/uaaj for submission instructions.

APPENDIX A. SUPPLEMENTARY INFORMATION FOR SECTION 3

This appendix section provides some definitions and proofs for Section 3.

A.1. Denseness Property

This subsection provides the proofs of Properties 1 and 2.

Proof of Property 1. The Gamma-LRMoE is the TG-LRMoE when $\gamma = 1$, so the class of TG-LRMoE contains the class of Gamma-LRMoE. Section 5.2.1 of Fung, Badescu, and Lin (2019b) shows that the class of Gamma-LRMoE is uniformly dense in a class of severity regression distributions. Therefore, the same denseness property applies to the class of TG-LRMoE. □

Proof of Property 2. Define a function $r_\gamma(u) = \big((1+u)^\gamma - 1\big)/\gamma$. Note that it is a monotone increasing function that continuously maps $(0,\infty)$ into $(0,\infty)$. From theorem 3.3 of Fung, Badescu, and Lin (2019b), the necessary and sufficient denseness condition for a class of LRMoE is that for all $q>0$, there exists a sequence of parameters $\{\boldsymbol{v}_q^{(l)}\}_{l=1,2,\ldots}$ such that $F\left(\cdot;\boldsymbol{v}_q^{(l)}\right)\xrightarrow{\mathcal{D}} q$ as $l\to\infty$, where F is the distribution function of the corresponding expert function. Note that Gamma-LRMoE is dense, so for any $q>0$, there exists a sequence of Gamma random variables $\{Y_q^{(l)}\}$ such that $Y_q^{(l)}\xrightarrow{\mathcal{D}} r_\gamma(q) > 0$. The continuous mapping theorem suggests that $r_\gamma^{-1}\left(Y_q^{(l)}\right)\xrightarrow{\mathcal{D}} r_\gamma^{-1}(r_\gamma(q)) = q$. Note that $r_\gamma^{-1}\left(Y_q^{(l)}\right)$ follows a TGD, so the result follows. □

A.2. Tail Heaviness

In this subsection, the formal definition to compare the tail heaviness of two severity random variables is first recalled in Definition A.1. The proofs of Properties 3 to 5 are then presented. Finally, a table that compares the tail of TGD (or TG-LRMoE) to various severity distributions is exhibited.

Definition A.1. *Consider two severity random variables Y_1 and Y_2, and the ratio limit $R(Y_1,Y_2) = \lim_{y\to\infty} f_{Y_1}(y)/f_{Y_2}(y)$, where f is the pdf. We say that Y_1 (or f_{Y_1}) has a heavier, similar, or lighter tail than Y_2 (or f_{Y_2}) if $R(Y_1,Y_2)=\infty, 0<R(Y_1,Y_2)<\infty$, or $R(Y_1,Y_2)=0$, respectively.*

Proof of Property 3. It is obvious that $R\big((Y_i|\boldsymbol{x}),Y^{(j)}\big) = \pi_j(\boldsymbol{x};\boldsymbol{\alpha})$ and $R\big((Y_i|\boldsymbol{x}^c),Y^{(j)}\big) = \tilde{\pi}_j(\boldsymbol{x}^c;\boldsymbol{\alpha})$, where $\tilde{\pi}_j(\boldsymbol{x}^c;\boldsymbol{\alpha})$ is defined under the proof of proposition 4.3 of Fung, Badescu, and Lin (2019b). are both positive and finite. □

Proof of Property 4. We order the TGD by $\tilde{f}(\cdot;m,\theta,\gamma) \prec \tilde{f}(\cdot;m^*,\theta^*,\gamma^*)$ when: (i) ($\gamma>\gamma^*$), (ii) ($\gamma=\gamma^*$ and $\theta<\theta^*$), or (iii) ($\gamma=\gamma^*$ and $\theta=\theta^*$ and $m<m^*$). Under each case, simple algebraic manipulations yield $\lim_{y\to\infty}(\tilde{f}(y;m^*,\theta^*,\gamma^*)/\tilde{f}(y;m,\theta,\gamma))=\infty$. □

Proof of Property 5. It is obvious from Equation (3.1) that

$$\lim_{y\to\infty}\frac{S(y\lambda)}{S(y)} = \lim_{y\to\infty}\frac{\lambda\tilde{f}(y\lambda;m,\theta,0)}{\tilde{f}(y;m,\theta,0)} = \lambda^{-1/\theta}.$$

□

A.3. Model Identifiability

Proof of Property 6. Firstly, from proposition 2 of Teicher (1963) and from the three-step approach adopted by the proof of theorem 3.1 of Fung, Badescu, and Lin (2019a), it immediately follows that the class of Gamma-LRMoE is identifiable up to translation and permutation.

Secondly, we aim to show that the class of TG-LRMoE with a fixed $\gamma>0$ is identifiable up to translation and permutation. Note from Equation (2.4) that if two pdfs of the TG-LRMoE (with a fixed $\gamma>0$) are equal; that is,

$$\sum_{j=1}^{g^*}\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}^*)f\big(\tilde{y}_i(\gamma);m_j^*,\theta_j^*\big)(1+y_i)^{\gamma-1} = \sum_{j=1}^{g}\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha})f\big(\tilde{y}_i(\gamma);m_j,\theta_j\big)(1+y_i)^{\gamma-1}, \qquad y_i>0,$$

then we have

$$\sum_{j=1}^{g^*}\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}^*)f\big(\tilde{y}_i(\gamma);m_j^*,\theta_j^*\big) = \sum_{j=1}^{g}\pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha})f\big(\tilde{y}_i(\gamma);m_j,\theta_j\big), \qquad y_i>0,$$

which is the pdf form of a Gamma-LRMoE. Because $y_i>0 \iff \tilde{y}_i(\gamma)>0$ and the Gamma-LRMoE is identifiable up to translation and permutation, we have $g^*=g$ and $\big(\boldsymbol{\alpha}_j^*,m_j^*,\theta_j^*\big) = \big(\boldsymbol{\alpha}_{c(j)}+\boldsymbol{\delta},m_{c(j)},\theta_{c(j)}\big)$ for $j=1,\ldots,g$, and hence the result follows.

Finally, it suffices to show that if two pdfs of the TG-LRMoE are equal, they must have the same γ. This is trivial, or otherwise the two TG-LRMoE will have different tail behaviors (see Property 4). □

TABLE A.1
Comparing the Tail of Various Severity Distributions to the TGD

Distribution	Parameters	pdf f	Tail comparison
Gamma	$(m,\ \theta)$	$\frac{y^{m-1}e^{-y/\theta}}{\Gamma(m)\theta^m}$	$TGD \sim f \iff \gamma = 1$
Weibull	$(\lambda,\ k)$	$\frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1} e^{-\left(\frac{y}{\lambda}\right)^k}$	$k \le 1 : TGD \sim f \iff (m,\theta,\gamma) = \left(1, \frac{\lambda^k}{k}, k\right)$ $k > 1 : TGD \prec (\succ) f \iff \gamma > (<)\ k$ or $(\gamma = k$ and $\theta \le (>) \frac{\lambda^k}{k})$
Inverse Gaussian	$(\mu,\ \lambda)$	$\sqrt{\frac{\lambda}{2\pi y^3}} e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}}$	$TGD \prec (\succ) f \iff \gamma > (<)\ 1$ or $\left(\gamma = 1$ and $\theta < (\ge) \frac{2\mu}{\lambda}\right)$
Lognormal	$(\mu,\ \sigma^2)$	$\frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\log(y)-\mu)^2}{2\sigma^2}}$	$TGD \prec (\succ) f \iff \gamma \ge (\to)\ 0$
Burr	$(\alpha,\ c,\ k)$	$\frac{\frac{kc}{\alpha}\left(\frac{y}{\alpha}\right)^{c-1}}{\left(1+\left(\frac{y}{a}\right)^c\right)^{k+1}}$	$TGD \sim f \iff \gamma \to 0$ and $(m,\theta) = \left(1, \frac{1}{ck}\right)$

Note: Here, the symbols "$\prec$", "$\sim$", and "$\succ$" respectively refer to a distribution having a lighter, similar, or heavier tail than the other distribution.

APPENDIX B. ECM ALGORITHM FOR CENSORED OBSERVATIONS

This appendix section derives the ECM algorithm for the censored version of TG-LRMoE with the restriction $\gamma = 1$ for computational ease, which is useful to fit the reporting delay data in Subsection 5.3.2. Notations are defined in the same way as in Section 4, except that we denote $\boldsymbol{y}^u = (y_1^u, ..., y_n^u)$ and $\boldsymbol{y}^l = \left(y_1^l, ..., y_n^l\right)$ respectively, as the upper and lower censored values. In other words, we know that the ith observation is between y_i^l and y_i^u. Using the reporting delay (Subsection 5.3.2) as an example, an observed three-day delay means that $y_i^l = 3$ and $y_i^u = 4$. The observed data posterior log-likelihood is given by

$$l^{\text{pos}}\left(\boldsymbol{\Phi}; \boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}\right) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{g} \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha})\left(F\left(y_i^u; m_j, \theta_j\right) - F\left(y_i^l; m_j, \theta_j\right)\right)\right] + \log p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) + \text{const.}, \tag{B.1}$$

where the prior distribution $p(\cdot)$ of the parameters is given by Equation (4.3). The complete data posterior log-likelihood is given by

$$l^{\text{pos}}(\boldsymbol{\Phi}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{Z}) = \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{g} Z_{ij}\left(\log \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) + \log f\left(y_i; m_j, \theta_j\right)\right) + \log p(\boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{\theta}) + \text{const.}, \tag{B.2}$$

where $\boldsymbol{y} = (y_1, ..., y_n)$ represents the uncensored observations.
In the lth iteration of the E step, the expectation of the complete data posterior log-likelihood given the observed data is

$$\begin{aligned}
&Q\left(\boldsymbol{\Phi}; \boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right) \\
&= E\left[l^{\text{pos}}\left(\boldsymbol{\Phi}; \boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}, \boldsymbol{Z}\right) | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}^{(l)}\left(\log \pi_j(\boldsymbol{x}_i;\boldsymbol{\alpha}) + (m_j - 1)\log \hat{y}_{ij}^{(1,l)} - \frac{\hat{y}_{ij}^{(2,l)}}{\theta_j} - m_j \log \theta_j - \log(m_j - 1)!\right) \\
&\quad - \sum_{j=1}^{g-1}\sum_{p=0}^{P} \frac{\alpha_{jp}^2}{2\sigma_{jp}^2} + \sum_{j=1}^{g}\left(\left(\nu_j^{(1)} - 1\right)\log m_j - \frac{m_j}{\lambda_j^{(1)}}\right) + \sum_{j=1}^{g}\left(\left(\nu_j^{(2)} - 1\right)\log \theta_j - \frac{\theta_j}{\lambda_j^{(2)}}\right) + \text{const.},
\end{aligned} \tag{B.3}$$

where $z_{ij}^{(l)}$ is given by

$$z_{ij}^{(l)} = E\left[Z_{ij}|\boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right] = \frac{\pi_j\left(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(l-1)}\right) F\left(y_i; \boldsymbol{\theta}_{jk}^{(l-1)}\right)}{\sum_{j'=1}^{g} \pi_{j'}\left(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(l-1)}\right) F\left(y_i; \boldsymbol{\theta}_{j'k}^{(l-1)}\right)}, \tag{B.4}$$

$\log \hat{y}_{ij}^{(1,l)}$ is computed by numerical integration as

$$\begin{aligned} \log \hat{y}_{ij}^{(1,l)} &= E\left[\log Y_i|\boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}, Z_{ij} = 1\right] \\ &= \int_{y_i^l}^{y_i^u} \log y \frac{f\left(y; m_j^{(l-1)}, \theta_j^{(l-1)}\right)}{F\left(y_i^u; m_j^{(l-1)}, \theta_j^{(l-1)}\right) - F\left(y_i^l; m_j^{(l-1)}, \theta_j^{(l-1)}\right)} dy, \end{aligned} \tag{B.5}$$

and $\hat{y}_{ij}^{(2,l)}$ is given by

$$\begin{aligned} \hat{y}_{ij}^{(2,l)} &= E\left[Y_i|\boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}, Z_{ij} = 1\right] \\ &= \int_{y_i^l}^{y_i^u} y \frac{f\left(y; m_j^{(l-1)}, \theta_j^{(l-1)}\right)}{F\left(y_i^u; m_j^{(l-1)}, \theta_j^{(l-1)}\right) - F\left(y_i^l; m_j^{(l-1)}, \theta_j^{(l-1)}\right)} dy \\ &= \frac{F\left(y_i^u; m_j^{(l-1)} + 1, \theta_j^{(l-1)}\right) - F\left(y_i^l; m_j^{(l-1)} + 1, \theta_j^{(l-1)}\right)}{F\left(y_i^u; m_j^{(l-1)}, \theta_j^{(l-1)}\right) - F\left(y_i^l; m_j^{(l-1)}, \theta_j^{(l-1)}\right)}. \end{aligned} \tag{B.6}$$

In the CM step, we update the parameters $\boldsymbol{\Phi}$ to increase $Q\left(\boldsymbol{\Phi}; \boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right)$. We adopt a similar decomposition strategy as Equations (4.7) to (4.9):

$$Q\left(\boldsymbol{\Phi}; \boldsymbol{y}^l, \boldsymbol{y}^u, \boldsymbol{x}, \boldsymbol{\Phi}^{(l-1)}\right) = Q^{(l)}(\boldsymbol{\alpha}) + \sum_{j=1}^{g} S_j^{(l)}\left(m_j, \theta_j\right), \tag{B.7}$$

where $Q^{(l)}(\boldsymbol{\alpha})$ is the same as Equation (4.8) and

$$\begin{aligned} S_j^{(l)}\left(m_j, \theta_j\right) = \sum_{i=1}^{n} z_{ij}^{(l)} &\left((m_j - 1) \log \hat{y}_{ij}^{(1,l)} - \frac{\hat{y}_{ij}^{(2,l)}}{\theta_j} - m_j \log \theta_j - \log (m_j - 1)! \right) \\ &+ \left(\nu_j^{(1)} - 1\right) \log m_j - \frac{m_j}{\lambda_j^{(1)}} + \left(\nu_j^{(2)} - 1\right) \log \theta_j - \frac{\theta_j}{\lambda_j^{(2)}}. \end{aligned} \tag{B.8}$$

Now, $\boldsymbol{\alpha}^{(l)}$ is obtained by maximizing $Q^{(l)}(\boldsymbol{\alpha})$ through the IRLS procedure in Equation (4.11). Then, the parameters $\left(\boldsymbol{m}^{(l)}, \boldsymbol{\theta}^{(l)}\right)$ are obtained through maximizing $S_j^{(l)}\left(m_j, \theta_j\right)$ with respect to $\left(m_j, \theta_j\right)$:

$$m_j^{(l)} = \underset{m_j > 0}{\operatorname{argmax}} \; S_j^{(l)}\left(m_j, \tilde{\theta}_j^{(l)}(m_j)\right); \qquad \theta_j^{(l)} = \tilde{\theta}_j^{(l)}\left(m_j^{(l)}\right), \tag{B.9}$$

where

$$\tilde{\theta}_j^{(l)}(m_j) = \frac{\lambda_j^{(2)}}{2} \left(\left(\nu_j^{(2)} - 1\right) - m_j \sum_{i=1}^{n} z_{ij}^{(l)} + \sqrt{\left(m_j \sum_{i=1}^{n} z_{ij}^{(l)} - \left(\nu_j^{(2)} - 1\right) \right)^2 + \frac{4}{\lambda_j^{(2)}} \sum_{i=1}^{n} z_{ij}^{(l)} \hat{y}_{ij}^{(2,l)}} \right). \tag{B.10}$$

Finally, the initialization of parameters and the adjustment of hyper-parameter g are no different from those described in Subsection 4.3.

APPENDIX C. FITTED PARAMETERS

C.1. Severity and Reporting Delay

In Subsections 5.3.1 and 5.3.2, we fitted the TG-LRMoE into the severity and reporting delay data. The fitted parameters and the related quantities are displayed in Tables C.1 to C.4. In the tables, the subgroup conditional (transformed) mean is given by $E[\tilde{Y}_i|Z_{ij}=1]=\hat{m}_j\hat{\theta}_j$, and the subgroup probability is estimated as $P(Z_{ij}=1)=\sum_{i=1}^{n}\pi_j(\boldsymbol{x}_i;\hat{\boldsymbol{\alpha}})/n$.

C.2. Frequency

In Subsection 5.3.3, we fitted a Poisson GLM to the claim frequency data. The regression parameters and test statistics are displayed in the left panel of Table C.5, showing that all pieces of policyholder information (covariates) have a significant influence on claim frequency.

To examine the time trend of claim frequencies, we fit a Poisson GLM that further includes the contract date of each policyholder as a covariate. For each policyholder, we calculate the contract date as the mid-point of the exposure period (i.e., average of contract start date and end date). The unit of contract date is year. The right panel of Table C.5 presents a summary of the fitted model. Despite the negative regression coefficient of contract date, its impact is too small to be statistically significant. In other words, the time effect on the exposure-adjusted claim frequencies is already sufficiently explained by the effects of other covariates.

TABLE C.1
Fitted Severity Model Parameters and the Related Quantities (AIC)

	Component j								
	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$
$\hat{\alpha}_{j0}$	3.614	0.795	–1.631	–3.069	–3.030	3.209	–0.186	1.355	0.000
$\hat{\alpha}_{j1}$	–0.100	0.002	–0.001	–0.028	–0.018	–0.035	0.011	0.001	0.000
$\hat{\alpha}_{j2}$	0.179	0.137	0.124	–0.029	0.175	–0.503	–0.072	–0.045	0.000
$\hat{\alpha}_{j3}$	0.030	–0.401	–0.334	0.735	–0.711	–0.050	–0.172	–0.095	0.000
$\hat{\alpha}_{j4}$	–1.279	–1.624	1.559	1.814	–2.203	–0.332	–0.114	–1.640	0.000
$\hat{\alpha}_{j5}$	–2.032	–0.963	1.180	1.688	–2.389	0.454	–2.111	–0.962	0.000
$\hat{\alpha}_{j6}$	0.678	–0.566	1.163	1.325	–1.203	0.691	0.036	–0.599	0.000
$\hat{\alpha}_{j7}$	–1.339	–1.156	2.245	2.555	–0.733	1.587	–1.066	–1.235	0.000
$\hat{\alpha}_{j8}$	–0.524	0.448	0.663	–0.970	–0.505	–1.750	–0.755	–0.182	0.000
$\hat{\alpha}_{j9}$	–1.021	0.248	0.456	0.241	0.034	0.251	0.010	–0.001	0.000
$\hat{\alpha}_{j10}$	–2.355	0.791	–0.267	–1.050	–0.012	–1.061	–0.869	0.302	0.000
$\hat{\alpha}_{j11}$	–0.211	–0.797	–0.032	1.058	0.984	0.135	0.424	–0.372	0.000
$\hat{m}_j$	114.253	77.849	21.701	17.304	8.074	34.003	109.477	66.706	214.113
$\hat{\theta}_j$	0.140	0.148	0.515	0.837	1.282	0.324	0.110	0.211	0.047
$\hat{\gamma}$	0.098								
$E[\tilde{Y}_i\|Z_j=1]$	16.015	11.497	11.181	14.485	10.347	11.026	12.068	14.095	10.010
$P(Z_j=1)$	0.024	0.214	0.161	0.059	0.019	0.152	0.096	0.147	0.129

TABLE C.2
Fitted Severity Model Parameters and the Related Quantities (BIC)

	Component j			
	$j=1$	$j=2$	$j=3$	$j=4$
$\hat{\alpha}_{j0}$	1.446	–2.792	–1.031	0.000
$\hat{\alpha}_{j1}$	–0.097	–0.011	–0.004	0.000
$\hat{\alpha}_{j2}$	0.186	0.054	0.003	0.000
$\hat{\alpha}_{j3}$	0.475	0.110	–0.016	0.000
$\hat{\alpha}_{j4}$	–0.558	1.335	1.081	0.000
$\hat{\alpha}_{j5}$	–1.130	1.394	0.983	0.000
$\hat{\alpha}_{j6}$	0.671	0.623	0.510	0.000
$\hat{\alpha}_{j7}$	–1.486	2.111	1.272	0.000
$\hat{\alpha}_{j8}$	–0.134	0.035	0.163	0.000
$\hat{\alpha}_{j9}$	–1.302	0.148	0.047	0.000
$\hat{\alpha}_{j10}$	–2.134	–0.547	–0.162	0.000
$\hat{\alpha}_{j11}$	–0.137	0.743	0.058	0.000
$\hat{m}_j$	123.777	11.887	81.090	45.445
$\hat{\theta}_j$	0.131	1.040	0.129	0.276
$\hat{\gamma}$	0.100			
$E[\tilde{Y}_i \mid Z_j = 1]$	16.209	12.363	10.421	12.540
$P(Z_j = 1)$	0.025	0.183	0.305	0.486

TABLE C.3
Fitted Reporting Delay AIC Model Parameters and the Related Quantities

	Component j							
	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$\hat{\alpha}_{j0}$	–1.227	0.763	–2.317	–1.097	–2.533	0.061	–0.364	0.000
$\hat{\alpha}_{j1}$	–0.042	–0.030	0.012	–0.003	0.003	0.002	–0.004	0.000
$\hat{\alpha}_{j2}$	0.023	–0.085	–0.076	0.064	0.176	0.005	–0.012	0.000
$\hat{\alpha}_{j3}$	0.914	0.070	0.047	–0.097	0.306	0.070	0.190	0.000
$\hat{\alpha}_{j4}$	1.840	1.871	0.535	0.266	–0.160	0.295	–0.870	0.000
$\hat{\alpha}_{j5}$	1.471	1.439	0.730	–0.058	0.608	0.430	–0.636	0.000
$\hat{\alpha}_{j6}$	–0.280	2.107	0.548	–0.441	–0.954	0.520	–0.164	0.000
$\hat{\alpha}_{j7}$	2.148	1.543	–0.681	0.644	1.240	0.964	–0.278	0.000
$\hat{\alpha}_{j8}$	–0.909	–1.461	0.462	–0.294	2.688	0.030	0.148	0.000
$\hat{\alpha}_{j9}$	–1.450	–2.364	1.071	–0.204	0.016	–0.390	–0.042	0.000
$\hat{\alpha}_{j10}$	0.046	–0.953	0.013	1.515	–1.874	0.227	0.134	0.000
$\hat{m}_j$	2.212	3.431	11.514	2.199	1.990	12.450	0.808	1.050
$\hat{\theta}_j$	28.984	1.027	0.257	1.468	1.592	0.089	226.712	11.164
$E[Y_i \mid Z_j = 1]$	64.112	3.523	2.963	3.227	3.169	1.106	183.126	11.722
$P(Z_j = 1)$	0.027	0.084	0.073	0.174	0.082	0.295	0.086	0.180

TABLE C.4
Fitted Reporting Delay BIC Model Parameters and the Related Quantities

	Component j			
	$j=1$	$j=2$	$j=3$	$j=4$
$\hat{\alpha}_{j0}$	0.321	0.618	0.883	0.000
$\hat{\alpha}_{j1}$	–0.005	–0.013	–0.006	0.000
$\hat{\alpha}_{j2}$	–0.004	–0.017	0.008	0.000
$\hat{\alpha}_{j3}$	–0.053	0.296	–0.048	0.000
$\hat{\alpha}_{j4}$	–0.098	–0.555	0.326	0.000
$\hat{\alpha}_{j5}$	–0.339	–0.711	0.136	0.000
$\hat{\alpha}_{j6}$	–0.508	–0.629	–0.095	0.000
$\hat{\alpha}_{j7}$	–0.838	–0.888	–0.263	0.000
$\hat{\alpha}_{j8}$	–0.159	–0.280	–0.079	0.000
$\hat{\alpha}_{j9}$	0.279	0.079	0.106	0.000
$\hat{\alpha}_{j10}$	–0.324	–0.259	–0.335	0.000
$\hat{m}_j$	0.909	0.508	2.776	12.189
$\hat{\theta}_j$	12.558	244.550	1.096	0.091
$E[Y_i\|Z_j=1]$	11.410	124.249	3.042	1.106
$P(Z_j=1)$	0.183	0.142	0.423	0.253

TABLE C.5
Estimated Regression Coefficients for Poisson GLM

	Without contract date			with contract date		
	Estimate	SE	p Value	Estimate	SE	p Value
Intercept	–2.0771	0.0491	.0000	–2.0554	0.0514	.0000
Policyholder age	–0.0085	0.0009	.0000	–0.0084	0.0009	.0000
Car age	–0.0157	0.0035	.0000	–0.0146	0.0036	.0001
Car fuel						
>Diesel	0.1389	0.0213	.0000	0.1402	0.0213	.0000
>Gasoline	–			–		
Geographical location						
>Region I	–0.0324	0.0285	.2570	–0.0365	0.0287	.2038
>Region II	0.1523	0.0283	.0000	0.1498	0.0283	.0000
>Region III	–0.0996	0.0362	.0060	–0.1003	0.0362	.0056
>Region IV	–0.1346	0.0331	.0000	–0.1381	0.0332	.0000
>Capital	–			–		
Car brand class						
>Class A	–0.0461	0.0268	.0860	–0.0456	0.0268	.0890
>Class B	–0.1216	0.0239	.0000	–0.1235	0.0240	.0000
>Class C	–			–		
Contract type						
>Renewal	–0.2083	0.0212	.0000	–0.2015	0.0218	.0000
>New	–			–		
Contract date	–			–0.0092	0.0065	.1565

Note: The left panel is the frequency model adopted for IBNR prediction. The right panel includes policyholder contract date as a covariate and aims to examine the time effect on claim frequencies.